Voxel-Wise Image Analysis
for White Matter Hyperintensity Segmentation

by

Jesse Knight

A Thesis
presented to
University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Applied Science
in
Engineering Systems and Computing

Guelph, Ontario, Canada
© Jesse Knight, December, 2017
White matter hyperintensities (WMH) are regions of increased pixel intensity in T2-weighted MRI which are correlated with several neurodegenerative diseases. Human segmentation of WMH is time consuming and inconsistent, motivating automation of WMH segmentation.

While many algorithms for this task have previously been proposed, few have been validated on MRI from different sources, despite the sensitivity of most algorithms to source-specific image features.

This thesis presents a segmentation algorithm called “Voxel-Wise Logistic Regression” (VLR), which provides both good interpretability and segmentation performance. VLR uses FLAIR MRI to estimate the WMH class probability image using spatially varying logistic parameters $\beta(x)$. These “parameter images” also concisely summarize the model class discrimination.

Additionally, a validation framework called “Leave-One-Source-Out Cross Validation” (LOSO-CV) is introduced, which provides more realistic estimation of model performance on “never-before-seen” MRI sources. Segmentation performance of the VLR model under LOSO-CV is presented using 96 open-source images from 7 MRI sources.
Acknowledgments

This research was supported in part by the Natural Science and Engineering Research Council of Canada (NSERC CGS-M) and by the Ontario Ministry of Advanced Education and Skills Development (OGS-M).

I also owe thanks to . . .
Jeremy, Danny, and Carly, for our cherished polemic forums;
Daniel, for sharing with me passions, pizza, and almost projects;
Thor, Colin, Terrance, and Dylan for our ‘deep’ conversations;
Brayden, Aaron, and Carson, for board games and Jamaican bacon;
Erika, Denise, Emily, and Zyra, for bunting, crosswords, and Harambe;
My parents and Ali, for all your love and support;
and
My supervisors, for sharing with me your time, expertise, and so many opportunities.

If you want the truth to stand clear before you,
never be for or against.
The struggle between ‘for’ and ‘against’
is the mind’s worst disease.

— Sent-ts’an c. 700 C.E.
Contents

1 Introduction 1
  1.1 Background ................................................................. 1
    1.1.1 Magnetic Resonance Imaging ........................................... 2
    1.1.2 White Matter Disease .................................................. 5
    1.1.3 MRI in White Matter Disease .......................................... 7
  1.2 Problem Statement ....................................................... 8
    1.2.1 Objective ............................................................... 9
    1.2.2 Challenges to Automatic Segmentation .............................. 9
  1.3 Prior Work ......................................................................... 11
    1.3.1 Segmentation Models & Features ....................................... 11
    1.3.2 Proposed Methods ....................................................... 13
    1.3.3 Limitations .................................................................. 17
  1.4 Proposed Algorithm ......................................................... 23
    1.4.1 Voxel-Wise Logistic Regression ...................................... 23
  1.5 Contributions .................................................................... 24

2 Pre-Processing 25
  2.1 Registration ................................................................. 25
  2.2 Bias Correction .............................................................. 28
  2.3 Graylevel Standardization ................................................ 29
    2.3.1 Quantifying Standardization ........................................... 32
    2.3.2 Supervised Standardization ............................................ 34
  2.4 Pre-Processing Summary .................................................. 34

3 Voxel-Wise Logistic Regression 35
  3.1 Model Fitting .................................................................... 35
3.1.1 Maximum Likelihood Estimation .............................................. 36
3.1.2 Iterative Updates .............................................................. 36
3.1.3 Simplification ................................................................. 38
3.1.4 Challenges ................................................................. 38
3.2 Regularization ................................................................. 40
  3.2.1 Data Augmentation ......................................................... 40
  3.2.2 Parameter Norm Penalties ............................................... 41
  3.2.3 Pseudo-Lesions ............................................................ 42
  3.2.4 Parameter Image Smoothing ........................................... 43
3.3 Post-Processing ................................................................. 44
  3.3.1 Thresholding ............................................................... 44
  3.3.2 Minimum Lesion Size .................................................... 45
3.4 Model Summary ............................................................... 45
  3.4.1 Tunable Parameters ...................................................... 46
4 Experiment & Results .............................................................. 48
  4.1 Data ............................................................................. 48
  4.2 Segmentation Performance Metrics ..................................... 49
  4.3 Cross Validation Frameworks ............................................ 50
    4.3.1 Leave-One-Source-Out CV .......................................... 52
  4.4 Graylevel Standardization ................................................ 53
  4.5 Regularization ............................................................... 56
    4.5.1 Toy Model ............................................................... 56
    4.5.2 Parameter Norm Penalties .......................................... 56
    4.5.3 Pseudo Lesion Regularization ...................................... 59
  4.6 Full Modal – Preliminaries ............................................... 61
    4.6.1 Convergence ............................................................ 61
    4.6.2 Cross Validation ....................................................... 62
    4.6.3 Baseline Model Performance ...................................... 64
  4.7 Full Model – Performance Results ..................................... 65
    4.7.1 Graylevel Standardization .......................................... 66
    4.7.2 Regularization ............................................................ 66
    4.7.3 Parameter Images ..................................................... 70
  4.8 Optimized Model Summary ............................................... 74
List of Figures

1.1 Visualization of T1 and T2 relaxation. .................................................. 2
1.2 RF and MR signal for a basic Spin Echo sequence. ................................. 4
1.3 Example MRI image set with WMH pathology; from [12]. ......................... 5
1.4 Distributions of TP, FP, and FN in 96 FLAIR MRI, following supervised optimal thres-
holding in MNI space. Best viewed in colour. ........................................... 20
1.5 The logistic regression model. ................................................................. 21
1.6 Overview of the necessary processing steps. ........................................... 23

2.1 Example set of FLAIR MRI before and after registration to the MNI brain space. From [12]. 26
2.2 Example bias correction. From [12]. Best viewed in colour. ....................... 29
2.3 Illustration of potential separability objective functions using three sets of sample data
(black circles). \( Z_\Delta \) counts the class changes in the sorted feature data (numbered red
lines). \( Z_\star \) computes the ratio of the overlap between class data distributions (red area) to
the total data distribution (blue areas). ...................................................... 33

3.1 Effect of varying the logistic model parameters. ...................................... 39
3.2 Challenges encountered during estimation of a logistic model. ................. 40
3.3 Tissue prior probability images in MNI space. Derived from [86]. Best viewed in colour. 43
3.4 Overview of the proposed algorithm. Typefaces – upright Roman: images in native space;
italic Roman: images in standard (MNI) space; calligraphic: a set of images from several
patients; bold: a set of images corresponding to different features; Variables – \( C(x) \): man-
ual segmentation; \( Y(x) \): FLAIR image; \( \beta(x) \): parameter image; \( \hat{C}(x) \): estimated lesion
segmentation. Best viewed in colour. ....................................................... 46

4.1 Average distribution of WMH in Dataset A. Best viewed in colour. .......... 49
4.2 Image intensity PMFs. Best viewed in colour. ......................................... 53
4.3 Image graylevel PMFs after the two best standardization operations. Best viewed in colour. 54
4.4 Spatial depiction of \( Z_\Delta(x) \) and \( Z_\star(x) \) comparing raw images and images standardized using
HM3. Best viewed in colour. ................................................................. 55
4.5 Synthetic data distributions (9 voxels) used in toy scenarios. .................. 57
4.6 Toy model likelihoods as a function of $\beta$: $P(\beta) = e^{\mathcal{P}(\beta)}$ and $J(\beta) = e^{\mathcal{J}(\beta)}$, for different $\lambda$, using scenario e. The optimum is shown as a blue star. Best viewed in colour.

4.7 Toy model MAP estimation results for 6 different scenarios and different $\lambda$. Best viewed in colour.

4.8 Toy model MAP estimation results for 9 different scenarios and different numbers of pseudo-lesions $V$, shown as coloured diamonds corresponding to the scenario (spread of diamonds is for visualization only). Best viewed in colour.

4.9 Convergence characteristics of $\beta(x)$: quantiles (0.00, 0.05, ..., 1.00) of the update magnitude $|\Delta \beta(t)|$ versus iteration $(t)$, using all augmented data. Convergence is apparent by the 15th iteration.

4.10 Number of images from each scanner in each KF-CV fold. Faded colours show the expected value (evenly distributed, non-whole numbers); full colours show the implementation (approximation, whole numbers). Best viewed in colour.

4.11 Comparison of the estimated model performance using different cross validation methods. Box plots show median (centre line), 25th and 75th percentiles (box), extreme values (whiskers), and outliers (+).

4.12 Fitted parameter images $T(x)$ and $S(x)$ from the first LOSO-CV fold of the baseline model. Obvious artifacts arise from inadequate regularization. Best viewed in colour.

4.13 Baseline model performance, stratified by LL tertiles.

4.14 Simulated FLAIR images after graylevel standardization using each technique under investigation.

4.15 Comparison of the optimized model employing each graylevel standardization technique.

4.16 Comparison of the baseline model under LOSO-CV incorporating each of the regularization strategies in isolation.

4.17 Comparison of the optimized model under LOSO-CV using different prior strengths $\lambda$.

4.18 Parameter images following different smoothing filters. Best viewed in colour.

4.19 Comparison of the optimized model under LOSO-CV using different $\beta(x)$ smoothing.

4.20 Spatial effect intercept parameter images $\beta^0(x)$ from the LPA and VLR algorithms. For visual comparison, image means and variances are matched. Best viewed in colour.

4.21 Fitted parameter images $T(x)$ and $S(x)$ from the first LOSO-CV fold of the final model. Best viewed in colour.

4.22 Example segmentation. Best viewed in colour.

4.23 Final model performance, stratified by LL tertiles.

4.24 Scatter plot of final model performance, with 3rd order trend line and 90% confidence interval shown in grey. Best viewed in colour.

4.25 Bland-Altman plot showing total LL agreement between manual and VLR-segmented WMH. Shown in Log-scale to better illustrate results for small LL.

4.26 Distribution of True Positives (TP), False Positives (FP), and False Negatives (FN) from all LOSO-CV folds of the final model. Best viewed in colour.
4.27 Effect of the threshold $\pi$ on performance ........................................ 78
4.28 Comparison of VLR model performance versus LPA .......................... 80
4.29 Results report for the submitted method provided by the WMH Segmentation Competition. 82

A.1 Simulated FLAIR images using scan parameters from the experimental database. Colourmap is arbitrary but consistent. Best viewed in colour. .............................. 102
A.2 Simulated T1 (TE/TR = 5/15 ms) and T2 (TE/TR = 100/5500 ms) images. Best viewed in colour. ................................................................. 103
A.3 PMF of each tissue from simulated FLAIR, T1, and T2 images. The PMF of the WMH class is scaled by 25 for visibility. Best viewed in colour. ......................... 104
A.4 Histogram matching of synthetic data to different target histograms. Quantiles show high agreement regardless of the target histogram. Best viewed in colour. ............... 106

B.1 3D Slicer user interface for performing in-house manual segmentations and revisions. The tools used are highlighted in yellow, while the in-progress segmentation is shown in blue. Best viewed in colour. ................................................................. 112
B.2 Example revisions to the manual segmentations for the MS 2008 challenge dataset. Best viewed in colour. ................................................................. 113
B.3 Manually refined brain mask in MNI space, overlaid on a simulated BrainWeb FLAIR image. Mask outline is highlighted in red; inclusions are shown in grayscale; exclusions tinted red. Best viewed in colour. .................................................. 114
B.4 Comparison of parameter images estimated at full and half-resolution. Best viewed in colour. 117
## List of Tables

1.1 T1 and T2 constants for brain tissues at 1.5 Tesla. .......................... 3
1.2 Mean inter-rater agreement measures for manual and semi-automated WMH segmentation reported in previous works. ................................. 9
1.3 Summary of previous approaches to WMH segmentation with respect to image variability and reported performance (SI). ................................. 14
1.4 Works demonstrating excellent validation of a WMH segmentation algorithm. ................................. 19
3.1 Image filters considered for smoothing the estimated parameter images. ............... 43
3.2 Model hyperparameters and baseline values. The values after optimization are given in Table 4.5 ................................................................. 47
4.1 Summary of experimental image database. ................................................ 49
4.2 Graylevel agreement objective functions (mean) for different standardization operations. .................. 54
4.3 Toy data definitions, with \( y_c \sim \mathcal{N}(\mu_c, \sigma_c) \). .......................... 56
4.4 Baseline model performance metrics (median) ............................................. 65
4.5 Model hyperparameters and optimized values. ........................................... 74
4.6 Final model performance metrics (median) ................................................ 75
4.7 Mean inter-rater agreement measures for manual WMH segmentation calculated for the available data. ................................. 79
A.1 Simulated FLAIR tissue intensities and WMH contrasts using scan parameters from the experimental database. Tissue intensities are normalized to the WM value. ........ 103
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM</td>
<td>grey matter</td>
</tr>
<tr>
<td>WM</td>
<td>white matter</td>
</tr>
<tr>
<td>CSF</td>
<td>cerebrospinal fluid</td>
</tr>
<tr>
<td>PD</td>
<td>proton density</td>
</tr>
<tr>
<td>FLAIR</td>
<td>fluid attenuation inversion recovery</td>
</tr>
<tr>
<td>WML</td>
<td>white matter lesion</td>
</tr>
<tr>
<td>WMH</td>
<td>white matter hyperintensity</td>
</tr>
<tr>
<td>DAWM</td>
<td>dirty appearing white matter</td>
</tr>
<tr>
<td>MS</td>
<td>Multiple Sclerosis</td>
</tr>
<tr>
<td>AD</td>
<td>Alzheimer’s Disease</td>
</tr>
<tr>
<td>PVE</td>
<td>partial volume effect</td>
</tr>
<tr>
<td>VLR</td>
<td>Voxel-Wise Logistic Regression</td>
</tr>
<tr>
<td>MLE</td>
<td>maximum likelihood estimation</td>
</tr>
<tr>
<td>MAP</td>
<td>maximum a posteriori</td>
</tr>
<tr>
<td>SI</td>
<td>similarity index</td>
</tr>
<tr>
<td>ICC</td>
<td>interclass correlation coefficient</td>
</tr>
<tr>
<td>LL</td>
<td>lesion load</td>
</tr>
<tr>
<td>CV</td>
<td>cross validation</td>
</tr>
<tr>
<td>LOO-CV</td>
<td>leave-one-out cross validation</td>
</tr>
<tr>
<td>KF-CV</td>
<td>k-fold cross validation</td>
</tr>
<tr>
<td>LOSO-CV</td>
<td>leave-one-source-out</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machine</td>
</tr>
<tr>
<td>K-NN</td>
<td>k-nearest neighbours</td>
</tr>
<tr>
<td>MRF</td>
<td>Markov random field</td>
</tr>
<tr>
<td>PMF</td>
<td>probability mass function</td>
</tr>
<tr>
<td>CDF</td>
<td>cumulative density function</td>
</tr>
</tbody>
</table>
### Notation

#### Variables

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>feature</td>
<td>$\in \mathbb{R}$</td>
</tr>
<tr>
<td>$\tilde{y}$</td>
<td>standardized feature</td>
<td>$\in \mathbb{R}$</td>
</tr>
<tr>
<td>$c$</td>
<td>true class</td>
<td>$\in {0, 1}$</td>
</tr>
<tr>
<td>$\hat{c}$</td>
<td>estimated class</td>
<td>$\in [0, 1]$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>logistic model feature weight</td>
<td>$\in \mathbb{R}$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>synthetic feature</td>
<td>$\in \mathbb{R}$</td>
</tr>
<tr>
<td>$\rho$</td>
<td>prior probability</td>
<td>$\in [0, 1]$</td>
</tr>
</tbody>
</table>

#### Indexing – e.g. arbitrary variable $a$

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>feature index</td>
<td>$\in {1, \ldots, K}$</td>
</tr>
<tr>
<td>$n$</td>
<td>subject index</td>
<td>$\in {1, \ldots, N}$</td>
</tr>
<tr>
<td>$t$</td>
<td>iteration index</td>
<td>$\in {1, \ldots, X}$</td>
</tr>
<tr>
<td>$x$</td>
<td>spatial location</td>
<td>$= [x_1, x_2, x_3]$</td>
</tr>
<tr>
<td>$a_{k}^{n}^{t}(x)$</td>
<td>$k^{\text{th}}$ feature; $n^{\text{th}}$ subject; $t^{\text{th}}$ iteration; location $x$</td>
<td></td>
</tr>
</tbody>
</table>

#### Images & sets – e.g. arbitrary variable $a$

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>one voxel, one feature, one subject</td>
</tr>
<tr>
<td>$a$</td>
<td>one voxel, all features, one subject</td>
</tr>
<tr>
<td>$A(x)$</td>
<td>image in native space</td>
</tr>
<tr>
<td>$A(x)$</td>
<td>image in standard space (MNI)</td>
</tr>
<tr>
<td>$A(x)$</td>
<td>image set: all features, one subject</td>
</tr>
<tr>
<td>$A(x)$</td>
<td>image set: one feature, all subjects</td>
</tr>
<tr>
<td>$A(x)$</td>
<td>image superset: all features, all subjects</td>
</tr>
<tr>
<td>$A$</td>
<td>full dataset: all features, subjects, voxels</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Digitization of medical imaging has facilitated innumerable advances in disease understanding and treatment. From multi-modal image fusion to image guided therapy, software tools now underpin research and clinical workflows in almost every domain of medical imaging.

This work concerns an unsolved segmentation problem in 3D brain magnetic resonance imaging (MRI), in which the objective is to automatically predict the class, or label, of every voxel (“volume pixel”) in the image. The objects of interest are white matter hyperintensities (WMH), non-cancerous brain lesions which are correlated with several neurodegenerative diseases. This chapter presents the motivation for automated WMH segmentation, gives a problem definition, explores the previously proposed solutions, and briefly introduces the algorithm proposed in this work.

1.1 Background

The brain is composed of three major classes of tissue: grey matter (GM), white matter (GM), and cerebrospinal fluid (CSF). Grey matter constitutes the peripheral surface of the brain – the cortex, approximately 5 mm thick – as well as some deeper structures called the basal ganglia. It contains neuronal cell bodies, and performs the bulk of neural processing. The white matter is composed primarily of myelinated axons, and functions to relay information between different GM structures in the brain. The brain is surrounded by CSF, which provides mechanical and immunological defence. It is produced by the choroid plexuses in the ventricles of the brain – a series of 4 connected cavities.
1.1.1 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) provides superior and flexible brain tissue contrast versus computed tomography (CT) imaging, and is the primary modality for imaging brain disease. Whereas CT measures tissue density via attenuation of transmitted X-rays, which does not vary significantly among brain tissues, MRI measures a mutable combination of 3 tissue characteristics: the proton density (PD),\(^1\) and T1 and T2 relaxation constants \([1]\). The physics of signal generation are described below.

In an MR scanner, a powerful magnetic field induces alignment of proton dipoles with the field. Only a tiny fraction of the total protons align, but they create a small magnetic field \(M_z\) which is distinct from the main field \([2]\). The aligned protons also rotate about the axis of alignment, imperfectly, like a spinning top; this is called precession, and the frequency of rotation is roughly homogeneous and proportional to the main field strength \([2]\). If a second magnetic field is applied which is 90° perpendicular to the first, and rotating at the precession frequency, the aligned protons can be forced into temporary alignment with this transverse rotating field, before decaying back towards their original state, as illustrated in Figure 1.1 \([2]\).

This transient applied magnetic field is induced by a radio frequency (RF) pulse, and the rate at which the original magnetization \(M_z\) is regained is described by the tissue-specific T1 relaxation constant,

\[
M_z = M_0 \left(1 - e^{-\left(\frac{t}{T_1}\right)}\right).
\]  (1.1)

The T1 constant is dictated by the ability of protons in the tissue to transfer energy to bonded atoms and surrounding molecules, since this energy transfer defines the transition from the high energy transverse state to the low energy original state \([2, 3]\). Large macromolecules, membranes, and lipids are generally able to facilitate this energy transfer more effectively than small molecules like water, producing a shorter

\(^1\) MRI can be used to image any nucleus with a net nuclear dipole, but proton (hydrogen) imaging is most common since hydrogen is biologically abundant and gives a strong signal intensity.

Figure 1.1: Visualization of T1 and T2 relaxation.
Table 1.1: T1 and T2 constants for brain tissues at 1.5 Tesla.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>T1 (ms)</th>
<th>T2 (ms)</th>
<th>$K[H]$ (a.u.)</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>WM</td>
<td>719 ± 33</td>
<td>73 ± 6</td>
<td>0.81 ± 0.03</td>
<td>[6]</td>
</tr>
<tr>
<td>GM</td>
<td>1165 ± 88</td>
<td>92 ± 11</td>
<td>0.98 ± 0.07</td>
<td>[6]</td>
</tr>
<tr>
<td>CSF</td>
<td>3337 ± 111</td>
<td>2562 ± 123</td>
<td>1.00 ± 0.07</td>
<td>[6]</td>
</tr>
<tr>
<td>WML</td>
<td>1124 ± 372</td>
<td>136 ± 79</td>
<td>−</td>
<td>[7]^a</td>
</tr>
</tbody>
</table>

^a Estimated from Fig 1 supratentorial data (numerical results not given); ± IQR, not SD; cf. § 1.1.2 for definition.

T1 [4]. For this reason, myelinated WM has a shorter T1 than GM, which in turn has a shorter T1 than CSF, which is mostly water [5].

The rate of decay of the transverse moment $M_{xy}$ is actually not equal to the rate of regeneration of $M_z$. Rather, this is governed by the T2 relaxation constant,

$$M_{xy} = M_0 \left( e^{-\left(\frac{t}{T_2}\right)} \right), \quad (1.2)$$

which is always shorter that T1. This is because, in addition to T1 effects, the net rotating moment $M_{xy}$ is eroded by proton dephasing. When precessing protons, having a net dipole, interact with other dipoles or charged particles, their rotational frequency can be increased or decreased, but overall less coherent, reducing the perceptible net magnetization $M_{xy}$ [2]. In highly structured tissues like GM and WM, these interactions are more variable, dephasing is faster, and T2 is shorter [5]. In fluid environments like CSF, proton interactions are more homogeneous, yielding longer T2 [5]. For this reason, T2-weighted images are especially useful in identifying pathologies which degrade tissue structure, since they will have abnormally high T2 [5]. Both relaxation constants depend in a small way on the main magnetic field strength, measured in Tesla (T). T1 and T2 values for various brain tissues at 1.5 T are summarized in Table 1.1.

Image acquisition involves sensing the transverse magnetization $M_{xy}$ following proton excitation by an RF pulse. The problem is that this small signal decays very quickly due to proton dephasing, which occurs even faster than $T2$ would predict due to a third factor, inhomogeneity in the main magnetic field [8]. The time constant for this decay is termed $T2^*$, and its effects are usually undesirable [8]. As a result, $M_{xy}$ is easily overpowered by the magnetic moment from the RF pulse, even after it is turned off, due to resonance. An important solution to this, called the spin-echo, was proposed by Erwin Hahn in 1950 [9]. If $T2^*$ for each proton is assumed to be constant, then reversing the direction of rotation at a time $t$ should cause all protons to align again at exactly $2t$. Therefore, at $2t$ the transverse magnetization $M_{xy}$ – the image signal – manifests again for sensing, no longer confounded by RF coil resonance [9].
This second signal is called the Spin Echo (SE), and the interval 2t is termed the echo time (TE). Reversing the direction of rotation can be achieved by a 180° RF pulse at time TE/2, in the same way the original excitation is achieved using a 90° RF pulse (amount of rotation is proportional to the energy of the pulse). Acquisition of an entire image requires repetitions of this sequence with an interval called the repetition time (TR). An example spin echo sequence, showing TE and TR, as well as T1 and T2 decay, is illustrated in Figure 1.2. Spatial encoding for creation of 2D and 3D images requires the use of additional electromagnetic gradients; however this topic is omitted here since it is quite complex, and not essential to the current work.\(^2\)

Using these principals, the nature of MR image contrast can finally be understood. That is, the signal intensity \(\Psi\) for a spin echo sequence at location \(x\) can be described by the following 3-term equation,

\[
\Psi_{\text{SE}}(x) = \left[K[H](x)\right] e^{-\left(\frac{TR}{T_2(x)}\right)} \left[1 - e^{-\left(\frac{TR}{T_1(x)}\right)}\right],
\]

where \(K\) is scaling factor, and \([H]\) denotes the proton density. If TR is chosen to be relatively long, then the longitudinal magnetization \(M_z\) is allowed to recover completely after each repetition, the third term tends towards 1 for all tissues, and differences in tissue specific T1 are nullified. Similarly, if TE is relatively short, then \(M_{xy}\) has little time to dephase, the second term is maintained close to 1, and differences in \(T_2\) are nullified. In order to emphasize differences in T1, therefore, TR can be chosen shorter; for T2-weighted contrast, TE can be chosen longer; and if differences in \([H]\) (proton density, PD) are to be emphasized, TR can be kept long and TE short. An example MRI slice using each of these image sequences is shown in Figure 1.3a, 1.3b, and 1.3c.

\(^2\) The interested reader is directed to this comprehensive resource on the topic: http://mri-q.com/
Figure 1.3: Example MRI image set with WMH pathology; from [12].

For identifying WML, T2-weighted images were conventionally used, since the lesions appear bright. However, CSF in the sulci and ventricles also appears bright on T2 images, making delineation of lesions—especially periventricular ones—difficult in T2 images (Figure 1.3b). To solve this problem, an adaptation of the spin echo RF pulse sequence can be used, called an inversion recovery (IR) [10]. In this sequence, an additional 180° inverting RF pulse is added before the 90° pulse, so that the longitudinal magnetization $M_z$ is inverted, then recovers to the original state, passing for a brief moment through zero net magnetization. The rate of recovery is governed by $T_1$, so it is tissue specific. Furthermore, if the 90° pulse is applied at the instant of zero net magnetization, no transverse moment will develop, nor the subsequent spin echo. Therefore this time interval, called the inversion time ($T_I$), can be chosen to null the signal from any tissue with a unique $T_1$. The equation governing the image signal simply adds an inversion term,

$$\Psi_{IR}(x) = K[H(x)]\left[ e^{-\left(\frac{TE}{T_2(x)}\right)} \right] \left[ 1 + e^{-\left(\frac{TR}{T_1(x)}\right)} - 2e^{-\left(\frac{T_I}{T_1(x)}\right)} \right].$$ (1.4)

This inversion principal is now often used to null the signal from CSF, especially for delineation of WMH, in a sequence called FLuid Attenuation Inversion Recovery (FLAIR) [11]. FLAIR images are usually T2-weighted. Figure 1.3d shows an example FLAIR image, where a WMH can be seen, posterior to the occipital horn of the left lateral ventricle, much more clearly than in the T2 image.

1.1.2 White Matter Disease

“Normal” ageing of the brain is characterized by a variety of physical and cognitive changes. Memory, synaptic plasticity, and brain volume decline, with observable effects on cognitive function [13, 14]. Brain ageing is also expedited in many patients by neurodegenerative diseases targeting the white matter, including Alzheimer’s disease (AD), cerebrovascular disease, and in rare cases Multiple Sclerosis (MS). While the etiologies of these diseases are not yet fully understood, there is considerable evidence to suggest
that the they are intertwined [15, 16, 17, 18].

Cerebrovascular disease describes changes to blood vessels in the brain which increase the risk of ischemic injury – a reduction in blood flow due to vessel occlusion or hemorrhage. Ischemic injuries include major events (stroke) [19], transient ischemic attacks [20], and chronic hypoperfusion due to small vessel disease [21]. In all such events, neuronal death occurs from insufficient nutrient supply [19]. Strokes involving major cerebral arteries can be fatal, and post-event quality of life in survivors is highly variable [22]. In the less dramatic courses, clinically quiet disease progression can lead to personality changes, memory loss, and reduced cognitive ability; such changes are termed vascular dementia [23].

Alzheimer’s Disease is another subclass of dementia with similar symptoms; in fact it is the most common type, affecting about 6% of the population over age 65 [24]. The cause of Alzheimer’s disease is hotly debated. Two 30-year-old theories linking the disease to the build up of amyloid \( \beta \) protein and misfolded protein \( \tau \) have been widely supported by correlational studies [25, 26, 27], but have lacked clear mechanisms of injury until recently. It is now thought that amyloid \( \beta \) oligomers interfere with neuronal mitochondria and synapse function, leading to cell death [28, 29], while aberrant \( \tau \) proteins disrupt microtubules necessary for intraneuronal transport [27]. During the search for these mechanisms, competing theories implicating vascular injury [18], immune response [17], and blood brain barrier disruption [30] have emerged, painting the picture of a more complex disease.

The pathophysiology of multiple sclerosis is similarly unclear, though genetics are a necessary factor, and it is known that symptoms arise from erosion of myelin – a fatty insulating layer surrounding axons which is critical for normal neuron firing [31]. Several theories hypothesize either that this damage is driven by autoimmune attack, followed by neuronal dysfunction and death, or that neurodegenerative changes stimulate recruitment of immune cells as part of the usual response to injury [31, 32]. Recent evidence favours the former mechanism, particularly with inflammatory injury as the initiating event [33, 34].

Connecting all these diseases are white matter lesions (WML, aka Leukoariosis), which represent the macroscopic changes to brain tissue in regions of white matter damage [15, 35, 36]. WML are very common in elderly populations, and a small volume of lesion does not necessarily implicate one of the above diseases; in one study of 1077 subjects aged 60-90, 95% had at least one WML [37]. WML appear as bright tissue regions in T2-weighted MRI due to some combination of inflammatory injury and degradation of tissue structure [35, 36]; in this imaging context, WML are often called white matter hyperintensities (WMH). Lesions are often focal, as opposed to diffuse, but there is evidence to suggest that surrounding regions of moderate hyperintensity, sometimes called “dirty appearing white matter (DAWM)”, are also related to the diseases [38]. As biomarkers of the most common WM diseases – conditions with unsolved
etiologies and inadequate treatments – WML are of special interest to many brain researchers. The next section discusses how they are used.

1.1.3 MRI in White Matter Disease

MR imaging plays important roles in diagnosis and research of white matter diseases. Typical MRI protocols include T1, T2, and FLAIR sequences, though only the latter two sequences depict WML as hyperintense [39, 40]. Depending on the disease and context, WMH can be quantified in several ways, including binary criteria (e.g. is there a lesion in a specific location) [41], rating scales (e.g. a summary of several criteria) [42], or explicit manual segmentation of the lesions by an expert [43].

WMH are arguably most important in MS. Particularly since WMH are more specific to this disease in younger patients, WMH have long been used in the diagnosis of MS, and can even be used to replace some clinical criteria, as in the 2010 McDonald Criteria [41]. MRI can also be used to discriminate between MS subtypes, which stratify disease aggression and course [41, 44, 45]. In numerous clinical trials, WMH have also been used as biomarkers of treatment efficacy [46, 47, 48], since WMH have been shown to be more sensitive to disease progression than clinical features in certain subtypes [49]. In fact, despite the central role of MRI in management and research of MS, there exists a so-called “clinico-radiological” paradox, which is the surprisingly limited correlation between WMH and clinical MS symptoms like physical and cognitive impairment [50]. However, this only strengthens the case for continued WMH research, particularly considering the recommendations by Mollison et al. in [50] to standardize image analysis in order to better understand the paradox.

In dementia (including vascular and AD), WMH are used to discriminate between disease subtypes during diagnosis. For example, the presence of at least one WML was deemed necessary for diagnosis of vascular dementia in 1993 [23], and subsequent revisions to these widely used criteria (NINCDS-ADRDA) have added this feature as an exclusionary criteria for AD [51]. While diagnosis of additional dementia subtypes may be improved using imaging [52, 53], diagnosis of the most prevalent – AD – continues to be based on clinical features alone [54]. As a result, WMH have not been used as an endpoint to any AD clinical trial. In fact, only recently have specific standards for use of WMH in vascular dementia studies been outlined [36, 40], with some subsequent uptake [55]. And yet, a 2010 meta-analysis found that WMH in brain MRI were independently correlated with stroke risk, dementia (including AD) and death [15], suggesting that much more can be done to make use of WMH as hallmarks of neurodegenerative disease.
1.2 Problem Statement

White matter hyperintensities, as ubiquitous biomarkers of several diseases with unsolved pathophysiology, are of great interest to brain researchers. Segmentation of WMH, compared to visual rating scales, provides a finer resolution for quantification of lesion load, and gives the explicit spatial distribution of pathology. This spatial information can be very useful, since diagnostic criteria often consider lesion location \[52\] and there are several correlations between lesion location and suspected etiology of WMH \[36, 56\].

Unfortunately, manual segmentation of WMH is laborious, and subject to large inter- and intra-rater variability, as reported in several works. Table 1.2 summarizes these reports, where similarity index (SI \(\in\ [0, 1]\)) is a measure of voxel-wise agreement, and interclass correlation coefficient (ICC \(\in\ [0, 1]\)) measures total volume agreement (cf. § 4.2 for definitions). Table 1.2 also gives the results using four semi-automated approaches, since these methods are reported to reduce variability and task time over strictly manual segmentation. Yet, for very large scale research studies, any approach requiring human intervention would be prohibitively time consuming and subjective.

Therefore, a fully automated algorithm to segment WMH in MRI is required. Such an algorithm would have, by construction, perfect repeatability and consistent bias – which is especially important for perceiving small changes in longitudinal studies \[57\]. Additionally, while an automated approach may not necessarily be faster than manual or semi-automatic segmentation on a per-case basis, it could be run on several computers in parallel continuously, yielding a significant overall speed up.

Furthermore, while T1, T2, and FLAIR sequences are typically recommended for both MS and dementia investigations \[39, 40, 45\], FLAIR sequences are at least as sensitive as T2 images for the detection of WML.\(^3\) As noted above, FLAIR images also have the advantage of easily distinguishing WMH from confounding CSF hyperintensity, which is important for highly prevalent periventricular lesions, and also for excluding lacunar infarcts \[60, 61\]. Consequently, it should be feasible to detect WMH using FLAIR MRI alone. This has several advantages, including minimizing the required MR sequences available during retrospective analyses, decreasing cost and scan time in prospective studies, and eliminating the need for intra-subject image registration if sequences are acquired at different resolutions (as is often the case).

---

\(^3\) Early studies exploring the utility of FLAIR sequences may contradict this claim \[58, 59\], but FLAIR imaging has since improved \[36\].
Table 1.2: Mean inter-rater agreement measures for manual and semi-automated WMH segmentation reported in previous works.

<table>
<thead>
<tr>
<th>Ref</th>
<th>Raters</th>
<th>Data</th>
<th>SI</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>[62]</td>
<td>5</td>
<td>10 images</td>
<td>0.64</td>
<td>—</td>
</tr>
<tr>
<td>[63]</td>
<td>2</td>
<td>6 images</td>
<td>0.75</td>
<td>—</td>
</tr>
<tr>
<td>[64]</td>
<td>2</td>
<td>120 slices</td>
<td>0.83</td>
<td>0.96</td>
</tr>
<tr>
<td>[43]</td>
<td>3</td>
<td>50 images</td>
<td>0.66</td>
<td>0.97</td>
</tr>
<tr>
<td>[65]</td>
<td>1</td>
<td>16 images</td>
<td>—</td>
<td>0.99</td>
</tr>
<tr>
<td>[66]</td>
<td>1</td>
<td>2 images</td>
<td>0.70</td>
<td>—</td>
</tr>
<tr>
<td>[67]</td>
<td>1</td>
<td>33 slices</td>
<td>0.78</td>
<td>—</td>
</tr>
<tr>
<td>[68]</td>
<td>2</td>
<td>30 images</td>
<td>0.78</td>
<td>—</td>
</tr>
</tbody>
</table>

1.2.1 Objective

The primary objective of this thesis is to develop an algorithm for fully automatic segmentation of WMH, using FLAIR MRI alone. Secondary objectives include:

- analysis of the limitations of prior work in this area;
- exploration and definition of appropriate cross validation techniques for the task;
- validation of the proposed algorithm on a large and heterogeneous database of FLAIR images.

1.2.2 Challenges to Automatic Segmentation

While fully automated segmentation of WMH is attractive, translation of expert knowledge into algorithmic constructs is difficult, and often requires assumptions which induce sensitivity of the model to seemingly extraneous image features. Moreover, human understanding of MR acquisition physics help radiologists to distinguish WML from image artifacts. Thus, there are several challenges to automatic segmentation. These can be summarized as follows:

1. **Noise & Partial volume effect:**

   The intensity of image voxels alone is not sufficient to determine their class; this is on account of two factors. First, the magnitude of magnetization sensed during image acquisition is extremely small. As a result, quasi-Gaussian additive noise from several sources corrupts image intensities throughout the image [69]. Second, with finite image resolution, voxels located on tissue boundaries will inevitably contain tissues of two or more tissues. This is known as partial volume effect (PVE), and the resulting signal intensity can be modelled as a linear mixture of the components [70, 71].
Niessen et al. [72] show that inadequate modelling of PVE can result in significant errors in tissue segmentation, though the widely reported 30% figure from this work is derived from unrealistic conditions.  

2. **Bias field:**

The most common image artifact in MRI is due to inhomogeneity in the main magnetic field or RF coil during acquisition, which is difficult to eliminate in strong electromagnets; this creates a low frequency variation in signal intensity over the imaged volume [73, 74]. The overall effect is that the same tissues may have different graylevels in different locations, further confounding the uniqueness of WMH graylevels [36].

3. **DAWM:**

Most of the inter-rater disagreement in manual segmentation of WMH is arguably due to ambiguity of pathological extent at the lesion borders, where the core lesion meets so-called DAWM [38]. If human judgement of this boundary is difficult, then programmatic definitions could be expected to be similarly challenged.

4. **Artifacts:**

Due to the complexity of signal acquisition, there are several artifacts which can manifest in MR images. Artifacts which appear hyperintense in T2 images (including FLAIR) are of particular importance to the current work, since these confound bright pathologies, and must therefore be excluded using other features; the most notable artifacts include [36]:

- CSF flow artifacts – ventricular hyperintensities resulting from movement of magnetically polarized CSF fluid during the inversion interval (cf. § 1.1.1) [75];
- Perivascular spaces – minuscule spaces adjacent to cerebral vessels whose properties differ from ventricular and sulcal CSF, and are therefore not attenuated in FLAIR images [36];
- Motion artifacts – artifacts which originate during frequency-domain encoding of spatial image content with subject motion, which is more common in MRI due to long acquisition times (several minutes); these typically manifest as high frequency “ringing” artifacts [76].

5. **Image variability:**

There are a large number variable characteristics of MR images; some of these can be selected at acquisition time based on time constraints, and physician preferences, while others are immutable. “Image variability” is taken to comprise:

---

Niessen et al. used morphological dilation of binary tissue masks in every direction, and compared the volumes of the resulting mask to the original. Practically, PVE modelling errors are much more likely to result in some areas of overestimation and some areas of underestimation, with overall effect closer to 4%, as the authors show.
• differences in image contrasts (and tissue graylevel distributions), due to selection of MRI parameters;
• differences in image resolution (voxel size);
• differences in MRI scanner, including field strength and proprietary image reconstruction;
• inter-subject anatomical variability and lesion heterogeneity.

Modelling this immense gamut of possible image characteristics (e.g. using parametric distributions or task-specific assumptions) represents perhaps the most challenging aspect to automated image analysis. Some specific impacts will be further discussed in § 1.3.3.

An optimal WMH segmentation algorithm will therefore consider and address each of these challenges.

1.3 Prior Work

This endeavour is far from original. Efforts to automate segmentation of WMH date back to 1990 [77], and the task has been the subject of several major reviews in 2012 [78, 79], 2013 [80], and 2015 [81]. The task has also been featured in four international competitions at the MICCAI (Medical Image Computing and Computer Assisted Intervention) Conference – 2008 [82], 2016 [83], and 2017 [12] – and the ISBI (International Symposium on Biomedical Imaging) Conference – 2015 [57] – in which researchers vie to produce the best segmentation algorithms.

This section reviews the approaches proposed in these competitions and other publications. The purpose is to provide specific criticisms of these methods, in order to motivate and provide context for the modelling decisions in the current work.

1.3.1 Segmentation Models & Features

Segmentation models represent a mapping from the content of an observed image – the features – to an image of labels or classes – in this case, tissues. The output class image comprises an estimated label for each observed voxel, or, in probabilistic models, the probability of each class for each voxel. As in many classification problems, models can be described as either supervised or unsupervised. Supervised models have relatively large capacity to model arbitrary mappings, but learn a mapping relevant to the current task using feedback from labelled examples (i.e. by a human). Unsupervised models, by contrast, are usually problem-specific, and leverage prior knowledge and the image features to predict the label image; they do not require labelled data for optimization, at least in principle. The core segmentation model
is usually constructed using prior knowledge, or wrapped in pre- and post-processing steps to create the overall algorithm.

Algorithm Types

Three general approaches have emerged for segmentation of brain MRI. The first and most popular is the pipeline, in which sub-tasks are completed in sequence, such as: pre-processing, classification, post-processing. This approach permits a flexible algorithm definition which can incorporate existing methods for individual steps. Most thresholding and classic supervised techniques are implemented in this way. The main drawback of this approach is that some steps could be improved by the results of downstream steps. For example, tissue classifying modules typically assume that the bias field is already corrected, but bias field estimation can be more accurate if the tissue segmentation is known.

The second paradigm, a unified generative model, aims to solve this chicken-egg problem. The segmentation is parameterized in one integrated probabilistic model, which often combines the input images with tissue prior probability images, a bias field model, and smoothness terms. Parameters of each sub-model are estimated using several expectation maximization (EM) iterations before the final segmentation is inferred. The challenges to this approach include balancing model complexity with estimability, robust convergence issues, and reduced ability to include external tools.

The final paradigm, deep learning, uses an optimization algorithm – e.g. stochastic gradient descent [84] – in conjunction with error back-propagation [85] to update thousands of model parameters in large cascaded layers to yield complex, relatively unstructured mappings from the input MRI to the output segmentation image. Using so-called “end-to-end” training, there is no guarantee that the usual sub-components (e.g. bias field, tissue graylevel distributions, etc.) will be estimated, though such elements could be expected to develop in the internal relationships if they are relevant to the task at hand. Most deep learning models still require standardization in space and graylevel, and more importantly, large labeled training datasets – which are rare in medical imaging.

Features

Features used in the above segmentation algorithms can be derived from individual voxels (e.g. graylevel), groups of voxels (e.g. local mean graylevel), the entire image (e.g. a histogram feature), spatial location (e.g. coordinates in a standardized space), or prior knowledge (e.g. class prior probability). It is often useful to imagine the space spanned by all possible values of all features; this is called the feature space.
Each observed voxel, having a unique value for each feature, therefore represents a unique location in this space. The task of segmentation is then to divide the feature space into subspaces corresponding to each class. In probabilistic models, these subspaces are better described as distributions of each class over the features.

Previous approaches to WMH segmentation have generally employed four types of features:

- **Graylevel**: graylevels of MRI sequences, often following standardization (e.g. T1, T2, PD, FLAIR);
- **Prior**: prior tissue probability, often derived from a coregistered prior image (e.g. ICBM [86]);
- **Spatial**: spatial location, often normalized to a common space (e.g. $x = x_1, x_2, x_3$).
- **Contextual**: local graylevel statistics or texture measures (e.g. local edge magnitude).

Additional features types are rarely used, since the combination of the above features are typically the only ones employed by human raters. At least one graylevel feature is always used, since it is the only voxel-specific information (i.e. the evidence).

### 1.3.2 Proposed Methods

For a more detailed understanding of the prior work, specific methods proposed for WMH segmentation are now reviewed. A summary of many of these works is also given in Table 1.3.

#### Thresholding Techniques

Since WMH are brighter than healthy brain tissue in FLAIR images, many unsupervised works have used thresholding of FLAIR intensities as the initial lesion segmentation. For example, in the works by Jack et al. [89], Boer et al. [63], and Smart et al. [111] optimal FLAIR thresholds are empirically estimated relative to histogram statistics, though Boer et al. use only estimated GM voxels in the histogram. Gibson et al. use a conservative FLAIR threshold initially, but then classify the remaining voxels using Fuzzy C Means clustering [106]. Samaille et al. use nonlinear diffusion filtering and watershed segmentation, before classifying candidate regions based on a FLAIR image threshold. Yoo et al. estimate the optimal threshold for FLAIR images using histogram statistics, derived from a regression model primarily considering the

---

5 This section adapted from a paper in submission [87]

6 A more detailed and interactive version of this table is available at [https://uoguelph-mlrg.github.io/vlr/wmh-table.html](https://uoguelph-mlrg.github.io/vlr/wmh-table.html)
Table 1.3: Summary of previous approaches to WMH segmentation with respect to image variability and reported performance (SI).

<table>
<thead>
<tr>
<th>#</th>
<th>Ref.</th>
<th>Year</th>
<th>Authors</th>
<th>MRI Sequences</th>
<th>U/S</th>
<th>S</th>
<th>I</th>
<th>SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[88]</td>
<td>2001</td>
<td>Van Leemput et al.</td>
<td>T1, T2, PD</td>
<td>U</td>
<td>20</td>
<td>1</td>
<td>0.51</td>
</tr>
<tr>
<td>2</td>
<td>[89]</td>
<td>2001</td>
<td>Jack et al.</td>
<td>FLAIR</td>
<td>U</td>
<td>39</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>[90]</td>
<td>2002</td>
<td>Zijdenbos et al.</td>
<td>T1, T2</td>
<td>S</td>
<td>10</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>4</td>
<td>[91]</td>
<td>2004</td>
<td>Anbeek et al.</td>
<td>T1, T2, PD, FLAIR, IR</td>
<td>S</td>
<td>20</td>
<td>1</td>
<td>0.61</td>
</tr>
<tr>
<td>5</td>
<td>[92]</td>
<td>2005</td>
<td>Anbeek et al.</td>
<td>T1, T2, PD, FLAIR, IR</td>
<td>S</td>
<td>10</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>6</td>
<td>[93]</td>
<td>2005</td>
<td>Admiraal-Behloul et al.</td>
<td>T2, PD, FLAIR</td>
<td>U</td>
<td>100</td>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>7</td>
<td>[94]</td>
<td>2006</td>
<td>Lao et al.</td>
<td>T1, T2, PD, FLAIR</td>
<td>S</td>
<td>45</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>8</td>
<td>[95]</td>
<td>2006</td>
<td>Wu et al.</td>
<td>T1, T2, PD</td>
<td>S</td>
<td>12</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>9</td>
<td>[96]</td>
<td>2006</td>
<td>Sajja et al.</td>
<td>T2, PD, FLAIR</td>
<td>S</td>
<td>23</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>10</td>
<td>[97]</td>
<td>2006</td>
<td>Harmouche et al.</td>
<td>T1, T2, PD</td>
<td>U</td>
<td>10</td>
<td>1</td>
<td>0.61</td>
</tr>
<tr>
<td>11</td>
<td>[98]</td>
<td>2006</td>
<td>Khayati et al.</td>
<td>FLAIR</td>
<td>U</td>
<td>20</td>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>12</td>
<td>[99]</td>
<td>2006</td>
<td>Herskovits et al.</td>
<td>T1, T2, PD, FLAIR</td>
<td>S</td>
<td>42</td>
<td>2</td>
<td>0.6</td>
</tr>
<tr>
<td>13</td>
<td>[100]</td>
<td>2007</td>
<td>Bricq et al.</td>
<td>T2, FLAIR</td>
<td>U</td>
<td>25</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>14</td>
<td>[101]</td>
<td>2007</td>
<td>Dyrby et al.</td>
<td>T1, T2, FLAIR</td>
<td>S</td>
<td>362</td>
<td>10</td>
<td>0.56</td>
</tr>
<tr>
<td>15</td>
<td>[102]</td>
<td>2007</td>
<td>Souplet et al.</td>
<td>T1, T2, FLAIR</td>
<td>U</td>
<td>25</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>16</td>
<td>[103]</td>
<td>2007</td>
<td>Boer et al.</td>
<td>T1, PD, FLAIR</td>
<td>S</td>
<td>20</td>
<td>2</td>
<td>0.72</td>
</tr>
<tr>
<td>17</td>
<td>[104]</td>
<td>2007</td>
<td>García-Lorenzo et al.</td>
<td>T1, T2, PD</td>
<td>S</td>
<td>10</td>
<td>1</td>
<td>0.63</td>
</tr>
<tr>
<td>18</td>
<td>[105]</td>
<td>2007</td>
<td>Akselrod-Ballin et al.</td>
<td>T1, T2, PD, FLAIR</td>
<td>S</td>
<td>41</td>
<td>1</td>
<td>0.53</td>
</tr>
<tr>
<td>19</td>
<td>[106]</td>
<td>2010</td>
<td>Schwarz et al.</td>
<td>T1, T2, PD</td>
<td>S</td>
<td>165</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>20</td>
<td>[107]</td>
<td>2010</td>
<td>Gibson et al.</td>
<td>T1, T2, FLAIR</td>
<td>U</td>
<td>18</td>
<td>1</td>
<td>0.81</td>
</tr>
<tr>
<td>21</td>
<td>[108]</td>
<td>2010</td>
<td>Shice et al.</td>
<td>T1, FLAIR</td>
<td>U</td>
<td>10</td>
<td>1</td>
<td>0.63</td>
</tr>
<tr>
<td>22</td>
<td>[109]</td>
<td>2010</td>
<td>Scully et al.</td>
<td>T1, T2, FLAIR</td>
<td>S</td>
<td>17</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>23</td>
<td>[110]</td>
<td>2011</td>
<td>García-Lorenzo et al.</td>
<td>T1, T2, FLAIR</td>
<td>U</td>
<td>10</td>
<td>1</td>
<td>0.65</td>
</tr>
<tr>
<td>24</td>
<td>[111]</td>
<td>2011</td>
<td>Geremia et al.</td>
<td>T1, T2, FLAIR</td>
<td>S</td>
<td>20</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>25</td>
<td>[112]</td>
<td>2011</td>
<td>Smart et al.</td>
<td>T1, FLAIR</td>
<td>U</td>
<td>30</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>26</td>
<td>[113]</td>
<td>2012</td>
<td>Samaille et al.</td>
<td>T1, FLAIR</td>
<td>U</td>
<td>67</td>
<td>6</td>
<td>0.72</td>
</tr>
<tr>
<td>27</td>
<td>[114]</td>
<td>2012</td>
<td>Schmidt et al.</td>
<td>T1, FLAIR</td>
<td>U</td>
<td>53</td>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>28</td>
<td>[115]</td>
<td>2012</td>
<td>Khademi et al.</td>
<td>FLAIR</td>
<td>U</td>
<td>24</td>
<td>1</td>
<td>0.83</td>
</tr>
<tr>
<td>29</td>
<td>[116]</td>
<td>2012</td>
<td>Abdullah et al.</td>
<td>T1, T2, FLAIR</td>
<td>S</td>
<td>61</td>
<td>3</td>
<td>—</td>
</tr>
<tr>
<td>30</td>
<td>[117]</td>
<td>2013</td>
<td>Sweeney et al.</td>
<td>T1, T2, PD, FLAIR</td>
<td>S</td>
<td>111</td>
<td>1</td>
<td>0.61</td>
</tr>
<tr>
<td>31</td>
<td>[118]</td>
<td>2013</td>
<td>Datta et al.</td>
<td>T1, T2, FLAIR</td>
<td>S</td>
<td>90</td>
<td>3</td>
<td>—</td>
</tr>
<tr>
<td>32</td>
<td>[119]</td>
<td>2013</td>
<td>Steenwijk et al.</td>
<td>T1, FLAIR</td>
<td>S</td>
<td>40</td>
<td>2</td>
<td>0.8</td>
</tr>
<tr>
<td>33</td>
<td>[120]</td>
<td>2014</td>
<td>Khademi et al.</td>
<td>FLAIR</td>
<td>U</td>
<td>25</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>34</td>
<td>[121]</td>
<td>2014</td>
<td>Ithapu et al.</td>
<td>T1, FLAIR</td>
<td>S</td>
<td>38</td>
<td>1</td>
<td>0.67</td>
</tr>
<tr>
<td>35</td>
<td>[122]</td>
<td>2014</td>
<td>Schmidt et al.</td>
<td>FLAIR</td>
<td>S</td>
<td>32</td>
<td>2</td>
<td>0.76</td>
</tr>
<tr>
<td>36</td>
<td>[123]</td>
<td>2015</td>
<td>Harmouche et al.</td>
<td>T1, T2, PD, FLAIR</td>
<td>U</td>
<td>100</td>
<td>35</td>
<td>0.56</td>
</tr>
<tr>
<td>37</td>
<td>[124]</td>
<td>2015</td>
<td>Guizard et al.</td>
<td>T1, T2, PD, FLAIR</td>
<td>S</td>
<td>108</td>
<td>32</td>
<td>0.6</td>
</tr>
<tr>
<td>38</td>
<td>[125]</td>
<td>2015</td>
<td>Jain et al.</td>
<td>T1, FLAIR</td>
<td>U</td>
<td>20</td>
<td>1</td>
<td>0.67</td>
</tr>
<tr>
<td>39</td>
<td>[126]</td>
<td>2015</td>
<td>Tomasc-Fernandez et al.</td>
<td>T1, T2, FLAIR</td>
<td>U</td>
<td>51</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>40</td>
<td>[127]</td>
<td>2015</td>
<td>Wang et al.</td>
<td>T1, T2, FLAIR</td>
<td>U</td>
<td>70</td>
<td>2</td>
<td>0.84</td>
</tr>
<tr>
<td>41</td>
<td>[128]</td>
<td>2015</td>
<td>Roy et al.</td>
<td>FLAIR</td>
<td>S</td>
<td>38</td>
<td>3</td>
<td>0.56</td>
</tr>
<tr>
<td>42</td>
<td>[129]</td>
<td>2015</td>
<td>Brosch et al.</td>
<td>T1, T2, FLAIR</td>
<td>S</td>
<td>20</td>
<td>2</td>
<td>0.36</td>
</tr>
<tr>
<td>43</td>
<td>[130]</td>
<td>2015</td>
<td>Partaria et al.</td>
<td>FLAIR</td>
<td>S</td>
<td>39</td>
<td>1</td>
<td>0.55</td>
</tr>
<tr>
<td>44</td>
<td>[131]</td>
<td>2015</td>
<td>Deshpanse et al.</td>
<td>T1, T2, PD, FLAIR</td>
<td>S</td>
<td>52</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>45</td>
<td>[132]</td>
<td>2015</td>
<td>Roura et al.</td>
<td>T1, FLAIR</td>
<td>U</td>
<td>20</td>
<td>2</td>
<td>0.34</td>
</tr>
<tr>
<td>46</td>
<td>[133]</td>
<td>2016</td>
<td>Knight et al.</td>
<td>FLAIR</td>
<td>U</td>
<td>15</td>
<td>3</td>
<td>0.7</td>
</tr>
<tr>
<td>47</td>
<td>[134]</td>
<td>2016</td>
<td>Mechrez et al.</td>
<td>T1, T2, FLAIR</td>
<td>S</td>
<td>20</td>
<td>2</td>
<td>0.31</td>
</tr>
<tr>
<td>48</td>
<td>[135]</td>
<td>2016</td>
<td>Strumia et al.</td>
<td>T1, FLAIR</td>
<td>U</td>
<td>20</td>
<td>3</td>
<td>0.52</td>
</tr>
<tr>
<td>49</td>
<td>[136]</td>
<td>2016</td>
<td>Griffanti et al.</td>
<td>T1, FLAIR</td>
<td>S</td>
<td>130</td>
<td>2</td>
<td>0.76</td>
</tr>
<tr>
<td>50</td>
<td>[137]</td>
<td>2016</td>
<td>Brosch et al.</td>
<td>T1, T2, PD, FLAIR</td>
<td>S</td>
<td>77</td>
<td>67</td>
<td>0.64</td>
</tr>
<tr>
<td>51</td>
<td>[138]</td>
<td>2017</td>
<td>Valverde et al.</td>
<td>T1, FLAIR</td>
<td>U</td>
<td>33</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>52</td>
<td>[139]</td>
<td>2017</td>
<td>Dadar et al.</td>
<td>T1, FLAIR</td>
<td>S</td>
<td>80</td>
<td>3</td>
<td>0.62</td>
</tr>
<tr>
<td>53</td>
<td>[140]</td>
<td>2017</td>
<td>Zhan et al.</td>
<td>T1, T2, FLAIR</td>
<td>S</td>
<td>50</td>
<td>2</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Abbreviations. I: number of MR image sets used for validation; S: number of MRI scanners used for validation; SI: reported validation similarity index; U/S: unsupervised/supervised.
total lesion load [119]. In works by Khademi et al., a peak in the conditional probability of edge content on graylevel is used to model partial volume averaging for unsupervised WML segmentation in FLAIR MRI for subjects with ischemic and MS diseases [71, 130, 138].

Mixture Models

Most other unsupervised approaches are probabilistic models, often framed as a mixture model. The work by Van Leemput et al. [88] uses a similar framework as the early work by Ashburner and Friston [139], later incorporated into the SPM Segment tool [140], which jointly estimates Gaussian graylevel distributions for each tissue class, and also bias field, using expectation maximization. In the model by Van Leemput et al., distribution parameters are estimated using outlier-insensitive estimators, and WMH are derived from model outliers using heuristic rules. The predicted classes are also smoothed spatially using a Markov Random Field (MRF).

Similar works by Bricq et al. [100], Schmidt et al. [114], Jain et al. [122], and Roura et al. [129] use parametric mixture models to predict WMH as model outliers, and all but [129] embed the model in a MRF. Khayati et al. [97] and Subbanna et al. [141] also use MRF-constrained mixture models, but model WMHs as a Gaussian-distributed tissue class, rather than as outliers. In the works by Harmouche et al., parametric distributions are also used to model lesions, but such distributions are parameterized independently per brain region, in order to reflect lobe heterogeneity; a MRF is again used for regularization [62, 120]. Schwarz et al. again employ a Bayesian MRF model, but use lognormal distributions for WM and WMH [105]. Souplet et al. use an augmented mixture model which includes partial volume averaging classes and an outlier class to perform initial brain tissue segmentation; WMH are subsequently classified using a FLAIR intensity threshold after contrast enhancement [102]. The work by Herskovits et al. is much the same, but uses statistical information from training data to classify lesions (i.e. it is supervised) [99]. More recently, Graph-Cuts have been used in conjunction with mixture models, as in the works by García-Lorenzo et al. [103], Tomas-Fernandez and Warfield [123], and Strumia et al. [132].

The Lesion-TOADS method by Shiee et al. [107], a lesion-specific adaptation of the TOADS algorithm [142], presents an entirely new non-Gaussian paradigm for modelling class distributions, and incorporates topological energies in the objective function. Other proposed unsupervised methods have used clustering by Fuzzy C-Means, including the works by Admiraal-Behloul et al. [93], Gibson et al. [106], and Valverde et al. [135].
Classic Supervised Methods

Many early supervised methods used K-Nearest Neighbours (K-NN) for voxel-wise WMH classification. Anbeek et al. used a K-NN model with features derived from spatial coordinates and voxel intensities from several modalities [91, 92]. In the works by Wu et al. [95], Steenwijk et al. [64], and Fartaria et al. [127], spatial coordinates are substituted for tissue priors as K-NN features. In the recently proposed BIANCA algorithm by Griffanti et al. [133], spatial coordinates are added back, along with some patch-based features.

Other works have also explored Support Vector Machines (SVM) for classification. The works by Lao et al. [94], Abdullah et al. [115], and Scully et al. [108] each use a selection of intensity features, neighbouring intensities, tissue priors, morphological, and texture features with an SVM classifier. Several more recent works have used decision tree-based classifiers, including Random Forest (RF) and AdaBoost. Akselrod-Ballin et al. [104] employ over 30 features for multi-scale image representation and classify voxels using RF. Both Geremia et al. [110] and Roy et al. [125] use a combination of intensity and tissue prior features to train a RF classifier, whereas Wels et al. [98] use a large number of Haar-like features to train an AdaBoost model. Ithapu et al. [118] explore the use of texton features in both SVM and RF models.

Logistic regression models have also gained popularity recently. In the OASIS model by Sweeney et al. [116], image intensities from T1, T2, PD, and FLAIR sequences are used individually, in multiplicative combination, and with Gaussian blurring as predictors for a global set of logistic regression parameters. In the work by Zhan et al. [137], a similar logistic model is fitted using only the raw T1, T2, and FLAIR intensities, while bias correction is performed as preprocessing and spatial smoothness using MRF post processing. In the work by Dadar et al. [136], spatial and intensity features from a flexible selection of MR sequences are used to train a linear regression model, the results of which are thresholded to give the lesion prediction. Still more works have proposed other supervised models, including nonparametric Parzen classifiers [96].

Deep Learning

A number of deep learning approaches have also been proposed, though their permeation in this problem space has been surprisingly limited until recently\(^7\). Both Zijdenbos et al. [90] and Dyrby et al. [101] train fully-connected voxel-wise Neural Networks with a selection of intensity, spatial, and tissue prior

\(^7\) The 2017 WMH Segmentation Competition, saw a massive increase, however, with 15/20 submitted methods using deep learning; cf. § 4.9.2 for more information.
features to predict the lesion class. In contrast, Brosch et al. [126, 134] construct a more modern deep convolutional model, which is capable of capturing both local and global dependencies.

### External Toolboxes

Many of the proposed methods use registration, brain extraction, bias field correction, and segmentation tools available in freely available toolkits; these include the SPM\(^8\) toolkit [96, 101, 104, 111, 114, 119, 118, 125, 135] and the FSL\(^9\) toolkit [64, 99, 106, 116, 117, 124, 125, 133, 137], as well as bias correction by the N3/4\(^10\) [143] algorithm [62, 90, 120, 121, 127, 131, 135, 136, 137].

### 1.3.3 Limitations

Despite over 50 proposed algorithms and several competitions, no WMH segmentation algorithm has clearly emerged the superior method,\(^11\) nor has any been taken up for use in the wider research community. This is contrasted with other neuroimaging tasks, where several robust tools noted above are now regularly used in analysis pipelines – e.g. N3/4 [143] for bias field correction, BET [144] for brain extraction, SPM Segment [140] / FSL FAST [145] for healthy brain segmentation, SPM Normalize [140] / FSL FNIRT [146] for registration.

Some general reasons for the lack of confidence placed in previous approaches will be explored in the next section. Then, specific limitations of the more promising models, on which this work is based, will also be discussed.

### Confidence Factors

In general, two hypotheses help explain the gap in widely used WMH segmentation tools.

First, very few of the proposed methods have been released as either open-source code or compiled applications. Researchers may not want to release source code for reasons related to intellectual property, or the additional work of ensuring robustness and writing documentation. Yet the field of deep learning illustrates how these practices can accelerate progress in the field enormously. Similarly, compiling applications for cross-platform compatibility is no small feat, though there are many examples for SPM

---

\(^8\) [http://www.fil.ion.ucl.ac.uk/spm/](http://www.fil.ion.ucl.ac.uk/spm/)

\(^9\) [https://fsl.fmrib.ox.ac.uk/fsl/](https://fsl.fmrib.ox.ac.uk/fsl/)

\(^10\) [https://www.slicer.org/wiki/Documentation/4.6/Modules/N4ITKBiasFieldCorrection](https://www.slicer.org/wiki/Documentation/4.6/Modules/N4ITKBiasFieldCorrection)

\(^11\) Things may have changed recently, cf. § 4.9.2.
extensions,\textsuperscript{12} as well as events by NA-MIC (National Alliance for Medical Image Computing) for development of 3D Slicer modules.\textsuperscript{13}

Second, very few of the WMH segmentation methods have been validated on large, multi-centre databases: of the 54 works reviewed (Table 1.3), less than half use more than one scanner for validation, and only 5 use more than three.\textsuperscript{14} As noted in § 1.2.2, there are several sources of image variability in MRI, and both supervised and unsupervised methods can be sensitive to these factors, as noted by several authors [78, 116, 136]. Therefore, while many of the proposed methods may be of use for in-house work (i.e. with images from a consistent source), there can be little confidence that they will perform as reported on data from new sources (generalization performance).

In supervised models, graylevel features must be standardized, since the MRI intensity scale is not consistent across scanners or scan parameters, due to the complexity of signal acquisition [147]. However, this is not an easy task. For example, Steenwijk et al. validate a supervised WMH segmentation algorithm using same-scanner training and testing for two different scanners independently (mean SI = 0.75, 0.84), after variance scaling of intensity features [64]. Yet, a follow-up experiment which saw the method trained on one scanner and tested on the other showed a precipitous drop in performance to mean SI = 0.50.

Unsupervised models also have parameters which can be inadvertently over-tuned to data from one or two sources. For example, mixture models which classify lesions as outliers often employ an outlier definition which depends on mixture model parameters (e.g. tissue graylevel mean and variance), which in turn are subject to MR slice thickness, noise level, and contrast [88, 102, 109, 129]. Graylevel thresholding techniques [71, 89, 111, 112, 114] are similarly affected by changes in image properties.

It is worth noting four works\textsuperscript{15} which run counter to this trend, demonstrating robust validation of their proposed methods. These works are summarized in Table 1.4. Perhaps not surprisingly, these papers report lower performance (mean SI \( \leq 0.64 \)) than other works; as a result, these algorithms would not likely be used for clinical research. Yet, even these works do not optimally estimate the model generalization performance for data from new scanners, as will be discussed in § 4.3. Moreover, the proposed algorithms in these papers require at least three MRI sequences, which fails the objectives of the current work.

\textsuperscript{12}http://www.fil.ion.ucl.ac.uk/spm/ext/

\textsuperscript{13}https://na-mic.org/wiki/Events

\textsuperscript{14}In some works, the algorithm is validated separately on two

\textsuperscript{15}The work by Samaille et al. (2012) [112] is also a good candidate, having used 6 scanners for validation; however, 43 of the 67 images (64\%) come only from one scanner, reducing the robustness of generalization results.
Table 1.4: Works demonstrating excellent validation of a WMH segmentation algorithm.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Year</th>
<th>Authors</th>
<th>I</th>
<th>S</th>
<th>SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>[101]</td>
<td>2008</td>
<td>Dyrby et al.</td>
<td>362</td>
<td>10</td>
<td>0.56</td>
</tr>
<tr>
<td>[121]</td>
<td>2015</td>
<td>Guizard et al.</td>
<td>108</td>
<td>32</td>
<td>0.60</td>
</tr>
<tr>
<td>[120]</td>
<td>2015</td>
<td>Harmouche et al.</td>
<td>100</td>
<td>35</td>
<td>0.56</td>
</tr>
<tr>
<td>[134]</td>
<td>2016</td>
<td>Brosch et al.</td>
<td>77</td>
<td>67</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Abbreviations. I: number of MR image sets used for validation; S: number of MRI scanner-parameter combinations used for validation; SI: reported validation similarity index.

FLAIR-Only Methods

Since the current work aims to develop a FLAIR-only segmentation method, special attention is given to the limitations of these approaches.

The majority of FLAIR-only WMH segmentation algorithms use a thresholding technique, mapping a single graylevel feature directly to a class or class probability (e.g. “healthy” or “lesion”). Such models often have complex methods of deriving this mapping (e.g. using mixture model parameters [129], histogram features [119], or conditional edge probability [148]), but the final rule is applied equally to the entire graylevel image. The most common challenge for these methods is a high number of false positives (cf. § 4.2 for definitions), since several artifacts and GM can overlap the WMH intensity distribution.

Preliminary investigations sought to characterize the spatial distribution of these common errors in order to understand the limitations of thresholding methods. Using a database of 96 FLAIR images and binary manual WMH segmentations, where 0 = healthy, and 1 = lesion (cf. § 4.1 for details), the optimal threshold for each image was calculated.\footnote{The \texttt{fminsearch} function in \textsc{matlab} was used [149].} The optimization maximized the Similarity Index (cf. § 4.2) between the thresholded FLAIR image and the manual segmentation. The resulting segmentations were spatially transformed (cf. § 2.1) to MNI space, and the average distribution of true positives (TP), false positives (FP), and false negatives (FN) were computed. These results shown in Figure 1.4, and for reference, the median optimal similarity index (SI) was 0.36.

The large proportion of FP and FN in these results suggests that even if an optimal threshold could be estimated for these data, the agreement with manual segmentation would be poor. For this reason, graylevels alone could not give a good estimation of the lesion segmentation, and some additional features should be used. It can be seen in Figure 1.4 that there are regular and distinct spatial distributions of the FP and FN errors. Moreover, it has been suggested that there is regional heterogeneity in relaxation...
Figure 1.4: Distributions of TP, FP, and FN in 96 FLAIR MRI, following supervised optimal thresholding in MNI space. Best viewed in colour.

rates of brain tissues [150], and that WML intensity depends in part on location [7, 120]. This implies that spatial coordinates could be helpful additional features used for this task, especially when incorporated into the main classification model (i.e. not as post-processing, as in so-called “false positive reduction”).

Spatial features have been used with FLAIR intensities in supervised classification models like K-NN, SVM, RF, etc. [91, 92, 101, 133, 136]. However, such models treat all features equally, which can lead to artifacts in the feature space decision boundary that contradict prior knowledge. For example, in spatial locations which do not observe any lesions during training, the optimized model may learn to never predict the lesion class, regardless of graylevel features. Similarly, there are several factors challenging the assumption that FLAIR graylevels will map monotonically to the lesion class. Therefore, spatial features should be treated differently. A model to do so will be developed throughout the remainder of this work.

Classic Logistic Regression Models

Logistic regression models have more recently gained popularity for WMH segmentation [116, 137, 151, 152], and have several advantages. First, model parameters are generally more interpretable than those in
other models, permitting better design of regularizations. Second, the simplicity of the model reduces its
capacity for over-fitting. Third, model foundations in statistical theory allow probabilistic interpretations
of the outputs, which may be helpful for quantifying marginally pathological tissues like DAWM.

Let \( c = 1 \) denote the lesion class, and \( c = 0 \) denote the healthy class. In the classic logistic model,
the probability of the lesion class, given a set of features \( y = [1, y^1, \ldots, y^k]^T \), is modelled by a logistic
function, parameterized by a vector of feature weights \( \beta = [\beta^0, \beta^1, \ldots, \beta^k]^T \),
\[
P(c = 1 \mid y; \beta) = \frac{1}{1 + e^{-\eta}}, \quad \eta = \beta^T y. \tag{1.5}
\]

This probability – the estimated lesion class probability – is denoted \( \hat{c} = P(c = 1 \mid y; \beta) \in [0, 1] \). Figure 1.5
shows an example class prediction for a single arbitrary input feature \( y_i \). Considering the spatial location
\( x = [x_1, x_2, x_3] \), the estimated label image becomes \( \hat{C}(x) = P(C(x) = 1 \mid Y(x); \beta) \).

In the OASIS algorithm \[116, 151\] and the algorithm by Zhan et al. \[137\], this model is used with a
number of different graylevel features. However, this ignores spatial location\(^{17}\) and therefore makes two
assumptions: first, that graylevels are monotonically related to the lesion class – i.e. that WMH are
either the brightest or darkest class in the image – and second, that the distribution of graylevels alone is
sufficient to discriminate classes. From the graylevel modelling results in § A.1, it can be shown that typical
selections of T1 and T2 imaging parameters do not create monotonic relationships between graylevel and
class label, contrast between GM and WMH graylevel distributions is often only around 40%, even in
FLAIR images. Moreover, these results are derived from ideal conditions; considering image noise, PVE,

\(^{17}\)The OASIS model also uses Gaussian-blurred images as features, which could add some spatial information, but this is
different from an explicit global context like \( x \).
imperfect bias field correction, and possible tissue heterogeneity, the plausibility of class separability by
graylevel alone is further diminished. Again, the potential utility of spatial features is highlighted.

Lesion Prediction Algorithm

An alternative approach by Schmidt aims to solve this issue by introducing a spatial effect parameter,
namely the intercept $\beta^0 \rightarrow \beta^0(x)$, which is defined uniquely for each location - i.e. voxel $x$ - in the imaged
volume. In this method, a Gaussian MRF model is used to estimate the spatial parameter $\beta^0(x)$, while
the other $\beta$ parameters - in fact there is only one: $\beta^1$, corresponding to the FLAIR graylevel - are fixed
for the entire image. A pre-trained version of this method was released as the Lesion Prediction Algorithm
(LPA) in the LST toolbox.\(^\text{18}\)

This particular parametrization has significant implications for model estimation, however, since $\beta^0(x)$
should be estimated uniquely for every spatial location, but $\beta^1$ should consider evidence from the entire
image volume. Efficient fitting of such models was the subject of major works by Schmidt et al. [152, 153], but several drawbacks remain. First, Markov Chain Monte Carlo estimation of the model appears
to create discontinuity artifacts in the spatial effect image $\beta^0(x)$ (cf. Figure 4.20, § 4.7.3). Second, MRF
modelling of the parameter images assumes that the missing data (i.e. WMH training examples in the
more superficial brain regions) can be interpolated spatially, but this may not be justified. Third, this
joint estimation procedure is computationally expensive (versus the methods proposed here), requiring
approximately two hours to estimate $\beta$ to only about 90% convergence [152]. Finally, it is not clear
whether any tied $\beta$ are necessary or advantageous in this context. These deficiencies then motivate
investigations into alternative solutions to the above challenges.

In addition to these potential modelling weaknesses, there were also several limitations to the validation
methodology for the LPA algorithm worth noting here. First, the “ground truth” segmentations were
generated using an automated algorithm – the Lesion Growth Algorithm (LGA) [114] of the same toolbox –
rather than a human expert. Second, the graylevel standardization procedure employed does not consider
the variance of image graylevels (only the mean is subtracted); this strongly assumes that user images
will have graylevels spanning a similar range. Third, all 53 training cases were obtained on the same MRI
scanner, which may limit generalization performance. Finally, no segmentation performance results are
given in either of the associated publications [152, 153]. Therefore, while the open-source release of the
LPA algorithm is greatly appreciated, significant improvements can be made to this algorithm.

\(^\text{18}\)\url{http://www.applied-statistics.de/lst.html}
Figure 1.6: Overview of the necessary processing steps.

1.4 Proposed Algorithm

This section presents a brief overview of the proposed WMH segmentation method.

For several reasons, the supervised pipeline approach was selected as the framework for this algorithm. First, preliminary work drawing on existing algorithms [71, 154] showed the feasibility of several relatively simple FLAIR-only methods, which confer robustness and interpretability through simplicity. Second, the parametric assumptions required by unified probabilistic models may be challenged by data from multiple sources [88]. Third, only a small number of training cases were available during initial development, which limited the feasibility of deep learning approaches. Finally, time constraints favoured the incorporation of existing tools to address challenges like bias correction and image registration, which would have been otherwise difficult to develop in a unified model.

The proposed pipeline can be summarized as follows. Pre-processing steps will aim to correct any bias field effect, standardize spatial coordinates, and also image graylevels, since the classification model assumes these features are drawn from a consistent distribution. The classification model will then employ the standardized FLAIR intensities and spatial features to give the initial segmentation. Finally, post-processing steps will generally aim to further improve segmentation performance. This pipeline is illustrated in Figure 1.6. Next, the proposed classification model is introduced.

1.4.1 Voxel-Wise Logistic Regression

The classification model proposed here is similar to the LPA model. However, several modifications are made to address the challenges outlined in § 1.3.3. Most significantly, spatial variation of all logistic parameters is now permitted, yielding a separate logistic regression model for each voxel – i.e. “Voxel-wise Logistic Regression” (VLR). Mathematically, this is

$$ P(c(x) = 1 \mid y(x); \beta(x)) = \frac{1}{1 + e^{-\eta(x)}}, \quad \eta = \beta(x)^T Y(x). $$

(1.6)

Training the VLR model then yields a complete and unique vector \( \beta \) for each voxel \( x \), or equivalently, one complete image for each parameter. Only one additional \( \beta \) is used here, again corresponding to the
FLAIR graylevel. This formulation allows completely independent estimation of the logistic model for each voxel \( x \), facilitating improved estimates. In turn, this permits essentially complete convergence in significantly less time, and does not require sampling approximations or smoothness assumptions (though methods of enforcing smoothness post hoc will be explored).

Overall, the VLR model solves the problem of unreliable separability by graylevel alone, and presents a method for differential treatment of spatial and graylevel features (versus K-NN, etc.). That is, the characteristics of the logistic regression model are maintained with respect to graylevel features, but spatial features can have more complex relationships (non-monotonic) with the output. The estimated VLR parameters are also highly interpretable, allowing prior knowledge to guide improved regularizations versus those used in the LPA algorithm. Different pre- and post-processing methods versus the LPA algorithm are also developed.

### 1.5 Contributions

This thesis aims to produce a WMH segmentation algorithm which can be used on FLAIR MRI from any source, and to characterize the expected performance on unseen data. The major contributions are as follows:

1. A review and critique of the previously proposed WMH segmentation algorithms, especially with respect to expected performance on unseen data;

2. Voxel-Wise Logistic Regression (VLR): a new FLAIR-only WMH segmentation algorithm;

3. Leave-One-Source-Out Cross Validation (LOSO-CV): a validation framework which accurately characterizes the generalization performance of medical image analysis methods;

4. Extensive validation of the proposed method and its components.

The remainder of this thesis is organized as follows: Chapter 2 explores the pre-processing steps required to satisfy the assumptions of the VLR model. Chapter 3 develops the voxel-wise logistic regression model, including expected challenges and regularization solutions with this approach; Chapter 4 explores optimization of model components through experimentation, and then presents segmentation performance results under various cross validation schemes; Chapter 5 draws conclusions about the work, and highlights avenues for future investigation.
Chapter 2

Pre-Processing

The proposed VLR classification model addresses several of the challenges outlined in § 1.2.2. The problem of overlapping tissue graylevel distributions is mostly solved through expansion of the feature space to include spatial features. CSF flow-through artifacts, which appear in roughly consistent locations, are similarly managed. Heterogeneity in the appearance of lesions is also considered by the spatial parametrization of logistic parameters. Finally, ambiguity regarding moderately hyperintense DAWM, and voxels affected by partial volume effect, is captured in the probabilistic output.

Several challenges, however, still remain. In particular, a number of assumptions were made about the input data for the VLR model which are likely invalid for raw images. These assumptions are that: 1. input MRI images are free of bias field artifact; 2. feature intensities are consistent across different subjects; 3. images are consistently sized and voxels represent the same anatomical regions across different subjects. Solving these challenges must therefore be accomplished by one or more pre-processing steps. This section explores these steps.

2.1 Registration

Image registration is the process of geometrically transforming a source image so that the image content is aligned per-voxel with a target image of the same subject. This process facilitates voxel-wise analysis of MRI from different subjects, such as “voxel-based morphometry” [155] and analysis of functional MRI data [156].\(^1\) Source images from multiple subjects are usually registered to the same target image; in

\(^1\) Incidentally, investigation of these topics were the motivations for developing of the SPM and FSL software packages, respectively.
In this context, it is useful to define a “native space” and a “standardized space”, denoting the original, subject-specific geometry, and the standardized target geometry. By convention, the target brain space is usually either the Talairach space [157] or the Montreal Neurological Institute (MNI) space [158], though any reasonable target image could be used. Figure 2.1 shows three FLAIR images before and after registration, with cross-hair shown to highlight differences in alignment corrected by the transformation.

Parameters defining registration transforms are fit by maximizing some measure of overlap between the images [159]. Simple registration methods employ only affine transformations, comprising a combination of translation, scaling, and shear transformations (e.g. the FLIRT tool [160] in FSL). The utility of these methods for neuroimage analysis is limited, since there are often significant differences in brain anatomy between subjects. Rather, most tools parametrize a spatial warping model, permitting local nonlinear deformations to better match brain structures (e.g. cubic B-splines in FSL FNIRT [146], discrete cosine transform in SPM Normalize [139, 140], a general diffeomorphism in the SyN algorithm [161]). While these models are more difficult to fit, several robust algorithms have been released for general use, with
widespread acceptance (e.g. the toolboxes mentioned above). A 2009 comparison of 14 different methods is a good resource on the subject [162], while a more recent (2014) study ranks the popular toolboxes SPM, FSL, and Brainsuite [163].

In the current work, registration is required during both training at testing. During training, the model parameters $\beta(x)$ are estimated using mutually aligned segmentation examples $\{Y, C\}$ in a standard brain space. At test time (or for actual use), these parameter images are transformed in the opposite direction from the standard space to the native space of the current subject. Since registration transforms are typically bijections – i.e. invertible – the second registration case can be estimated using the same method as the first, minimizing bias.

Unlike other applications, it is not essential that perfect image registration is achieved here. As noted by Harmouche et al. [120], in smoothly varying models, small registration errors can be expected to have a negligible impact. Therefore, registration was not a primary focus of optimization in this work.

While the registration component of the SPM8 Segment feature [140] was not among the top performing in the 2009 study [162], subsequent implementation revisions in SPM12\(^2\) called New Segment have apparently improved results. In the 2014 study [163], the SPM12 New Segment method achieved the highest Similarity Index of all methods on real (IBSR [164]) data.

Furthermore, the SPM module has several other features amenable to this work. First, unlike many other registration algorithms, the objective function does not require source and target images to have the same contrast. This is helpful, since no suitable FLAIR template image is available for use as a target.\(^3\) Second, it is simple to invoke the SPM modules via the command line or MATLAB scripts; this facilitates smooth integration of this tool in the pipeline. Third, it is possible to save previously estimated registration transformations, and apply them in the forward or reverse directions efficiently. This can save significant time during cross validation, since the estimated parameter images $\beta(x)$ must eventually be transformed to the native space of every subject. Finally, the SPM New Segment model additionally estimates the bias field during execution with high accuracy, saving an additional step (cf. § 2.2, below). For all these reasons, the registration performed by SPM New Segment was used throughout this work.

After satisfactory visual inspection of all training set images, no other registration tools were investigated. The default brain space of this module is MNI.

The algorithm underlying New Segment combines several models in a unified probabilistic framework [140].

\(^2\) [http://www.fil.ion.ucl.ac.uk/spm/software/spm12/SPM12_Release_Notes.pdf](http://www.fil.ion.ucl.ac.uk/spm/software/spm12/SPM12_Release_Notes.pdf)

\(^3\) One FLAIR template is available in [165]; however, this was generated using SPM5, so using it would compound any registration biases associated with the older method.
Graylevels are modelled using a mixture of Gaussians, parameterized by means $\mu$ and covariances $\Sigma$ (for-multi-image compatibility); potential bias field is modelled using a discrete cosine transform, parameterized by a vector $b$; spatial tissue priors in MNI space are deformed according to a similar discrete cosine transform, parameterized by a vector $a$ [166]. The parameters of each sub-model are optimized using expectation maximization – i.e. iteratively maximizing the likelihood of each parameter set, given the data (input images), while fixing the others, until convergence. The deformation field defined by the optimal $a$ then gives the transformation from MNI space to native space, which can be inverted as needed.

2.2 Bias Correction

As noted in § 1.2.2, bias field (aka intensity inhomogeneity), is a smoothly varying intensity variation artifact common in MRI. The sources of bias field artifact include inhomogeneities in the magnetic field and RF coils used for pulse transmission and signal sensing, as well as non-ideal magnetic properties of the imaged object [74]. The field and coil related sources are more significant at clinical field strengths. These can be corrected prospectively, though techniques for doing so are limited, and often a small bias field continues to corrupt acquired images [74]. Therefore, retrospective correction has been the subject of much research.

Similar to image registration, several widely accepted algorithms for estimating and correcting bias field have emerged. The N3 algorithm [167], subsequently updated to N4(ITK) [143] and integrated in the FreeSurfer toolbox4 is perhaps the most popular, as it makes minimal assumptions about the image. This method aims to sharpen the image histogram by dividing the image by an estimated bias field, parameterized by B-splines for smoothness. As noted above, the SPM New Segment model [140] includes integrated bias field estimation and correction, and models the field using a discrete cosine transform. The FSL Segment feature also estimates bias field in a similar overall model to SPM New Segment, except the tissue prior probability maps in SPM are replaced with a Markov Random Field Model, as described in [145]. Again, two reviews with quantitative performance comparisons provide a good reference of other proposed algorithms, including a comprehensive review in 2006 [168] and a comparison of mainly popular methods in 2016 [169].

In the 2016 comparison [169], the authors note that the SPM and FSL models – which include segmentation – outperformed the N3 algorithm [167] and another non-segmenting method [170]. These results are

4 https://surfer.nmr.mgh.harvard.edu/
consistent with the advantages of unified generative models described in § 1.3.1 and discussed in “Unified segmentation.” [140]. Specifically, the estimation of both bias field and tissue segmentation are each improved if the other is already known. Rolling these tasks into a single EM-fitted model allows alternating conditional estimates to converge on better results overall.

Bias field estimation was not a primary focus of this work. Therefore, due to the better performance over N3/4, and the advantages already afforded by SPM New Segment for registration noted above, this model was employed for bias correction throughout this work. No other bias field correction tools were investigated. Figure 2.2 shows a FLAIR image with conspicuous bias field, the bias field estimated by SPM, and the corrected image.

2.3 Graylevel Standardization

The flexibility of image contrast in MRI is a double-edged sword. This feature, in addition to properties of spatial encoding during acquisition, preclude direct interpretation of image graylevels as a tissue property. As a result, the same brain region in the same subject may be assigned a different graylevel depending on the MRI sequence time constants, scanner, and spatial acquisition protocol. For automated analysis of MRI images, therefore, standardization of image graylevels is required.

Graylevel standardization can be achieved using a univariate transformation \( \tau : y \mapsto \tilde{y} \), defined as

\[
\tilde{y} = \tau(y), \quad (2.1)
\]

where \( \tau \) is monotonic, and considers characteristics of the input image, such as basic graylevel statistics
or the histogram. The histogram of an image represents the number of occurrences of each graylevel in the image. Normalizing the histogram by the total number of voxels $X$ yields the probability mass function (PMF) $f_Y(y)$,

$$f_Y(y) = \frac{1}{X} h_Y(y) = \frac{1}{X} \sum_{i=1}^{X} \begin{cases} 1 & Y(x_i) = y \\ 0 & Y(x_i) \neq y. \end{cases} \quad (2.2)$$

The cumulative density function (CDF) $F_Y(y)$ is the cumulative sum of $f_Y(y)$,

$$F_Y(y) = \sum_{\gamma=y_{\text{min}}}^{y} f_Y(\gamma). \quad (2.3)$$

The PMF of an MRI can be decomposed into the contributions of each constituent tissue, as shown in Figure A.3. This is the fundamental principle underlying mixture models. The goal of standardization is therefore to align these sub-distributions as closely as possible.

Previously proposed methods of standardizing MRI intensities include the following:

- **Range Matching:** The simplest approach to standardization involves rescaling the data using the minimum and maximum intensities,

$$\tau(y) = \frac{y - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}}. \quad (2.4)$$

Naively, this method is very susceptible to corruption by outliers. For more robustness, $y_{\text{min}}$ and $y_{\text{max}}$ can be defined using intensity quantiles – e.g. $[\epsilon_1, 1 - \epsilon_2]$. However, selection of an appropriate $\epsilon_2$ for WMH segmentation is difficult, since WMH typically constitute only the top 1% of the total brain volume. Furthermore, differences in image contrasts are not considered by this approach.

- **Statistical standardization:** Another simple but popular approach uses the first and second order moments of the PMF,

$$\tau(y) = \frac{y - \mu_Y}{\sigma_Y}. \quad (2.5)$$

As with range matching, variable image contrasts are not well modelled by this method.

- **Histogram equalization:** Histogram equalization transforms the image PMF to a uniform dis-

---

5 This section considers only univariate standardization methods, since only FLAIR intensities are used in this work.
distribution, thereby distributing image intensities equally across the available range. The desired transform is defined as the CDF of the input image (cf. [171] for derivation),

\[ \tau(y) = F_Y(y). \]  \hspace{1cm} (2.6)

The chief assumption of histogram equalization for standardization is that input images contain consistent amounts of each tissue class. In MRI with WMH, this assumption may not be valid; however, this technique may still have value.

- **Histogram matching**: Histogram matching is similar to histogram equalization, except that the output PMF is not uniform, but some other specified distribution, \( f_\tilde{Y} \). This transform is defined as the function composition of the input CDF and the inverse target CDF,

\[ \tau(y) = F_\tilde{Y}^{-1}(F_Y(y)). \]  \hspace{1cm} (2.7)

While histogram matching yields images with different contrast characteristics than histogram equalization, these two methods are equivalent in their ability to standardize image graylevels (cf. A.2 for an illustration and experimental evidence).

- **Nyul standardization**: In [147, 172], Nyul and Udupa proposed a method for intensity standardization which has subsequently been used in other works. This method defines \( \tau \) with piecewise linear segments connecting the \( Q \) quantiles of the input PMF \( q_i \), with quantiles of a target PMF \( r_i \),

\[ \tau(y) = r_i + (y - q_i) \left( \frac{r_{i+1} - r_i}{q_{i+1} - q_i} \right), \quad y \in [q_i, q_{i+1}]. \]  \hspace{1cm} (2.8)

However, it can be shown that this transformation is a non-uniform trapezoidal Riemann approximation of true histogram matching, which performs worse in terms of intensity standardization.6

- **Regional characteristics**: Decorrelating variation in intensities from variability in anatomical content is a central challenge in intensity standardization. One solution is to define the imagespecific transformation using characteristics from a more anatomically consistent brain region. This is the approach employed by Shinohara et al. in the so-called white stripe method [174]. In the current work, this region, denoted \( \mathcal{X}_\tau \), can be defined in MNI space using tissue priors (Figure 3.3), anatomical label maps, or any other method of selecting a subset of voxels.

---

6 This result is presented and supported with experiments in [173].
2.3.1 Quantifying Standardization

While the major advantages and challenges to several graylevel standardization methods have been briefly noted, it remains to explore the utility of each experimentally. In the current work, the goal of this step is to maximize the separation of the two classes. This can even be maximized voxel-wise, due to the characteristics of the VLR model. Therefore, an intermediate objective function $Z$ should be defined which quantifies the degree of separation of the two classes.\(^7\) This objective function can then be used to optimize any tunable parameters in each of the transforms – e.g. $\epsilon$, $Q$, $\chi$ – and also to select the best overall method.

Two such functions are proposed. The first is discrete, and inspired by the Zero-Crossing Rate [175]. It measures the number of class transitions in the sorted feature data $\tilde{Y}_s$, as shown in Figures 2.3a and 2.3c. With $C_s$ as the class labels after sorting by the feature $\tilde{Y} = \tau(Y)$, the objective function $Z_{\Delta}$ is defined as

$$Z_{\Delta} = \sum_{n=1}^{N-1} \begin{cases} 1 & C_s^n \neq C_s^{n+1} \\ 0 & C_s^n = C_s^{n+1}. \end{cases}$$  \hfill (2.9)

This function is discrete and bounded, as in $Z_{\Delta} \in \mathbb{Z} \left[1, \left\lfloor \frac{N}{2} \right\rfloor \right]$\(^8\) and the lower bound is optimal – i.e. $Z_{\Delta}$ should be minimized.

The second function is continuous, and inspired by probability theory. It measures the overlap of the two class distributions, $p(\tilde{y} \mid c = 1)$ and $p(\tilde{y} \mid c = 0)$, estimated using kernel smoothing, relative to the total distribution of the data, as shown in Figures 2.3b and 2.3d. This objective function $Z_*$ can be defined as

$$Z_* = \int_{\tilde{y}_{\min}}^{\tilde{y}_{\max}} \min \left\{ \frac{p(\psi \mid c = 1), p(\psi \mid c = 0)}{\max \{ p(\psi \mid c = 1), p(\psi \mid c = 0) \}} \right\} \delta(\psi - \tilde{y}) \ast G_{\sigma}(\psi), \quad p(\psi \mid c) \approx \sum_{\tilde{y} \in \{\tilde{Y} \mid c\}} \delta(\psi - \tilde{y}) \ast G_{\sigma}(\psi).$$  \hfill (2.10)

where $G_{\sigma}(\psi)$ is a Gaussian convolution kernel with width $\sigma$. This function is continuous and bounded, as in $Z_* \in [0, 1]$, and the lower bound is again optimal, since data overlap between the classes is undesirable.

\(^7\) Alternatively, the entire pipeline can be executed under cross validation and overall performance compared between standardization methods. However this does not consider potential interactions between the standardization method and tunable downstream parameters.

\(^8\) The lower bound can be zero if one class is not observed.
Figure 2.3: Illustration of potential separability objective functions using three sets of sample data (black circles). $Z_\Delta$ counts the class changes in the sorted feature data (numbered red lines). $Z_\star$ computes the ratio of the overlap between class data distributions (red area) to the total data distribution (blue areas).
2.3.2 Supervised Standardization

It is not hard to see that it should be possible to estimate an optimal graylevel standardization transform using the training data. That is, a \textit{supervised graylevel standardization}.\footnote{To the best of this author’s knowledge, such a technique has never been proposed.} If $Z$ is differentiable, then this optimization can be performed using gradient descent, or similar methods. Unfortunately, neither of the above objective functions $Z$ were reasonably differentiable, but many other optimization paradigms which do not require differentiable objective functions could be used – e.g. reinforcement learning \cite{176} or Bayesian Optimization \cite{177}. Due to time constraints, these methods were not explored.

2.4 Pre-Processing Summary

In summary, to train the VLR model, a set of labelled training images must first be registered to a standard brain space (MNI). This is achieved using the SPM \textit{New Segment} tool, which also produces bias corrected images. Next, image intensities are standardized using a graylevel transformation, to be determined in Chapter 4. Training then proceeds to fit the VLR model parameters, as described in Chapter 3.

At test time, SPM \textit{New Segment} is used again to correct the bias field and estimate the registration to MNI space for a given input image. However, the inverse transform is now used to warp the parameter images $\beta(x)$ from MNI space to the native space. This transformation of the smooth parameter images prior to inference is preferable to transforming the detailed label image afterwards. The VLR model then predicts the WMH class label, followed by the necessary post-processing steps.
Chapter 3

Voxel-Wise Logistic Regression

This section presents the proposed classification model – Voxel-wise Logistic Regression (VLR) – in more detail, and explores the specific parameters and regularization strategies requiring optimization.

To review, the predicted lesion class image $\hat{C}(x)$ is defined using the subject-specific features $Y(x) = [1, Y^1(x), \ldots, Y^K(x)]^T$ and the corresponding model weights $\beta(x) = [\beta^0(x), \beta^1(x), \ldots, \beta^K(x)]^T$,

$$\hat{C}(x) = \frac{1}{1 + e^{-\eta(x)}}, \quad \eta = \beta(x)^T Y(x).$$ (3.1)

While the implementation used for experimentation in this work uses only one feature image ($K = 1$), the FLAIR graylevel, the derivations and discussions below will maintain generality for any selection of features.

3.1 Model Fitting

Fitting the VLR model involves estimating $\beta$ for each voxel $x$. This requires some training data: feature vectors from a population of $N$ observations $Y(x) = \{Y_1(x), \ldots, Y_N(x)\}$, and the corresponding labels $C(x) = \{C_1(x), \ldots, C_N(x)\}$. As in many probabilistic models, parameter estimation involves maximizing the likelihood of the model, given this data – i.e. maximum likelihood estimation (MLE).
3.1.1 Maximum Likelihood Estimation

Each parameter vector \( \beta(x) \) is considered completely independent, and doing so greatly simplifies model fitting. In this section, ML estimation of independent parameter vectors \( \beta \) is developed. For clarity, only a single voxel is considered – i.e. \( y \) from \( Y(x) \), etc.

As noted above, the optimal \( \beta \) for each independent voxel can be resolved using MLE. If the training data – features \( \mathcal{Y} = \{ y_1, \ldots, y_n \} \) and labels \( \mathcal{C} = \{ c_1, \ldots, c_n \} \) – are also assumed to be independently observed, then the likelihood (conditioned on the data) is defined from binomial theory as

\[
L(\beta \mid \mathcal{C}, \mathcal{Y}) = \prod_{n=1}^{N} P(c = 1 \mid y_n; \beta)^{c_n} (1 - P(c = 1 \mid y_n; \beta))^{1-c_n} \\
= \prod_{n=1}^{N} \left[ \hat{c}_n (1 - \hat{c}_n)^{1-c_n} \right]. \tag{3.2}
\]

For computational reasons, it is simpler and asymptotically equivalent to maximize the log-likelihood, denoted \( \mathcal{L}(\beta) \), with the conditioning on \( \mathcal{C} \) and \( \mathcal{Y} \) omitted for clarity,

\[
\mathcal{L}(\beta) = \log \prod_{n=1}^{N} \left[ \hat{c}_n (1 - \hat{c}_n)^{1-c_n} \right] \\
= \sum_{n=1}^{N} \left[ c_n \log \hat{c}_n + (1 - c_n) \log(1 - \hat{c}_n) \right] \\
= \sum_{n=1}^{N} \left[ c_n \beta^T y_n - \log(1 + e^{\beta^T y_n}) \right]. \tag{3.3}
\]

The optimal \( \beta \) is therefore resolved by maximizing the log-likelihood,

\[
\beta^* = \arg \max_{\beta} \mathcal{L}(\beta) \\
= \arg \max_{\beta} \sum_{n=1}^{N} \left[ c_n \beta^T y_n - \log(1 + e^{\beta^T y_n}) \right]. \tag{3.4}
\]

3.1.2 Iterative Updates

Estimation of \( \beta^* \) can be performed using iterative optimization, using an initial estimate \( \beta^{(0)} \) and an update term \( \Delta \beta^{(t)} \),

\[
\beta^{(t+1)} \leftarrow \beta^{(t)} + \alpha \Delta \beta^{(t)}, \tag{3.5}
\]
where $\alpha$ is a small valued learning rate parameter. There are many possible definitions of $\Delta \beta$, including simply the gradient of $\mathcal{L}(\beta)$, denoted $\nabla \beta \mathcal{L}$. However, it can be shown that $\mathcal{L}(\beta)$ is convex, so higher order update equations can be used. The work by Minka [178] compares several options, including Newton’s method (and variants), conjugate gradient, iterative scaling (and variants), and dual optimization.\(^1\) For small feature dimensionality ($K$), performance differences among the options were small. Classic Newton updates gave a good balance between memory requirements and computational order, so they are used. However, since computation of the exact Hessian matrix is $O(K^2)$, addition of a few additional features can significantly increase the memory and computational costs of this approach.

If the gradient $\nabla \beta \mathcal{L}$ and Hessian matrix $\nabla^2 \beta \mathcal{L}$ are defined as

\[
\nabla \beta \mathcal{L} = \begin{bmatrix}
\frac{\partial \mathcal{L}}{\partial \beta^1} \\
\vdots \\
\frac{\partial \mathcal{L}}{\partial \beta^k}
\end{bmatrix},
\]  

(3.6)

\[
\nabla^2 \beta \mathcal{L} = \begin{bmatrix}
\frac{\partial^2 \mathcal{L}}{\partial \beta^1 \partial \beta^1} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \beta^1 \partial \beta^k} \\
\vdots & \ddots & \vdots \\
\frac{\partial^2 \mathcal{L}}{\partial \beta^k \partial \beta^1} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \beta^k \partial \beta^k}
\end{bmatrix},
\]  

(3.7)

then the Newton update is given by

\[
\Delta \beta = -\nabla^2 \beta \mathcal{L}^{-1} \nabla \beta \mathcal{L}.
\]  

(3.8)

In the current model, the gradient is given by

\[
\nabla \beta \mathcal{L} = \sum_{n=1}^{N} y_n (c_n - \hat{c}_n),
\]  

(3.9)

and the Hessian by

\[
\nabla^2 \beta \mathcal{L} = \sum_{n=1}^{N} y_n y_n^T (c_n - \hat{c}_n).
\]  

(3.10)

Substituting (3.9) and (3.10) into (3.8), the explicit update $\Delta \beta$ for (3.5) is obtained. At each iteration, $\Delta \beta^{(t)}$ is re-computed, and the process continues until some convergence criterion is satisfied.

\[^1\text{MATLAB code available at https://github.com/tminka/logreg/}\]
3.1.3 Simplification

It is not necessary to complete the above procedure for all voxels in the standardized space. Instead, only voxels in the expected location of the brain need to be computed; such voxels can be selected using a binary “brain mask”, denoted \( M(x) \). More details about the brain mask used in this work can be found in § B.2.2. Moreover, since the parameters of each voxel are estimated independently, this can also be computed in parallel. The details of this implementation are presented in § B.3.1, after incorporation of the regularizations described in the next section.

Finally, the model has so far been derived in general terms, so that any choice of feature set \( y \) can be used. However, with only one feature – the FLAIR graylevel \( y = y^1 \) – it is possible to reparameterize the sigmoid argument as

\[
\beta^T y = \beta^0(1) + \beta^1(y) = s(y - \tau) \begin{cases} 
  s = \beta^1 \\
  \tau = -\frac{\beta^0}{\beta^1}. 
\end{cases}
\] (3.11)

In this form, the parameters \( \tau \) and \( s \) emerge as a graylevel threshold and a slope parameter, respectively. Specifically, when \( y = \tau \), the predicted probability of lesion is \( \hat{c} = \frac{1}{1 + e^{-\tau}} = 0.5 \), so \( \tau \) controls the location of the class discrimination, as shown in Figure 3.1a. Similarly, the \( s \) parameter defines the sensitivity of the logistic function to \( y \), as shown in Figure 3.1b. By contrast, varying the original parameter \( \beta^1 \) with \( \beta^0 \) constant (Figure 3.1d) results in correlation of these characteristics. These new parameters, and the corresponding images \( \mathcal{T}(x) \) and \( \mathcal{S}(x) \), are therefore informative descriptors of the predictive model.

3.1.4 Challenges

Three major challenges emerge during model fitting. These challenges involve contradictions between prior knowledge and the fitted model using the available training data. That is, these challenges could all be overcome by a more complete training set, but this is rarely available. The three challenges are:

1. **Separable classes**: When data from two classes are perfectly separable, the MLE-fitted logistic model can approach a step-function – i.e. \( \beta^k \to +\infty \). This implies that on either side of a specific graylevel threshold, the model is either 100% confident in predicting the healthy class, or 100% confident in predicting the lesion class. In fact, no threshold is ever so perfect, and instead a level of uncertainty is desirable around the decision boundary. Figure 3.2a illustrates these two cases.
2. **Sparsely observed lesion class:** Since WML are often distributed in consistent locations, many brain regions contain no lesions across the entire training dataset. These voxels will be termed “healthy training” voxels, and denoted $X_h$. In some locations, this is expected (e.g. the GM, since by definition WMH manifest in the WM), while in others, prior knowledge predicts lesions will eventually be observed (e.g. the rest of the WM). As illustrated in Figure 3.2b, the MLE-fitted model may not maintain the ability to predict $\hat{c} = 1$ in such locations, regardless of the features. However, the ability to predict lesions should be maintained in many of these locations.

3. **Smooth parameter images:** It is assumed that similar locations will contain similar training data, yielding smooth parameter images. If this assumption is sometimes invalid, parameter images could contain noise or discontinuities, creating artifacts in estimated lesion class images.
3.2 Regularization

Regularizations are methods of controlling the capacity of a model, so that it does not overfit the training data (cf. § 4.3). Often, this involves injecting prior knowledge about the expected model into the optimization. Assuming voxel-wise independence of model parameters requires the use of regularization strategies to solve the challenges outlined in § 3.1. Several regularization methods are explored below.

3.2.1 Data Augmentation

Noting the central role of training data in each of the challenges, methods of artificially increasing the training dataset size may be particularly useful in solving them. Data augmentation has long been used in machine learning tasks with limited training data, and there are several methods of generating synthetic data. In low dimensional input/output spaces, random sampling of fitted class-conditional posterior distributions can produce reasonable samples with known labels [179]. In higher dimensional problem spaces, however, imputation is more difficult [180]. For example, the space of potential $100 \times 100 \times 100$-sized images has $100^3$ dimensions (one per voxel), yet only a small subspace represents plausible images. Generating synthetic examples in this space is therefore challenging, especially for segmentation tasks, where the outputs have dimensionality roughly equal to the input. Recent work [181, 182] has explored interpolation in a lower dimensional embedding of the data, which has yielded both empirically plausible training examples and generalization performance improvements. However, in the current work, no such representations are readily available, since the segmentation model does not leverage deep learning methods.
Alternatively, simple image manipulations can still afford model improvements [183]. In segmentation tasks, both the input image(s) and the corresponding label images can be translated, reflected, rotated, and perhaps resized, thereby avoiding the generation of genuinely synthetic examples. In the current work, reflections and small (one-voxel) translations can be applied to the label and FLAIR images following registration to the MNI brainspace. The potential benefits of this augmentation are explored in § 4.7.2.

### 3.2.2 Parameter Norm Penalties

The separable classes challenge is well-known in regression problems, and a good solution is to penalize the magnitude of model parameters using the $L_p$-norm: $\lambda \|\beta\|_p$ [184]. It can be shown that $L_1$ regularization corresponds to a Laplacian prior on elements of $\beta$, with scale parameter inversely proportional to $\lambda$ (equivalently, this assumes that the model error follows this distribution). Similarly, $L_2$ regularization implies a Gaussian prior, with standard deviation inversely proportional to $\lambda$ [184]. Model fitting which includes this prior-derived term is called maximum a posteriori (MAP) estimation, and the penalty can be appended to the objective function (3.4), as in

$$
\beta^* = \arg \max_\beta \mathcal{J}(\beta) = \arg \max_\beta \mathcal{L}(\beta) - \lambda \|\beta\|_p
= \arg \max_\beta \sum_{n=1}^N \left[ c_n \beta^T y_n - \log(1 + e^{\beta^T y_n}) \right] - \lambda \|\beta\|_p .
$$

(3.12)

Due to its relatively large gradient near zero, $L_1$ regularization is typically used to encourage sparsity in the feature weights (i.e. $\beta^k \to 0$) [185]. This is not desirable in the current model, since the feature (FLAIR graylevel) is known to be discriminative. Moreover, the expansion of the $\|\beta\|_1$ term in the gradient of the objective function is not straightforward, since it is non-differentiable at zero [185, 186]. Conversely, $L_2$ regularization is more effective at limiting parameter magnitude – which is the current aim – and the first and second order gradients of (3.12) derive easily [178]. For these reasons, only $L_2$ regularization is considered, yielding the following change to the Newton update expression (3.8),

$$
\Delta \beta = -\nabla^2_{\beta} \mathcal{J}^{-1} \nabla_{\beta} \mathcal{J} = -\left( \nabla^2_{\beta} \mathcal{L} - \lambda I \right)^{-1} \left( \nabla_{\beta} \mathcal{L} - \lambda \beta \right).
$$

(3.13)

What remains is to select an appropriate value of $\lambda$. This is explored experimentally in § 4.5.2 using a toy model.
3.2.3 Pseudo-Lesions

The sparsely observed lesion class challenge is less common, since discriminative models are rarely fit in the absence of one class altogether. This occurs here because all voxels are modelled independently. It is therefore tempting to simply sample features from the lesion class at other spatial locations in order to fit the logistic model in the healthy training voxels, similar to the approach by Schmidt. However, as noted in § 1.2.2 and § 1.3.3, WML are thought to have different intensities in different brain regions [7, 150], and some locations will likely never contain any WMH. Considering these facts, the use of deterministic synthetic lesion-class samples, or “pseudo-lesions”, could instead permit better use of prior knowledge about WMH. These synthetic observations could be appended to the training data for each voxel so as to minimally balance the training classes, and act as a prior on the distribution of lesion-class features.

If the same number of synthetic observations are appended to the training data for each voxel, this is equivalent to appending a number of synthetic images to the training set. The synthetic feature data are denoted $\mathcal{V}(x) = \{\gamma_1(x), \ldots, \gamma_V(x)\}$. It is assumed that the labels of all synthetic data are “lesion”, so the set of synthetic label images is simply denoted $\mathbf{1}(x)$. The updated training set is therefore $\mathcal{Y}_{\gamma}(x) = \{\mathcal{Y}(x), \mathcal{V}(x)\}$, and $\mathcal{C}_{\gamma}(x) = \{\mathcal{C}(x), \mathbf{1}(x)\}$.

Design of the synthetic image set $\mathcal{V}(x)$ should be guided by prior knowledge. For the same reasons as described above, it is not possible to derive this knowledge from the training set. Unfortunately, few other sources of structured information are available. One reasonable approach could make use of healthy tissue prior probability images (Figure 3.3), denoted $\rho(x)$. Specifically, the expected intensity for the lesion class in each tissue can be multiplied by the tissue probability image, and the results summed to give the overall synthetic image,

$$\gamma(x) = \gamma_{\text{GM}} \cdot \rho_{\text{GM}}(x) + \gamma_{\text{WM}} \cdot \rho_{\text{WM}}(x) + \gamma_{\text{CSF}} \cdot \rho_{\text{CSF}}(x).$$  (3.14)

While WMH are not possible in either the GM or the CSF, it is necessary to select a FLAIR graylevel – perhaps the maximum possible intensity – to complete this model. Additionally, such parameters will inevitably play a role for subjects with imperfect registration or outlier anatomy. The contributions of pseudo-lesions to model fitting are explored experimentally in § 4.5.3.
Figure 3.3: Tissue prior probability images in MNI space. Derived from [86]. Best viewed in colour.

Table 3.1: Image filters considered for smoothing the estimated parameter images.

<table>
<thead>
<tr>
<th>Name</th>
<th>Parameters</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>width $\sigma$</td>
<td>no artifacts</td>
<td>blurs edges</td>
<td>[171]</td>
</tr>
<tr>
<td>Median</td>
<td>width $w$</td>
<td>preserves edges</td>
<td>square artifacts</td>
<td>[171]</td>
</tr>
<tr>
<td>Bilateral</td>
<td>$\sigma_y, \sigma_x$</td>
<td>balanced blurring and detail</td>
<td>Expensive</td>
<td>[187]</td>
</tr>
</tbody>
</table>

3.2.4 Parameter Image Smoothing

Finally, independent model fitting in every voxel risks yielding noisy parameter images. The simplest solution to this problem involves filtering the reconstructed parameter images after estimation. A wide range of possible filters for this task exist, and a selection of these are summarized in Table 3.1, including the mutable parameters for each.

An alternative solution might involve modelling the parameter images as a spatial function (e.g. band-limited discrete cosine / Fourier transform). However, there are two challenges with this approach. First, deriving the update gradients for such a model would be challenging, and their computation could significantly increase training time. Second, such an encoding may introduce artifacts in the resulting parameter images. Moreover, it is well known that frequency domain band-limiting can be equivalently...
achieved by convolution (i.e. filtering) in the spatial domain [171]. Therefore, only conventional filtering is explored in this work (cf. § 4.7.3).

3.3 Post-Processing

At this point, the motivation and details of the proposed VLR model have been presented, in addition to the preprocessing steps required to satisfy its assumptions. The last component of a segmentation pipeline typically includes post-processing. In principle, this step aims to incorporate any additional knowledge of the problem which has not been considered in upstream elements. Here, these include the connected morphology of WML, and the minimum lesion size. Discussions of these topics, however, would assume that the label image is already binary, whereas the output from the VLR model is probabilistic. Therefore, the first post-processing step thresholds the WMH class probability at a value $\pi$ to give a “hard” classification: $\hat{c} \rightarrow \hat{c}^\pi$ This also facilitates comparison with manual segmentation masks, which are usually binary.

3.3.1 Thresholding

If the assumptions of any probabilistic model are valid, then the “hard” classification between any set of classes $c$ is straightforward:

$$\hat{c}^\pi = \arg \max_c p(c \mid y; \beta). \quad (3.15)$$

In a 2-class logistic regression model, this simplifies to thresholding at $p$:

$$\hat{c}^\pi = \begin{cases} 0 & \hat{c} < \pi \\ 1 & \hat{c} \geq \pi, \end{cases} \quad (3.16)$$

with $\pi = 0.5$. However, since these assumptions are often only partially true, most models are able to achieve better agreement with manual segmentations using a different threshold $\pi$ for the WMH class. In fact several of the freely available toolboxes permit a user-specified threshold which can be optimized for the user’s data. During model validation, this parameter should be optimized using the training data for each cross validation fold. It is also prudent to illustrate the sensitivity of the model to this parameter, using either a plot of performance versus threshold [64], or a precision-recall (PR) curve [188]. In the current work, both these techniques are employed: $\pi$ is optimized using the training data in native space, and a PR curve is later given (Figure 4.27a).
3.3.2 Minimum Lesion Size

With finite image resolution and appreciable noise in MRI, lesions appearing as only a few connected voxels are indistinguishable from image noise, even by human experts. Such potential lesions are therefore not included in radiologists’ assessment of WML. Accordingly, most WMH segmentation algorithms employ a minimum-connected-voxels exclusion criterion during post-processing. Connectedness can be defined in 2D or 3D, and consider only direct adjacency or diagonal connections too. Most works employ the most liberal definition: 26-connectedness, which considers all $3 \times 3 \times 3 - 1 = 26$ candidates surrounding a given voxel in 3D. Ideally, the number of required connected voxels will adapt to the image resolution, and correspond to a minimum lesion volume. Typical volumes range from about $x_{\text{min}}^c = 3.5 \text{mm}^3$ in [64, 127], to $9.0 \text{mm}^3$ in [119, 189].

In the current work, the inclusion of a minimum lesion size rule is explored. The optimal value for $x_{\text{min}}^c$ is resolved experimentally during each cross validation fold, and the resulting values compared with the above conventions. Additionally, the gains in performance afforded by this step are quantified.

3.4 Model Summary

In summary, the proposed algorithm uses graylevel features to train a logistic regression model for each voxel independently – Voxel-Wise Logistic Regression. In order to train the VLR model, a set of labelled training images must first be registered to a standard brain space (MNI). This is achieved using the SPM New Segment tool, which also produces bias corrected images. Next, image intensities are standardized using a graylevel transformation, to be determined in the next section. The parameter images $\beta(x)$ are then computed using iterative MAP estimation, with Newton updates and an augmented dataset. These images are smoothed to reflect prior knowledge. Finally, the optimal probability threshold $\pi$ and minimum lesion size $x_{\text{min}}^c$ are estimated using the training data. This completes the training phase.

At test time, SPM New Segment is used again to correct the bias field and estimate the registration to MNI space for a given input image. However, the inverse transform is now used to warp the parameter images $\beta(x)$ from MNI space to the native space. This transformation of the smooth parameter images prior to inference is preferable to transforming the detailed label image afterwards. The probability of the WMH class is computed by evaluating the independent logistic models at every voxel. This initial estimate $\hat{C}(x)$ is then thresholded using $\pi$, and binary objects smaller than $x_{\text{min}}^c$ are removed. The resulting label image is the final output. These training and testing phases are illustrated in Figure 3.4.
3.4.1 Tunable Parameters

In order to achieve the best possible model performance, it is prudent to track tunable model parameters (AKA hyperparameters) which are distinct from those fitted during each cross validation fold – i.e. $\beta(x)$ and $\pi$. Considering both the main VLR model and the pre- and post-processing aspects, the parameters of the proposed algorithm are summarized in Table 3.2. The optimization of these model components will be the subject of the next chapter.
Table 3.2: Model hyperparameters and baseline values. The values after optimization are given in Table 4.5

<table>
<thead>
<tr>
<th>Stage</th>
<th>Parameter</th>
<th>Notation</th>
<th>Type</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Processing</td>
<td>Reflect Augmentation</td>
<td>$a_R$</td>
<td>$\mathbb{B}$</td>
<td>false</td>
</tr>
<tr>
<td></td>
<td>Shift Augmentation</td>
<td>$a_S$</td>
<td>$\mathbb{N}_p$</td>
<td>$N_0$</td>
</tr>
<tr>
<td></td>
<td>Graylevel Transform</td>
<td>$\tau_y$</td>
<td>$f : \mathbb{R} \mapsto \mathbb{R}$</td>
<td>$\tau_{RM3}$</td>
</tr>
<tr>
<td></td>
<td>Transform Mask</td>
<td>$\mathcal{X}_\tau$</td>
<td>$\mathbb{B}(x)$</td>
<td>$\mathcal{X}_{brain}$</td>
</tr>
<tr>
<td>VLR Fitting</td>
<td>Iterations</td>
<td>$T$</td>
<td>$\mathbb{Z}$</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Initial $\beta$</td>
<td>$\beta^{(0)}$</td>
<td>$\mathbb{R}^2$</td>
<td>$[0, 0]$</td>
</tr>
<tr>
<td></td>
<td>Estimation Scale $^a$</td>
<td>$r$</td>
<td>$\mathbb{R}$</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Learning Rate</td>
<td>$\alpha$</td>
<td>$\mathbb{R}$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Regularization</td>
<td>$\lambda$</td>
<td>$\mathbb{R}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Pseudo-Lesions</td>
<td>$\mathcal{V}(x)$</td>
<td>${ \cdot \in \mathbb{R} }$</td>
<td>${ }$</td>
</tr>
<tr>
<td></td>
<td>$\beta$ Filter</td>
<td>$F_\beta$</td>
<td>$f : \mathbb{R}(x) \mapsto \mathbb{R}(x)$</td>
<td>$\tilde{\beta}(x) = \beta(x)$</td>
</tr>
<tr>
<td>Post-Processing</td>
<td>Min Lesion Size</td>
<td>$x^c_{\min}$</td>
<td>$\mathbb{R}$ (mm$^3$)</td>
<td>0</td>
</tr>
</tbody>
</table>

Notation. $\mathbb{B}$: boolean value; $\mathbb{Z}$: integer value; $\mathbb{R}$: real value; $\mathbb{R}^n$: vector; $\mathbb{B}(x)$: image; $\mathbb{N}_p$: nearest $p$ voxel neighbourhood. $^a$ cf. § B.3.3.
Chapter 4

Experiment & Results

Having defined each of the algorithm components, and derived the estimation procedures, this section explores model validation and optimization. Performance of model components is characterized with respect to intermediate objectives, including graylevel standardization and regularization, in toy scenarios. The segmentation performance of the full model is then presented under several cross validation frameworks, and compared to a similar algorithm.

4.1 Data

For the several reasons (cf. § 4.3) it was important to collect a large and diverse database of FLAIR images for model validation. A total of 129 FLAIR images from from 10 different scanners were collected; the number of images and scan parameters are summarized in Table 4.1. Except for the MS 2008\(^1\) and In-House datasets, all of the data are freely available as part of the segmentation competitions. Since direct comparison of results on equal datasets is important for establishing state-of-the-art, all results are presented using only these freely available data (“Dataset A”), though all the available data are summarized for reference. The average distribution of WMH in Dataset A is shown in Figure 4.1.

Regarding simulated MR images, none are used for validation of segmentation performance in this work. While such data (e.g. BrainWeb [191]) are useful for evaluation of whole-brain segmentation methods, only three examples of WMH are currently available, and there is little documentation as to how these were generated.\(^2\) Furthermore, several works [192, 193] have noted significant discrepancies in estimated

---

1 The manual segmentations used in this dataset were revised in-house, as described in § B.2.1.
Table 4.1: Summary of experimental image database.

<table>
<thead>
<tr>
<th>Img (#)</th>
<th>Database</th>
<th>Ref.</th>
<th>Scanner</th>
<th>TE (ms)</th>
<th>TR (ms)</th>
<th>TI (ms)</th>
<th>Voxel Size (mm)</th>
<th>Manuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>WMH 2017 (1)</td>
<td>[12]</td>
<td>3T Philips Achieva</td>
<td>125</td>
<td>11000</td>
<td>2800</td>
<td>0.96 × 0.96 × 3.00</td>
<td>1 a</td>
</tr>
<tr>
<td>20</td>
<td>WMH 2017 (2)</td>
<td>[12]</td>
<td>3T Siemens TrioTim</td>
<td>82</td>
<td>9000</td>
<td>2500</td>
<td>1.00 × 1.00 × 3.00</td>
<td>1 a</td>
</tr>
<tr>
<td>20</td>
<td>WMH 2017 (3)</td>
<td>[12]</td>
<td>3T GE Signa HDxt</td>
<td>126</td>
<td>8000</td>
<td>2340</td>
<td>0.98 × 1.20 × 3.00</td>
<td>1 a</td>
</tr>
<tr>
<td>5</td>
<td>MS 2016 (1)</td>
<td>[83]</td>
<td>3T Philips Ingenia</td>
<td>360</td>
<td>5400</td>
<td>1800</td>
<td>0.50 × 1.10 × 0.50</td>
<td>7 b</td>
</tr>
<tr>
<td>5</td>
<td>MS 2016 (2)</td>
<td>[83]</td>
<td>1.5T Siemens Aera</td>
<td>336</td>
<td>5400</td>
<td>1800</td>
<td>1.04 × 1.25 × 1.04</td>
<td>7 b</td>
</tr>
<tr>
<td>5</td>
<td>MS 2016 (3)</td>
<td>[83]</td>
<td>3T Siemens Verio</td>
<td>399</td>
<td>5000</td>
<td>1800</td>
<td>0.74 × 0.70 × 0.74</td>
<td>7 b</td>
</tr>
<tr>
<td>21</td>
<td>ISBI MS 2015</td>
<td>[57]</td>
<td>3T Philips</td>
<td>68</td>
<td>11000</td>
<td>2800</td>
<td>0.43 × 0.43 × 3.00</td>
<td>2 c</td>
</tr>
</tbody>
</table>

*Manuals were generated following the standards outlined in [81], and were subsequently reviewed by a second rater, only WMH labels were included; b Manuals were fused using the LOP-STAPLE method [190]; c Manuals were fused using logical ‘and’; d Manuals were generated in-house; e Manuals were revised in-house (cf. § B.2.1).*

Figure 4.1: Average distribution of WMH in Dataset A. Best viewed in colour.

segmentation performance between BrainWeb and real data (ISBR [164]).

4.2 Segmentation Performance Metrics

Quantifying the performance of a model is essential to optimizing its design. Typically, WMH segmentation performance is characterized in two respects: voxel-wise agreement and total lesion load (LL) volume agreement. When comparing the estimated class \( \hat{c} \) to the ground truth class \( c \), each individual voxel can occupy one of four states (colours shown for future reference):

- **True Positive (TP):** \( c = 1 \) and \( \hat{c} = 1 \), correctly predicted “lesion”.
- **False Positive (FP):** \( c = 0 \) and \( \hat{c} = 1 \), incorrectly predicted “lesion”.
- **False Negative (FN):** \( c = 1 \) and \( \hat{c} = 0 \), incorrectly predicted “healthy”.
- **True Negative (TN):** \( c = 0 \) and \( \hat{c} = 0 \), correctly predicted “healthy”.

49
Summing the number of voxels in each state over an entire image volume, voxel-wise agreement can then be quantified using the following measures:

- **Similarity Index (SI)** (aka Dice Similarity Coefficient, F1-Score)
  Measures overall segmentation performance:
  \[
  SI = \frac{2TP}{2TP + FP + FN}.
  \]  
  (4.1)

- **Precision (Pr)** (aka Overlap Fraction, Positive Predictive Value)
  Fraction of predicted predicted positives which are true positives:
  \[
  Pr = \frac{TP}{TP + FP}.
  \]  
  (4.2)

- **Recall (Re)** (aka Sensitivity, True Positive Rate)
  Fraction of true positives which are predicted positive:
  \[
  Re = \frac{TP}{TP + FN}.
  \]  
  (4.3)

Each measure is \(\in [0, 1]\), where higher is better. Note that typical performance metrics like accuracy and specificity are avoided, since they include the \(TN\) count in the numerator, which is typically much larger than \(TP + FP + FN\) combined – i.e. \(c = 1\) is a rare event.

Overall volume agreement between segmentations is characterized using the 2-way mixed-effects single-rater absolute intraclass correlation coefficient (ICC)\(^3\) [194], while trends in over/undersegmentation with lesion load are illustrated using Blant-Altman plots [195].

### 4.3 Cross Validation Frameworks

Supervised segmentation models require the capacity to model complex relationships between the input image(s) and output label images. When models with large capacity are trained on a dataset which does not represent the full gamut of potential input data, they risk **overfitting**: acquiring a bias towards the training data [196]. The main problem associated with overfitting is decreased performance on new data (aka generalization performance) [196]. Popular techniques for characterizing this expected decrease

\(^3\) Option ‘A-1’ in the MATLAB function ICC from [https://www.mathworks.com/matlabcentral/fileexchange/22099](https://www.mathworks.com/matlabcentral/fileexchange/22099)
include cross validation (CV) procedures. These involve splitting the N available examples into training (r) and testing (e) subsets, where the training data are used to fit the model parameters, and the test data are used to approximate the expected generalization performance; the data splits are usually repeated, randomly or exhaustively, to ensure robust results [197]. The most popular CV frameworks include:

- **LOO – Leave-One-Out**: Use all images except one as the training set; use it as the test case \( N_r = N - 1; N_e = 1 \); repeat \( N \) times.
  
  \textit{Benefit}: Close approximation of the expected generalization performance
  
  \textit{Drawback}: Expensive to compute – \( \mathcal{O}(N) \)

- **KFCV – K-Fold Cross Validation**: Use all images except a random batch of \( B = N/K \) images as the training set; use these as the test set \( N_r = N - B; N_e = B \); repeat \( K \) times (without replacement).
  
  \textit{Benefit}: Less expensive to compute – \( \mathcal{O}(N/K) \)
  
  \textit{Drawback}: Worse approximation of the expected generalization performance

Many authors also validate their model using LOO-CV, but only use images from a consistent source during training, repeating the process for all sources. This only approximates performance on images from a consistent source, and therefore assumes that labelled training data will always be available from the use-case scanner. This framework can be summarized as follows:

- **OSAAT – One-Scanner-At-A-Time**: Use all images from a single scanner \( N_s \) except one as the training set; use it as the test case \( N_r = N_s - 1; N_e = 1 \); repeat \( N \) times.
  
  \textit{Benefit}: Estimates source-specific generalization performance
  
  \textit{Drawback}: Does not approximate generalization performance for new image sources

It is worth noting one additional framework which does not estimate model generalization performance, but which lends insights into the capacity of the algorithm to model the desired relationship. This is actually to not use any cross validation at all:

- **No CV – No Cross Validation**: Train and test the model on all available data \( N_r = N_e = N \); no repetition.
  
  \textit{Benefit}: Characterize model limitations; least expensive to compute – \( \mathcal{O}(1) \)
  
  \textit{Drawback}: Not a valid approximation of the expected generalization performance
These results can be seen as a cap on model performance – the estimated generalization performance should never exceed the performance under No-CV. In models with high capacity, No-CV results would be expected near perfect, due to obvious overfitting. However, in models with stronger priors or imperfect assumptions, No-CV results illustrate the best possible performance achievable through optimization regularization components alone.

4.3.1 Leave-One-Source-Out CV

The choice of cross validation framework can have significant impacts on the reported model performance (see [197] for an in-depth review), and there is at least one assumption of the above methods which is not always valid: that examples are independent and identically distributed (iid). This is not true for data originating from multiple sources with different underlying distributions (e.g. MRI with different scan-parameter combinations) [198]. In fact, Geras and Sutton show that in multi-source problems where the expected use case involves data from entirely new sources, random KFCV (and therefore also LOO, as a special case of KFCV with $B = 1$) significantly overestimates the generalization performance. This is because random training fold selection allows the model to perceive source-specific characteristics of the test examples, which cannot be repeated for truly new examples. In such scenarios, the author proposes the following:

- **LOSO – Leave-One-Source-Out:**\(^4\) Withhold all examples from source $s \in 1 \ldots S$ from the training set, and use these as the test set ($N_r = N - N_s; N_e = N_s$); repeat $S$ times.

  **Benefit:** Best approximation of the expected generalization performance in multi-source problems

  **Drawback:** Still only an approximation

As noted in the introduction (cf. § 1.3.3), there has been surprisingly limited use of data from multiple sources for validation of WMH segmentation algorithms. Moreover, CV frameworks vary widely among papers, and to the best of this author’s knowledge, no WMH algorithm has yet been validated using LOSO CV. This represents a significant caveat to reported performances, since MRI have many sources of variability (cf. § 1.2.2), including scanner manufacturer, field strength, sequence parameters, resolution, anatomical and disease variability. As the aim of this work is to develop a segmentation algorithm which will perform well on any given FLAIR MRI, the LOSO framework was initially developed without knowledge of the work by Geras and Sutton. However, this paper happily corroborates the importance

\(^4\) The original name used by Geras and Sutton was “Multi-Source Cross Validation”
of LOSO CV to the current work. In this case, one data source is defined as a unique scanner-parameter combination.

### 4.4 Graylevel Standardization

The objective of graylevel standardization in this work is relatively simple: voxel-wise separation of the lesion class from healthy tissues. Two methods of quantifying this were proposed: Equations (2.9) and (2.10). Therefore, using these metrics, the graylevel standardization techniques defined in § 2.3 were compared for the FLAIR intensities in Dataset A, and only those voxels in the brain mask.

Since the truly raw image intensities range from $[0, 129]$ to $[0, 77537]$, some minimal standardization is required as a baseline and to allow visualization. For this, range matching with $\epsilon = [10^{-4}, 1 - 10^{-4}]$ is used. Figure 4.2a illustrates the “raw” image PMFs from Dataset A following this transformation. Note that these PMFs differ appreciably from those simulated in Figure A.3, particularly in the separation of tissue-distribution peaks. This may be due to several of the challenges noted in § 1.2.2, but overall highlights the difficulties of segmenting real versus simulated images.

Next, each of the graylevel standardization techniques described in § 2.3 were applied to the FLAIR intensity data. For transforms with tunable parameters, several selections were made. The target PMFs for histogram matching operations are also shown in Figure 4.2b. Both standardization objective functions were then computed for all voxels, and averaged across the image. Comparison of these metrics, shown in Table 4.2, permits selection of the best graylevel standardization technique.
Table 4.2: Graylevel agreement objective functions (mean) for different standardization operations.

<table>
<thead>
<tr>
<th>τ</th>
<th>Parameters</th>
<th>$E[Z_\Delta]$</th>
<th>$E[Z_*]$</th>
<th>FFI(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RM1</td>
<td>$\epsilon = [10^{-4}, 1 - 10^{-4}]$</td>
<td>16.1</td>
<td>7.42</td>
<td></td>
</tr>
<tr>
<td>RM2</td>
<td>$\epsilon = [10^{-3}, 1 - 10^{-3}]$</td>
<td>16.7</td>
<td>7.70</td>
<td></td>
</tr>
<tr>
<td>RM3</td>
<td>$\epsilon = [10^{-2}, 1 - 10^{-2}]$</td>
<td>15.5</td>
<td>7.52</td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>—</td>
<td>12.2</td>
<td>5.77</td>
<td>*</td>
</tr>
<tr>
<td>HE</td>
<td>—</td>
<td>10.0</td>
<td>8.16</td>
<td>*</td>
</tr>
<tr>
<td>HM1</td>
<td>$f_{\tilde{y}} = \mathcal{N}(\frac{1}{2}, \frac{1}{5})$</td>
<td>11.5</td>
<td>6.86</td>
<td>*</td>
</tr>
<tr>
<td>HM2</td>
<td>$f_{\tilde{y}} = \gamma^5 - \gamma^6$</td>
<td>12.0</td>
<td>9.11</td>
<td>*</td>
</tr>
<tr>
<td>HM3</td>
<td>$f_{\tilde{y}} = (1 - \gamma)^5 - (1 - \gamma)^6$</td>
<td>10.2</td>
<td>6.44</td>
<td>*</td>
</tr>
<tr>
<td>NY</td>
<td>$Q = [0, \frac{1}{16}, \ldots, 1]$</td>
<td>12.0</td>
<td>8.98</td>
<td></td>
</tr>
</tbody>
</table>

FFI: For further investigation.

Figure 4.3: Image graylevel PMFs after the two best standardization operations. Best viewed in colour.

From these results, it can be seen that three transformations provide good reductions in class graylevel overlap: statistical standardization (SS), histogram equalization (HE), and the 3rd histogram matching operation (HM3). While the HM3 operation is not optimal in either metric, it achieved second place in both. The worst results are given by all three range-matching operations (RM), the Nyul standardization method (NY) and the 2nd histogram matching operation (HM2); therefore these will not be subject to further investigation. Finally, since the graylevel agreement measures $Z_\Delta(x)$ and $Z_*(x)$ are computed voxel-wise, it is possible to show their distribution spatially. This is illustrated for the raw images and the best performing transformation:
Figure 4.4: Spatial depiction of $Z_{\Delta}(x)$ and $Z_{\star}(x)$ comparing raw images and images standardized using HM3. Best viewed in colour.
4.5 Regularization

This section explores the optimization of regularization strategies using a toy model, in order to reduce the complexity of experimentation. In particular, the value of $\lambda$, and the definition of pseudo-lesions $\{Y_\gamma, C_\gamma\}$ are explored, since these are implemented voxel-wise. The segmentation performance of the full model under LOSO-CV are later used to validate these results, in addition to exploration of the other regularizations: data augmentation and parameter image smoothing techniques.

4.5.1 Toy Model

The toy model used here represents a single voxel during training, with synthetic observations. Regularizations are then chosen to maintain desired characteristics in the fitted functions. No specific objective function is defined for this purpose; rather, the expected characteristics of the logistic function illustrated in Figure 3.2 are used to empirically drive parameter selection.\(^5\) In order to explore the various problem scenarios, 9 sets of synthetic data are generated with the PMF shown in Table 4.3, with the resulting distributions shown in Figure 4.5.

4.5.2 Parameter Norm Penalties

Before exploring the 9 different scenarios specified above, it is worth illustrating the effect of $L_2$ regularization on the MAP objective function $J(\beta)$ in the 2D plane composed of $\beta = [\beta^0, \beta^1]$. Using synthetic dataset e, which would be expected to experience overfitting, $J(\beta)$ was computed over a grid of different

---

Table 4.3: Toy data definitions, with $y_c \sim N(\mu_c, \sigma_c)$.

<table>
<thead>
<tr>
<th>#</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$N$</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.3</td>
<td>0.12</td>
<td>100</td>
<td>0.7</td>
<td>0.12</td>
<td>100</td>
</tr>
<tr>
<td>b</td>
<td>0.3</td>
<td>0.12</td>
<td>100</td>
<td>0.7</td>
<td>0.12</td>
<td>10</td>
</tr>
<tr>
<td>c</td>
<td>0.3</td>
<td>0.24</td>
<td>100</td>
<td>0.7</td>
<td>0.24</td>
<td>10</td>
</tr>
<tr>
<td>d</td>
<td>0.3</td>
<td>0.06</td>
<td>100</td>
<td>0.7</td>
<td>0.06</td>
<td>10</td>
</tr>
<tr>
<td>e</td>
<td>0.3</td>
<td>0.03</td>
<td>100</td>
<td>0.7</td>
<td>0.03</td>
<td>10</td>
</tr>
<tr>
<td>f</td>
<td>0.6</td>
<td>0.08</td>
<td>100</td>
<td>0.3</td>
<td>0.08</td>
<td>10</td>
</tr>
<tr>
<td>g</td>
<td>0.4</td>
<td>0.10</td>
<td>100</td>
<td>—</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>h</td>
<td>0.6</td>
<td>0.08</td>
<td>100</td>
<td>—</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>i</td>
<td>0.8</td>
<td>0.06</td>
<td>100</td>
<td>—</td>
<td>—</td>
<td>0</td>
</tr>
</tbody>
</table>

---

\(^5\) This popular technique is also called “hand waving”.

56
\( \beta^0 \) and \( \beta^1 \) values. The \( \beta \) prior function \( \mathcal{P}(\beta) = ||\beta||_2 \) is computed similarly. These functions are then exponentiated, as in \( J = e^J \) and \( P = e^P \) – i.e. the likelihood, as opposed to the log-likelihood. For \( \lambda = 0 \), the prior is a uniform distribution (\( \mathcal{P}(\beta) = 1 \), Figure 4.6a), and the MAP objective equates the MLE objective (\( J(\beta) = L(\beta) \), Figure 4.6b). For nonzero \( \lambda = [10^{-3}, 10^{-2}, 10^{-1}] \), the MAP likelihood \( J(\beta) \) can be defined as the product of \( \mathcal{P}(\beta | \lambda) \) (Figures 4.6c, 4.6e, and 4.6g) and \( L(\beta) \) (Figure 4.6b), yielding Figures 4.6d, 4.6f, and 4.6h. Thus, as expected, increasing \( \lambda \) reduces the magnitudes of fitted \( \beta \), thereby limiting the slope parameter \( s = \beta^1 \), as desired.

Next, the appropriate \( \lambda \) is determined by fitting the logistic model for the first 6 toy scenarios, since the last 3 have no lesion class examples. For each scenario, each of the same four \( \lambda = [0, 10^{-3}, 10^{-2}, 10^{-1}] \) are used to regularize the estimated parameters. These results are shown in Figure 4.7, where the final
Figure 4.6: Toy model likelihoods as a function of $\beta$: $P(\beta) = e^{P(\beta)}$ and $J(\beta) = e^{J(\beta)}$, for different $\lambda$, using scenario e. The optimum is shown as a blue star. Best viewed in colour.
Figure 4.7: Toy model MAP estimation results for 6 different scenarios and different $\lambda$. Best viewed in colour.

state in each condition is shown in a different colour, and the progression of the fitted logistic function is depicted from light to dark. Note that a heuristic rule is used to ignore lesion observations which are below the mean graylevel of the non-lesion class – i.e. the data $\{y_{c=1} | y < E[y_{c=0}]\}$ are ignored; this is demonstrated in scenario f.

When the data from both classes overlap (scenarios a–c), the results with and without regularization are roughly the same, except for the strongest $\lambda = 10^{-1}$, which tends to have too much impact. This implies that regularization in these scenarios is unnecessary, and that the deviation from the MLE-fitted case ($\lambda = 0$) should be minimized, so as to avoid biasing the model. When the data from both classes do not overlap (scenarios d and e), $\lambda$ plays an important role in limiting the magnitude of $\beta$. Overall, $\lambda \in [10^{-3}, 10^{-2}]$ gives a good trade-off of reduction in logistic slope in scenarios d and e, and minimal impact in scenarios a–c.

4.5.3 Pseudo Lesion Regularization

Next, pseudo-lesion regularizations are explored, namely selection of the number of synthetic lesions $V$. Similar to above, only a subset of the scenarios are originally problematic, and in need of this
regularization; these are the last four: f–i, where no typical lesions have been observed. As before, the impact of the regularization should therefore be minimal on the other scenarios, a–e. Four selections of $V = [0, 1, 3, 9]$ are used to train different models, all with constant $V = y_{\text{max}} = 1$ and $\lambda = 10^{-3}$. These results, presented in the same way as before, are shown in Figure 4.8.

It can be seen that the inclusion of pseudo-lesions has no appreciable impact on the first 5 scenarios, as desired. In the problematic scenarios f–i, the most significant change occurs with the inclusion of the first pseudo-lesion ($V = 0 \rightarrow 1$). Larger values of $V$ have little impact, but act to move $\tau$ slightly lower. Therefore, the simple inclusion of one pseudo-lesion may be all that is required. Further investigations will explore this regularization with segmentation performance results using the full model.

Figure 4.8: Toy model MAP estimation results for 9 different scenarios and different numbers of pseudo-lesions $V$, shown as coloured diamonds corresponding to the scenario (spread of diamonds is for visualization only). Best viewed in colour.
4.6 Full Modal – Preliminaries

Before exploring the segmentation performance of the full algorithm, it is necessary to ensure that the model is converging during training, and that the cross validation framework is appropriate. It will also be helpful to establish a baseline model performance, for comparison with model variants later. These are the objectives of this section.

4.6.1 Convergence

The rate of convergence in each voxel will be unique. During parallel fitting, it is prudent to stop training after a maximum number of iterations, $t_{\text{max}}$, rather than wait for all voxels to achieve a certain stopping criterion, in case a few aberrant voxels do not converge. In order to determine this number, the model was fitted using all available data from Dataset A, including augmentations (reflection: $a_r = \text{true}$ and shift one voxel in each dimension: $a_s = N_6$) starting from the default initialization $\beta = [0, 0]^T$ for all voxels. The magnitude of $\Delta \beta$ (5% quantiles) was recorded for each fitting iteration and plotted, as shown in Figure 4.9.

Evidently, the majority of convergence occurs before the 15th iteration. Therefore, using a two-fold factor of safety, $t_{\text{max}}$ was defined as 30 for all subsequent experiments.
4.6.2 Cross Validation

In § 4.3, it was argued that the LOSO-CV framework gives a better estimation of generalization performance. Practically speaking, this usually equates to a lower estimated performance, since the other CV frameworks described above allow perception of test scanner characteristics within the training set, facilitating better performance. In this section, the full model is trained and tested under each of the described CV frameworks, in order to validate this assertion. For a fair comparison with LOSO-CV, the KF-CV condition was implemented using the same numbers of images in each fold. Moreover, to avoid variance associated with random image selection, the number of images assigned to each fold was guided by the expected value, as illustrated in Figure 4.10. The parameter selections for this version are summarized in Table 4.5. Finally, it will be assumed that the images have already been registered and transformed to MNI brain space (cf. § B.3.2 for details about this workflow).

The three performance metrics for each condition are summarized using box plots in Figure 4.11. The general trend in reported performance is as expected: LOSO-CV < KF-CV ≈ LOO-CV ≈ OSAAT-CV < No-CV. Ignoring the LL groupings (i.e. $N = 96$), a paired non-parametric statistical test (signrank in MATLAB) was used to test for significant differences among these conditions. The No-CV condition reported significantly higher performance in both $SI$ and $Re$, versus LOO-CV, KF-CV, and LOSO-CV, (6 of 6 comparisons). This demonstrates the capacity of this model to overfit, since training and testing on the same data yields better results than any scenario where the test data are not seen during training. The OSAAT condition gave consistently higher $Pr$, but lower $Re$ versus all other conditions, yielding
only significant differences in \( SI \) with LOSO-CV. This is most likely attributable to the smaller number of training examples, since the training set in each fold comprises only same-scanner images.

More importantly, the reported performance metrics were significantly higher under LOO-CV and KF-CV than under LOSO-CV, in all comparisons except \( Re \) in LOO-CV vs LOSO-CV (5 of 6 comparisons). This illustrates the potential overestimation of generalization performance using classic CV techniques, wherein scanner-specific characteristics of images in the test set are perceived during training. In reality, images from the use-case scanner are often not available for training, so the proposed LOSO-CV framework should be used to provide a better estimate of expected performance.

It is worth noting that the trend in differences is most significant among Precision results (Figure 4.11b). This implies that differences arise primarily from the number of false positives (cf. Equation (4.2)). One explanation for this result is that each scanner has a characteristic spatial distribution of hyperintense artifacts, which can be ignored once it is perceived during training.

In sum, these results support the discussion presented in § 4.3, which states that the LOSO-CV framework is the most challenging cross validation paradigm, giving the most realistic estimate of expected generalization performance on data from new scanners. Therefore, this framework as used throughout the remaining analysis of segmentation performance.
4.6.3 Baseline Model Performance

The next section will explore model variants which yield performance improvements; therefore, results from a minimal working algorithm are first presented for sake of comparison. The parameters of this version (“base”) is summarized in Table 3.2.

The fitted parameter images from the baseline model are shown in Figure 4.12. While the threshold image contains reasonable values (near $y_{max}$) in the regions typically containing WMH (Figure 4.1), there are obvious artifacts throughout both images, corresponding to locations where no lesions were observed in the training set. These voxels do not exhibit stable convergence properties without regularization, hence necessity of a static $t_{max}$. Classification results in any of these voxels will almost certainly be wrong; therefore, overcoming these artifacts is a priority during investigation of regularization strategies.

After training and testing this version of the model under LOSO-CV, the median performance metrics are summarized in Table 4.4; the same metrics are illustrated in Figure 4.13, stratified by LL tertiles. The overall median SI was 0.64, Precision 0.70, and Recall 0.70. Considering the artifacts in Figure 4.12, these results are surprisingly good, rivalling many of the reported performances in Table 1.4, which were obtained using much less challenging validation conditions. As is often the case, performance is correlated with LL, since voxel misclassifications have a larger effect when the number of positive examples is small.

With $Pr > Re$ for high LL, it can be inferred that the model incurs more FN than FP at high LL – i.e. it is more specific than sensitive. This would therefore predict an underestimation of the total LL in these subjects. The performance among the first two MS 2016 scanners is also low. This may be attributable
### Table 4.4: Baseline model performance metrics (median)

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.66</td>
<td>0.75</td>
<td>0.68</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.77</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.67</td>
<td>0.67</td>
<td>0.77</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.49</td>
<td>0.86</td>
<td>0.50</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.38</td>
<td>0.51</td>
<td>0.32</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.74</td>
<td>0.50</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.62</td>
<td>0.58</td>
<td>0.76</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td><strong>12</strong></td>
<td><strong>0.64</strong></td>
<td><strong>0.70</strong></td>
<td><strong>0.70</strong></td>
</tr>
</tbody>
</table>

![Box plots](image)

**Figure 4.13:** Baseline model performance, stratified by LL tertiles.

4.7 Full Model – Performance Results

Next, the full model is trained and tested under a large variety of different conditions. The results above are validated in terms of segmentation performance, and optimization of additional model components is explored similarly. Except where specified, the model will be trained and tested as per the LOSO-CV framework, using Dataset A. Additionally, while the above performance measures serve as a baseline, an optimal combination of parameters was eventually resolved; these parameters are summarized in Table 4.5, § 4.8. In many cases, the optimized parameters are essential for good model performance, so these are
used during exploration of other model components. This parametrization is denoted “default”, when compared against other model variants.

4.7.1 Graylevel Standardization

Five graylevel standardization techniques with promising results predicted by the objective functions were identified in § 4.4 (cf. Table 4.2). Each of these methods was applied to Dataset A, yielding the contrast characteristics shown in Figure 4.14. Next, the VLR model was trained and tested using these data, and the segmentation performance results were compared. Figure 4.15 compares the results under each condition, again using box plots stratified by LL tertiles.

While statistical standardization (SS) outperforms all other techniques for subjects with high LL, limitations in Re at low and medium LL resulted in worse performance overall. Histogram equalization (HE) was similarly afflicted by poor Pr at low and medium LL, yielding suboptimal SI performance. Two histogram matching operations, (HM1 and HM3) having higher contrast at the upper end of the graylevel range, were more successful in terms of segmentation performance. Considering both the slight advantage of HM3 over HM1 here, and the objective function results from § 4.4, HM3 was selected as the optimal method going forward.

These results can be rationalized using Figure 4.14, where differences in image contrast produce predictable trade-offs between Pr and Re. For example, while histogram equalization (HE) gives good lesion contrast, a large number of false positives are also typically incurred in the GM, decreasing Precision. Conversely, the optimal HM3 method maintains good lesion contrast, while minimizing GM/WM contrast.

4.7.2 Regularization

In this section, each of the regularization strategies presented in Chapter 2 are explored, particularly with respect to their impact on segmentation performance. These include:

- Pseudo-lesion regularization: V – number to include
- Parameter norm penalties: λ
- Data augmentation: ar – reflection; as – shift.

Parameter image smoothing is further explored in 4.7.3, though optimization of both components is a chicken-and-egg problem, since good regularizations are necessary to produce plausible parameter images (cf. obvious artifacts in Figure 4.12), while parameter image smoothing is similarly important. Therefore,
Figure 4.14: Simulated FLAIR images after graylevel standardization using each technique under investigation.
Figure 4.15: Comparison of the optimized model employing each graylevel standardization technique.
In order to characterize the contributions of each regularization technique independently, each was added, one-at-a-time, to the baseline model. This investigation did not use parameter image smoothing. The performance metrics under each condition are summarized in Figure 4.16.

Each of the regularization techniques yielded improvements in overall performance, as measured by Similarity Index. However, as conjectured in 3.2.1, data augmentation strategies were most successful in boosting performance, especially the shift augmentation. Recall was most improved (fewer FN) through inclusion of pseudo-lesions; this is as expected, since the no-lesion training voxels illustrated in Figure 3.2b maintain the ability to predict $\hat{c} > 0$ under this condition. This improvement came at the expense of a slight decrease in Precision. Conversely, Precision was greatly improved through parameter norm penalties, with an associated decrease in Recall. This implies that overfitting associated with MLE estimation most often results in False Positives, which are minimized through the use of $\lambda$.

The results in § 4.5.3 demonstrated that use of additional pseudo lesions ($V > 1$) did not have appreciable impacts on the fitted parameters (cf. Figure 4.8), so additional selections of $V$ are not presented here. Similarly, the reflection data augmentation is always helpful to include, but no further investigations are needed. Additional spatial data augmentations were similarly omitted for exploration, since neighbourhoods larger than $N_6$ (shifts larger than 1 voxel) become less plausible as training images with small registration errors. Therefore, only the $\lambda$ parameter was subject to further formal investigation in terms of segmentation performance.
Parameter Norm Penalties

Selection of the appropriate $\lambda$ was already partially explored in § 4.5.2, where it was determined that $\lambda \in [10^{-3}, 10^{-2}]$ provided a good trade-off between limiting the magnitude of $\beta$ and maintaining MLE characteristics. A similar range of $\lambda$ was explored in the full model: $[10^{-5}, 10^{-1}]$. Exploration for each selection this time employed the final optimized model parameters summarized in Table 4.5, in order to consider interactions between the different regularization strategies. Performance metric results are again summarized using box plots in Figure 4.17.

From these results, it can be seen that overall SI performance is surprisingly robust to the definition of $\lambda$. This may be attributable to the effects of other regularizations, especially the data augmentations and parameter image smoothing. However, a maximum in Similarity Index is achieved using $\lambda = 10^{-3}$; therefore, this value was selected as the optimal $\lambda$.

4.7.3 Parameter Images

Noting the obvious artifacts in Figure 4.12, generation of more plausible parameter images was a priority. This section explores additional filtering operations applied to the MAP estimated parameter images, whose aim was to both improve the segmentation performance and improve the qualitative plausibility of the resulting parameter images.
Figure 4.18: Parameter images following different smoothing filters. Best viewed in colour.

Smoothing

The regularizations described in § 4.7.2 were surprisingly effective at achieving these objectives, yielding the raw fitted parameter image shown in the left most column of Figure 4.18. However, several artifacts are visible, including voxels with very high magnitude in the sensitivity image $S(x)$, and large discontinuities between the voxels which do and do not observe lesion examples during training in the threshold image $T(x)$. Therefore, the smoothing filters proposed in Table 3.1 were each applied to the fitted parameter images in an attempt to correct these problems, yielding the remaining panels in Figure 4.18.

Performance differences among the different filters (Figure 4.19) were not large in magnitude. However, the Gaussian filter with $\sigma = 2$ MNI voxels (3 mm) achieved statistically higher performance than all other conditions except $G_{\sigma 1}$. This method also has the advantage of producing exceedingly smooth parameter images, which are less likely to contain artifacts associated with the training set. This advantage is contrasted with many other image filtering tasks in medicine, where maintenance of image edges or other details is often a priority.

Considering this result, one additional modification was made to the model estimation procedure. These details are presented in § B.3.3.

Interpretation

The final parameter images provide concise descriptions of the VLR model. The threshold image $T(x)$ indicates the graylevels corresponding to a 50% probability of the lesion class $\hat{c}$, while the sensitivity
image $S(x)$ describes the rate of change in predicted probability near the threshold. The regions of low threshold appear to align with the typical distribution of lesions (cf. Figure 4.1), permitting even small hyperintensities in these areas to be recognized as WMH. Conversely, lower threshold values are observed throughout the GM, and in areas of common false positives, facilitating their exclusion. The sensitivity image reflects the confidence of the model in the current prediction, and is often lower in regions of TP and FP overlap. For example, the border of the ventricles may contain hyperintensities due to WMH or flow through artifacts in dilated ventricles, and similarly, the corpus callosum is often bright, but inconsistently included by manual raters in the WMH segmentation.

Note that $S(x)$ is significantly less important for segmentation performance. One investigation which replaced this parameter image with its mean value saw only a 0.24 decrease in median $SI$. In fact, this approach mirrors the model proposed by Schmidt et al. in the LPA algorithm, since only the $\beta^0$ term is parameterized spatially. This partly validates the modelling decisions by Schmidt et al., though the advantages in estimability and performance afforded by the current approach are significant.

Comparison with LPA Spatial Effect Parameter

The inspiration for the current algorithm came from the LPA algorithm by Schmidt et al. In this method, the logistic regression is parametrized by only one spatial effect term: $\beta^0(x)$. This parameter was extracted from the toolbox\textsuperscript{6} and reconstructed in MNI space, for comparison with the equivalent VLR-fitted

\textsuperscript{6} The variable sp\_mni2\_Bf2 in the LST\_lpa\_stuff.mat datafile from the toolbox.
In order to facilitate visual comparison, the image means and variances were matched, yielding the results shown in Figure 4.20.

The two parameter images appear overall similar, with areas of larger magnitude reflecting the usual distribution of lesions, as in the regions of lower threshold in $T(x)$. The LPA parameter image is less detailed in most aspects, but occasionally contains sharp artifacts from the estimation procedure, which employs random sampling of spatial locations. The VLR parameter image is more detailed, perhaps due to weaker assumptions about smoothness, despite significant image filtering.

Another notable difference is that the VLR $\beta^0(x)$ is decreased in regions of the typical GM, particularly in the insula and along the mid-line, while the LPA image is not. This is because the LPA model uses SPM-estimated GM and WM tissue segmentations to apply tissue-specific graylevel standardization (type SS), and therefore assumes that all standardized graylevels in the image are derived from a single normal distribution. This approach was avoided in the VLR implementation, since the SPM tissue segmentation almost always misclassifies WMH as GM, resulting in erroneous standardization which decreases WMH contrast. WMH contrast decreases because both the mean and variance of the GM class are typically larger than that of the WM class; subtracting the larger mean and dividing by the larger standard deviation yields decreased WMH graylevels.

The seven VLR model $\beta^0(x)$ images from LOSO-CV folds using SS standardization were averaged, since this most closely approximates the standardization employed by the LPA algorithm.
Table 4.5: Model hyperparameters and optimized values.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Parameter</th>
<th>Notation</th>
<th>Type</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Processing</td>
<td>Reflect Augmentation</td>
<td>$a_r$</td>
<td>$\mathbb{B}$</td>
<td>true</td>
</tr>
<tr>
<td></td>
<td>Shift Augmentation</td>
<td>$a_s$</td>
<td>$\mathbb{N}_p$</td>
<td>$N_6$</td>
</tr>
<tr>
<td></td>
<td>Graylevel Transform</td>
<td>$\tau_g$</td>
<td>$f: \mathbb{R} \mapsto \mathbb{R}$</td>
<td>$\tau_{RM3}$</td>
</tr>
<tr>
<td></td>
<td>Transform Mask</td>
<td>$\mathcal{X}_\tau$</td>
<td>$\mathbb{B}(x)$</td>
<td>$\mathcal{X}_{brain}$</td>
</tr>
<tr>
<td>VLR Fitting</td>
<td>Iterations</td>
<td>$T$</td>
<td>$\mathbb{Z}$</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Initial $\beta$</td>
<td>$\beta(0)$</td>
<td>$\mathbb{R}^2$</td>
<td>$[0, 0]$</td>
</tr>
<tr>
<td></td>
<td>Estimation Scale</td>
<td>$r$</td>
<td>$\mathbb{R}$</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Learning Rate</td>
<td>$\alpha$</td>
<td>$\mathbb{R}$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Regularization</td>
<td>$\lambda$</td>
<td>$\mathbb{R}$</td>
<td>$1 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>Pseudo-Lesions</td>
<td>$\mathcal{V}(x)$</td>
<td>${ \cdot \in \mathbb{R} }$</td>
<td>${y_{max}}$</td>
</tr>
<tr>
<td></td>
<td>$\beta$ Filter</td>
<td>$F_{\beta}$</td>
<td>$f: \mathbb{R}(x) \mapsto \mathbb{R}(x)$</td>
<td>$\tilde{\beta}(x) = G_{\sigma_2}(\beta(x))$</td>
</tr>
<tr>
<td>Post-Processing</td>
<td>Min Lesion Size</td>
<td>$x_{\min}$</td>
<td>$\mathbb{R}$ (mm$^3$)</td>
<td>0</td>
</tr>
</tbody>
</table>

Notation. $\mathbb{B}$: boolean value; $\mathbb{Z}$: integer value; $\mathbb{R}$: real value; $\mathbb{R}^n$: vector; $\mathbb{R}(x)$: image; $\mathbb{N}_p$: nearest $p$ voxel neighbourhood.

### 4.8 Optimized Model Summary

Considering all experimental results, the optimal model hyperparameters were selected. These values are summarized in Table 4.5. Fitted parameter images from one LOSO-CV fold are also shown in Figure 4.21, and an example segmentation is shown in Figure 4.22.

#### 4.8.1 Segmentation Performance

This section explores more detailed segmentation performance results associated with the final model definition. Median overall $SI$ performance was 0.69, a reasonable improvement over the baseline of 0.64, and only 0.02 lower than the maximum possible performance of 0.71 using no cross validation. As with the baseline model, results are broken down by scanner in Table 4.6, where it can be seen that data from ISBI 2015 and MS 2016 (1) have been the major beneficiaries of model improvements. The same overall trends in scanner performance persist, however, likely due to representation imbalances, since there are only 5 images from each of the three MS 2016 scanners. The model is also still more precise than sensitive, particularly for high LL, as shown in Figures 4.23b and 4.24b. Several subjects even reach near 100% Precision. Conversely, little overall improvements in Recall were made during model optimization ($Re = 0.70$ to 0.71), and Recall performance even decreases for high LL. This is likely attributable to the histogram matching operation, which begins to attenuate the WMH in images with high LL, due to an implicit assumption that a consistent volume of hyperintensities will appear in the image.
Figure 4.21: Fitted parameter images $\mathcal{T}(x)$ and $\mathcal{S}(x)$ from the first LOSO-CV fold of the final model. Best viewed in colour.

Figure 4.22: Example segmentation. Best viewed in colour.

Table 4.6: Final model performance metrics (median)

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.69</td>
<td>0.87</td>
<td>0.65</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.81</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.68</td>
<td>0.70</td>
<td>0.77</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.55</td>
<td>0.89</td>
<td>0.47</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.60</td>
<td>0.32</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.85</td>
<td>0.47</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.70</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td>ALL</td>
<td>12</td>
<td>0.69</td>
<td>0.75</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Figure 4.23: Final model performance, stratified by LL tertiles.

(a) Similarity Index (SI)  
(b) Precision (Pr)  
(c) Recall (Re)

Figure 4.24: Scatter plot of final model performance, with 3rd order trend line and 90% confidence interval shown in grey. Best viewed in colour.

(a) Similarity Index (SI)  
(b) Precision (Pr)  
(c) Recall (Re)

Figure 4.25 shows the volume agreement between the manually segmented WMH and the VLR-estimated WMH. As predicted by the $Pr$ and $Re$ results, the model tends to underestimate the LL, with underestimation getting worse for very high LL. Again, this trend is likely a result of the histogram-matching graylevel standardization, since this histogram-based nonlinear transformation, acts to equalize the proportion of hyperintensity in every image. Therefore, in subjects with more hyperintensities, lesion contrast is reduced. This unfortunately led to a poor overall volume agreement, as measured by ICC: 0.71.

Finally, reflecting on the original motivations for including spatial features in the model (cf. Figure 1.4 in § 1.3.3), the distributions of TP, FP and FN from the LOSO-CV segmentations of the VLR model are presented again here in Figure 4.26. The distribution of FP is now limited to the same spatial regions as TP, implying that practically all the problematic regions of FP shown in Figure 1.4 have been managed though the spatial parametrization. In fact, the distributions of FP and FN are now very similar, suggesting that little more can be done with the current model to distinguish these classes. This conclusion is also corroborated by the No-CV segmentation performance results, which indicate a maximum possible performance of the current model. Potential methods of augmenting the model to solve this problem will be explored in the next chapter.
Figure 4.25: Bland-Altman plot showing total LL agreement between manual and VLR-segmented WMH. Shown in Log-scale to better illustrate results for small LL.

\[ A = 0.159 + 0.857 \cdot M \]
\[ R^2 = 0.923 \]

Figure 4.26: Distribution of True Positives (TP), False Positives (FP), and False Negatives (FN) from all LOSO-CV folds of the final model. Best viewed in colour.
Threshold Optimization

The initial output from the VLR model is a probabilistic lesion class image, which is thresholded at a probability $\pi$ for comparison with manual segmentation in performance analysis. Varying this threshold can trade sensitivity for specificity, and the relationship between these three parameters can be summarized in a Precision-Recall Curve. Figure 4.27a gives this result, where each profile represents the performance metrics for an image in the dataset, while varying the value of $\pi$ from 0 to 1. The area under this curve gives an indication of the overall performance of the model, and in this case, the value is 0.80. Similarly, Figure 4.27b illustrates the effect of threshold value $\pi$ on the overall Similarity Index, and implies an optimal value of $\pi$ where $SI$ is maximized. It is worth noting that the left-right skew of these profiles is attributable to LL, where lower LL favours a higher threshold to avoid false positives, whereas higher LL favours a lower threshold to avoid false negatives.

“Turing Test”

In order to determine whether the VLR algorithm produces results which are indistinguishable from other human raters, it was first necessary to establish a measure of human performance. To do so, the inter-rater agreement was calculated for those datasets having multiple manual segmentations: MS 2016 [83] and ISBI 2015 [57]. Since $SI$ is a true metric, the direction of comparison (test-to-standard or standard-to-test) does not matter. Therefore, the $SI$ can be computed between any two human raters.

The inter-rater $SI$ was calculated in all possible pair-wise comparisons among the 7 raters (7-choose-2 =
Table 4.7: Mean inter-rater agreement measures for manual WMH segmentation calculated for the available data.

<table>
<thead>
<tr>
<th>Ref</th>
<th>Dataset</th>
<th>Raters</th>
<th>Data</th>
<th>SI</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>[83]</td>
<td>MSSEG 2016</td>
<td>7</td>
<td>15 images</td>
<td>0.63±0.16</td>
<td>0.91</td>
</tr>
<tr>
<td>[57]</td>
<td>MS 2015 ISBI</td>
<td>2</td>
<td>21 images</td>
<td>0.73±0.10</td>
<td>0.98</td>
</tr>
</tbody>
</table>

21 total), then averaged, for all 15 available images in the MS 2016 dataset. The same procedure was repeated for the two raters, for all 21 images in the ISBI 2015 dataset. The ICC (cf. § 4.2) between segmented lesion volumes was also calculated. These results, summarized Table 4.7, are consistent with other reports in the literature (Table 1.2).

Next, non-parametric unpaired tests (ranksum in MATLAB) compared the human inter-rater $SI$ values ($n = 21, n = 315$) to the VLR-vs-human $SI$ values ($n = 96$), to test for significant differences. Differences were significant in the MS 2015 ISBI comparison ($p = 0.037$), but not for MS 2016 comparison ($p = 0.086$). This implies that the VLR algorithm was indistinguishable from the human raters in the MS 2016 dataset.

4.9 Comparison with Other Methods

While the selection of freely available data for most of this work permits direct replication of the validation conditions by future works, several existing algorithms have already been deployed for use by other researchers. It is therefore possible to compare the performance of these methods with the proposed VLR model directly.

4.9.1 Lesion Prediction Algorithm (LPA)

The LPA algorithm is the only freely available FLAIR-only WMH segmentation tool, and has been used in several comparisons with other methods [43, 133, 134].\(^8\) For this reason, the segmentation performance of the LPA algorithm was compared to the proposed algorithm under LOSO-CV. In order to binarize the probabilistic class images produced by the LPA algorithm, a user-defined threshold can be used. To reduce the bias of the comparison, this threshold is optimized for each of the same LOSO-CV folds as for the VLR algorithm. No other LPA parameters can be specified by the user.

Box plots comparing the segmentation performance metrics, stratified by LL are given in Figure 4.28,
and the Sign Rank test was again used to test for significant differences. The VLR algorithm easily outperforms its LPA forerunner in $SI$ at small and medium LL ($p < 0.001$), while differences at large LL were not significant. Overall, median $SI$ were 0.69 and 0.58, respectively, also significantly different ($p < 0.001$). The VLR algorithm was also more precise at high lesion loads than the LPA algorithm ($p < 0.001$), but overall had lower recall, mainly due to significant differences at high LL ($p < 0.001$).

These results might be explained by the parameter images shown in Figure 4.20: while the VLR algorithm learns to exclude potential FP in the GM by spatial location alone, the LPA algorithm maintains sensitivity to hyperintensities in these areas. As a result, the VLR algorithm is very precise, at the expense of sensitivity, while the LPA algorithm is able to detect more peripheral lesions, sacrificing precision.

### 2017 WMH Segmentation Challenge Results

The proposed VLR method was submitted to the WMH Segmentation Challenge at MICCAI 2017. The available training data included T1 and FLAIR MRI from 60 subjects and 3 different scanners (20+20+20), while the testing data comprised 110 total subjects from 5 different scanners (30+30+30+10+10). A total of 20 teams participated, and teams were scored using a combination of the following 5 performance metrics:

- Similarity Index – $dsc$
- Hausdorff distance (modified, 95th percentile) $[199]$ – $h95$
- Percent volume difference – $avd$
- Recall for individual lesions – $recall$
- Similarity Index for individual lesions – $f1$
Scores for “individual lesions” count each set of connected ($N_{26}$) voxels – i.e. one “lesion” – as an single observation, which can be classified as TP if there is at least one voxel of overlap with the manual segmentation, FP if a predicted lesion has no corresponding voxels in the manual segmentation, and FN if a manual lesion has no corresponding voxels in the predicted lesion. Each of the 5 metrics are averaged across all 110 test subjects, and the overall score considers an average of all 5 metrics after scaling by the range of minimum and maximum scores achieved by the 20 challengers; this score is $\in [0,1]$ where lower is better.\(^9\)

The VLR method achieved an average SI performance of 0.70 on the test data, which is actually slightly higher than the LOSO-CV predicted performance in § 4.8. Using the challenge metric scaling, this represents a relative score of 82.3%, ranking the VLR method 8\(^{th}\) in this dimension. Other metrics did not look so favourably on the proposed method, including lesion-wise recall ($\text{recall} = 0.25$), where VLR ranked dead last. Considering these metrics in the overall ranking, the VLR method ranked only 15\(^{th}\) of 20 teams, with a score of 0.4159. The challenge performance report provided by the competition organizers is given in Figure 4.29.\(^{10}\)

At the MICCAI 2016 MSSEG Competition, only 4 of the 15 submitted methods used deep learning, while the 2017 competition saw 15 of 20 methods use this approach, including the top performing 13 methods. This impressive and sudden display of model dominance should not be taken for granted, especially considering the large number of previously proposed non-deep WMH segmentation methods (cf. § 1.3.2). The top performing non-deep method (team “tig” 14\(^{th}\) place, overall score of 0.3858) uses an adaptation of the unsupervised unified mixture model described in [200]. This method outperforms the VLR submission in both lesion-wise metrics ($\text{recall} = 0.38$ vs 0.25, and $f1 = 0.42$ vs 0.35), but performs worse in mean SI (0.60 vs 0.70), and additionally requires both T1 and FLAIR images.

These results also highlight a major and perhaps flawed assumption used throughout the current work: that optimizing “segmentation performance” is equivalent to maximizing the Similarity Index with manual segmentations. In fact, diagnostic criteria considering WML often focus instead on identification of new lesions in different spatial locations [41, 52]. Limitations such as this will be further discussed in the next chapter.

\(^9\) For more information, see [http://wmh.isi.uu.nl/evaluation/](http://wmh.isi.uu.nl/evaluation/).
\(^{10}\) Detailed results and competitor methods descriptions are available at: [http://wmh.isi.uu.nl/results/](http://wmh.isi.uu.nl/results/).
Figure 4.29: Results report for the submitted method provided by the WMH Segmentation Competition.
Chapter 5

Conclusion

This chapter concludes the thesis. A summary of the major contributions will be given. Then, an analysis of limitations of this work will help define a roadmap for follow-on research.

5.1 Summary of Contributions

This thesis explored the task of automated white matter hyperintensity segmentation, which aims to improve the speed and precision over manual analysis. After reviewing the motivation for this task, and the previously proposed solutions, a novel segmentation algorithm – Voxel-Wise Logistic Regression – was proposed in Chapter 1. This algorithm, the necessary regularizations, and pre- and post-processing components were then developed throughout Chapters 2 and 3. Chapter 4 then presented extensive validation of the model and parameter optimization, including a critical analysis of the popular cross validation frameworks. In this section, the major contributions and conclusions are summarized.

5.1.1 Algorithm Validation

One of the major limitations of previous works in this area is the use of validation conditions which overestimate the segmentation performance on images from sources that were not seen during training. These conditions include a small number of different image sources, and the use of training data which comes from the same source as the test data, an unrealistic condition for many naive algorithm use cases. In fact, this criticism likely applies to validation of solutions in many different image analysis tasks, especially MRI. However, this criticism does assume that the goal of these projects is to develop algorithms for
general-purpose use – i.e. for segmentation of images from any given source. It may, in some cases, be possible and preferable to leverage existing training data for a constant image source to optimize the algorithm parameters more specifically to these characteristics.

In the current work, the Leave-One-Source-Out Cross Validation (LOSO-CV) framework was presented, a rediscovery of the “Multi-Source” Cross Validation procedure described by Geras and Sutton in [198]. Experimental results showed how other frameworks like Leave-One-Out (LOO) and K-Fold (KF) Cross Validation estimate higher segmentation performance than LOSO-CV. Imprudent use of such frameworks could therefore lead to premature adoption of particular automated WMH segmentation algorithms, or overconfidence in their results.

Another fault in previously employed validation frameworks is the small number of images and different image sources. Only 5 of the 54 reviewed methods use more than 3 scanners for validation. Considering the many aspects of image variability, including MR imaging parameters, voxel size, subject anatomy, scanner field strength, noise and bias field characteristics (cf. § 1.2.2), validation of any brain MRI analysis algorithm should employ data from more sources which better reflect this variation. Fortunately, as illustrated here, there are now at least 96 free labelled FLAIR (and T1) image sets (Table 4.1) which can be used for validation of WMH segmentation algorithms. Future works in this area are encouraged to follow the lead of the machine learning community and adopt standardized datasets, facilitating direct comparison of reported performance across publications. While the numbers of images from the 7 scanners used in this work are imbalanced (20, 20, 20, 5, 5, 5, 21), these data still serve as a good starting point for establishing such a dataset.

5.1.2 Voxel-Wise Logistic Regression

Many of the previously proposed WMH segmentation algorithms use some sort of thresholding technique, mapping a single graylevel feature to the class probabilities. This approach is attractive for FLAIR-only methods, since it mimics the basic decision making process by human experts. However, preliminary investigations (Figure 1.4) showed how such techniques can incur a large number of False Positives and False Negatives, even if the optimal threshold is used.1

One circumvention to this problem proposed by Schmidt was to include a “spatial effect” parameter $\beta^0(x)$, which expands the feature space from one to four dimensions: FLAIR graylevel $y$, and spatial coordinates $x_1, x_2, x_3$. However, fitting this model was computationally expensive, and required various

---

1 The threshold which maximizes Similarity Index with the manual segmentation.
sampling approximations and assumptions about smoothness. The Voxel-Wise Logistic Regression model overcomes these challenges by relaxing the assumption that all voxels are equally sensitive to the FLAIR feature – i.e. that $\beta^1$ is constant. By doing so, all parameters may vary spatially: $\beta \rightarrow \beta(x)$, yielding independent parameters for all voxels, and the following predictive model:

$$P(c(x) = 1 \mid y(x), \beta(x)) = \frac{1}{1 + e^{-\eta(x)}}, \quad \eta = \beta(x)^T Y(x). \quad (5.1)$$

In this work, a Newtonian algorithm for estimating the parameters $\beta(x)$ is presented, maintaining generality for any number of input features. However, the implementation explored here uses only the FLAIR graylevels, yielding only two parameter images, and facilitating the efficient parallel fitting described in § B.3.1. These images can then be reparametrized via (3.11) to yield a threshold image $\mathcal{T}(x)$ and a sensitivity image $\mathcal{S}(x)$.

As explored in § 4.7.3, these parameter images agree with prior knowledge about the expected class discrimination function. In particular, the threshold image illustrates a spatially varying graylevel decision boundary corresponding to $\hat{c}(x) = 0.5$. As one might expect, this boundary is higher in spatial regions of common false positives (e.g. the cortical gray matter and corpus callosum) and lower in regions of common false negatives (e.g. the periventricular white matter). Similarly, the sensitivity image illustrates the separability of the classes, where lower values arise in voxels with overlapping class distributions.

**Pre-Processing**

The VLR model assumes that both spatial and graylevel features are comparable across input images, which is not true for raw MRI. Standardization of these features, therefore, is the main objective of pre-processing steps. Graylevel standardization is achieved using a histogram matching operation. Specifically, it was found that a target histogram with high contrast in the upper intensity range helped improve segmentation performance. For standardization of spatial coordinates during model estimation, images are registered and resampled to the MNI template brainspace using the *New Segment* feature in SPM. This confers the additional benefit of estimating and correcting any bias field artifact. At test time, the inverse transform is applied to map the estimated parameter images to the native subject space for inference.
Regularization

Estimating independent models for every voxel drastically reduces the number of observations available to estimate each parameter. This yields several challenges, namely MLE-fitted parameters which contradict prior knowledge about the problem, as discussed in cf. 3.1.4. To address these challenges, several regularization strategies were explored. The most successful technique, as in many other problem domains, was data augmentation (cf. § 4.7.2). In this work, this consisted of reasonable image transformations applied to training image sets, including reflection about the midline, and shifting by one voxel (1.5 mm) in MNI space. In addition, a zero-mean Gaussian prior on the values of $\beta$ was used, with variance proportional to a tunable parameter $\lambda$. In order to deal with the unique challenge of an unobserved class (the $c = 1$ lesion class) in many of the peripheral voxels, pseudo-lesions – deterministic synthetic data points – were appended to the training set. This approach is really just dataset balancing, and can be employed in other supervised segmentation algorithms. Finally, parameter images can be smoothed after estimation, using any simple filter; performance was not particularly sensitive to the selection (cf. Figure 4.19 in § 4.7.3).

5.2 Future Work

Despite significant segmentation performance gains over its LPA forbearer (cf. 4.9.1), and performance reaching human level (cf. 4.8.1), there are several limitations to the proposed VLR model. Chiefly, when training and testing the proposed algorithm on the exact same data, Similarity Index still only reaches 0.71 (cf. § 4.6.2). This demonstrates a significant ceiling on performance which likely cannot be overcome by the regularizations explored in this work: L2-priors on $\beta$, pseudo-lesions, and parameter image smoothing.

5.2.1 Model Variants

Potential solutions to this challenge, which might be the subject of future work, include improved graylevel standardization, improved image registration, and inclusion of additional features (graylevel or otherwise). Improved graylevel standardization is perhaps the most promising, since there is evidence to suggest that the histogram matching method employed here was suboptimal, particularly for subjects with large LL. For example, Recall consistently declined for very high LL (e.g. Figure 4.24c), as discussed in § 4.8.1. The extent to which registration inaccuracies affected the performance of the VLR model was not significantly explored. However, as suggested in § 2.1 and [162], many image registration algorithms have been proposed which may align anatomical structures better than the SPM New Segment method used here.
As outlined in Chapter 1, additional graylevel features were not explored in this work because the aim was to develop a FLAIR-only WMH segmentation algorithm. Considering the performance limitations with the current model, it might be tempting to include graylevel features from other MRI sequences, such as T1 and T2. However, as noted in § 1.3.3, the WMH class is not monotonic in T1 or T2 intensities, which makes their inclusion on the logistic regression model improper. Similarly, texture features, derived from raw MRI sequences could possibly be explored, but the inclusion of many features significantly erodes the model interpretability, and there is little evidence to suggest that WMH have unique texture characteristics.

Alternatively, the classification model (currently “logistic regression”) could be replaced by a more complex, non-linear model in order to incorporate additional features, such as a multilayer perceptron [201]. The contributions of the VLR paradigm to such a model would then derive from explicit spatial parametrization of the model parameters, though this would imply a drastic increase in their number. This approach is contrasted with methods which employ spatial features (e.g. coordinate values) alongside graylevel features within the segmentation model [135], and could afford improvements through differential regularization of conventional and spatial features.

### 5.2.2 Open Sourcing

The VLR model may be of interest to other researchers, either as a clinical research tool for segmenting WMH in large numbers of FLAIR images, or for future development and improvement, as outlined above. For this reason, the VLR algorithm will be released in two forms. First, all associated MATLAB code has been released at the following public GitHub repository: https://github.com/uoguelph-mlrg/vlr. This will permit inspection, re-training, and modification of the model by other medical imaging researchers. Second, the inference component, including “pre-trained” parameter images from all available data described here, will be packaged as part of a 3D Slicer module, providing a graphical user-interface for running the algorithm without programming experience or access to MATLAB.

### 5.2.3 Deep Learning

Finally, the poor performance of the proposed model relative to the deep learning approaches from the 2017 WMH Segmentation Competition (cf. § 4.9.2), especially the U-Net architecture, should not be ignored. Convolutional neural networks and the VLR model really occupy opposite ends of the model
complexity spectrum. The VLR model achieves good performance with relatively little fine-tuning, since prior knowledge of the problem is incorporated in the model itself. This also provides the main advantage of the VLR model – that it is not a “black box”, that the parameters of the model illustrate exactly how the classification probability is assigned. However, as shown here, these “strongly-biased” models often eventually reach a limit in performance due to invalid assumptions. For example, the VLR model assumes that in a given spatial location, FLAIR graylevel alone is sufficient to discriminate the WMH class from healthy tissues; however, in spatial locations where both bright GM and WMH are sometimes observed due to anatomical variability, this assumption fails. Instead, more complex contextual information may actually be needed to make this discrimination.

Deep learning approaches have the capacity to learn such features much more efficiently than human-guided methods, even at several scales. Issues associated with learning the complex image-to-image mappings with such models have recently been overcome with considerable success, and these methods have clearly emerged as dominant in other image analysis tasks [202]. The results of the 2017 WMH Segmentation Challenge suggest that this task too might be added to their palmares.

Therefore, the proposed VLR method may just go the way of the dodo. The VLR method is, in fact, quite inflexible to application in any other task, due to constraints about feature monotonicity with the lesion class, and only two modelled classes. This is not to say that the work presented here is all for naught. Many of the regularization strategies explored here, including pseudo-lesion synthetic data, parameter constraints (typically already used), and preprocessing methods for ensuring consistent distributions of input data, might improve performance of convolutional neural networks. Indeed, all of the top-performing methods in the 2017 WMH Challenge already employed standardization of images in space and graylevel dimensions. Moreover – perhaps more importantly – validation of these networks can and should make use of the LOSO-CV framework presented here.
Bibliography


Mark Jenkinson et al. “Improved optimization for the robust and accurate linear registration and motion correction.”


Massachusetts General Hospital Center for Morphometric Analysis. *Internet Brain Segmentation Repository (v2.0).* Massachusetts, 2003. URL: www.nitrc.org/projects/ibsr.


Appendix A

Additional Results

This section presents various results which are not essential to the thesis.

A.1 FLAIR MRI Intensity Modelling

While MR imaging is both complex and mutable, simulation of the expected signal intensities is possible using the relaxometry data in Table 1.1 and sequence signal equations – e.g. (1.3) and (1.4). In practice, this simulation helps select appropriate acquisition parameters $TE/TR/TI$ for the desired contrast; however, these characteristics can also be later considered as covariates in performance analysis of segmentation tools. To this end, Equations (1.3) and (1.4) were used with the relaxometry data from Table 1.1 to calculate expected tissue intensities and $WMH$ contrasts. Nine sets of acquisition parameters were taken from the FLAIR image database (Table 4.1), in addition to one simulated T1 image ($TE/TR = 5/15\text{ ms}$) and one simulated T2 image ($TE/TR = 100/5500\text{ ms}$). These results are summarized in Table A.1. In Figure A.1, an example image is shown for each parameter set using the tissue maps from the BrainWeb database [191], while in Figure A.3, the PMF for each tissue class from the same data are given.

1 http://brainweb.bic.mni.mcgill.ca/
Figure A.1: Simulated FLAIR images using scan parameters from the experimental database. Colourmap is arbitrary but consistent. Best viewed in colour.
Table A.1: Simulated FLAIR tissue intensities and WMH contrasts using scan parameters from the experimental database. Tissue intensities are normalized to the WM value.

<table>
<thead>
<tr>
<th>Scanner</th>
<th>GM</th>
<th>WM</th>
<th>CSF</th>
<th>WMH</th>
<th>WMH_{GM}</th>
<th>WMH_{WM}</th>
<th>WMH_{CSF}</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>1.21</td>
<td>1.00</td>
<td>0.26</td>
<td>1.67</td>
<td>1.38</td>
<td>1.67</td>
<td>6.32</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>1.11</td>
<td>1.00</td>
<td>0.05</td>
<td>1.28</td>
<td>1.16</td>
<td>1.28</td>
<td>25.29</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>1.14</td>
<td>1.00</td>
<td>0.01</td>
<td>1.49</td>
<td>1.31</td>
<td>1.49</td>
<td>105.74</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>1.32</td>
<td>1.00</td>
<td>0.00</td>
<td>3.40</td>
<td>2.57</td>
<td>3.40</td>
<td>∞</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>1.28</td>
<td>1.00</td>
<td>0.00</td>
<td>3.05</td>
<td>2.38</td>
<td>3.05</td>
<td>∞</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>1.38</td>
<td>1.00</td>
<td>0.08</td>
<td>4.08</td>
<td>2.96</td>
<td>4.08</td>
<td>48.09</td>
</tr>
<tr>
<td>MS 2008 UNC</td>
<td>1.21</td>
<td>1.00</td>
<td>0.53</td>
<td>1.67</td>
<td>1.38</td>
<td>1.67</td>
<td>3.17</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>1.13</td>
<td>1.00</td>
<td>0.29</td>
<td>1.30</td>
<td>1.15</td>
<td>1.30</td>
<td>4.51</td>
</tr>
<tr>
<td>In-House</td>
<td>1.21</td>
<td>1.00</td>
<td>0.53</td>
<td>1.67</td>
<td>1.38</td>
<td>1.67</td>
<td>3.17</td>
</tr>
<tr>
<td>T1 e.g.</td>
<td>0.68</td>
<td>1.00</td>
<td>0.21</td>
<td>0.54</td>
<td>0.79</td>
<td>0.54</td>
<td>2.50</td>
</tr>
<tr>
<td>T2 e.g.</td>
<td>1.29</td>
<td>1.00</td>
<td>3.25</td>
<td>1.89</td>
<td>1.47</td>
<td>1.89</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Figure A.2: Simulated T1 (TE/TR = 5/15 ms) and T2 (TE/TR = 100/5500 ms) images. Best viewed in colour.
Figure A.3: PMF of each tissue from simulated FLAIR, T1, and T2 images. The PMF of the WMH class is scaled by 25 for visibility. Best viewed in colour.
A.2 Histogram Matching vs Histogram Equalization

In § 2.3 it was argued that histogram matching is equivalent to histogram equalization in terms of effectiveness at standardizing graylevels in heterogeneous input images. This is because the histogram matching operation is defined as the function composition of the histogram equalization transform of the input image, \( F_y \), and the inverse equalization transform for the desired output histogram, \( F_{\tilde{y}}^{-1} \),

\[
\tau(y) = F_{\tilde{y}}^{-1}(F_y(y)).
\]  

(A.1)

The second transformation in the cascade does not depend on \( Y \), and so it is applied equally to images. Therefore the choice of target PMF is not important for the objective of graylevel standardization. This result is verified experimentally using synthetic images. Four 100 × 100 × 100 images were created to have the following density functions \( f_y \),

- **Uniform:** \( y \sim \mathcal{U}(y_{\text{min}} = 0, y_{\text{max}} = 1) \)
- **Unimodal:** \( y \sim \mathcal{N}(\mu = 0.5, \sigma = 0.08) \)
- **Bimodal:** \( y \sim (0.5 \mathcal{N}(\mu = 0.3, \sigma = 0.05) + 0.5 \mathcal{N}(\mu = 0.7, \sigma = 0.05)) \)
- **Trimodal:** \( y \sim (0.3 \mathcal{N}(\mu = 0.25, \sigma = 0.05) + 0.4 \mathcal{N}(\mu = 0.5, \sigma = 0.05) + 0.3 \mathcal{N}(\mu = 0.75, \sigma = 0.05)) \)

All four images were then histogram-matched to each of the respective density functions, with the aim of increasing agreement of image intensities. This agreement can be approximated by the alignment of intensity quantiles, since this is the target of histogram matching operations. As shown in Figure A.4, the quantiles agree almost perfectly in every output image, regardless of the choice of output PMF.
Figure A.4: Histogram matching of synthetic data to different target histograms. Quantiles show high agreement regardless of the target histogram. Best viewed in colour.
## A.3 Segmentation Performance Data

This section presents the raw performance data for all model iterations used in the thesis, which compare various regularization strategies and cross validation frameworks. Median values shown throughout.

**Baseline**

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.66</td>
<td>0.75</td>
<td>0.68</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.77</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.67</td>
<td>0.67</td>
<td>0.77</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.49</td>
<td>0.86</td>
<td>0.50</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.38</td>
<td>0.51</td>
<td>0.32</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.74</td>
<td>0.50</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.62</td>
<td>0.58</td>
<td>0.76</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.64</td>
<td>0.70</td>
<td>0.70</td>
</tr>
</tbody>
</table>

**Final**

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.69</td>
<td>0.87</td>
<td>0.65</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.81</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.68</td>
<td>0.70</td>
<td>0.77</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.55</td>
<td>0.89</td>
<td>0.47</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.60</td>
<td>0.32</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.85</td>
<td>0.47</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.70</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.69</td>
<td>0.75</td>
<td>0.71</td>
</tr>
</tbody>
</table>

**$\lambda = 10^{-3}$**

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.66</td>
<td>0.77</td>
<td>0.66</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.77</td>
<td>0.80</td>
<td>0.69</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.64</td>
<td>0.69</td>
<td>0.73</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.61</td>
<td>0.92</td>
<td>0.47</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.42</td>
<td>0.68</td>
<td>0.31</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.59</td>
<td>0.84</td>
<td>0.46</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.66</td>
<td>0.69</td>
<td>0.68</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.66</td>
<td>0.74</td>
<td>0.66</td>
</tr>
</tbody>
</table>

**$V = 1$**

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.66</td>
<td>0.75</td>
<td>0.67</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.76</td>
<td>0.77</td>
<td>0.75</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.65</td>
<td>0.63</td>
<td>0.79</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.51</td>
<td>0.81</td>
<td>0.51</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.38</td>
<td>0.48</td>
<td>0.34</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.62</td>
<td>0.73</td>
<td>0.51</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.62</td>
<td>0.56</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.64</td>
<td>0.67</td>
<td>0.71</td>
</tr>
</tbody>
</table>

**$a_R = 1$**

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.68</td>
<td>0.77</td>
<td>0.68</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.78</td>
<td>0.77</td>
<td>0.76</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.67</td>
<td>0.68</td>
<td>0.78</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.50</td>
<td>0.86</td>
<td>0.52</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.39</td>
<td>0.50</td>
<td>0.33</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.75</td>
<td>0.49</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.64</td>
<td>0.59</td>
<td>0.79</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.66</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Scanner</td>
<td>LL</td>
<td>SI</td>
<td>Pr</td>
<td>Re</td>
</tr>
<tr>
<td>---------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.69</td>
<td>0.84</td>
<td>0.66</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.80</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.68</td>
<td>0.71</td>
<td>0.79</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.54</td>
<td>0.89</td>
<td>0.50</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.55</td>
<td>0.32</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.62</td>
<td>0.81</td>
<td>0.50</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.66</td>
<td>0.65</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.67</td>
<td>0.73</td>
<td>0.69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.63</td>
<td>0.92</td>
<td>0.50</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.77</td>
<td>0.92</td>
<td>0.70</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.69</td>
<td>0.80</td>
<td>0.71</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.51</td>
<td>0.90</td>
<td>0.35</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.40</td>
<td>0.79</td>
<td>0.29</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.55</td>
<td>0.90</td>
<td>0.39</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.74</td>
<td>0.83</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.70</td>
<td>0.86</td>
<td>0.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.61</td>
<td>0.94</td>
<td>0.47</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.78</td>
<td>0.93</td>
<td>0.67</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.73</td>
<td>0.87</td>
<td>0.66</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.46</td>
<td>0.94</td>
<td>0.30</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.92</td>
<td>0.27</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.58</td>
<td>0.88</td>
<td>0.44</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.74</td>
<td>0.89</td>
<td>0.65</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.70</td>
<td>0.89</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.79</td>
<td>0.73</td>
<td>0.77</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.78</td>
<td>0.76</td>
<td>0.81</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.64</td>
<td>0.61</td>
<td>0.83</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.52</td>
<td>0.82</td>
<td>0.57</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.71</td>
<td>0.29</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.58</td>
<td>0.51</td>
<td>0.56</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.51</td>
<td>0.94</td>
<td>0.34</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.61</td>
<td>0.77</td>
<td>0.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.69</td>
<td>0.87</td>
<td>0.65</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.81</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.68</td>
<td>0.70</td>
<td>0.77</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.55</td>
<td>0.89</td>
<td>0.47</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.60</td>
<td>0.32</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.85</td>
<td>0.47</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.70</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.69</td>
<td>0.75</td>
<td>0.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.65</td>
<td>0.93</td>
<td>0.52</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.77</td>
<td>0.91</td>
<td>0.71</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.69</td>
<td>0.81</td>
<td>0.71</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.52</td>
<td>0.91</td>
<td>0.35</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.71</td>
<td>0.29</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.54</td>
<td>0.90</td>
<td>0.38</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.73</td>
<td>0.82</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.70</td>
<td>0.86</td>
<td>0.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.65</td>
<td>0.93</td>
<td>0.53</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.78</td>
<td>0.91</td>
<td>0.72</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.71</td>
<td>0.80</td>
<td>0.72</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.54</td>
<td>0.91</td>
<td>0.37</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.42</td>
<td>0.72</td>
<td>0.30</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.56</td>
<td>0.90</td>
<td>0.41</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.74</td>
<td>0.82</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.71</td>
<td>0.86</td>
<td>0.65</td>
</tr>
<tr>
<td>Scanner</td>
<td>LL</td>
<td>SI</td>
<td>Pr</td>
<td>Re</td>
</tr>
<tr>
<td>---------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.69</td>
<td>0.89</td>
<td>0.62</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.80</td>
<td>0.84</td>
<td>0.77</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.67</td>
<td>0.68</td>
<td>0.77</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.52</td>
<td>0.86</td>
<td>0.47</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.59</td>
<td>0.33</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.60</td>
<td>0.84</td>
<td>0.46</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.69</td>
<td>0.70</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.69</td>
<td>0.76</td>
<td>0.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.69</td>
<td>0.87</td>
<td>0.65</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.81</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.68</td>
<td>0.70</td>
<td>0.77</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.55</td>
<td>0.89</td>
<td>0.47</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.60</td>
<td>0.32</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.85</td>
<td>0.47</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.70</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.69</td>
<td>0.75</td>
<td>0.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.69</td>
<td>0.87</td>
<td>0.65</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.81</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.68</td>
<td>0.70</td>
<td>0.77</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.55</td>
<td>0.89</td>
<td>0.47</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.60</td>
<td>0.32</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.85</td>
<td>0.47</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.70</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.69</td>
<td>0.75</td>
<td>0.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.68</td>
<td>0.88</td>
<td>0.66</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.80</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.68</td>
<td>0.70</td>
<td>0.78</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.55</td>
<td>0.89</td>
<td>0.48</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.42</td>
<td>0.60</td>
<td>0.33</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.85</td>
<td>0.47</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.71</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.69</td>
<td>0.76</td>
<td>0.70</td>
</tr>
</tbody>
</table>
\[
\lambda = 10^{-1}
\]

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.64</td>
<td>0.77</td>
<td>0.68</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.79</td>
<td>0.80</td>
<td>0.76</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.64</td>
<td>0.67</td>
<td>0.76</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.65</td>
<td>0.93</td>
<td>0.48</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.42</td>
<td>0.69</td>
<td>0.31</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.90</td>
<td>0.46</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.70</td>
<td>0.75</td>
<td>0.76</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.68</td>
<td>0.75</td>
<td>0.70</td>
</tr>
</tbody>
</table>

\[
\sigma = 1
\]

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.69</td>
<td>0.86</td>
<td>0.61</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.80</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.68</td>
<td>0.72</td>
<td>0.80</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.51</td>
<td>0.87</td>
<td>0.50</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.40</td>
<td>0.55</td>
<td>0.32</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.80</td>
<td>0.48</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.66</td>
<td>0.62</td>
<td>0.79</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.67</td>
<td>0.74</td>
<td>0.71</td>
</tr>
</tbody>
</table>

\[
\sigma = 2
\]

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.70</td>
<td>0.84</td>
<td>0.63</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.80</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.68</td>
<td>0.71</td>
<td>0.80</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.52</td>
<td>0.88</td>
<td>0.49</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.40</td>
<td>0.56</td>
<td>0.32</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.61</td>
<td>0.81</td>
<td>0.48</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.67</td>
<td>0.64</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.68</td>
<td>0.74</td>
<td>0.70</td>
</tr>
</tbody>
</table>

\[
\sigma = 3
\]

<table>
<thead>
<tr>
<th>Scanner</th>
<th>LL</th>
<th>SI</th>
<th>Pr</th>
<th>Re</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMH 2017 (1)</td>
<td>24</td>
<td>0.68</td>
<td>0.85</td>
<td>0.63</td>
</tr>
<tr>
<td>WMH 2017 (2)</td>
<td>17</td>
<td>0.79</td>
<td>0.82</td>
<td>0.76</td>
</tr>
<tr>
<td>WMH 2017 (3)</td>
<td>6</td>
<td>0.65</td>
<td>0.66</td>
<td>0.77</td>
</tr>
<tr>
<td>MS 2016 (1)</td>
<td>29</td>
<td>0.53</td>
<td>0.88</td>
<td>0.45</td>
</tr>
<tr>
<td>MS 2016 (2)</td>
<td>5</td>
<td>0.41</td>
<td>0.60</td>
<td>0.31</td>
</tr>
<tr>
<td>MS 2016 (3)</td>
<td>10</td>
<td>0.59</td>
<td>0.86</td>
<td>0.45</td>
</tr>
<tr>
<td>ISBI MS 2015</td>
<td>5</td>
<td>0.70</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>12</td>
<td>0.69</td>
<td>0.76</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Appendix B

Implementation

This appendix contains implementation details.

B.1 Computing

All computation in the current work was performed using the following workstation and software:

- **CPU:** Intel Core i7-6700K 4.00 GHz
- **RAM:** 16.0 GB DDR4
- **GPU:** NVIDIA GeForce GTX 980 Ti
- **OS:** Windows 10
- **Code:** MATLAB R2011a

B.2 Manual Segmentations

It was necessary to create and edit a small number of binary segmentation masks during this work. To do this, the Editor module from the 3D Slicer imaging platform [203] was used,\(^1\) including the Wand, Paint, and Erase functions. Figure B.1 shows the user interface during a lesion segmentation.

\(^1\) 3D Slicer Editor tool documentation is available here: https://www.slicer.org/wiki/Documentation/4.6/Modules/Editor.
Figure B.1: 3D Slicer user interface for performing in-house manual segmentations and revisions. The tools used are highlighted in yellow, while the in-progress segmentation is shown in blue. Best viewed in colour.

B.2.1 MS 2008 WMH Masks

Since the reported performance of an automatic segmentation algorithm depends on the manual segmentations to which it is compared, it is important to obtain good manual segmentations. Unfortunately, the original manuals in the MS 2008 Segmentation Challenge contained obvious artifacts and inconsistencies, as shown at left in Figure B.2. Therefore, it was deemed necessary to redo these manuals. The resulting revisions are shown at right in Figure B.2.

B.2.2 Brain Mask

In order to vectorize image data for parallel processing, a binary mask selecting voxels of interest in standardized space is also required. Since only voxels in the brain are of interest, this mask is called a “brain mask”. The brain mask used here was derived from the ICBM tissue prior images [86] in MNI space: after initial thresholding of the combined GM + WM + CSF probabilities at 0.5, manual refinements were completed and symmetry was enforced. The mask is slightly small on purpose, since tissues outside the brain are frequently bright in FLAIR images, and can be mistaken for lesions by naive models. The resulting mask is shown in Figure B.3.
Figure B.2: Example revisions to the manual segmentations for the MS 2008 challenge dataset. Best viewed in colour.
Figure B.3: Manually refined brain mask in MNI space, overlaid on a simulated BrainWeb FLAIR image. Mask outline is highlighted in red; inclusions are shown in grayscale; exclusions tinted red. Best viewed in colour.

B.3 Acceleration

Speed of model fitting is a significant factor during development, particularly considering optimization of hyperparameters. Faster training yields more model iterations, which inevitably bear improvements. This section summarizes the implementation decisions specifically taken to accelerate training and testing of the model.

B.3.1 Parallel Model Estimation

While the estimation procedure outlined in § 3.1 must be repeated for all standardized voxels in the brain mask, it is possible to do this in parallel, since every estimation is independent. To do so, the training data must first be vectorized with respect to spatial location $x$, and matrix operations expanded explicitly to accommodate the new dimension.

To begin, the standardized training data from all subjects – features $\hat{Y}_\gamma(x)$, with $\hat{Y}_\gamma^0 = 1$, and labels $C_\gamma(x)$ – are sampled from nonzero locations in the brain mask $M(x)$. These data are stored in two matrices $\mathbb{Y}$ and $\mathbb{C}$, with dimensions $[X, N, K + 1]$ and $[X, N, 1]$, respectively, where $X$ is the total number of nonzero voxels in the brain mask, $N$ is the number of subjects, and $K$ is the number of features. A similar matrix is constructed for the initial parameters $\beta^{(0)}(x)$, denoted $\mathbb{B}^{(0)}$, with dimensions $[X, 1, K + 1]$. Let $\mathbb{Y}_n^k$ denote the vector of data from all voxels for the $k^{\text{th}}$ feature from the $n^{\text{th}}$ subject, and so on for $\mathbb{C}$ and $\mathbb{B}$.

In order to simplify subsequent calculations, the feature data are rectified according to the class labels,
before the first iteration, as in

\[
Y^k_n = \begin{cases} 
+Y^k_n, & C_n \geq 0.5 \\
-Y^k_n, & C_n < 0.5
\end{cases}, \quad \forall k \in \{1, \ldots, K\}. \quad (B.1)
\]

Next, for a given iteration \(t\), the following vector-compatible expansions of Equations (3.9), (3.10), and (3.13) yield the desired update matrix \(\Delta \mathbb{B}(t)\). Regarding notation: 1. the iteration index \((t)\) is omitted for clarity, 2. element-wise multiplication is denoted by \(\circ\), and 3. the variable \(K\) is now defined as 1, since this is essential to the simplification.

\[
S = \frac{1}{1 + e^{+\eta}}, \quad \eta = \mathbb{B}^0 + (\mathbb{B}^1 \circ Y^1) \quad (B.2)
\]

\[
A = S \circ (1 - S) \quad (B.3)
\]

\[
G = \nabla_{\mathbb{B}} \mathcal{L} - \lambda \mathbb{B}
\]

\[
= \begin{bmatrix}
G^0 \\
G^1
\end{bmatrix}
\]

\[
= \begin{bmatrix}
\sum_{n=1}^{N} (Y^0_n \circ S) \\
\sum_{n=1}^{N} (Y^1_n \circ S)
\end{bmatrix} - \lambda \begin{bmatrix}
\mathbb{B}^0 \\
\mathbb{B}^1
\end{bmatrix} \quad (B.4)
\]

\[
H = \nabla_{\mathbb{B}}^{1} \mathcal{L} - \lambda \mathbb{I}
\]

\[
= \begin{bmatrix}
H^{0,0} & H^{0,1} \\
H^{1,0} & H^{1,1}
\end{bmatrix}
\]

\[
= \begin{bmatrix}
\sum_{n=1}^{N} (A \circ Y^0_n \circ Y^0_n) & \sum_{n=1}^{N} (A \circ Y^1_n \circ Y^0_n) \\
\sum_{n=1}^{N} (A \circ Y^0_n \circ Y^1_n) & \sum_{n=1}^{N} (A \circ Y^1_n \circ Y^1_n)
\end{bmatrix} - \lambda \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix} \quad (B.5)
\]

\[
\mathbb{D} = \text{det} \mathbb{H}
\]

\[
= \left( H^{0,0} \circ H^{1,1} \right) - \left( H^{0,1} \circ H^{1,0} \right) \quad (B.6)
\]

\[
\Delta \mathbb{B} = -H^{-1}G
\]

\[
= \frac{1}{\mathbb{D}} \left[ \begin{bmatrix}
H^{1,1} \circ G^0 - H^{1,0} \circ G^1 \\
H^{0,1} \circ G^1 - H^{0,0} \circ G^0
\end{bmatrix} \right]^T \quad (B.7)
\]
B.3.2 Image Deformations

During cross validation, it is eventually necessary to transform each available image to the MNI brain space for training, and also to warp the fitted parameter images \( \beta(x) \) to the native space of every subject for inference. The image registration need only be estimated by the SPM Segment algorithm once, since this procedure is computationally expensive.

For maximum efficiency, the following image outputs from this procedure are saved for future use:

- the bias-corrected FLAIR image in native space;
- the bias-corrected FLAIR image in MNI space;
- the registration transformation, as a discrete diffeomorphism,

The SPM function \texttt{spm_diffeo} can then be used to apply the transformation to any new image, in either the forward or reverse direction. The only downside to this workflow is that \texttt{spm_diffeo} uses only \texttt{.nii} files for all input and output data flows, so the estimated \( \beta(x) \) images must be written from \texttt{MATLAB} to file before transformation, and loaded from file into \texttt{MATLAB} afterwards.

B.3.3 Half Resolution Model Estimation

Following the results from \S 4.7.3, it was observed that a minimum amount of parameter image smoothness is always desirable. An alternative method to enforcing parameter image smoothness is to estimate \( \beta(x) \) at a lower resolution, followed by interpolative upsampling. This has the additional advantage of requiring significantly less time for model estimation at \( O(n^3) \) for isotropic resizing).

To implement this approach, all training images and manual segmentations were resized by the scale factor \( r \) after application of graylevel standardization (in case resizing affects the graylevel statistics). The parameter images are then fitted using the methods described in \S B.3.1, yielding low resolution parameter images. Next, these images, \( \beta_r(x) \), were linearly interpolated to the original resolution \( (r = 1) \), before application of the smoothing filter to yield the final \( \beta(x) \). For reference, an example parameter image at each resolution is shown in Figure B.4.

All results were computed with this adaptation (specifically \( r = 0.5 \)) except for the parameter image smoothing experiments described in \S 4.7.3, which justify this modification.
Figure B.4: Comparison of parameter images estimated at full and half-resolution. Best viewed in colour.
Appendix C

Code

The MATLAB code used for all aspects of this thesis can be found at:

https://github.com/uoguelph-mlrg/vlr

and additional resources can be found at:

https://uoguelph-mlrg.github.io/vlr/

If these links die, please contact the author at: jesse.x.knight@gmail.com