Voxel-Wise Image Analysis for White Matter Hyperintensity Segmentation

by

Jesse Knight

Submitted to the School of Engineering
in partial fulfillment of the requirements for the degree of

Master of Applied Science in Engineering Systems and Computing

at the

UNIVERSITY OF GUELPH

2017

Voxel-Wise Image Analysis

for White Matter Hyperintensity Segmentation

by

Jesse Knight

Submitted to the School of Engineering
on 2017, in partial fulfillment of the
requirements for the degree of
Master of Applied Science in Engineering Systems and Computing

Abstract

This masters thesis has been examined by
a Committee of the School of Engineering as follows:

Julie Vale .....
Chair, Thesis Committee
Associate Professor
University of Guelph

April Khademi .....
Thesis Co-Supervisor
Assistant Professor
Ryerson University

Graham Taylor .....
Thesis Co-Supervisor
Assistant Professor
University of Guelph

Bob Dony .....
Member, Thesis Committee
Professor
University of Guelph

Acknowledgments

I 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, for all your support through the years;
and
Ali, for everything.

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

List of Figures

List of Tables

Abbreviations

GM	Grey matter
WM	White matter
CSF	Cerebrospinal fluid
PD	Proton density
FLAIR	Fluid attenuation inversion recovery
WML	White matter lesion
WMH	White matter hyperintensity
DAWM	Dirty appearing WM
MS	Multiple Sclerosis
AD	Alzheimer's Disease
PVE	Partial volume effect
VLR	Voxel-Wise Logistic Regression
MLE	Maximum likelihood estimation
MAP	Maximum a posteriori
SI	Similarity index
ICC	Interclass Correlation Coefficient
LL	Lesion load
CV	Cross validation
LOO-CV	Leave-one-out CV
KF-CV	K-fold CV
LOSO-CV	Leave-one-source-out
SVM	Support vector machine
K-NN	K-nearest neighbours
MRF	Markov random field
FPR	False positive reduction
PMF	Probability mass function
CDF	Cumulative density function

Notation

Variables		
y	feature	$\in \mathbb{R}$
$\tilde{y}$	standardized feature	$\in \mathbb{R}$
c	true class	$\in \{0, 1\}$
$\hat{c}$	estimated class	$\in [0, 1]$
β	logistic model feature weight	$\in \mathbb{R}$
γ	synthetic feature	$\in \mathbb{R}$
ρ	prior probability	$\in [0, 1]$
Indexing – e.g. arbitrary variable a		
k	feature index	$\in \{1, \ldots, K\}$
n	subject index	$\in \{1, \ldots, N\}$
t	iteration index	$\in \{1, \ldots, X\}$
x	spatial location	$= [\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3]$
$a_n^{k\,(t)}(x)$	k^{th} feature; n^{th} subject; t^{th} iteration; location x	
Images & sets – e.g. arbitrary variable a		
a	one voxel, one feature, one subject	
$\boldsymbol{a}$	one voxel, all features, one subject	
$\mathrm{A}(x)$	image in native space	
$A(x)$	image in standard space (MNI)	
$\boldsymbol{A}(x)$	image set: all features, one subject	
$\mathcal{A}(x)$	image set: one feature, all subjects	
$\boldsymbol{\mathcal{A}}(x)$	image superset: all features, all subjects	
$\mathbb{A}$	full dataset: all features, subjects, voxels	

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 5mm 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,

$$M_z = M_0 \left(1 - e^{-\left(\frac{t}{T1}\right)}\right). \tag{1.1}$$

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.

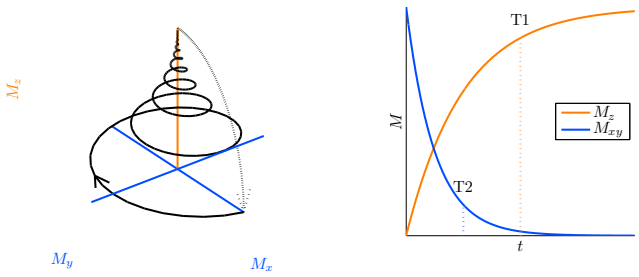


Figure 1.1: Visualization of T1 and T2 relaxation.

Table 1.1: T1 and T2 constants for brain tissues at 1.5 Tesla.

Tissue	T1 (ms)	T2 (ms)	$K[H]$ (a.u.)	Ref
WM	719 ± 33	73 ± 6	0.81 ± 0.03	[6]
GM	1165 ± 88	92 ± 11	0.98 ± 0.07	[6]
CSF	3337 ± 111	2562 ± 123	1.00 ± 0.07	[6]
WML	1124 ± 372	136 ± 79	–	[7][a]

[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,

$$M_{xy} = M_0 \left(e^{-\left(\frac{t}{T2}\right)} \right), \tag{1.2}$$

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 1.5T 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 decays 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].

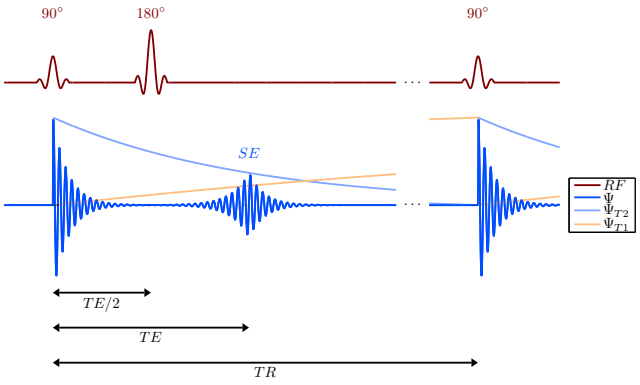


Figure 1.2: RF and MR signal for a basic Spin Echo sequence.

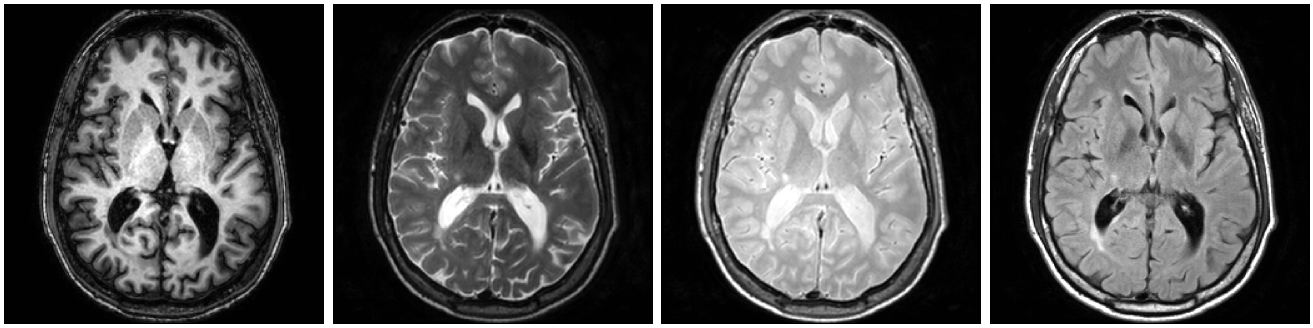
This second signal is called the Spin Echo (SE), and the interval $2t$ is termed the echo time (TE). Reversing the direction of rotation can be achieved by a 180° RF pulse at time $TE/2$, in the same way the original excitation is achieved using a 90° RF pulse (amount of rotation is proportional to the 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 Ψ for a spin echo sequence at location x can be described by the following 3-term equation,

$$\Psi_{SE}(x) = \Big[K[H](x)\Big]\Big[e^{-\left(\frac{TE}{T2(x)}\right)}\Big]\Big[1 - e^{-\left(\frac{TR}{T1(x)}\right)}\Big], \tag{1.3}$$

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 $T2$ 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/



(a) T1 (b) T2 (c) PD (d) FLAIR

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 $T1$, 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 (TI), can be chosen to null the signal from any tissue with a unique $T1$. The equation governing the image signal simply adds an inversion term,

$$\Psi_{IR}(x) = \left[K\left[H(x)\right]\right]\left[e^{-\left(\frac{TE}{T2(x)}\right)}\right]\left[1 + e^{-\left(\frac{TR}{T1(x)}\right)} - 2e^{-\left(\frac{TI}{T1(x)}\right)}\right]. \tag{1.4}$$

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 β protein and misfolded protein τ 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 β oligomers interfere with neuronal mitochondria and synapse function, leading to cell death [28, 29], while aberrant τ 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 [32, 31]. 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 [40, 36], 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 [56, 36].

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.

	Ref	Raters	Data	SI	ICC
Manual	[62]	5	10 images	0.64	—
	[63]	2	6 images	0.75	—
	[64]	2	120 slices	0.83	0.96
	[43]	3	50 images	0.66	0.97
Semi-Automated	[65]	1	16 images	—	0.99
	[66]	1	2 images	0.70	—
	[67]	1	33 slices	0.78	—
	[68]	2	30 images	0.78	—

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.[4]

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:

[4] 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.

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 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 error back-propagation to update thousands of parameters in a large, relatively unstructured model mapping 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 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 [84]);
- **Spatial:** spatial location, often normalized to a common space (e.g. 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.[5] A summary of many of these works is also given in Table 1.3.[6]

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. [87], Boer et al. [63], and Smart et al., [109] 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 [104]. 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 total lesion load [117]. 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, 136, 128].

[5] This section adapted from a paper in submission [85]
[6] A more detailed and interactive version of this table is available at www.uoguelph.ca/~jknigh04/wmlseg/table.html

Table 1.3: Summary of previous approaches to WMH segmentation with respect to image variability and reported performance (SI).

#	Ref.	Year	Authors	MRI Sequences	I	S	SI
1	[86]	2001	Van Leemput et al.	T1, T2, PD	20	1	0.51
2	[87]	2001	Jack et al.	FLAIR	39	1	
3	[88]	2002	Zijdenbos et al.	T1, T2	10	1	0.6
4	[89]	2004	Anbeek et al.	T1, T2, PD, FLAIR, IR	20	1	0.61
5	[90]	2005	Anbeek et al.	T1, T2, PD, FLAIR, IR	10	1	0.78
6	[91]	2005	Admiraal-Behloul et al.	T2, PD, FLAIR	100	1	0.75
7	[92]	2006	Lao et al.	T1, T2, PD, FLAIR	45	1	
8	[93]	2006	Wu et al.	T1, T2, PD	12	1	
9	[94]	2006	Sajja et al.	T2, PD, FLAIR	23	1	0.78
10	[62]	2006	Harmouche et al.	T1, T2, PD	10	1	0.61
11	[95]	2008	Khayati et al.	FLAIR	20	1	0.75
12	[96]	2008	Wels et al.	T1, T2, FLAIR	6	1	0.57
13	[97]	2008	Herskovits et al.	T1, T2, PD, FLAIR	42	2	0.6
14	[98]	2008	Bricq et al.	T2, FLAIR	25	2	
15	[99]	2008	Dyrby et al.	T1, T2, FLAIR	362	10	0.56
16	[100]	2008	Souplet et al.	T1, T2, FLAIR	25	2	
17	[63]	2009	Boer et al.	T1, PD, FLAIR	20	2	0.72
18	[101]	2009	García-Lorenzo et al.	T1, T2, PD	10	1	0.63
19	[102]	2009	Akselrod-Ballin et al.	T1, T2, PD, FLAIR	41	1	0.53
20	[103]	2009	Schwarz et al.	T1, T2, PD	165	2	
21	[104]	2010	Gibson et al.	T1, T2, FLAIR	18	1	0.81
22	[105]	2010	Shiee et al.	T1, FLAIR	10	1	0.63
23	[106]	2010	Scully et al.	T1, T2, FLAIR	17	1	
24	[107]	2011	García-Lorenzo et al.	T1, T2, FLAIR	10	1	0.65
25	[108]	2011	Geremia et al.	T1, T2, FLAIR	20	2	
26	[109]	2011	Smart et al.	T1, FLAIR	30	1	
27	[110]	2012	Samaille et al.	T1, FLAIR	67	6	0.72
28	[111]	2012	Khademi et al.	FLAIR	24	1	0.83
29	[112]	2012	Schmidt et al.	T1, FLAIR	53	1	0.75
30	[113]	2012	Abdullah et al.	T1, T2, FLAIR	61	3	
31	[114]	2013	Sweeney et al.	T1, T2, PD, FLAIR	111	1	0.61
32	[115]	2013	Datta et al.	T1, T2, FLAIR	90	3	
33	[64]	2013	Steenwijk et al.	T1, FLAIR	40	2	0.8
34	[71]	2014	Khademi et al.	FLAIR	25	1	0.78
35	[116]	2014	Ithapu et al.	T1, FLAIR	38	1	0.67
36	[117]	2014	Yoo et al.	FLAIR	32	2	0.76
37	[118]	2015	Harmouche et al.	T1, T2, PD, FLAIR	100	35	0.56
38	[119]	2015	Guizard et al.	T1, T2, PD, FLAIR	108	32	0.6
39	[120]	2015	Jain et al.	T1, FLAIR	20	1	0.67
40	[121]	2015	Tomas-Fernandez et al.	T1, T2, FLAIR	51	2	
41	[122]	2015	Wang et al.	T1, T2, FLAIR	70	2	0.84
42	[123]	2015	Roy et al.	FLAIR	38	3	0.56
43	[124]	2015	Brosch et al.	T1, T2, FLAIR	20	2	0.36
44	[125]	2015	Fartaria et al.	FLAIR	39	1	0.55
45	[126]	2015	Deshpande et al.	T1, T2, PD, FLAIR	52	1	0.5
46	[127]	2015	Roura et al.	T1, FLAIR	20	2	0.34
47	[128]	2016	Knight et al.	FLAIR	15	3	0.7
48	[129]	2016	Mechrez et al.	T1, T2, FLAIR	20	2	0.31
49	[130]	2016	Strumia et al.	T1, FLAIR	20	3	0.52
50	[131]	2016	Griffanti et al.	T1, FLAIR	130	2	0.76
51	[132]	2017	Valverde et al.	T1, FLAIR	33	2	
52	[133]	2016	Brosch et al.	T1, T2, PD, FLAIR	77	67	0.64
53	[134]	2017	Dadar et al.	T1, FLAIR	80	3	0.62
54	[135]	2017	Zhan et al.	T1, T2, FLAIR	50	2	0.76

Abbreviations. I: number of MR image sets used for validation; S: number of MRI scanners used for validation; SI: reported validation similarity index.

Mixture Models

Most other unsupervised approaches are probabilistic models, often framed as a mixture model. The work by Van Leemput et al. [86] uses a similar framework as the early work by Ashburner and Friston [137], later incorporated into the SPM "segment" tool [138], 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. [98], Schmidt et al. [112], Jain et al. [120], and Roura et al. [127] use parametric mixture models to predict WMH as model outliers, and all but [127] embed the model in a MRF. Khayati et al. [95] and Subbanna et al. [139] 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, 118]. Schwarz et al. again employ a Bayesian MRF model, but use lognormal distributions for WM and WMH [103]. 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 [100]. 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) [97]. More recently, Graph-Cuts have been used in conjunction with mixture models, as in the works by García-Lorenzo et al. [101], Tomas-Fernandez and Warfield [121], and Strumia et al. [130].

The Lesion-TOADS method by Shiee et al. [105], a lesion-specific adaptation of the TOADS algorithm [140], 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. [91], Gibson et al. [104], and Valverde et al. [132].

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 [89, 90]. In the works by Wu et al. [93], Steenwijk et al. [64], and Fartaria et al. [125],

spatial coordinates are substituted for tissue priors as K-NN features. In the recently proposed BIANCA algorithm by Griffanti et al. [131], 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. [92], Abdullah et al. [113], and Scully et al. [106] 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. [102] employ over 30 features for multi-scale image representation and classify voxels using RF. Both Geremia et al. [108] and Roy et al. [123] use a combination of intensity and tissue prior features to train a RF classifier, whereas Wels et al. [96] use a large number of Haar-like features to train an AdaBoost model. Ithapu et al. [116] 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. [114], 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. [135], 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. [134], 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 [94].

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. [88] and Dyrby et al. [99] train fully-connected voxel-wise Neural Networks with a selection of intensity, spatial, and tissue prior features to predict the lesion class. In contrast, Brosch et al. [124, 133] construct a more modern deep convolutional model, which is capable of capturing both local and global dependencies.

[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.

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 [94, 99, 102, 109, 112, 117, 116, 123, 132] and the FSL[9] toolkit [97, 104, 115, 64, 114, 123, 122, 131, 135], as well as bias correction by the N3/4[10] [141] algorithm [88, 62, 125, 119, 118, 129, 132, 134, 135].

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 [141] for bias field correction, BET for brain extraction [142], SPM Segment [138] / FSL FAST [143] for healthy brain segmentation, SPM Normalize [138] / FSL FNIRT [144] 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 extensions,[12] as well as events by NA-MIC (National Alliance for Medical Image Computing) for development of 3D Slicer modules.[13]

[8] http://www.fil.ion.ucl.ac.uk/spm/
[9] https://fsl.fmrib.ox.ac.uk/fsl/
[10] https://www.slicer.org/wiki/Documentation/4.6/Modules/N4ITKBiasFieldCorrection
[11] Things may have changed recently, cf. § 4.9.2.
[12] http://www.fil.ion.ucl.ac.uk/spm/ext/
[13] https://na-mic.org/wiki/Events

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 4 use more than three. 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, 114, 134]. 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 [145]. 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 [86, 100, 107, 127]. Graylevel thresholding techniques [87, 109, 110, 112, 71] are similarly affected by changes in image properties.

It is worth noting four works[14] 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.

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

[14] The work by Samaille et al. (2012) [110] 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.

Ref.	Year	Authors	I	S	SI
[99]	2008	Dyrby et al.	362	10	0.56
[119]	2015	Guizard et al.	108	32	0.60
[118]	2015	Harmouche et al.	100	35	0.56
[133]	2016	Brosch et al.	77	67	0.64

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.

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 [127], histogram features [117], or conditional edge probability [146]), 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.[15] 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, is it has been suggested that there is regional heterogeneity in relaxation rates of brain tissues [147], and that WML intensity depends in part on location [7, 118]. 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. [89, 90, 99, 131, 134]. However, such models treat all features equally, which can lead

[15] The `fminsearch` function in Matlab was used.

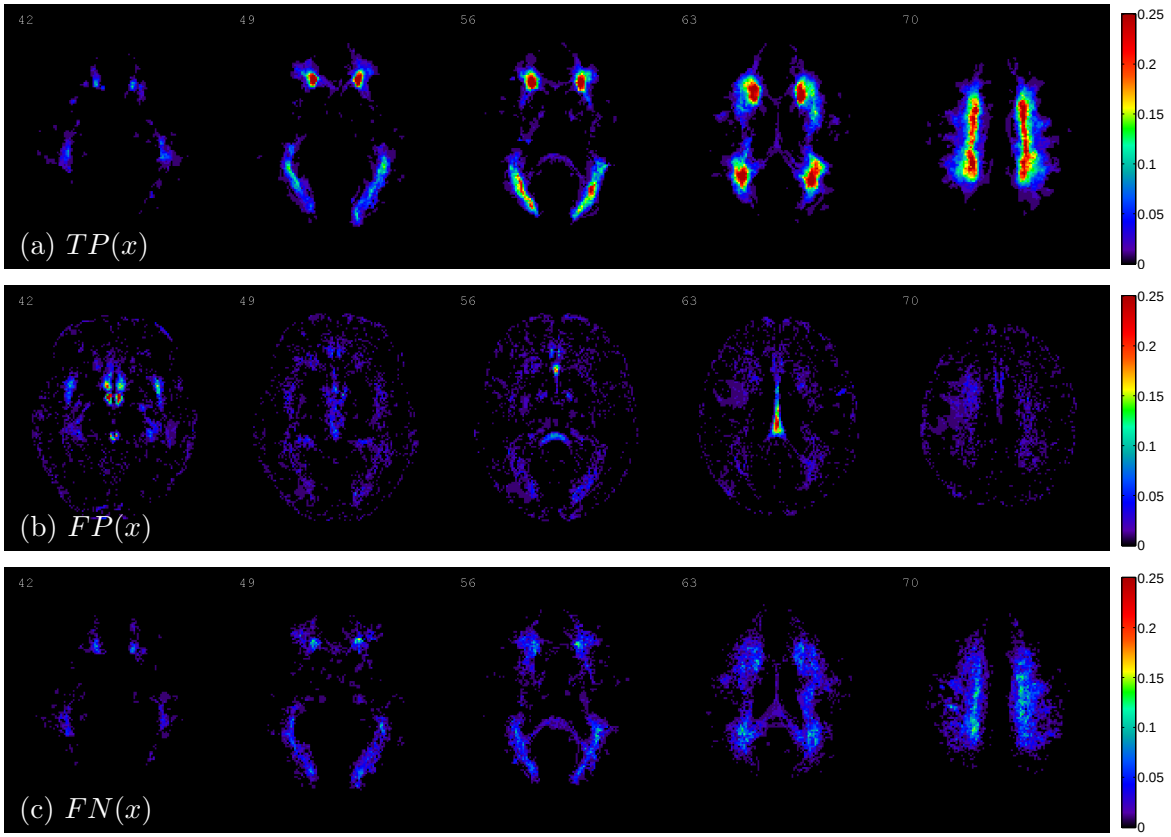


Figure 1.4: Distributions of TP, FP, and FN in 96 FLAIR MRI, following supervised optimal thresholding in MNI space.

to artifacts in the 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 [148, 114, 149, 135], 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 $\boldsymbol{y} = [1, y^1, \ldots, y^K]^T$, is modelled by a logistic

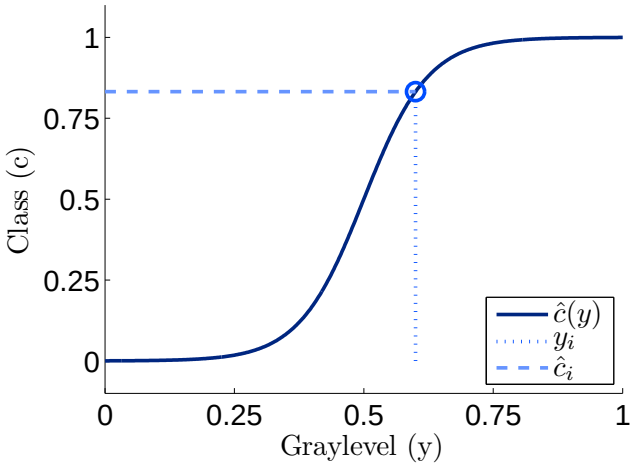


Figure 1.5: The logistic regression model.

function, parameterized by a vector of feature weights $\boldsymbol{\beta} = [\beta^0, \beta^1, \ldots, \beta^{\mathrm{K}}]^T$,

$$P\big(c = 1 \mid \boldsymbol{y}, \boldsymbol{\beta}\big) = \frac{1}{1 + e^{-\eta}}, \qquad \eta = \boldsymbol{\beta}^T \boldsymbol{y}. \tag{1.5}$$

This probability – the estimated lesion class label – is denoted $\hat{c} = P(c = 1 \mid \boldsymbol{y}, \boldsymbol{\beta}) \in [0, 1]$. Figure 1.5 shows an example class prediction for an arbitrary input graylevel y_i. Considering the spatial location $x = [\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3]$, the estimated label image becomes $\hat{C}(x) = P(C(x) = 1 \mid \boldsymbol{Y}(x), \boldsymbol{\beta})$.

In the OASIS algorithm [148, 114] and the algorithm by Zhan et al. [135], this model is used with a number of different graylevel features. However, this ignores spatial location[16] 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, 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.

[16] 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.

Lesion Prediction Algorithm

An alternative approach by Schmidt aims to solve this issue by introducing a spatial effect parameter, namely the intercept $\beta^0 \rightarrow \beta^0(x)$. In this method, a Gaussian MRF model is used to estimate the spatial parameter $\beta^0(x)$, while the other β parameters – in fact there is only one: β^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.[17]

This particular parametrization has significant implications for model estimation, however, since $\beta^0(x)$ should be estimated uniquely for every spatial location, but β^1 should consider evidence from the entire image volume. Efficient fitting of such models was the subject of major works by Schmidt et al. [150, 149], 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 $\boldsymbol{\beta}$ to only about 90% convergence [149]. Finally, it is not clear whether any tied β 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) [112] 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 [150, 149]. Therefore, while the open-source release of the LPA algorithm is greatly appreciated, significant improvements can be made to this algorithm.

1.4 Proposed Algorithm

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

[17] http://www.applied-statistics.de/lst.html

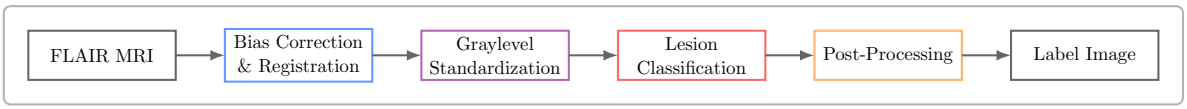


Figure 1.6: Overview of the necessary processing steps.

For several reasons, the supervised pipeline approach was selected as the framework for this algorithm. First, preliminary work drawing on existing algorithms [71, 151] 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 [86]. 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

$$P\big(c(x) = 1 \mid \boldsymbol{y}(x), \boldsymbol{\beta}(x)\big) = \frac{1}{1 + e^{-\eta(x)}}, \qquad \eta = \boldsymbol{\beta}(x)^T \boldsymbol{Y}(x). \tag{1.6}$$

Training the VLR model then yields a complete and unique vector $\boldsymbol{\beta}$ for each voxel, or equivalently, one complete image for each parameter. Only one additional β 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 estimability. 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 experiment, 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" [152] and analysis of functional MRI data [153].[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.

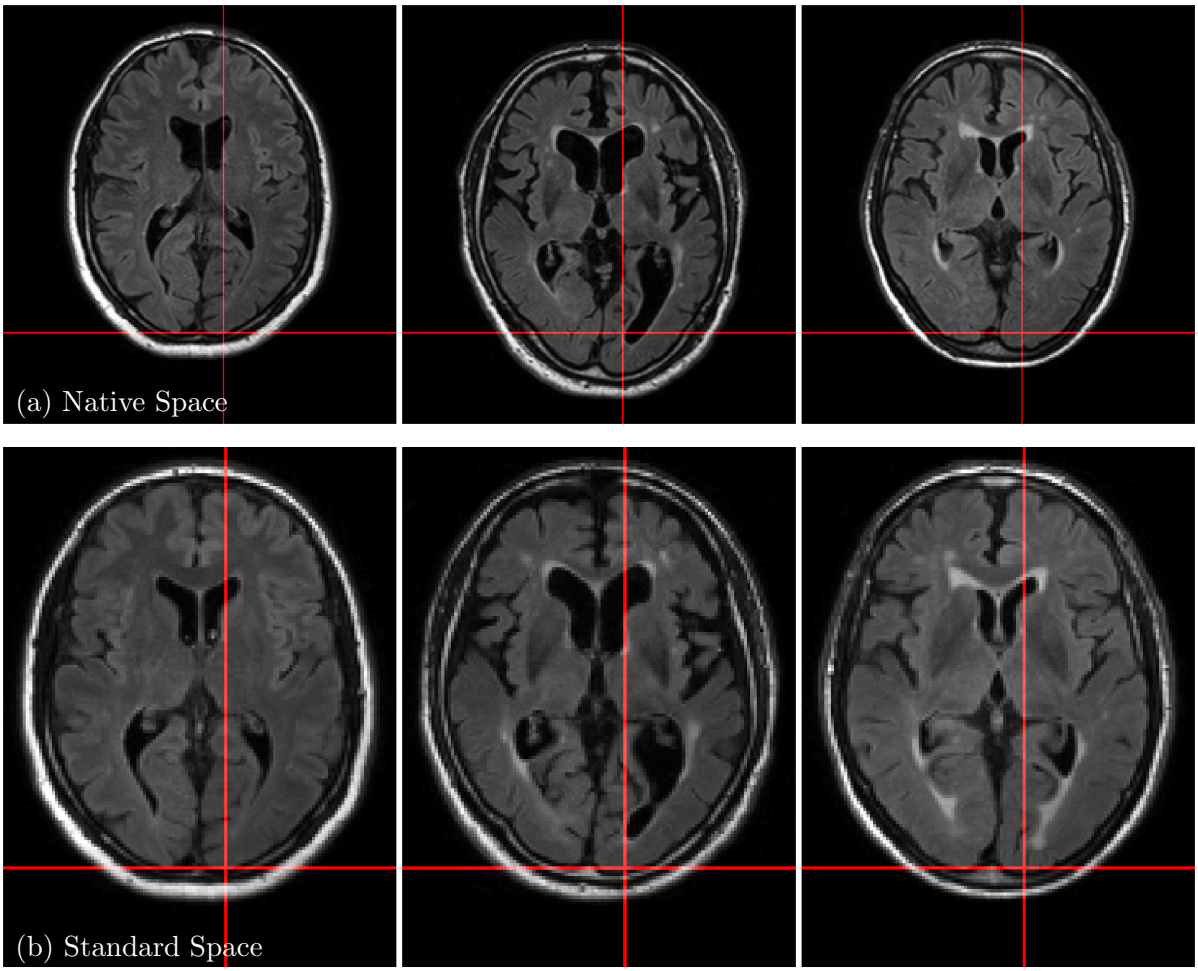


Figure 2.1: Example set of FLAIR MRI before and after registration to the MNI brain space. From [12].

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 [154] or the Montreal Neurological Institute (MNI) space [155], 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 fitted by maximizing some measure of overlap between the images [156]. Simple registration methods employ only affine transformations, comprising a combination of translation, scaling, and shear transformations (e.g. the FLIRT tool [157] 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 [144], discrete cosine transform in SPM Normalize [137, 138], a general diffeomorphism in the SyN algorithm [158]). 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 [159], while a more recent (2014) study ranks the popular toolboxes SPM, FSL, and Brainsuite [160].

In the current work, registration is required during both training at testing. During training, the model parameters $\boldsymbol{\beta}(x)$ are estimated using mutually aligned segmentation examples $\{\boldsymbol{\mathcal{Y}}, \mathcal{C}\}$ 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. [118], 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 [138] was not among the top performing in the 2009 study [159], subsequent implementation revisions in SPM12[2] called "New Segment" have apparently improved results. In the 2014 study [160], the SPM12 New Segment method achieved the highest Similarity Index of all methods on real (IBSR [161]) 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 $\boldsymbol{\beta}(x)$ 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.

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

[3] One FLAIR template is available in [162]; however, this was generated using SPM5, so using it would compound any registration biases associated with the older method.

2.2 Bias Correction

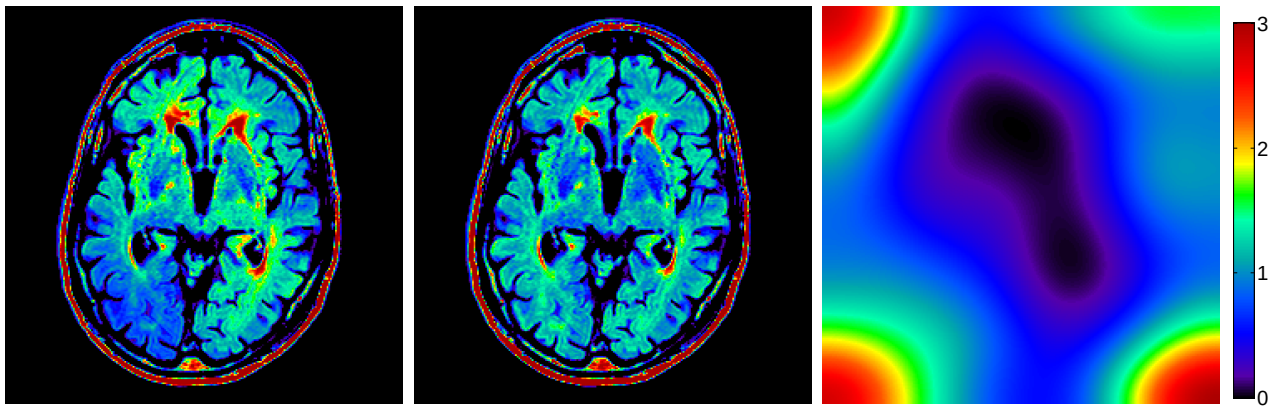
As noted in § 1.2.2, bias field (AKA intensity inhomogeneity), is a smoothly varying intensity variation artifact common in MRI. The sources of bias field artifact include inhomogeneities in the magnetic field and RF coils used for pulse transmission and signal sensing, as well as non-ideal magnetic 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 [163], subsequently updated to N4(ITK) [141] and integrated in the FreeSurfer toolbox[4] is perhaps the most popular, 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 Segment model [138] includes integrated bias field estimation and correction. The bias field model in this work is parameterized by the discrete cosine transform, as described in in a previous work by Ashburner and Friston [164]. The FSL Segment feature also estimates bias field in a similar overall model to SPM Segment, except the tissue prior probability maps in SPM are replaced with a Markov Random Field Model, as described in [143]. Again, two reviews with quantitative performance comparisons provide a good reference of other proposed algorithms, including a comprehensive review in 2006 [165] and a comparison of mainly popular methods in 2016 [166].

In the 2016 comparison [166], the authors note that the SPM and FSL models – which include segmentation – outperformed the N3 algorithm [163] and another non-segmenting method [167]. These results are consistent with the advantages of unified generative models described in § 1.3.1 and discussed in "Unified segmentation." [138]. 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 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.

[4] https://surfer.nmr.mgh.harvard.edu/



(a) Raw FLAIR (b) Bias-corrected FLAIR (c) Estimated bias field

Figure 2.2: Example bias correction. From [12].

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

$$\tilde{y} = \tau(y), \tag{2.1}$$

where τ 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 the probability mass function (PMF) $f_{\mathrm{Y}}(y)$,

$$\begin{aligned} f_{\mathrm{Y}}(y) &= \frac{1}{X} h_{\mathrm{Y}}(y) \\ &= \frac{1}{X} \sum_{i=1}^{X} \begin{cases} 1 & Y(x_i) = y \\ 0 & Y(x_i) \neq y \end{cases} \end{aligned} \tag{2.2}$$

The cumulative density function (CDF) $F_{\mathrm{Y}}(y)$ is the cumulative sum of $f_{\mathrm{Y}}(y)$,

$$F_{\mathrm{Y}}(y) = \sum_{\gamma=y_{\min}}^{y} f_{\mathrm{Y}}(\gamma) \tag{2.3}$$

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 (cf. § **??**). 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:[5]

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

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

  Naively, this method is very susceptible to corruption by outliers. For more robustness, $y_{\min}$ and $y_{\max}$ can be defined using intensity quantiles – e.g. $[\epsilon_1, 1 - \epsilon_2]$. However, selection of an appropriate ϵ_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,

$$\tau(y) = \frac{y - \mu_{\mathrm{Y}}}{\sigma_{\mathrm{Y}}}. \tag{2.5}$$

  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 distribution, thereby distributing image intensities equally across the available range. The desired transform is defined as the CDF of the input image (cf. [168] for derivation),

$$\tau(y) = F_{\mathrm{Y}}(y). \tag{2.6}$$

  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, $f_{\tilde{\mathrm{Y}}}$. This transform is defined as

[5] This section considers only univariate standardization methods, since only FLAIR intensities are used in this work.

the function composition of the input CDF and the inverse target CDF,

$$\tau(y) = F_{\tilde{Y}}^{-1}\big(F_{Y}(y)\big) \tag{2.7}$$

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.2.1 for an illustration and experimental evidence).

- **Nyul standardization:** In [145, 169], Nyul and Udupa proposed a method for intensity standardization which has subsequently been used in other works. This method defines τ with piecewise linear segments connecting the Q quantiles of the input PMF q, with quantiles of a target PMF r,

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

  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 image-specific 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 [171]. 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.

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 $\mathcal{Z}$ 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. $\boldsymbol{\epsilon}$, Q, $\mathcal{X}_\tau$ – and also to select the best

[6] This result is presented and supported with experiments in [170].

[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.

overall method.

Two such functions are proposed. The first is discrete, and inspired by the Zero-Crossing Rate [172]. It measures the number of class transitions in the sorted feature data $\tilde{\mathcal{Y}}_s$, as shown in Figures 2.3a and 2.3c. With $\mathcal{C}_s$ as the class labels after sorting by the feature $\tilde{\mathcal{Y}} = \tau(\mathcal{Y})$, the objective function $\mathcal{Z}_\Delta$ is defined as

$$\mathcal{Z}_\Delta = \sum_{n=1}^{N-1} \begin{cases} 1 & \mathcal{C}_s^n \neq \mathcal{C}_s^{n+1} \\ 0 & \mathcal{C}_s^n = \mathcal{C}_s^{n+1} \end{cases}. \tag{2.9}$$

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

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

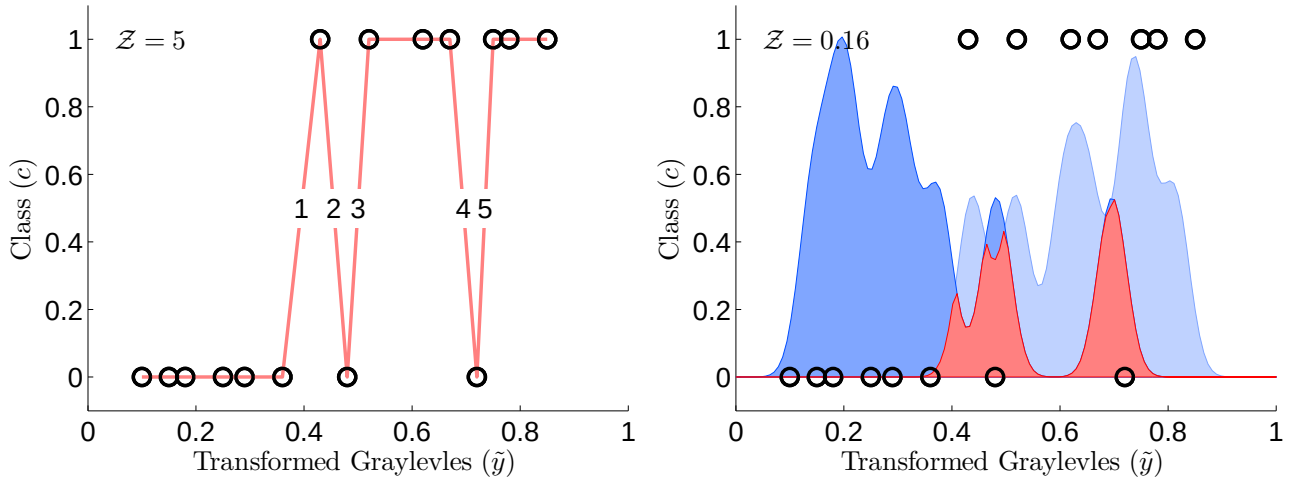
$$\mathcal{Z}_\star = \int_{\tilde{y}_{\min}}^{\tilde{y}_{\max}} \frac{\min\{p(\psi \mid c=1), p(\psi \mid c=0)\}}{\max\{p(\psi \mid c=1), p(\psi \mid c=0)\}} \partial\psi, \qquad p(\psi \mid c) \approx \sum_{\tilde{y} \in \{\tilde{\mathcal{Y}} | c\}} \delta(\psi - \tilde{y}) \star G_\sigma(\psi) \tag{2.10}$$

where $G_\sigma(\psi)$ is a Gaussian convolution kernel with width σ. This function is continuous and bounded, as in $\mathcal{Z}_\Delta \in [0, 1]$, and the lower bound is again optimal.

2.3.2 Supervised Standardization

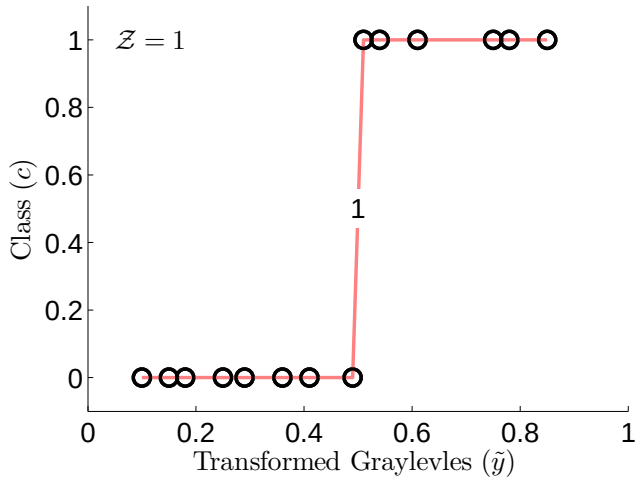
It is not hard to see that it should be possible to estimate an optimal graylevel standardization transform using the training data. That is, a *supervised graylevel standardization*.[9] If $\mathcal{Z}$ is differentiable, then this optimization can be performed using gradient descent, or similar methods. Unfortunately, neither of the above objective functions $\mathcal{Z}$ were reasonably differentiable, but many other optimization paradigms which do not require gradients could be used.

[8] The lower bound can be zero if one class is not observed.
[9] To the best of this author's knowledge, such a technique has never been proposed.

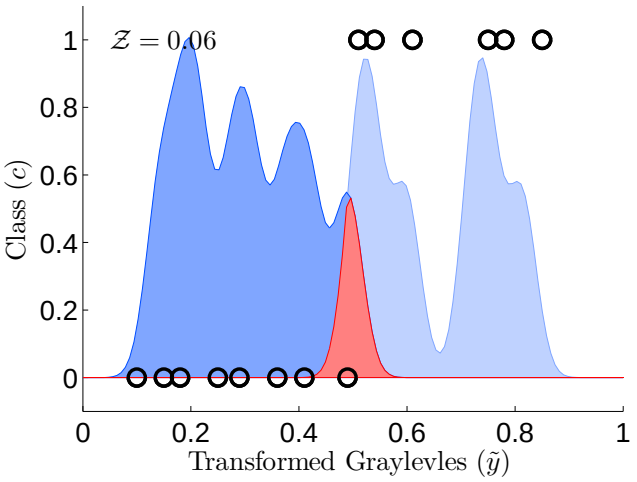


(a) $\mathcal{Z}_\Delta$ with overlapping classes

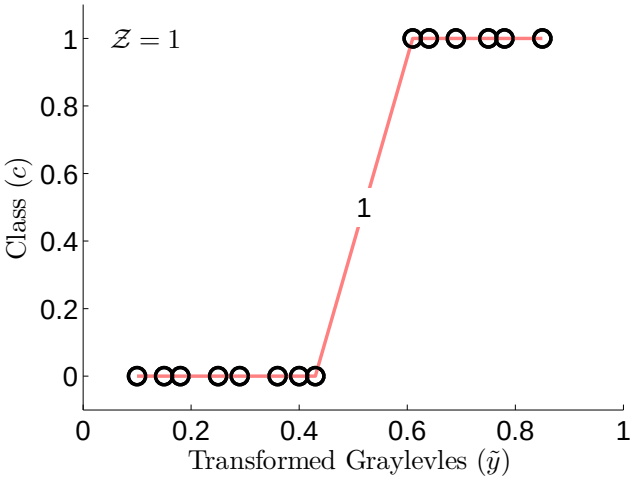
(b) $\mathcal{Z}_\star$ with overlapping classes



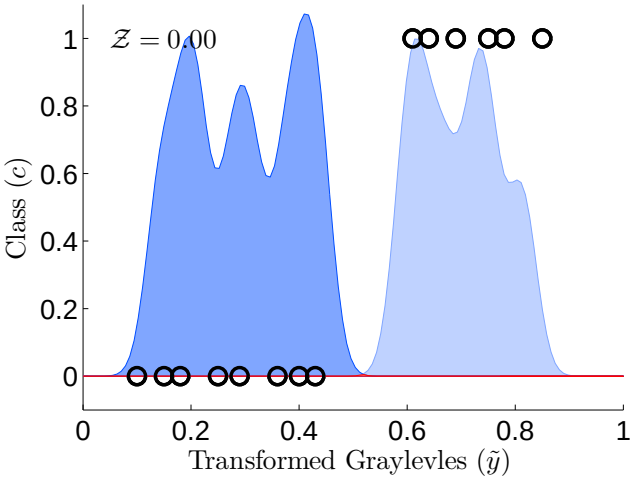
(c) $\mathcal{Z}_\Delta$ with barely separable classes



(d) $\mathcal{Z}_\star$ with barely separable classes



(e) $\mathcal{Z}_\Delta$ with easily separable classes



(f) $\mathcal{Z}_\star$ with easily separable classes

Figure 2.3: Illustration of potential separability objective functions.

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 Segment tool, which also produces bias corrected images. Next, image intensities are standardized using a graylevel transformation, to be determined in the Chapter 4. Training then proceeds to fit the VLR model parameters, as described in Chapter 3.

At test time, SPM 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 $\boldsymbol{\beta}(x)$ 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 label image $\hat{C}(x)$ is defined using the subject-specific features $\boldsymbol{\mathcal{Y}}(x) = [1, Y^1(x), \ldots, Y^{\mathrm{K}}(x)]^T$ and the corresponding model weights $\boldsymbol{\beta}(x) = [\beta^0(x), \beta^1(x), \ldots, \beta^{\mathrm{K}}(x)]^T$,

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

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 $\boldsymbol{\beta}$ for each voxel x. This requires some training data: feature vectors from a population of N observations $\boldsymbol{\mathcal{Y}}(x) = \{\boldsymbol{\mathcal{Y}}_1(x), \ldots, \boldsymbol{\mathcal{Y}}_{\mathrm{N}}(x)\}$, and the corresponding labels $\mathcal{C}(x) = \{C_1(x), \ldots, C_{\mathrm{N}}(x)\}$. 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 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 should be maintained around the decision boundary. These two cases are illustrated in Figure 3.1a.

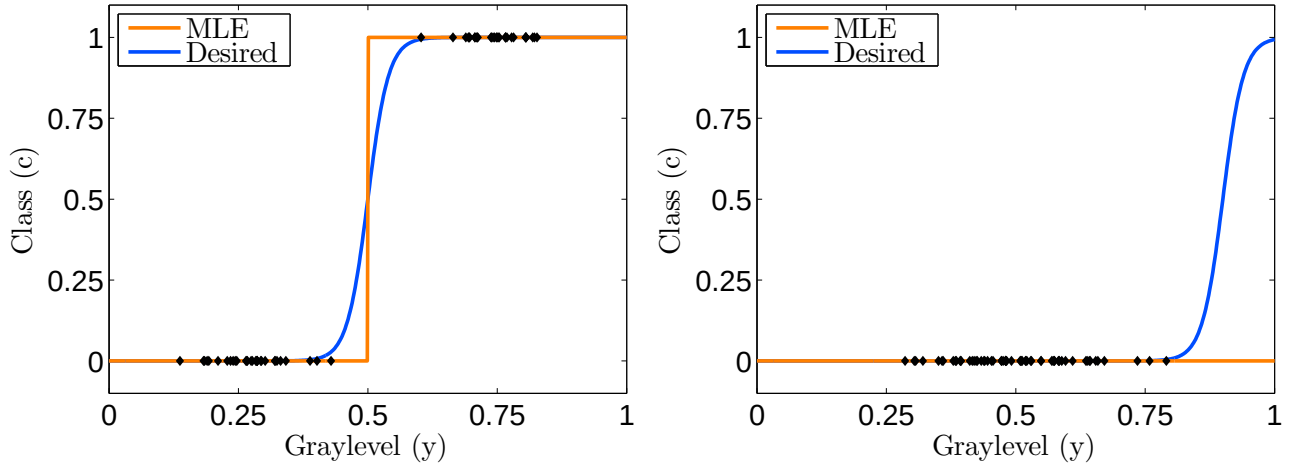
2. **Sparsely observed lesion class:** Since WML are often distributed in consistent locations, many brain regions contain no lesions across the entire training dataset. These voxels will be termed "healthy training" voxels, and denoted $\mathcal{X}_h$. In some locations, this is expected (e.g. the GM, 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 3.1b, 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 lesion 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.

3.1.2 Maximum Likelihood Estimation

Each parameter vector $\boldsymbol{\beta}(x)$ is considered completely independent. Doing so greatly simplifies model fitting, but does not address the challenges described above. These must instead be addressed using regularization strategies, as discussed below in § 3.2. In this section, ML estimation of independent parameter vectors $\boldsymbol{\beta}$ is developed. For clarity, only a single voxel is considered – i.e. y from $Y(x)$, etc.

As note above, the optimal $\boldsymbol{\beta}$ for each independent voxel can be resolved using MLE. If the training data are also assumed to be independently observed, then the likelihood (conditioned on the data) is defined



(a) Separable classes

(b) No lesions

Figure 3.1: Challenges encountered during estimation of a logistic model.

from binomial theory as

$$\begin{aligned} L(\boldsymbol{\beta} \mid \mathcal{C}, \boldsymbol{\mathcal{Y}}) &= \prod_{n=1}^{N} P(c = 1 \mid \boldsymbol{y}_n, \boldsymbol{\beta})^{c_n} \left(1 - P(c = 1 \mid \boldsymbol{y}_n, \boldsymbol{\beta})^{1-c_n}\right) \\ &= \prod_{n=1}^{N} \left[\hat{c}_n{}^{c_n} \left(1 - \hat{c}_n{}^{1-c_n}\right)\right]. \end{aligned} \tag{3.2}$$

For computational reasons, it is simpler and asymptotically equivalent to maximize the log-likelihood,

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}) &= \log \prod_{n=1}^{N} \left[\hat{c}_n{}^{c_n} \left(1 - \hat{c}_n{}^{1-c_n}\right)\right] \\ &= \sum_{n=1}^{N} \left[c_n \log \hat{c}_n + (1 - c_n) \log(1 - \hat{c}_n)\right] \\ &= \sum_{n=1}^{N} \left[c_n \boldsymbol{\beta}^T \boldsymbol{y}_n - \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{y}_n})\right]. \end{aligned} \tag{3.3}$$

The optimal $\boldsymbol{\beta}$ is therefore resolved by maximizing the log-likelihood,

$$\begin{aligned} \boldsymbol{\beta}^* &= \arg\max_{\boldsymbol{\beta}} \ \mathcal{L}(\boldsymbol{\beta}) \\ &= \arg\max_{\boldsymbol{\beta}} \ \sum_{n=1}^{N} \left[c_n \boldsymbol{\beta}^T \boldsymbol{y}_n - \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{y}_n})\right] \end{aligned} \tag{3.4}$$

3.1.3 Iterative Updates

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

$$\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)} + \alpha\, \Delta\boldsymbol{\beta}^{(t)}, \tag{3.5}$$

where α is a small valued learning rate parameter. There are many possible definitions of $\Delta\boldsymbol{\beta}$, including simply the gradient of $\mathcal{L}(\boldsymbol{\beta})$, denoted $\nabla_{\boldsymbol{\beta}}\mathcal{L}$. However, it can be shown that $\mathcal{L}(\boldsymbol{\beta})$ is convex, so higher order update equations can be used. The work by Minka [173] 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.

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

$$\nabla_{\boldsymbol{\beta}}\mathcal{L} = \begin{bmatrix} \frac{\partial L}{\partial \beta^1} \\ \vdots \\ \frac{\partial L}{\partial \beta^K} \end{bmatrix}, \tag{3.6}$$

$$\nabla^2_{\boldsymbol{\beta}}\mathcal{L} = \begin{bmatrix} \frac{\partial^2 L}{\partial \beta^1 \partial \beta^1} & \cdots & \frac{\partial^2 L}{\partial \beta^1 \partial \beta^K} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 L}{\partial \beta^K \partial \beta^1} & \cdots & \frac{\partial^2 L}{\partial \beta^K \partial \beta^K} \end{bmatrix}, \tag{3.7}$$

then the Newton update is given by

$$\Delta\boldsymbol{\beta} = -\nabla^2_{\boldsymbol{\beta}}\mathcal{L}^{-1}\nabla_{\boldsymbol{\beta}}\mathcal{L}. \tag{3.8}$$

In the current model, the gradient is given by

$$\nabla_{\boldsymbol{\beta}}\mathcal{L} = \sum_{n=1}^{N} \boldsymbol{y}_n \left(c_n - \hat{c}_n\right), \tag{3.9}$$

and the Hessian by

$$\nabla^2_{\boldsymbol{\beta}}\mathcal{L} = \sum_{n=1}^{N} \boldsymbol{y}_n \boldsymbol{y}_n{}^T \left(c_n - \hat{c}_n\right). \tag{3.10}$$

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

[1] Matlab code available at https://github.com/tminka/logreg/

3.1.4 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 $\boldsymbol{y}$ can be used. However, with only one feature – the FLAIR graylevel – it is possible to reparameterize the sigmoid argument as

$$\begin{aligned} \boldsymbol{\beta}^T \boldsymbol{y} &= \beta^0 + \beta^1 y^1 \\ &= s(y - \tau) \qquad \begin{cases} s = \beta^1 \\ \tau = -\frac{\beta^0}{\beta^1} \end{cases} . \end{aligned} \tag{3.11}$$

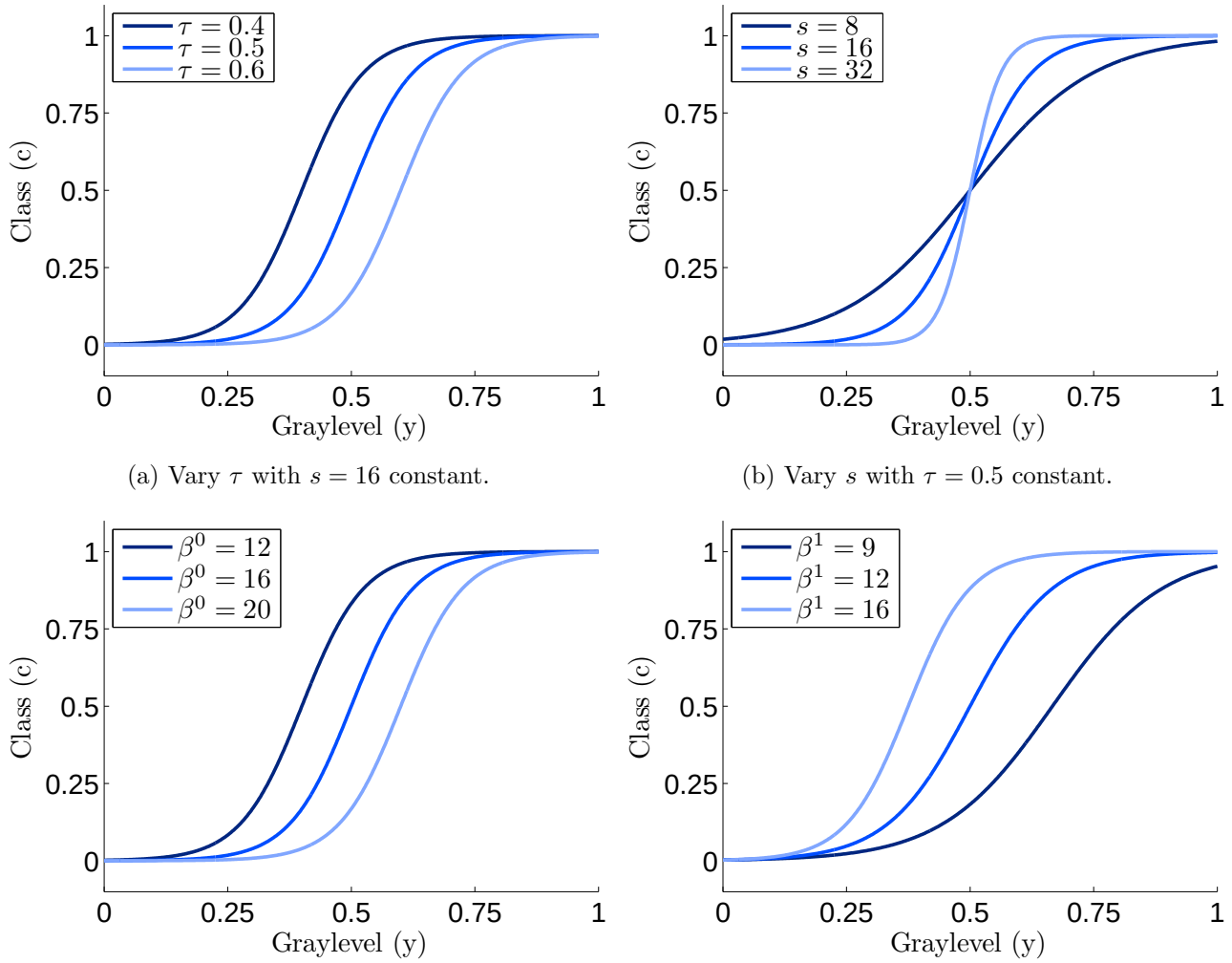
In this form, the parameters τ and s emerge as a graylevel threshold and a slope parameter, respectively. Specifically, when $y = \tau$, the predicted probability of lesion is $\hat{c} = \frac{1}{1+e^0} = 0.5$, so τ controls the location of the class discrimination, as shown in Figure 3.2a. Similarly, the s parameter defines the sensitivity of the logistic function to y, as shown in Figure 3.2b. By contrast, varying the original parameter β^1 with β^0 constant (Figure 3.2d) results in correlation of these characteristics. These new parameters, and the corresponding images $\mathcal{T}(x)$ and $\mathcal{S}(x)$, are therefore salient descriptors of the predictive model.

3.2 Regularization

Regularizations are methods of 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



(a) Vary τ with $s = 16$ constant.

(b) Vary s with $\tau = 0.5$ constant.

(c) Vary β^0 with $\beta^1 = 16$ constant.

(d) Vary β^1 with $\beta^0 = -6$ constant.

Figure 3.2: Effect of varying the logistic model parameters.

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 [174]. In higher dimensional problem spaces, however, imputation is more difficult [175]. 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.

Alternatively, simple image manipulations can still afford model improvements [176]. 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 § **??**.

3.2.2 Classic Regularization

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: $\lambda \left|\left|\boldsymbol{\beta}\right|\right|_p$ [177]. It can be shown that L_1 regularization corresponds to a Laplacian prior on elements of $\boldsymbol{\beta}$, with scale parameter inversely proportional to λ (equivalently, this assumes that the model error follows this distribution). Similarly, L_2 regularization implies a Gaussian prior, with standard deviation inversely proportional to λ [177]. 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

$$\begin{aligned}
\boldsymbol{\beta}^* &= \arg\max_{\boldsymbol{\beta}} \ \mathcal{J}(\boldsymbol{\beta}) \\
&= \arg\max_{\boldsymbol{\beta}} \ \mathcal{L}(\boldsymbol{\beta}) - \lambda \left|\left|\boldsymbol{\beta}\right|\right|_p \\
&= \arg\max_{\boldsymbol{\beta}} \ \sum_{n=1}^{N} \left[c_n \boldsymbol{\beta}^T \boldsymbol{y}_n - \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{y}_n})\right] - \lambda \left|\left|\boldsymbol{\beta}\right|\right|_p
\end{aligned} \tag{3.12}$$

Due to its relatively large gradient near zero, L_1 regularization is typically used to encourage sparsity in the feature weights (i.e. $\beta^k \to 0$) [178]. This is not desirable in the current model, since the feature (FLAIR graylevel) is known to be discriminative. Moreover, the expansion of the $\left|\left|\boldsymbol{\beta}\right|\right|_1$ term in the gradient of the objective function is not straightforward, since it is non-differentiable at zero [178, 179]. 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 [173]. For these reasons, only L_2 regularization is considered, yielding the following change to the Newton update expression (3.8),

$$\begin{aligned}
\Delta\boldsymbol{\beta} &= -\nabla^2_{\boldsymbol{\beta}} \mathcal{J}^{-1} \nabla_{\boldsymbol{\beta}} \mathcal{J} \\
&= -\left(\nabla^2_{\boldsymbol{\beta}} \mathcal{L} - \lambda I\right)^{-1} \left(\nabla_{\boldsymbol{\beta}} \mathcal{L} - \lambda \boldsymbol{\beta}\right)
\end{aligned} \tag{3.13}$$

What remains is to select an appropriate value of λ. This is explored experimentally in § **??** 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 [147, 7], 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 $\boldsymbol{\mathcal{V}}(x) = \{\boldsymbol{\gamma}_1(x), \ldots, \boldsymbol{\gamma}_{\mathrm{V}}(x)\}$. 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 $\boldsymbol{\mathcal{Y}}_\gamma(x) = \{\boldsymbol{\mathcal{Y}}(x), \boldsymbol{\mathcal{V}}(x)\}$, and $\boldsymbol{\mathcal{C}}_\gamma(x) = \{\boldsymbol{\mathcal{C}}(x), \mathbf{1}(x)\}$.

Design of the synthetic image set $\boldsymbol{\mathcal{V}}(x)$ 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,

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

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 § **??**.

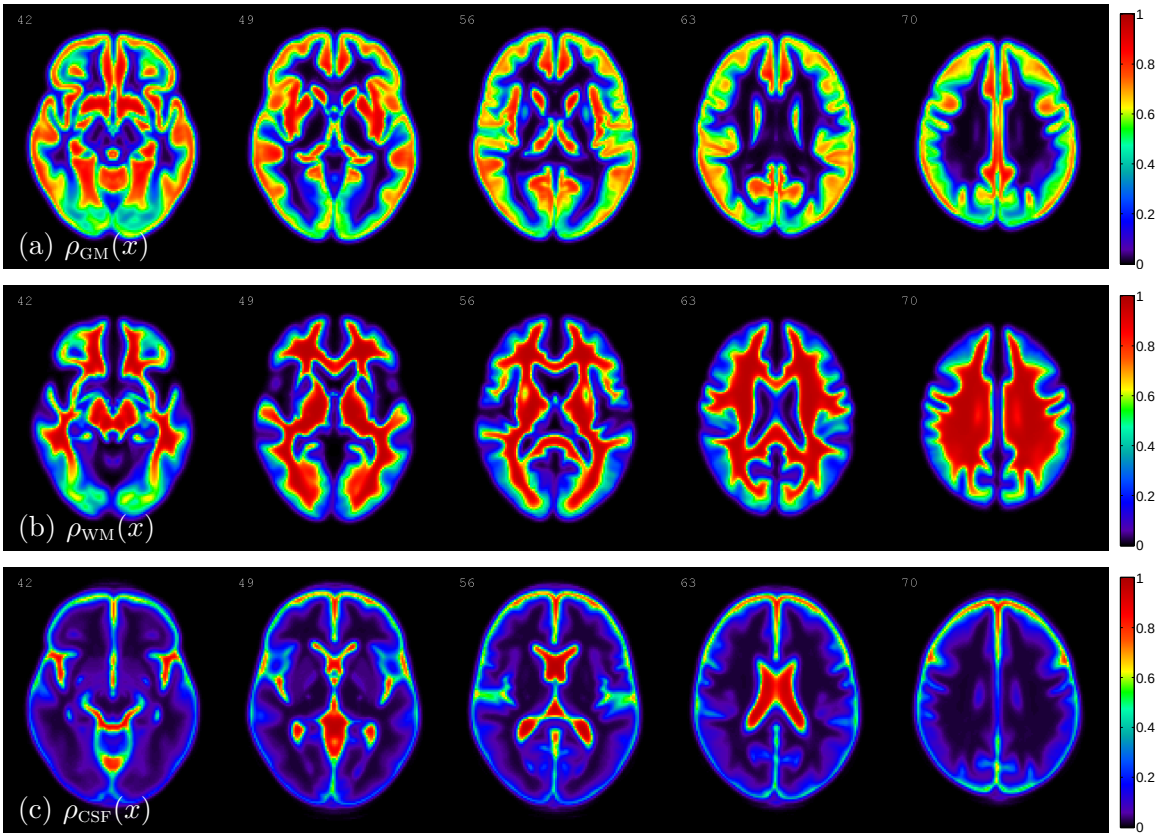


Figure 3.3: Tissue prior probability images in MNI space. Derived from [84].

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

Name	Parameters	Advantages	Disadvantages	Ref.
Gaussian	width σ	no artifacts	blurs edges	[168]
Median	width w	preserves edges	square artifacts	[168]
Bilateral	σ_y,σ_x	balanced blurring and detail	Expensive	[180]

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 [168]. Therefore, only conventional filtering is explored in this work (cf. § **??**).

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 to give a "hard" classification: $\hat{c} \rightarrow \hat{c}^{\circ}$ 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 is straightforward:

$$\hat{c}^{\circ} = \arg\max_{\omega} \; p(\omega \mid \boldsymbol{y}, \boldsymbol{\beta}) \quad \forall \omega. \tag{3.15}$$

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

$$\hat{c}^{\circ} = \begin{cases} 0 & \hat{c} < \pi_c \\ 1 & \hat{c} \geq \pi_c \end{cases}, \tag{3.16}$$

with $\pi_c = 0.5$. However, since these assumptions are often only partially true, most models are able to achieve better agreement with manual segmentations using a different threshold π_c for the WMH class. In fact several of the freely available toolboxes (cf. § **??**) 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 [181]. In the current work, both these techniques are employed: π_c is optimized on the fly, and a PR curve is given after.

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 $\mathrm{x}^c_{\mathrm{min}} = 3.5$ mm^3 in [64, 125], to 9.0 mm^3 in [117, 182].

In the current work, the inclusion of a minimum lesion size rule is explored. The optimal value for $\mathrm{x}^c_{\mathrm{min}}$ 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 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 $\boldsymbol{\beta}(x)$ 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 π_c and minimum lesion size $\mathrm{x}^c_{\mathrm{min}}$ are estimated using the training data. This completes the training phase.

At test time, SPM 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 $\boldsymbol{\beta}(x)$ 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 π_c, and binary objects smaller than $\mathrm{x}^c_{\mathrm{min}}$ are removed. The resulting label image is the final output. These training and testing phases are illustrated in Figure 3.4.

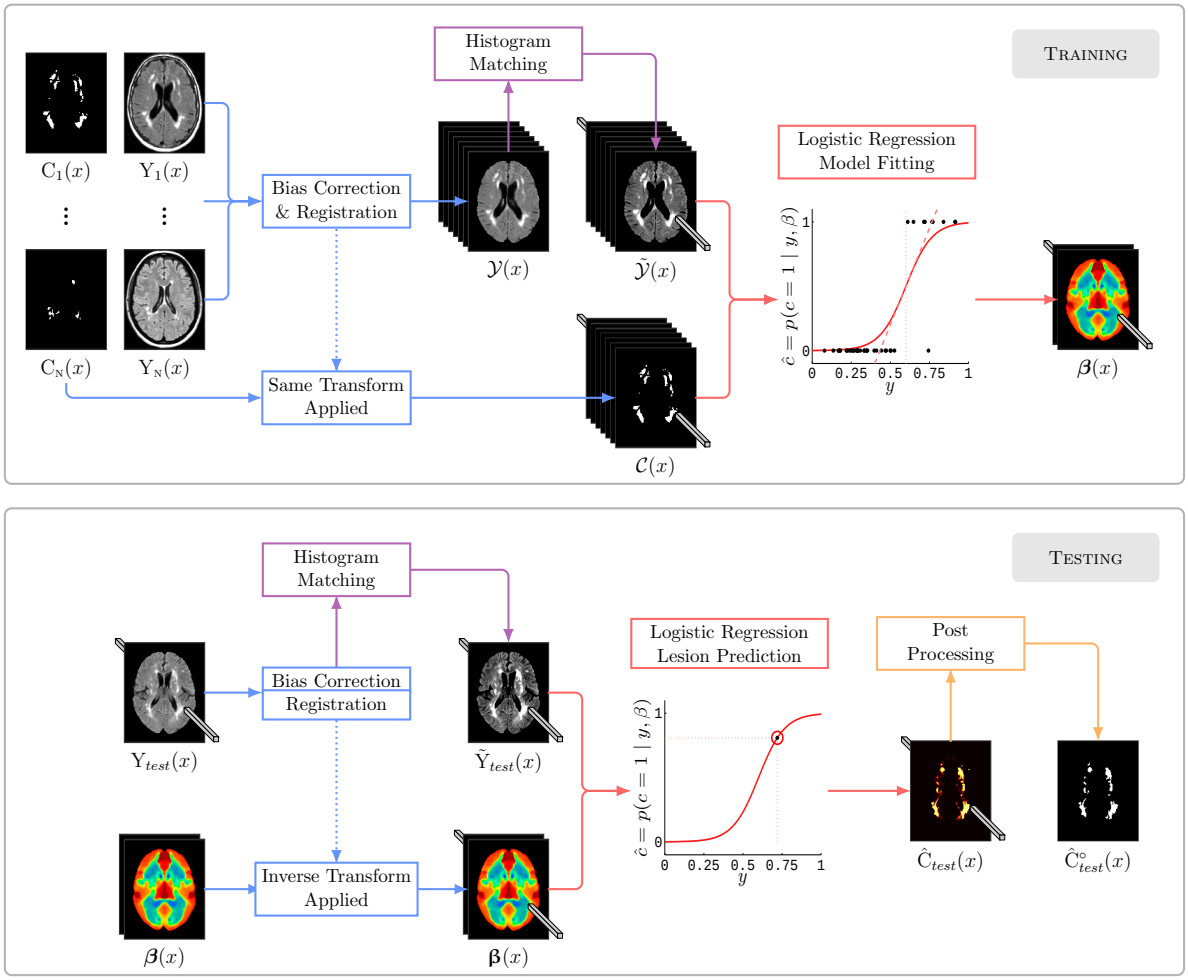


Figure 3.4: Overview of the proposed algorithm. Typefaces – upright Roman: images in native space; italic Roman: images in standard (MNI) space; calligraphic: a set of images from several patients; bold: a set of images corresponding to different features; Variables – $C(x)$: manual segmentation; $Y(x)$: FLAIR image; $\beta(x)$: parameter image; $\hat{C}(x)$: estimated lesion segmentation.

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. $\boldsymbol{\beta}(x)$ and π_c. 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 and also programmatically in **??**. The optimization of these model components will be the subject of the next chapter.

Table 3.2: Model hyperparameters and baseline values.

Stage	Parameter	Notation	Type	Baseline
Pre-Processing	Reflect Augmentation	$\mathrm{a_R}$	$\mathbb{B}$	`false`
	Shift Augmentation	$\mathrm{a_S}$	$\mathbf{N}_p$	$\mathbf{N}_0$
	Graylevel Transform	τ_y	$f : \mathbb{R} \mapsto \mathbb{R}$	$\boldsymbol{\tau}_{\mathbf{RM3}}$
	Transform Mask	$\mathcal{X}_\tau$	$\mathbb{B}(x)$	$\mathcal{X}_{\text{brain}}$
VLR Fitting	Iterations	T	$\mathbb{Z}$	30
	Initial $\boldsymbol{\beta}$	$\boldsymbol{\beta}^{(0)}$	$\mathbb{R}^2$	$[0, 0]$
	Estimation Scale[a]	r	$\mathbb{R}$	0.5
	Learning Rate	α	$\mathbb{R}$	1
	Regularization	λ	$\mathbb{R}$	0
	Pseudo-Lesions	$\boldsymbol{\mathcal{V}}(x)$	$\{\cdot \in \mathbb{R}\}$	$\{\}$
	$\boldsymbol{\beta}$ Filter	$F_{\boldsymbol{\beta}}$	$f : \mathbb{R}(x) \mapsto \mathbb{R}(x)$	$\tilde{\boldsymbol{\beta}}(x) = \boldsymbol{\beta}(x)$
Post-Processing	Min Lesion Size	$\mathrm{x}^c_{\text{min}}$	$\mathbb{R}$ (mm^3)	0

Notation. $\mathbb{B}$: boolean value; $\mathbb{Z}$: integer value; $\mathbb{R}$: real value; $\mathbb{R}^n$: vector; $\mathbb{R}(x)$: image; $\mathbf{N}_p$: 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. Two datasets were defined using 129 FLAIR images from from 10 different scanners. 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, results are primarily presented using only these freely available data ("Dataset A"), though all the available data ("Dataset B") are also used for some results. The average distribution of WMH in Dataset A is shown in Figure 4.1 for reference.

Regarding simulated MR images, none are used for validation of segmentation performance in this work. While such data (e.g. BrainWeb [184]) 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

[1] The manual segmentations used in this dataset were generated in-house, as described in § B.2.1.

Table 4.1: Summary of experimental image database.

Img (#)	Database	Ref.	Scanner	TE (ms)	TR (ms)	TI (ms)	Voxel Size (mm)	Manuals (#)
20	WMH 2017 (1)	[12]	3T Philips Achieva	125	11000	2800	$0.96 \times 0.96 \times 3.00$	1 [a]
20	WMH 2017 (2)	[12]	3T Siemens TrioTim	82	9000	2500	$1.00 \times 1.00 \times 3.00$	1 [a]
20	WMH 2017 (3)	[12]	3T GE Signa HDxt	126	8000	2340	$0.98 \times 1.20 \times 3.00$	1 [a]
5	MS 2016 (1)	[83]	3T Philips Ingenia	360	5400	1800	$0.50 \times 1.10 \times 0.50$	7 [b]
5	MS 2016 (2)	[83]	1.5T Siemens Aera	336	5400	1800	$1.04 \times 1.25 \times 1.04$	7 [b]
5	MS 2016 (3)	[83]	3T Siemens Verio	399	5000	1800	$0.74 \times 0.70 \times 0.74$	7 [b]
21	ISBI MS 2015	[57]	3T Philips	68	11000	2800	$0.43 \times 0.43 \times 3.00$	2 [c]
13	In-House	—	3T Philips Achieva	125	9000	2800	$1.00 \times 1.00 \times 1.00$	1 [d]
10	MS 2008 CHB	[82]	—	—	—	—	$0.50 \times 0.50 \times 0.50$	1 [e]
10	MS 2008 UNC	[82]	3T Siemens Allegra	125	9000	2800	$0.50 \times 0.50 \times 0.50$	1 [f]

[a] 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 [183]; [c] Manuals were fused using logical 'and'; [d] Manuals were generated in-house; [e] [X]; [f] [X].

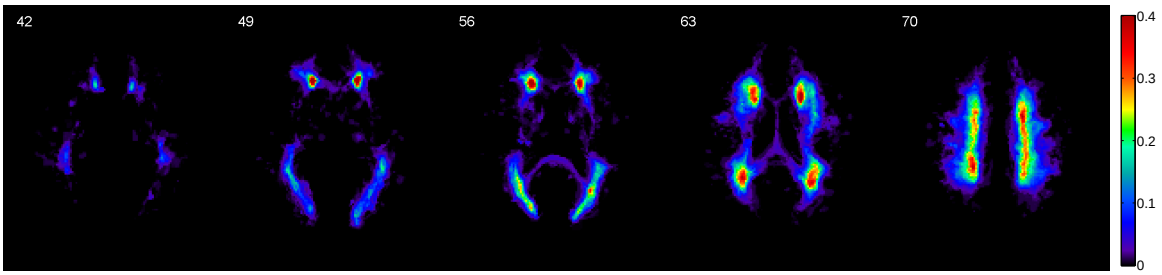


Figure 4.1: Average distribution of WMH in Dataset A.

were generated.[2] Furthermore, several works [185, 186] have noted significant discrepancies in estimated segmentation performance between BrainWeb and real data (ISBR [161]).

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".

[2] Documentation found here: http://brainweb.bic.mni.mcgill.ca/brainweb/selection_ms.html.

- False Negative (FN): $c = 1$ and $\hat{c} = 0$, incorrectly predicted "healthy".
- True Negative (TN): $c = 0$ and $\hat{c} = 0$, correctly predicted "healthy".

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.

$$SI = \frac{2TP}{2TP + FP + FN} \tag{4.1}$$

- **Precision (Pr)** (AKA Overlap Fraction, Positive Predictive Value)
  Fraction of predicted predicted positives which are true positives.

$$Pr = \frac{TP}{TP + FP} \tag{4.2}$$

- **Recall (Re)** (AKA Sensitivity, True Positive Rate)
  Fraction of true positives which are predicted positive.

$$Re = \frac{TP}{TP + FN} \tag{4.3}$$

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] [187], while trends in over/undersegmentation with lesion load are illustrated using Blant-Altman plots [188].

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

[3] Option `'A-1'` in the Matlab function `ICC` from https://www.mathworks.com/matlabcentral/fileexchange/22099

not represent the full gamut of potential input data, they risk *overfitting*: acquiring a bias towards the training data [189]. The main problem associated with overfitting is decreased performance on new data (AKA generalization performance) [189]. Popular techniques for characterizing this expected decrease 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 [190]. 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.
  Benefit: Close approximation of the expected generalization performance
  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; and these as the test set ($N_r = N - B$; $N_e = B$); repeat K times (without replacement).
  Benefit: Less expensive to compute – $\mathcal{O}(N/K)$
  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 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.
  Benefit: Estimates source-specific generalization performance
  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.
  Benefit: Characterize model limitations; least expensive to compute – $\mathcal{O}(1)$
  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.

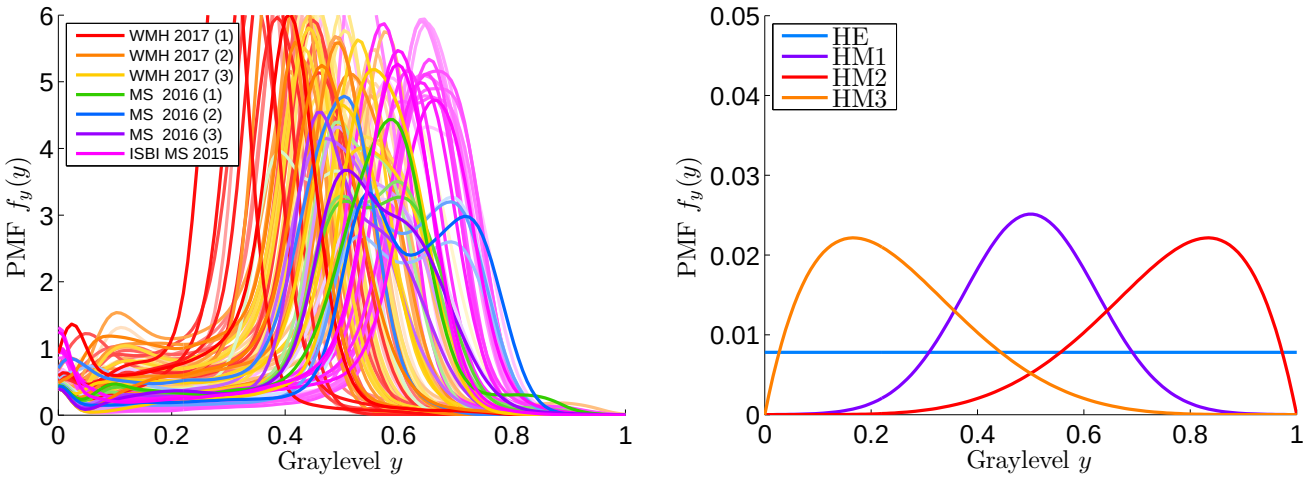
4.3.1 Leave-One-Source-Out CV

The choice of cross validation framework can have significant impacts on the reported model performance (see [190] 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) [191]. 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 authors propose the following:

- **LOSO – Leave-One-Source-Out:**[4] Withhold all examples from source $s \in 1 \dots S$ 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"



(a) Raw image PMFs before standardization.

(b) Target PMFs for histogram matching operations.

Figure 4.2: Image intensity PMFs.

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.

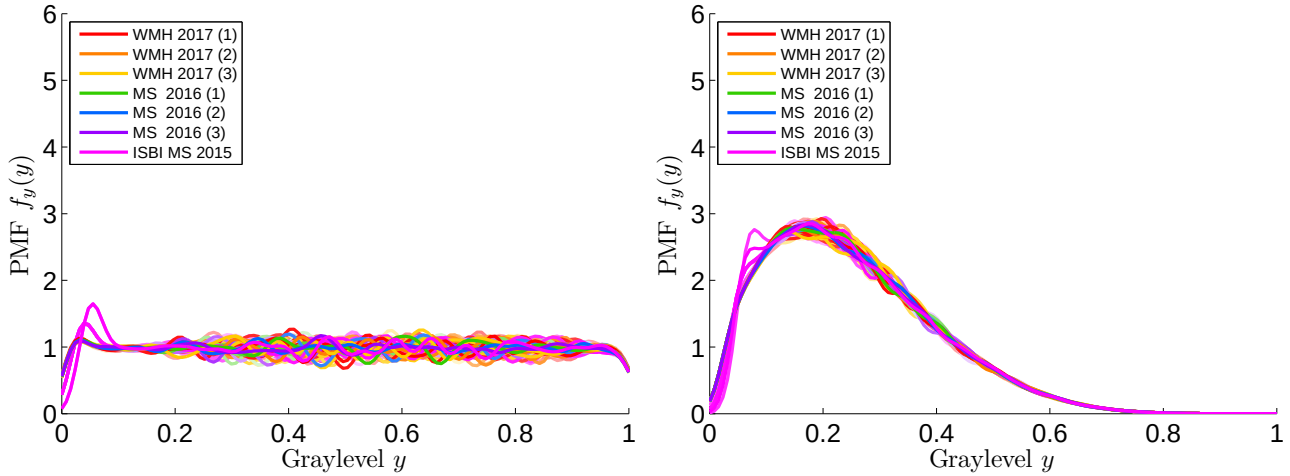
Since the truly raw image intensities range from $[0, 129]$ to $[0, 77537]$, some minimal standardization is required as a baseline and to allow visualization. For this, range matching with $\boldsymbol{\epsilon} = [10^{-4}, 1 - 10^{-4}]$ is used. Figure 4.2a illustrates the "raw" image PMFs from Dataset 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.

τ	Parameters	$\mathbb{E}[\mathcal{Z}_\Delta]$	$\mathbb{E}[\mathcal{Z}_\star]$	FFI[a]
RM1	$\boldsymbol{\epsilon} = [10^{-4}, 1-10^{-4}]$	16.1	7.42	
RM2	$\boldsymbol{\epsilon} = [10^{-3}, 1-10^{-3}]$	16.7	7.70	
RM3	$\boldsymbol{\epsilon} = [10^{-2}, 1-10^{-2}]$	15.5	7.52	
SS	—	12.2	**5.77**	$\star$
HE	—	**10.0**	8.16	$\star$
HM1	$f_{\tilde{y}} = \mathcal{N}(\frac{1}{2}, \frac{1}{8})$	11.5	6.86	$\star$
HM2	$f_{\tilde{y}} = \gamma^5 - \gamma^6$	12.0	9.11	$\star$
HM3	$f_{\tilde{y}} = (1-\gamma)^5 - (1-\gamma)^6$	**10.2**	**6.44**	$\star$
NY	$Q = [0, \frac{1}{16}, \ldots, 1]$	12.0	8.98	

FFI: For further investigation.



(a) Histogram Equalization (HE).

(b) Histogram Matching 2: "Left Skew" (HM3).

Figure 4.3: Image graylevel PMFs after the two best standardization operations.

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 3[rd] 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 2[nd] histogram matching operation (**HM2**); therefore these will not be subject to further investigation. Finally, since the graylevel agreement measures $Z_\Delta(x)$ and $Z_\star(x)$ are computed voxel-wise, it is possible to show their distribution spatially. This is illustrated for the raw images and the best performing transformation:

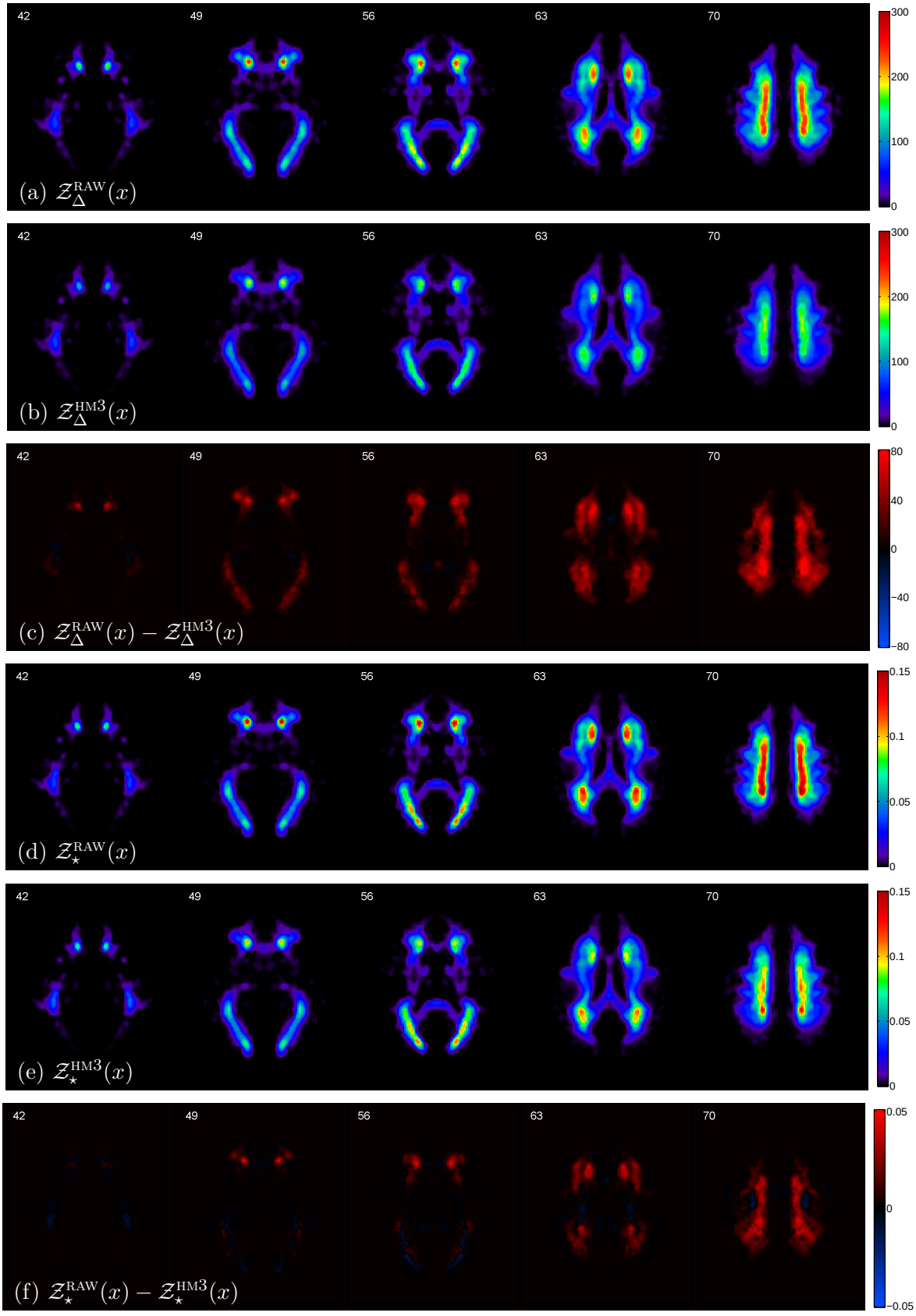


Figure 4.4: Spatial depiction of $\mathcal{Z}_{\Delta}(x)$ and $\mathcal{Z}_{\star}(x)$ comparing RM1 and HM2.

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

	$c=0$			$c=1$		
#	μ	σ	N	μ	σ	N
a	0.3	0.12	100	0.7	0.12	100
b	0.3	0.12	100	0.7	0.12	10
c	0.3	0.24	100	0.7	0.24	10
d	0.3	0.06	100	0.7	0.06	10
e	0.3	0.03	100	0.7	0.03	10
f	0.6	0.08	100	0.3	0.08	10
g	0.4	0.10	100	—	—	0
h	0.6	0.08	100	—	—	0
i	0.8	0.06	100	—	—	0

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 λ, and the definition of pseudo-lesions $\{\boldsymbol{\mathcal{Y}}_\gamma, \boldsymbol{\mathcal{C}}_\gamma\}$ 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 3.1 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 Classic Regularization $\boldsymbol{\lambda}$

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

[5] This popular technique is also called "hand waving".

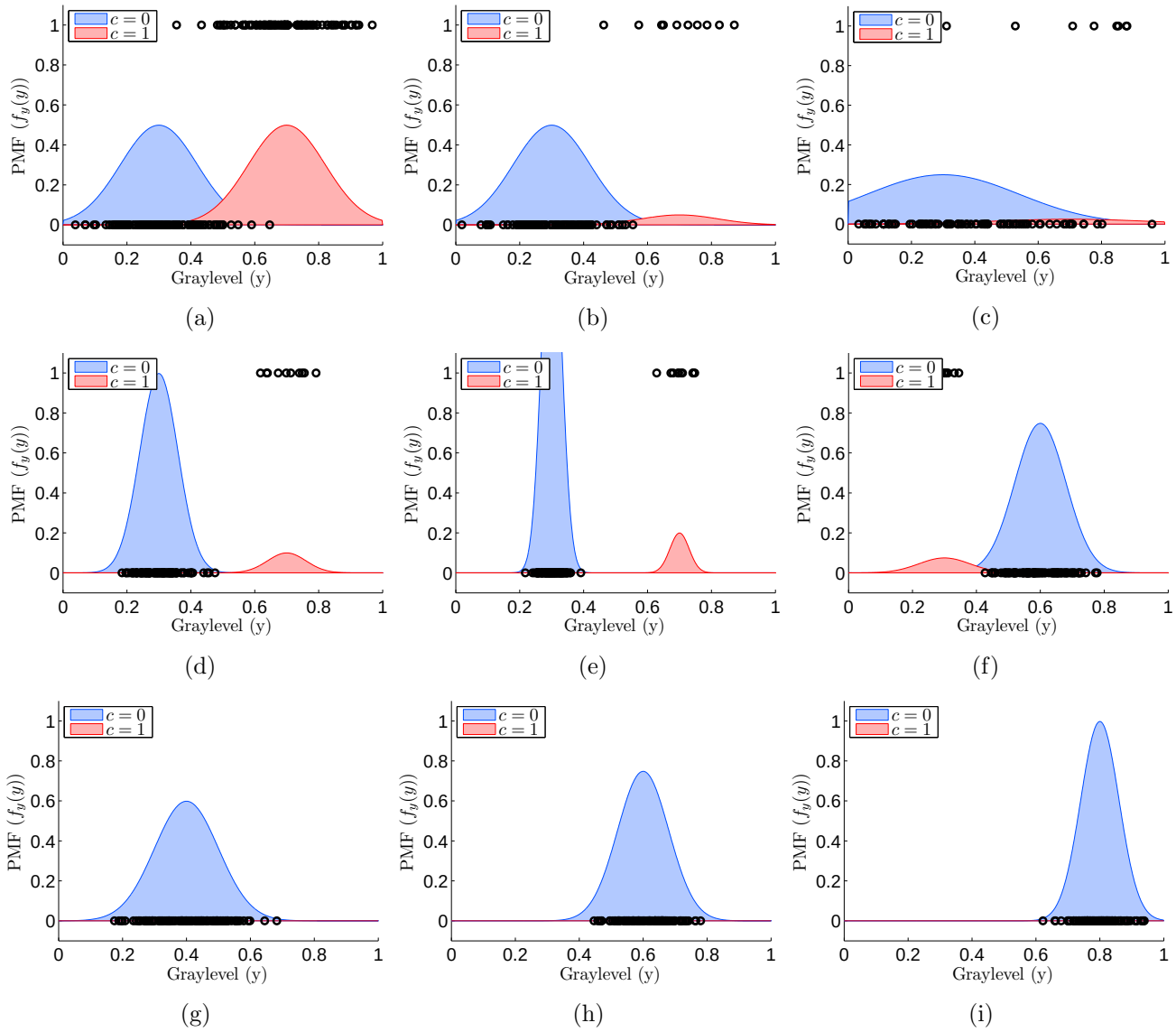
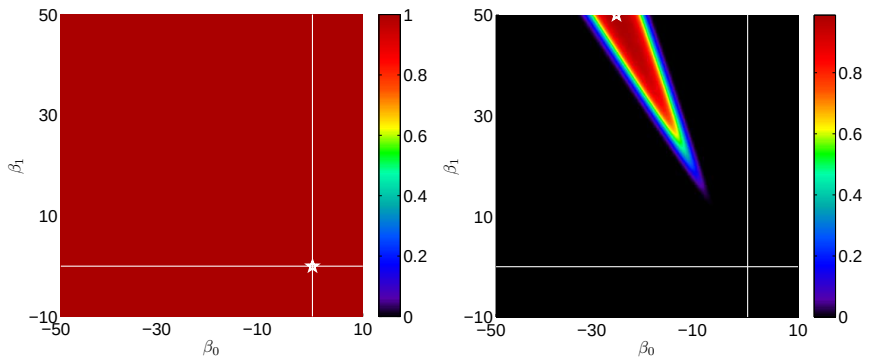


Figure 4.5: Synthetic data distributions (9 voxels) used in toy scenarios.

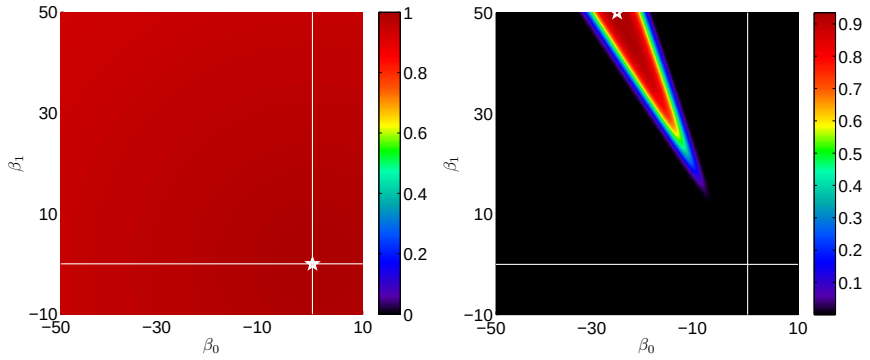
β^0 and β^1 values. The $\boldsymbol{\beta}$ prior function $\mathcal{P}(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2$ is computed similarly. These functions are then exponentiated, as in $J = e^{\mathcal{J}}$ and $P = e^{\mathcal{P}}$ – i.e. the likelihood, as opposed to the log-likelihood. For $\lambda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

Next, the appropriate λ 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



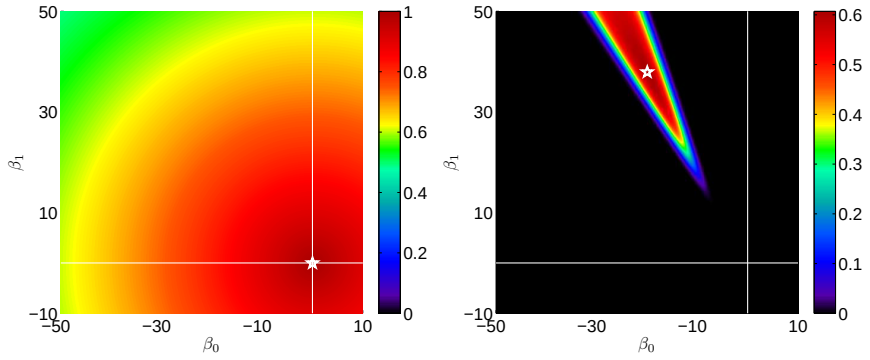
(a) $P(\boldsymbol{\beta}) = 1, \quad \lambda = 0$

(b) $J(\boldsymbol{\beta}) = L(\boldsymbol{\beta}), \quad \lambda = 0$



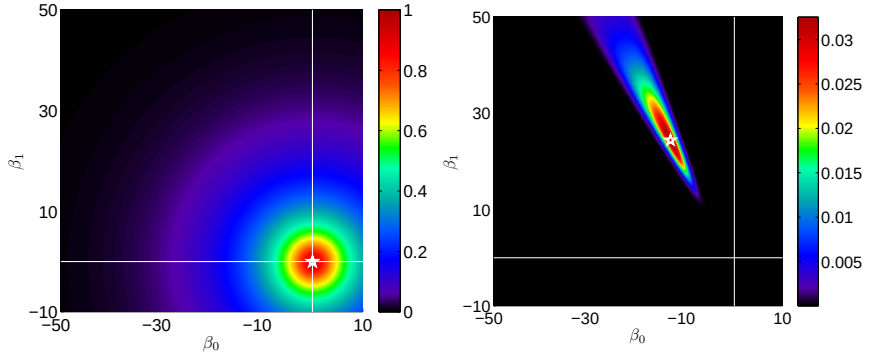
(c) $P(\boldsymbol{\beta}), \quad \lambda = 10^{-3}$

(d) $J(\boldsymbol{\beta}), \quad \lambda = 10^{-3}$



(e) $P(\boldsymbol{\beta}), \quad \lambda = 10^{-2}$

(f) $J(\boldsymbol{\beta}), \quad \lambda = 10^{-2}$



(g) $P(\boldsymbol{\beta}), \quad \lambda = 10^{-1}$

(h) $J(\boldsymbol{\beta}), \quad \lambda = 10^{-1}$

Figure 4.6: Toy model likelihoods as a function of $\boldsymbol{\beta}$: $P(\boldsymbol{\beta}) = e^{\mathcal{P}(\boldsymbol{\beta})}$ and $J(\boldsymbol{\beta}) = e^{\mathcal{J}(\boldsymbol{\beta})}$, for different λ, using scenario e. The optimum is shown as a white star.

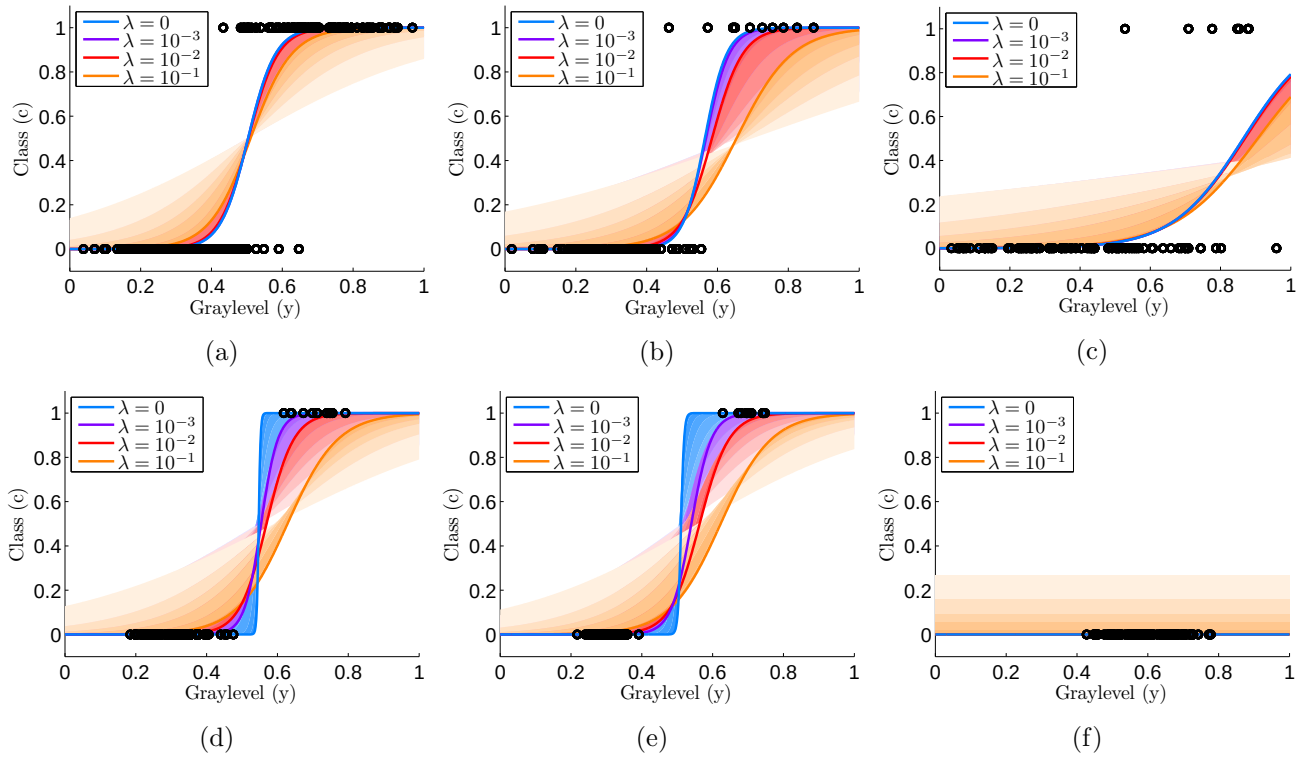


Figure 4.7: Toy model MAP estimation results for 6 different scenarios and different λ.

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 $\{\mathcal{Y}_{c=1} \mid y < \mathbb{E}[\mathcal{Y}_{c=0}]\}$ 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), λ plays an important role in limiting the magnitude of $\boldsymbol{\beta}$. 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 $\mathcal{V} = y_{\max} = 1$ 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 τ 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.

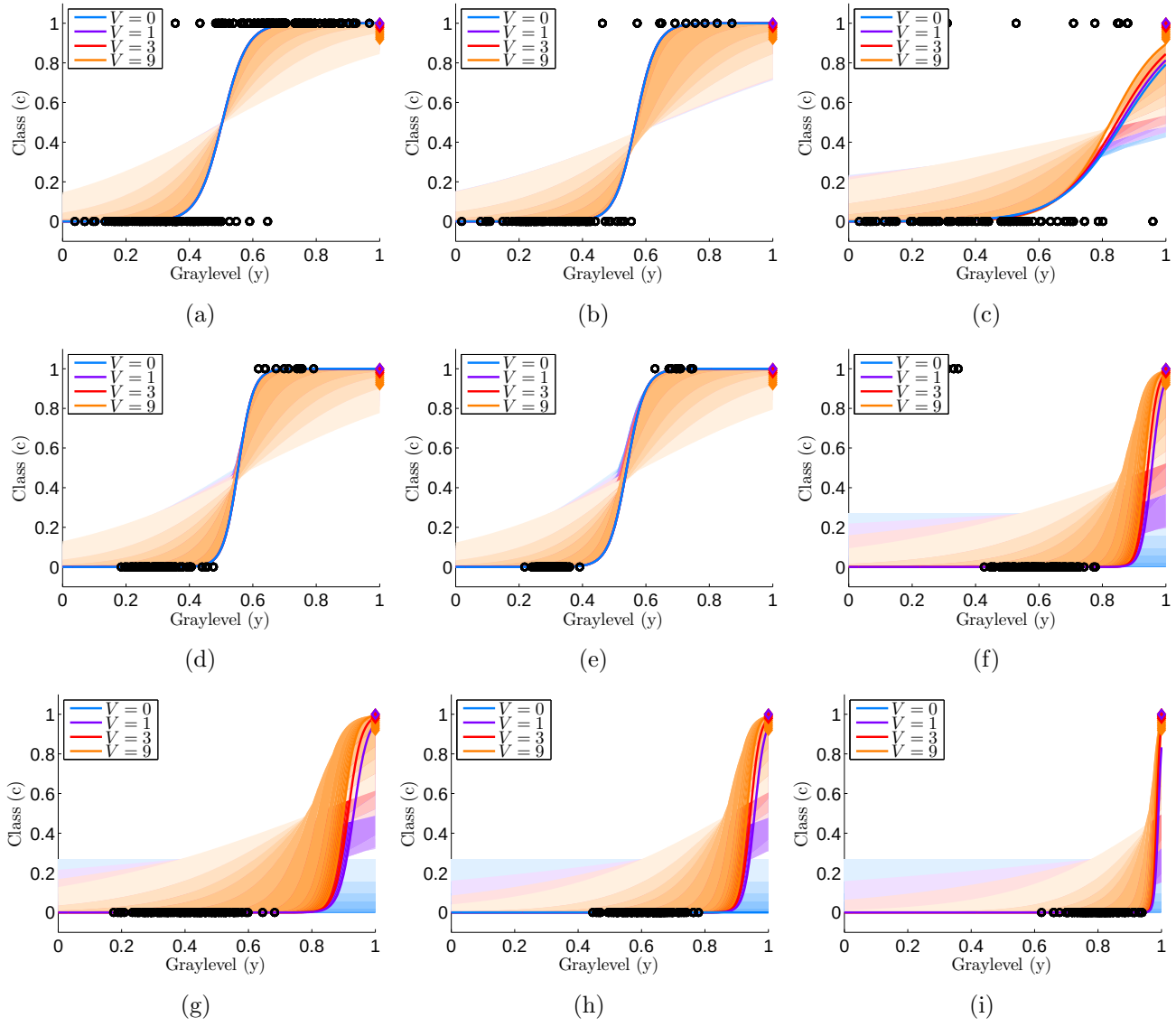
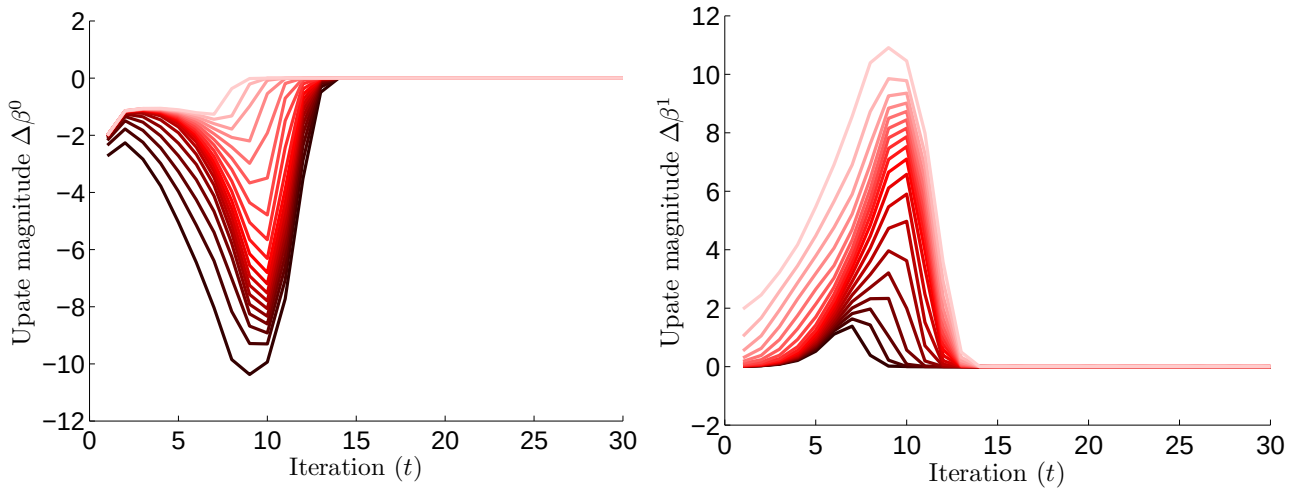


Figure 4.8: Toy model MAP estimation results for 9 different scenarios and different numbers of pseudo-lesions V, shown as coloured diamonds corresponding to the scenario (spread of diamonds is for visualization only).



(a) β^0 (b) β^1

Figure 4.9: Convergence characteristics of $\beta(x)$: update magnitude $\Delta\beta^{(t)}(x)$ quantiles $(0.05, \ldots, 0.95)$ versus iteration (t), using all augmented data from Dataset A. Convergence is apparent by the 15^{th} iteration.

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, $t_{\max}$, 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: $\mathrm{a_R} = \texttt{true}$ and shift one voxel in each dimension: $\mathrm{a_S} = \mathbf{N}_6$) starting from the default initialization $\boldsymbol{\beta} = [0, 0]^T$ for all voxels. The magnitude of $\Delta\beta$ (5^{th} to 95^{th} quantiles) was recorded for each fitting iteration and plotted, as shown in Figure 4.9.

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

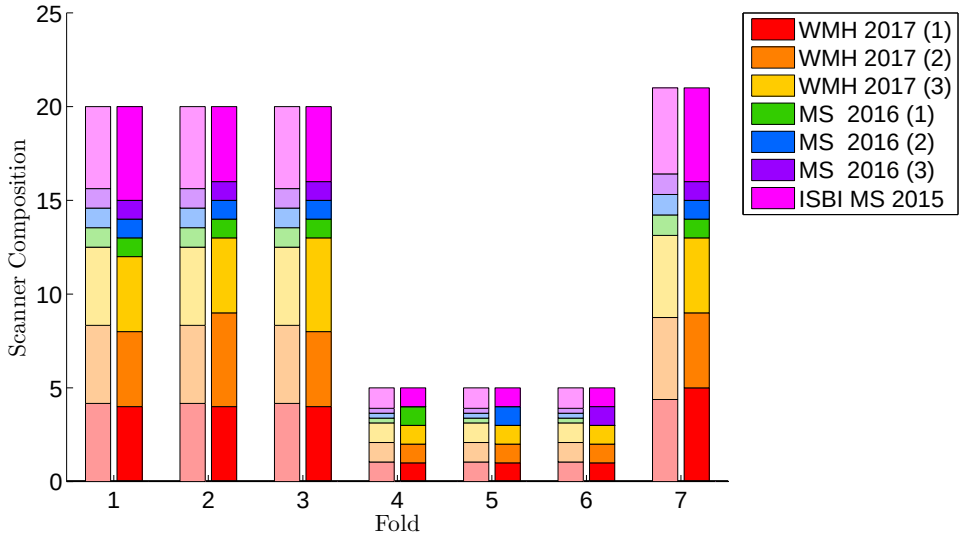
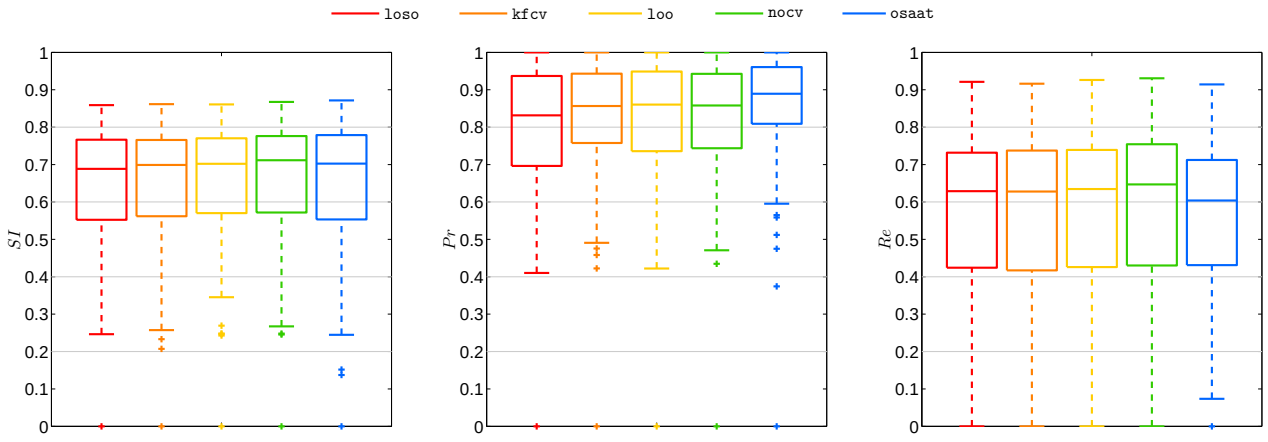


Figure 4.10: Number of images from each scanner in each KF-CV fold. Faded colours show the expected value (evenly distributed, non-whole numbers); full colours show the implementation (approximation, whole numbers).

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, 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 $\approx$ LOO-CV $\approx$ 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



(a) SI (b) Pr (c) Re

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

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.

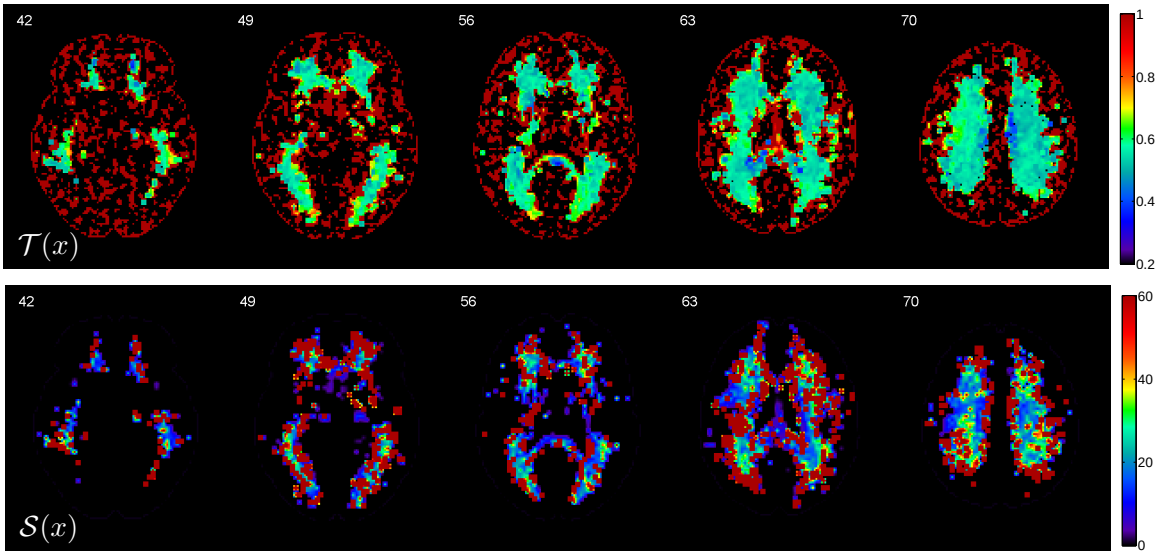


Figure 4.12: Fitted parameter images $\mathcal{T}(x)$ and $\mathcal{S}(x)$ from the first LOSO-CV fold of the baseline model. Obvious artifacts arise from inadequate regularization.

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.

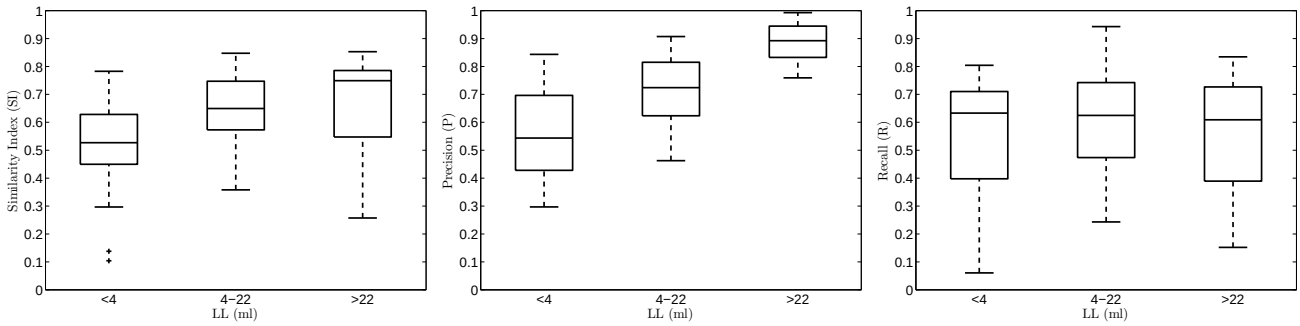
The fitted parameter images from the baseline model are shown in Figure 4.12. While the threshold image contains reasonable values (near $y_{\max}$) in the regions typically containing WMH (Figure 4.1), there are obvious artifacts throughout both images, corresponding to locations where no lesions were 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 0.63, Precision 0.77, and Recall 0.63. 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 $Pr > Re$, it can be inferred that the model incurs more FN than FP – i.e. it is more specific than sensitive. This would therefore predict an underestimation of the total LL. The performance among the three MS 2016 scanners is also low. This may be attributable to their notably different TE/TR/TI

Table 4.4: Baseline model performance metrics (median)

Scanner	LL	SI	Pr	Re
WMH 2017 (1)	24	0.65	0.80	0.53
WMH 2017 (2)	17	0.75	0.85	0.68
WMH 2017 (3)	6	0.70	0.72	0.72
MS 2016 (1)	29	0.46	0.84	0.38
MS 2016 (2)	5	0.37	0.53	0.29
MS 2016 (3)	10	0.54	0.77	0.40
ISBI MS 2015	5	0.61	0.62	0.68
ALL	12	0.63	0.76	0.63



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

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

parameters (Table 4.1), which are simulated in Figure A.1. In particular, a large performance drop is observed for the data from Scanner 2 in the MS 2016 dataset; two factors may help explain these results. First, this is the only 1.5T scanner in the dataset, which implies higher levels of noise due to smaller MR signal magnitude during image acquisition. Second, the median LL for these subjects was only 5 mL, which is expected to correlate with decreased performance, as noted above.

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 the near equivalence of these two methods (paired SI test: $p = 0.4923$), the objective function results from § 4.4 were used to select **HM3** 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
- Classic regularization: λ
- Data augmentation: a_R – reflection; a_S – 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,

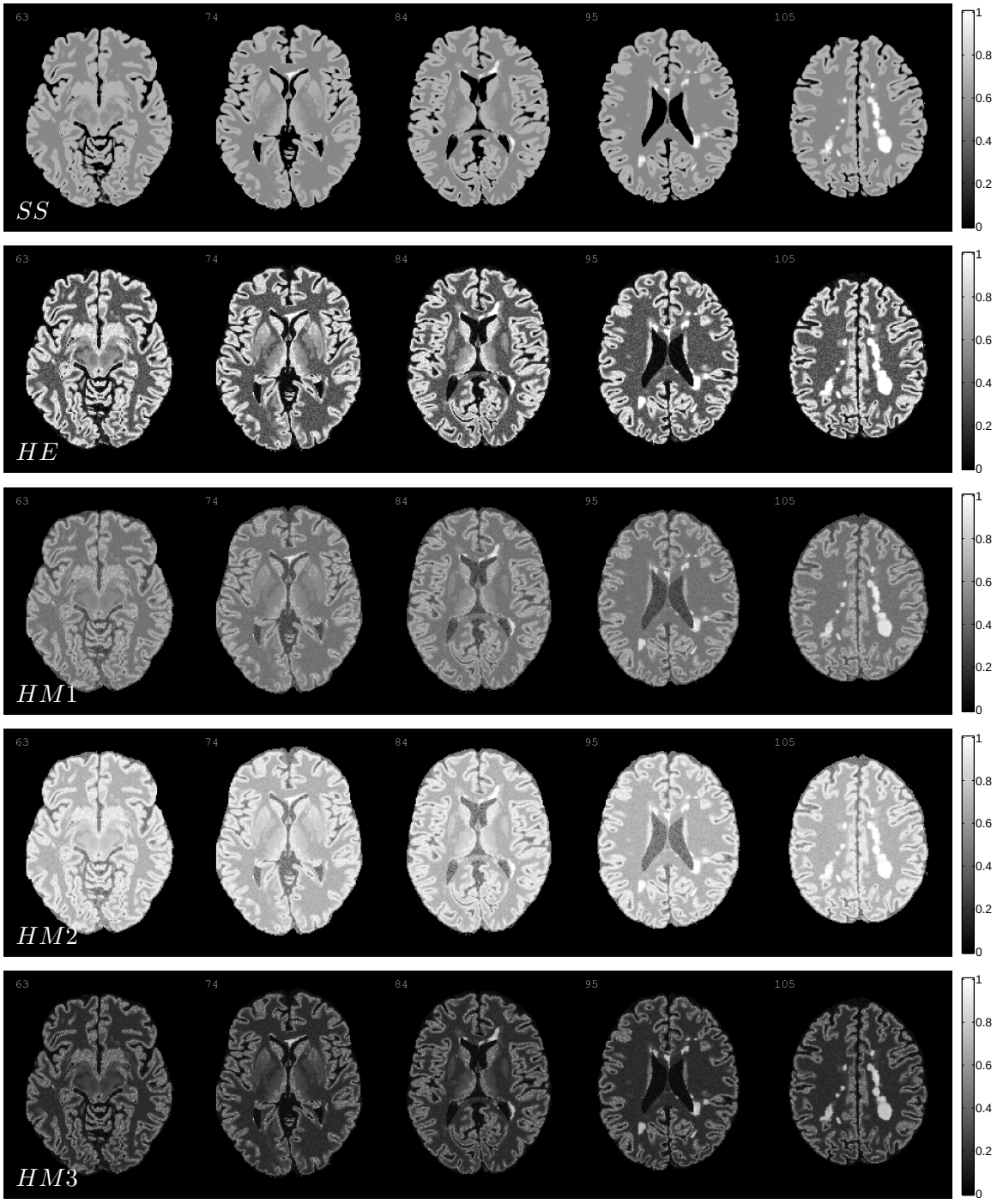
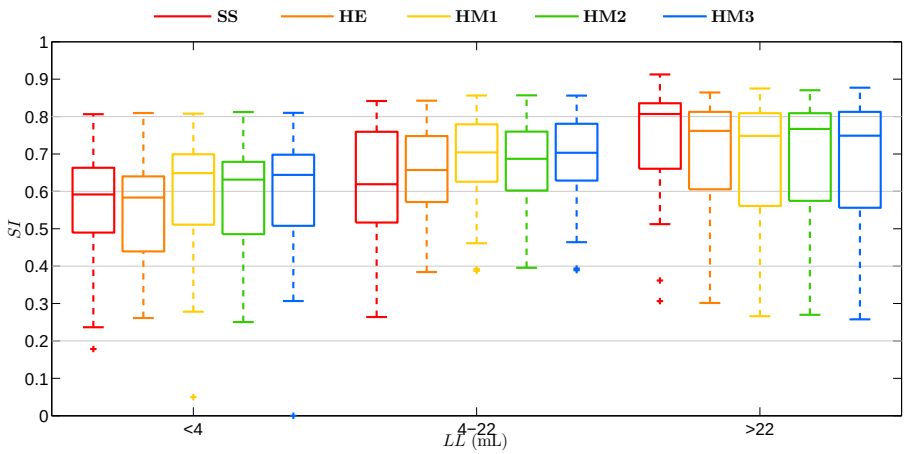
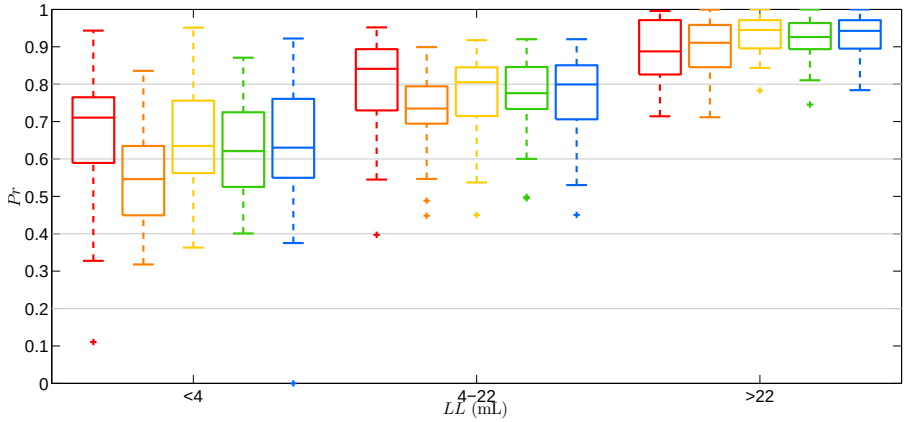


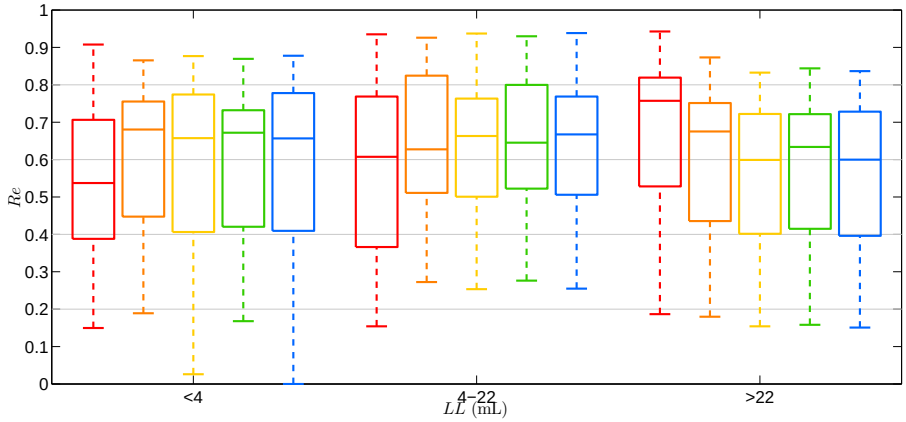
Figure 4.14: Simulated FLAIR images after graylevel standardization using each technique under investigation.



(a) SI

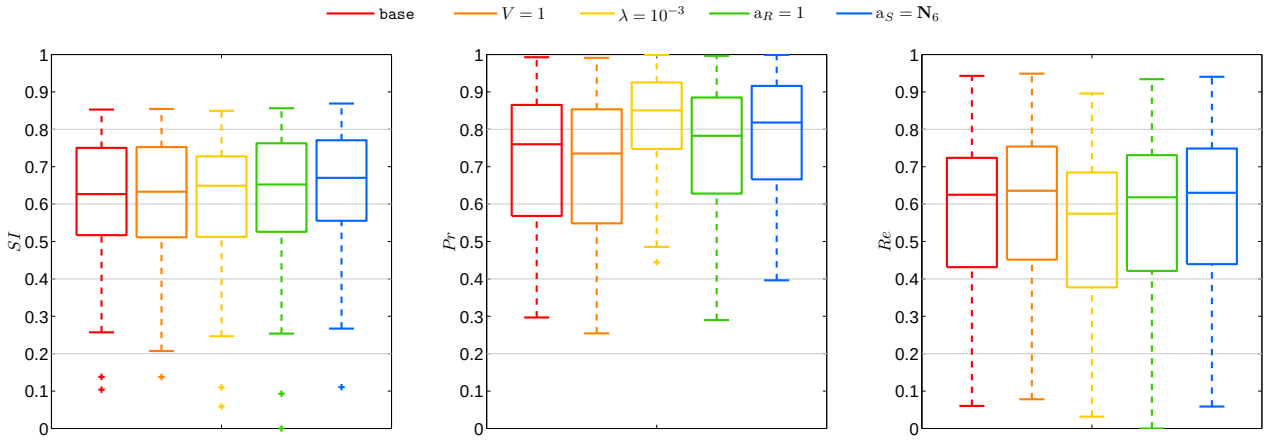


(b) Pr



(c) Re

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



(a) SI (b) Pr (c) Re

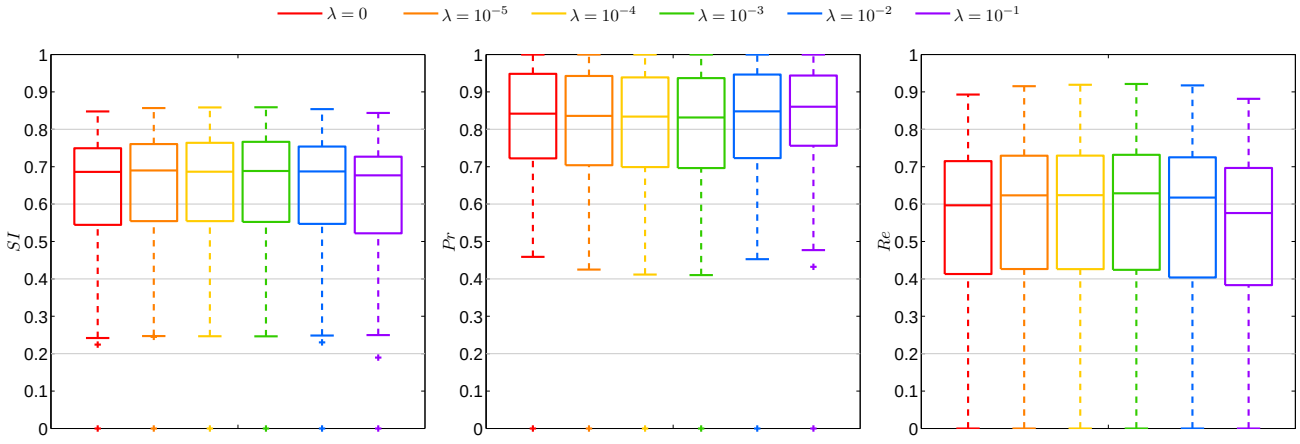
Figure 4.16: Comparison of the baseline model under LOSO-CV incorporating each of the regularization strategies in isolation.

unless otherwise specified, the default non-baseline parameter image smoothing: $G_{\sigma 2}$ was throughout this section.

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 3.1b 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 classic regularization, 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 λ.

The results in § 4.5.3 demonstrated that use of additional pseudo lesions 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 λ parameter was subject to further formal investigation in terms of segmentation performance.



(a) SI (b) Pr (c) Re

Figure 4.17: Comparison of the optimized model under LOSO-CV using different prior strengths λ.

Classic Regularization

Selection of the appropriate λ 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 $\boldsymbol{\beta}$ and maintaining MLE characteristics. A similar range of λ was explored in the full model: $[10^{-5}, 10^{-1}]$. Exploration of the full model 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 overal SI performance is surprisingly robust to the definition of λ. This may be attributable to the effects of other regularizations, especially the data augmentations and parameter image smoothing. However, a maximum in Recall is achieved using $\lambda = 10^{-3}$. This is desirable, since balancing Pr and Re should give more accurate total LL estimates, and Recall is often much lower than Precision. Therefore, $\lambda = 10^{-3}$ was selected as the optimal value.

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.

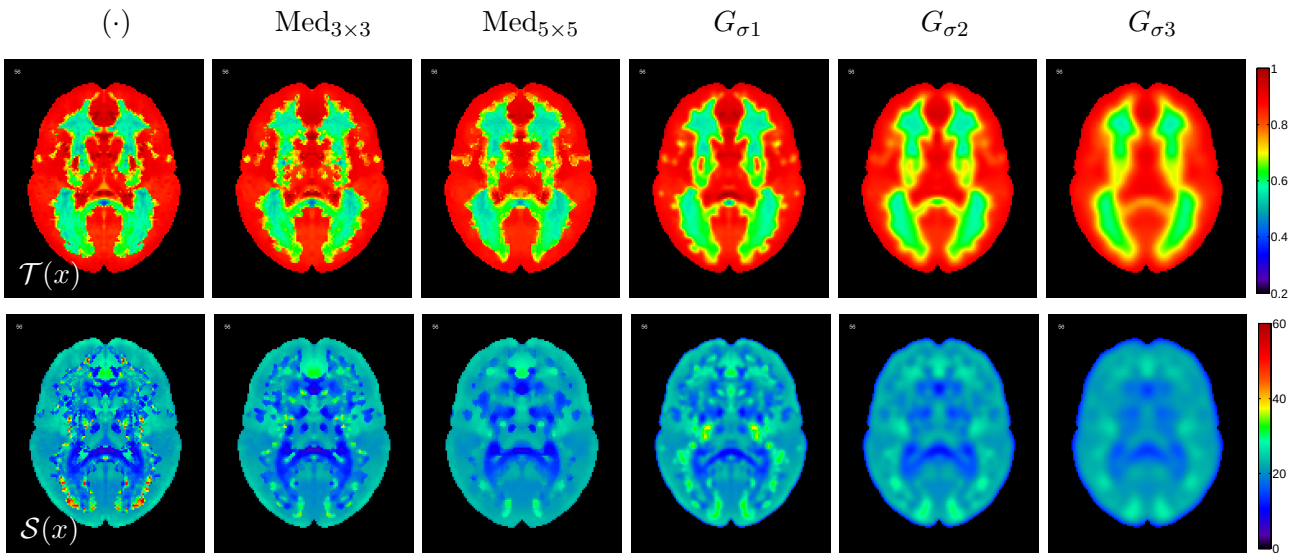


Figure 4.18: Parameter images following different smoothing filters.

Smoothing

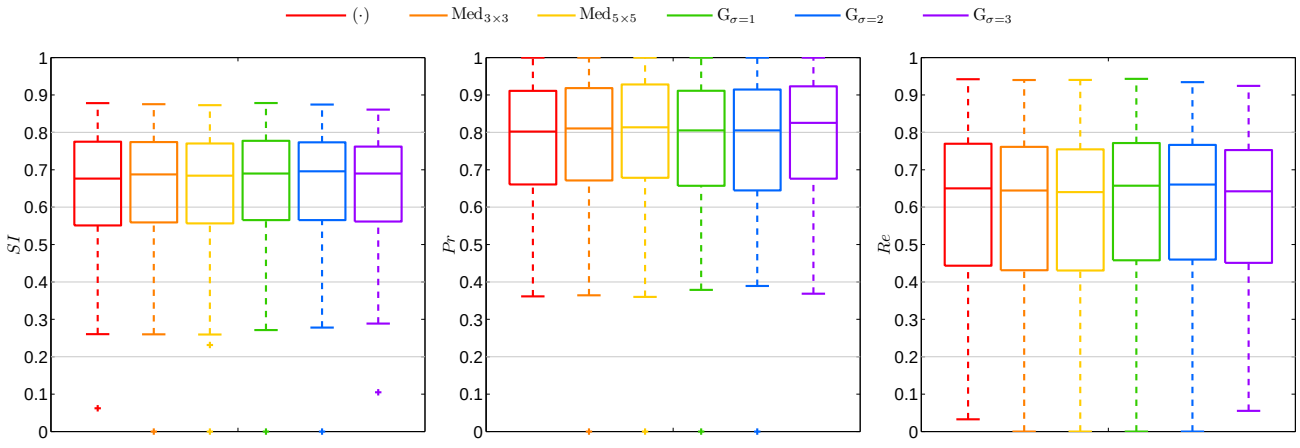
The regularizations described in § 4.7.2 were surprisingly effective at achieving these objectives, yielding the raw fitted parameter image shown in the left most column of Figure 4.18. However, several artifacts are visible, including voxels with very high magnitude in the sensitivity image $\mathcal{S}(x)$, and large discontinuities between the voxels which do and do not observe lesion examples during training in the threshold image $\mathcal{T}(x)$. 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 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 $\mathcal{T}(x)$ indicates the graylevels corresponding to a 50% probability of the lesion class $\hat{c}$, while the sensitivity



(a) SI (b) Pr (c) Re

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

image $\mathcal{S}(x)$ 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 boarder 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 $\mathcal{S}(x)$ is significantly less important for segmentation performance. One investigation which replaced this parameter image with it's 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 β^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[6] and reconstructed in MNI space, for comparison with the equivalent VLR-fitted

[6] `sp_mni2_Bf2` in the `LST_lpa_stuff.mat` datafile from the toolbox.

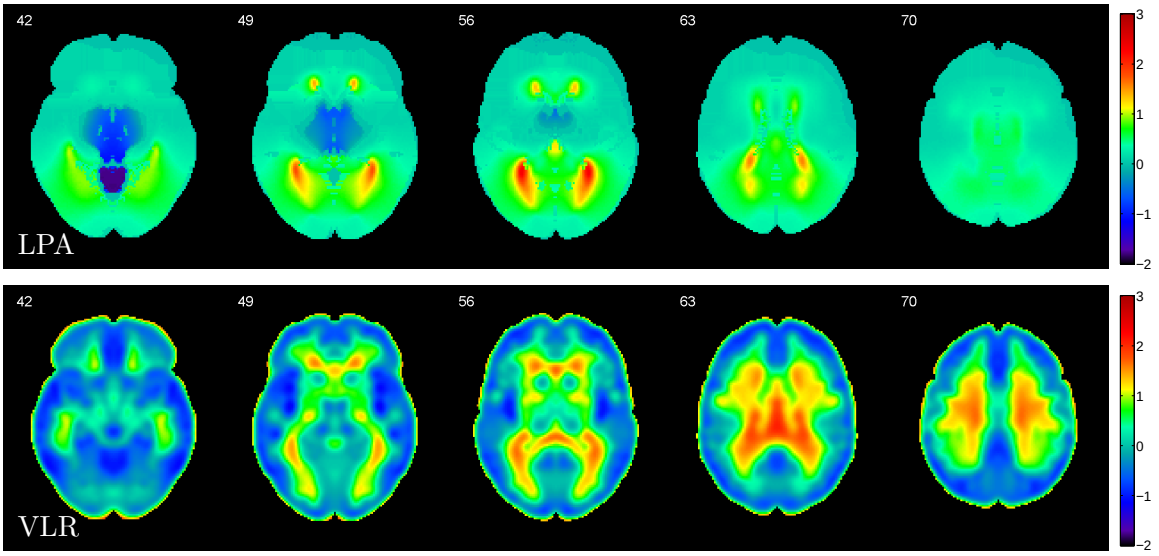


Figure 4.20: Spatial effect intercept parameter images $\beta^0(x)$ from the LPA and VLR algorithms. For visual comparison, image means and variances are matched.

parameter image.[7] 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 $\mathcal{T}(x)$. 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 the 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.

[7] 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.

Stage	Parameter	Notation	Type	Default
Pre-Processing	Reflect Augmentation	$\mathrm{a_R}$	$\mathbb{B}$	`true`
	Shift Augmentation	$\mathrm{a_S}$	$\mathbf{N}_p$	$\mathbf{N}_6$
	Graylevel Transform	τ_y	$f : \mathbb{R} \mapsto \mathbb{R}$	$\tau_{\mathbf{RM3}}$
	Transform Mask	$\mathcal{X}_\tau$	$\mathbb{B}(x)$	$\mathcal{X}_{\text{brain}}$
VLR Fitting	Iterations	T	$\mathbb{Z}$	30
	Initial $\boldsymbol{\beta}$	$\boldsymbol{\beta}^{(0)}$	$\mathbb{R}^2$	$[0, 0]$
	Estimation Scale	r	$\mathbb{R}$	0.5
	Learning Rate	α	$\mathbb{R}$	1
	Regularization	λ	$\mathbb{R}$	1×10^{-3}
	Pseudo-Lesions	$\boldsymbol{\mathcal{V}}(x)$	$\{\cdot \in \mathbb{R}\}$	$\{y_{\max}\}$
	$\boldsymbol{\beta}$ Filter	$F_{\boldsymbol{\beta}}$	$f : \mathbb{R}(x) \mapsto \mathbb{R}(x)$	$\tilde{\boldsymbol{\beta}}(x) = G_{\sigma 2}(\boldsymbol{\beta}(x))$
Post-Processing	Min Lesion Size	$\mathrm{x}^c_{\min}$	$\mathbb{R}$ (mm^3)	1

Notation. $\mathbb{B}$: boolean value; $\mathbb{Z}$: integer value; $\mathbb{R}$: real value; $\mathbb{R}^n$: vector; $\mathbb{R}(x)$: image; $\mathbf{N}_p$: 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 0.63, 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 (2) 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 significantly 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, no overall improvements in Recall were made during model optimization ($Re = 0.63$ again), 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.

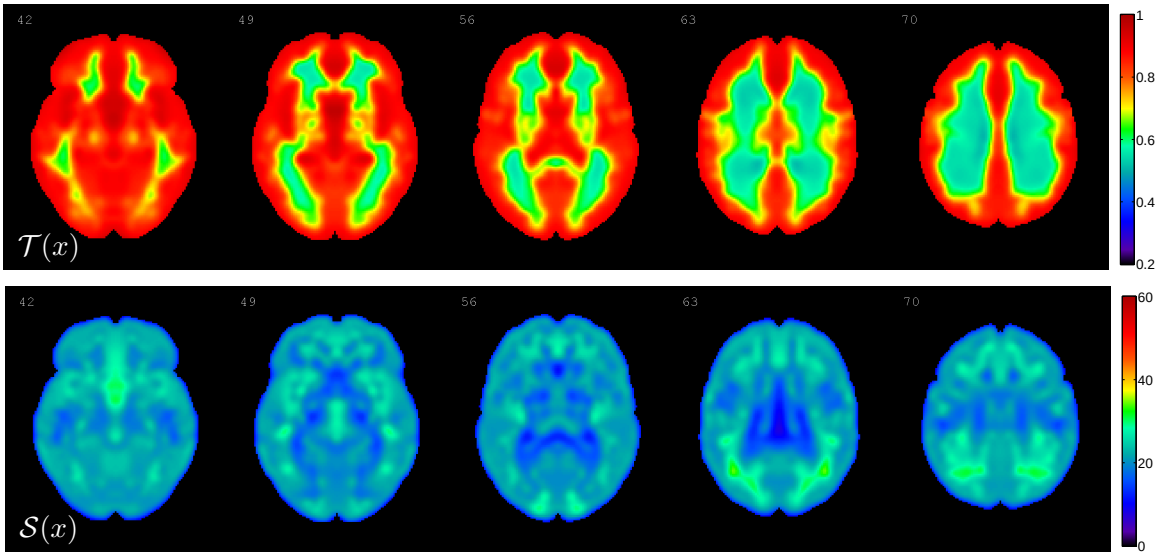
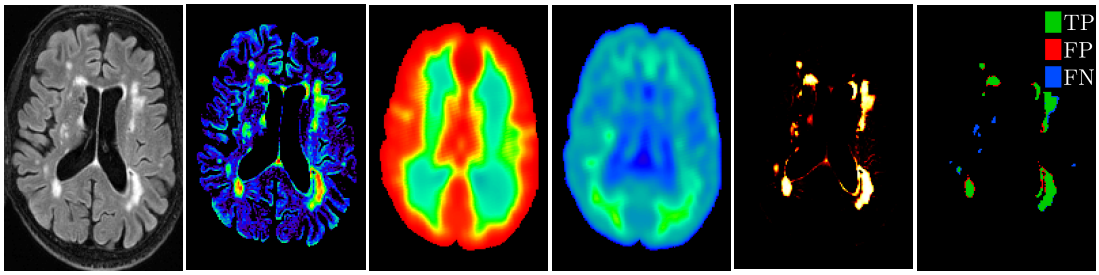


Figure 4.21: Fitted parameter images $\mathcal{T}(x)$ and $\mathcal{S}(x)$ from the first LOSO-CV fold of the final model.

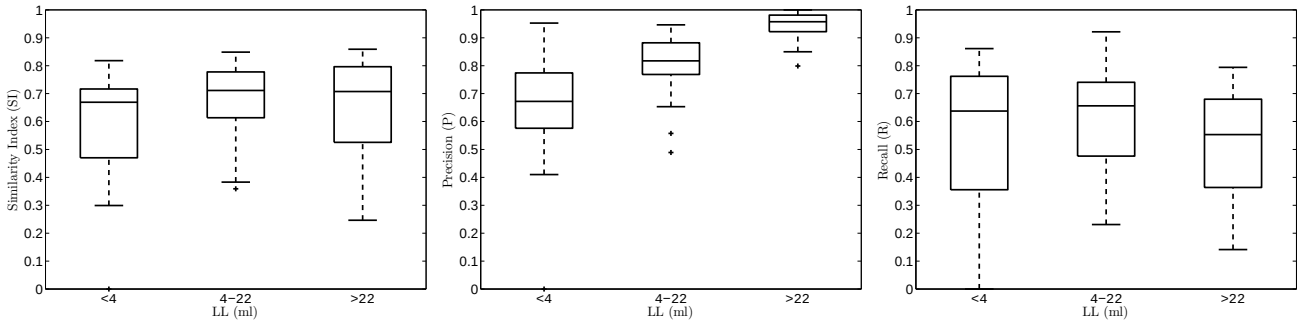


(a) $\mathrm{Y}(x)$ (b) $\tilde{\mathrm{Y}}(x)$ (c) $\mathrm{T}(x)$ (d) $\mathrm{S}(x)$ (e) $\hat{\mathrm{C}}(x)$ (f) $\cdot(x)$

Figure 4.22: Example segmentation.

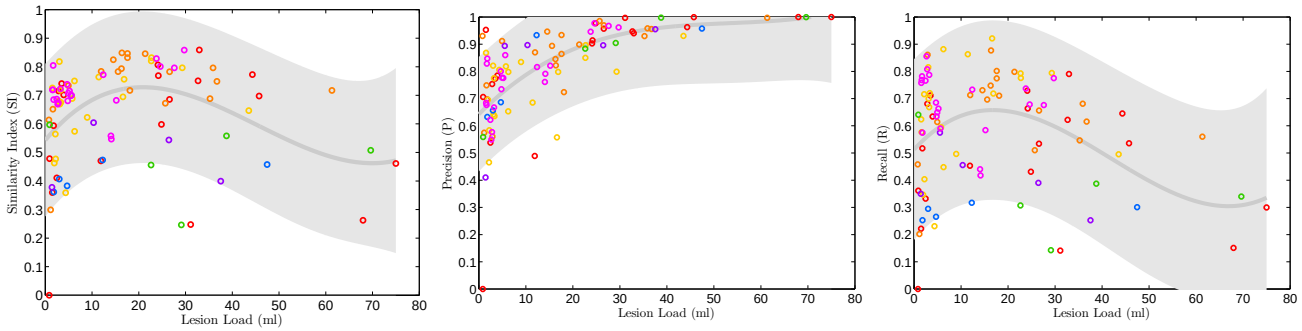
Table 4.6: Final model performance metrics (median)

Scanner	LL	SI	Pr	Re
WMH 2017 (1)	24	0.64	0.93	0.53
WMH 2017 (2)	17	0.77	0.91	0.69
WMH 2017 (3)	6	0.69	0.80	0.71
MS 2016 (1)	29	0.51	0.90	0.34
MS 2016 (2)	5	0.41	0.69	0.29
MS 2016 (3)	10	0.54	0.90	0.39
ISBI MS 2015	5	0.72	0.79	0.69
ALL	12	0.69	0.83	0.63



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

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 pof inal model performance, with 3$^{\text{rd}}$ order trend line and 90% confidence interval shown in grey.

It is worth drawing attention to one consistent outlier from the WMH 2017 (1) scanner, for which no TP were predicted (resulting in zero for all metrics). The cause of this error was imperfect registration to MNI space by SPM, resulting in bright skull tissue within the static brain mask. While this did not yield any FP, due to the ability of the VLR model to exclude these regions, the large number of other hyperintensities in the image affected the histogram matching operation used for standardization, resulting in WMH which were much darker than usual, as noted above.

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. This trend is likely a result of the histogram-matching graylevel standardization, since with histogram-based nonlinear transformations, more hyperintensities result in reduced contrast. This unfortunately led to a poor overall volume agreement, as measured by ICC: 0.58.

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

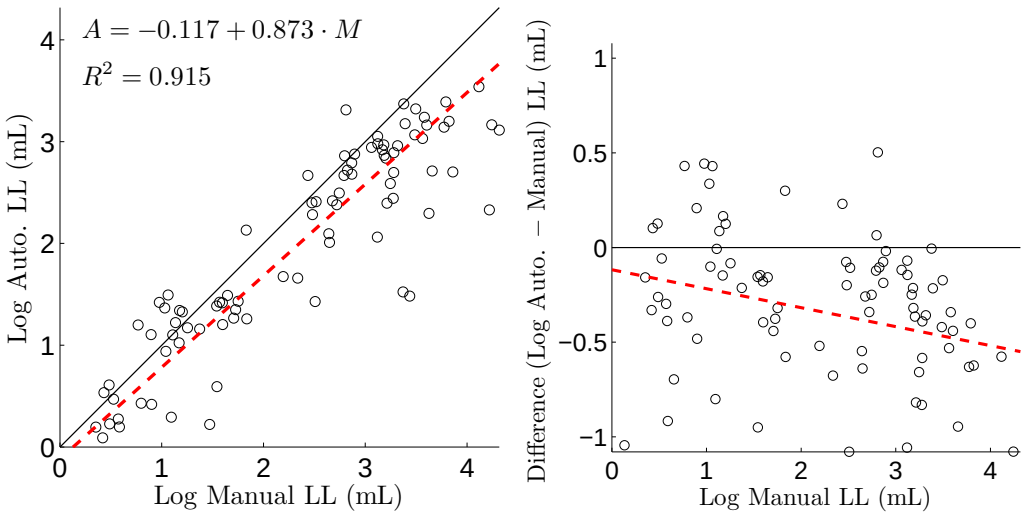


Figure 4.25: Bland-Altman plot showing total LL agreement between manual and VLR-segmented WMH. Shown in Log-scale to better illustrate results for small LL.

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 essentially identical, 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.

"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 = 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

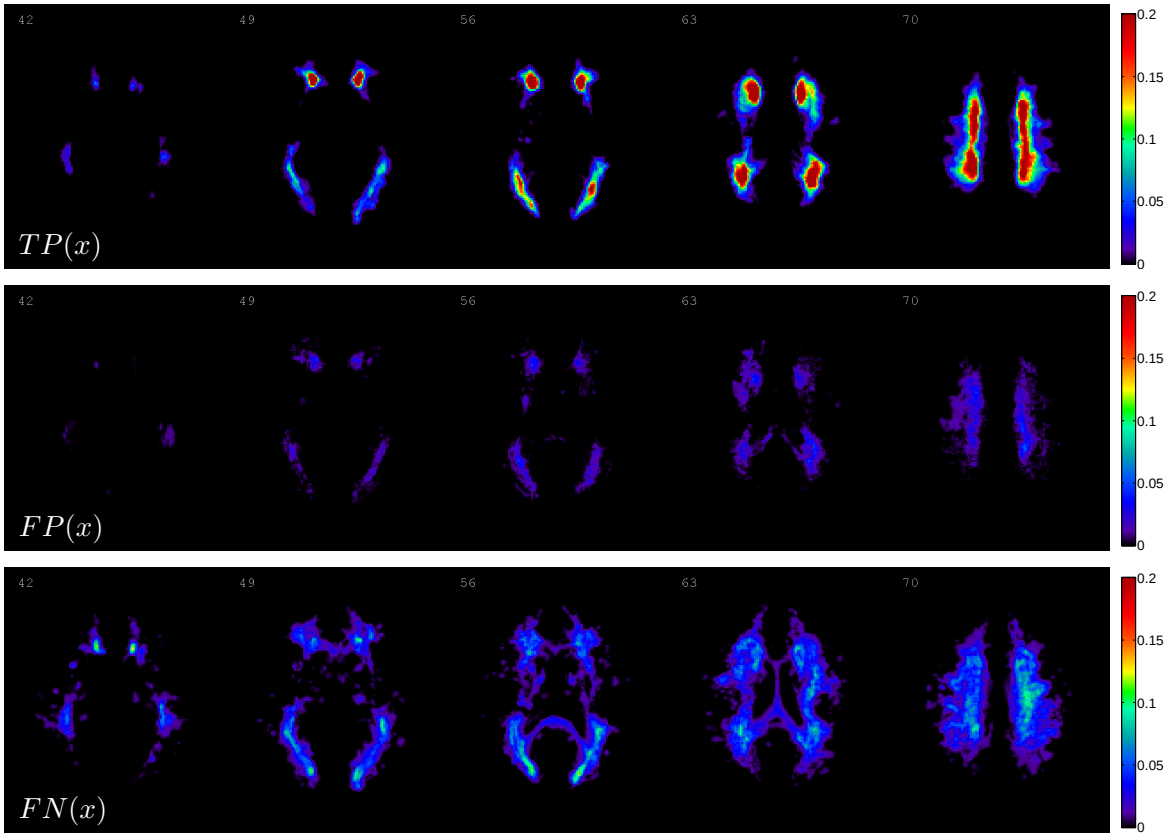


Figure 4.26: Distribution of True Positives (TP), False Positives (FP), and False Negatives (FN) from all LOSO-CV folds of the final model.

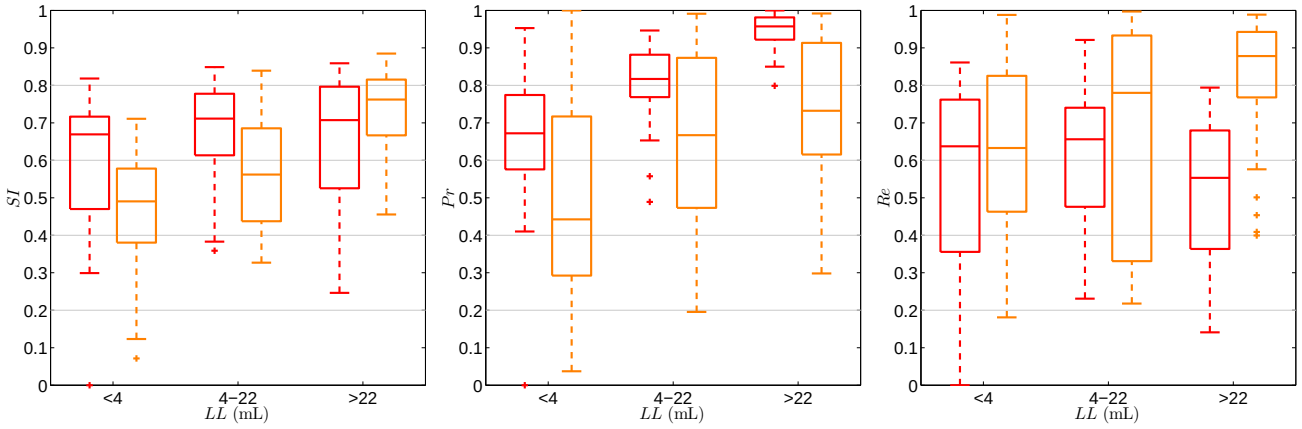
Table 4.7: Mean inter-rater agreement measures for manual WMH segmentation calculated for the available data.

Ref	Dataset	Raters	Data	SI	ICC
[83]	MSSEG 2016	7	15 images	0.63±0.16	0.91
[57]	MS 2015 ISBI	2	21 images	0.73±0.10	0.98

($n = 21$, $n = 315$) to the VLR-vs-human SI values ($n = 96$), to test for significant differences. Differences were significant in the MS 2015 ISBI comparison, but not for MS 2016 comparison. 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.



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

Figure 4.27: Comparison of VLR model performance versus

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 [131, 133, 43][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 4.27, 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 all lesion loads than the LPA algorithm ($p \leq 0.001$), but overall had lower recall, mainly due to 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.

[8] These methods cite [112], which is the reference requested for citations related to the LST toolbox. The PhD thesis describing the LPA algorithm was also only published in 2017.

4.9.2 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, 95^{th} percentile) [192] – `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 ($\mathbf{N}_{26}$) 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 (`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 4.28.[10]

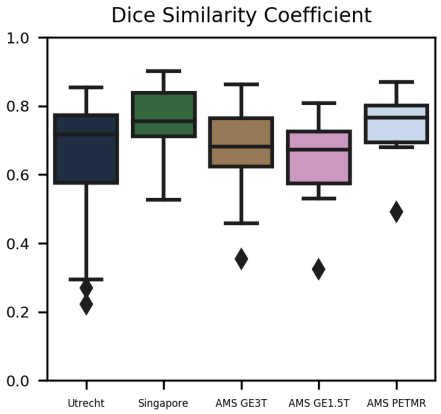
At the MICCAI 2016 MSSEG Competition, only [X] of the 15 submitted methods used deep learning, while the 2017 competition saw 15 of 20 methods use this approach, *including the top performing 13 methods*. This impressive and sudden display of model dominance should not be taken for granted,

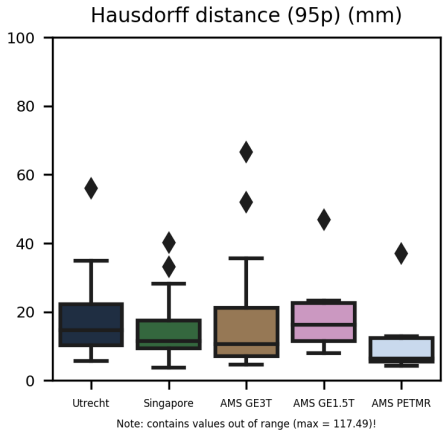
[9] For more information, see http://wmh.isi.uu.nl/evaluation/.
[10] Detailed results and competitor methods descriptions are available at: http://wmh.isi.uu.nl/results/.

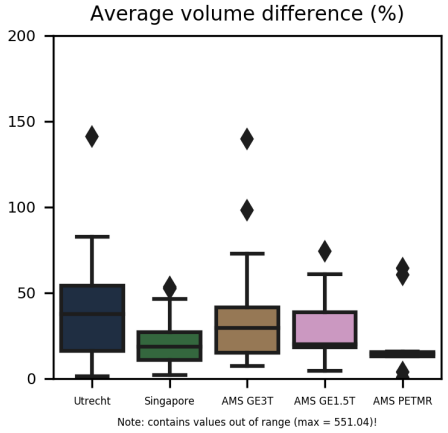
Team: knight, rank: 0.416 (15th place)

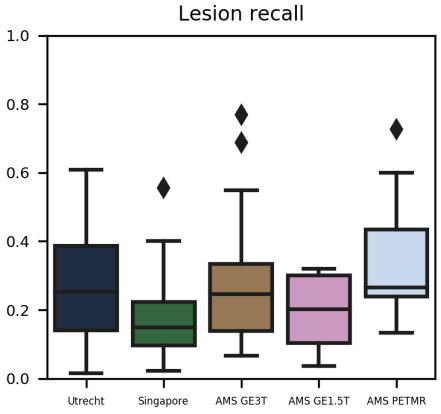
Overview of all metrics per scanner.
Total is the weighted average for all 110 scans.
Rank ranges from 0=best to 1=worst method.

	dsc	h95	avd	recall	f1
Utrecht (n=30)	0.66	18.22	73.28	0.27	0.34
Singapore (n=30)	0.75	14.38	21.81	0.17	0.28
AMS GE3T (n=30)	0.68	16.80	34.39	0.28	0.40
AMS GE1.5T (n=10)	0.64	28.29	30.00	0.20	0.32
AMS PETMR (n=10)	0.74	10.81	21.40	0.35	0.45
weighted average	0.70	17.03	39.99	0.25	0.35
rank [0..1]	0.177	0.235	0.027	1.000	0.641









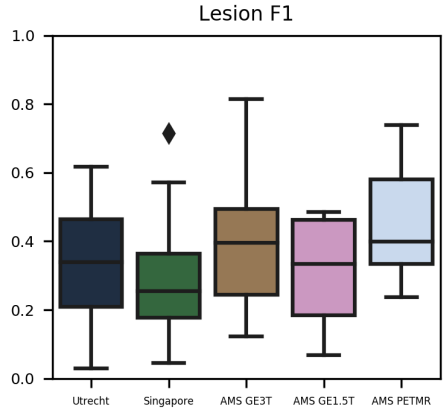


Figure 4.28: Results report for the submitted method provided by the WMH Segmentation Competition.

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$^{\text{th}}$ place, overall score of 0.3858) uses an adaptation of the unsupervised unified mixture model described in [193]. This method outperforms the VLR submission in both lesion-wise metrics (`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.

Chapter 5

Conclusion

This chapter concludes the thesis. A summary of the major contributions will be given. Then, a roadmap for research which builds on this work will be presented, following an analysis of limitations of this work.

5.1 Summary

This thesis explored the task of automated white matter hyperintensity segmentation, which aims to improve the speed and precision over manual analysis.

5.1.1 Algorithm Validation

One of the major limitations of previous works in this area is the use of validation conditions which over-estimate the segmentation performance on images from sources 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 most naive algorithm use cases. In fact, this criticism likely applies to validation of solutions in many different image analysis tasks.

In the current work, the Leave-One-Source-Out Cross Validation (LOSO-CV) framework is presented, a rediscovery of the "Multi-Source" Cross Validation procedure described by Geras and Sutton in [191]. Experimental results show 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 lead to premature adoption of particular automated WMH segmentation algorithms, or overconfidence in their results.

5.2 Future Work

There are several limitations to the current work.

In § 4.6.2 it was shown how the current model can be trained and tested using the exact same data, and Similarity Index still only reaches 0.71. This demonstrates a significant ceiling to performance which perhaps cannot be overcome through changes to pre / post-processing or regularization alone.

Bibliography

[1] Robert A. Pooley. "Fundamental Physics of MR Imaging". In: *RadioGraphics* 25.4 (July 2005), pp. 1087–1099. DOI: 10.1148/rg.254055027.

[2] F Bloch, W W Hansen, and Martin Packard. "Nuclear Induction". In: *Phys. Rev.* 69.3-4 (Feb. 1946), p. 127. DOI: 10.1103/PhysRev.69.127.

[3] Robert G Bryant and Jean Pierre Korb. "Nuclear magnetic resonance and spin relaxation in biological systems". In: *Magnetic Resonance Imaging*. Vol. 23. 2 SPEC. ISS. 2005, pp. 167–173. DOI: 10.1016/j.mri.2004.11.026.

[4] S. H. Koenig et al. "Relaxometry of brain: Why white matter appears bright in MRI". In: *Magnetic Resonance in Medicine* 14.3 (June 1990), pp. 482–495. DOI: 10.1002/mrm.1910140306.

[5] T. P L Roberts and David Mikulis. "Neuro MR: Principles". In: *Journal of Magnetic Resonance Imaging* 26.4 (Oct. 2007), pp. 823–837. DOI: 10.1002/jmri.21029.

[6] Peter Schmitt et al. "Inversion Recovery TrueFISP: Quantification of T1, T 2, and Spin Density". In: *Magnetic Resonance in Medicine* 51.4 (Apr. 2004), pp. 661–667. DOI: 10.1002/mrm.20058.

[7] V.L L. Stevenson et al. "Variations in T1 and T2 relaxation times of normal appearing white matter and lesions in multiple sclerosis". In: *Journal of the neurological sciences* 178.2 (Sept. 2000), pp. 81–87. DOI: 10.1016/S0022-510X(00)00339-7.

[8] Govind B. Chavhan et al. "Principles, Techniques, and Applications of T2*-based MR Imaging and Its Special Applications". In: *RadioGraphics* 29.5 (Sept. 2009), pp. 1433–1449. DOI: 10.1148/rg.295095034.

[9] E. L. Hahn. "Spin echoes". In: *Physical Review* 80.4 (1950), pp. 580–594. DOI: 10.1103/PhysRev.80.580.

[10] G M Bydder and I R Young. "MR imaging: clinical use of the inversion recovery sequence." In: *Journal of computer assisted tomography* 9.4 (1985), pp. 659–675.

[11] J V Hajnal et al. "Use of fluid attenuated inversion recovery (FLAIR) pulse sequences in MRI of the brain." In: *Journal of computer assisted tomography* 16.6 (1992), pp. 841–844. DOI: 10.1097/00004728-199211000-00001.

[12] Hugo J. Kuijf et al. *WMH Segmentation Challenge*. 2017.

[13] Margaret M. Esiri. "Ageing and the brain". In: *Journal of Pathology* 211.2 (Feb. 2007), pp. 181–187. DOI: 10.1002/path.2089.

[14] C.D. Good et al. "A voxel-based morphometric study of ageing in 465 normal adult human brains". In: *5th IEEE EMBS International Summer School on Biomedical Imaging, 2002.* IEEE, 2002, II˙5˙1–II˙5˙16. DOI: 10.1109/SSBI.2002.1233974.

[15] S. Debette and H. S. Markus. "The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis". In: *BMJ* 341 (July 2010). DOI: 10.1136/bmj.c3666.

[16] John Conklin et al. "Are acute infarcts the cause of leukoaraiosis? Brain mapping for 16 consecutive weeks." eng. In: *Annals of neurology* 76.6 (Dec. 2014), pp. 899–904. DOI: 10.1002/ana.24285.

[17] Frank L Heppner, Richard M Ransohoff, and Burkhard Becher. "Immune attack: the role of inflammation in Alzheimer disease". In: *Nat Rev Neurosci* 16.6 (June 2015), pp. 358–372.

[18] Heather M. Snyder et al. "Vascular contributions to cognitive impairment and dementia including Alzheimer's disease". In: *Alzheimer's and Dementia* 11.6 (2015), pp. 710–717. DOI: 10.1016/j.jalz.2014.10.008.

[19] H. Bart van der Worp and Jan van Gijn. "Acute Ischemic Stroke". In: *New England Journal of Medicine* 357.6 (Aug. 2007), pp. 572–579. DOI: 10.1056/NEJMcp072057.

[20] Gregory W. Albers et al. "Transient Ischemic Attack - Proposal for a New Definition". In: *New England Journal of Medicine* 347.21 (Nov. 2002), pp. 1713–1716. DOI: 10.1056/NEJMsb020987.

[21] Leonardo Pantoni. "Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges". In: *The Lancet Neurology* 9.7 (July 2010), pp. 689–701. DOI: 10.1016/S1474-4422(10)70104-6.

[22] Shyam Prabhakaran, Ilana Ruff, and Richard A. Bernstein. "Acute Stroke Intervention A Systematic Review Clinical Review & Education Review". In: *JAMA* 313.14 (Apr. 2015), pp. 1451–1462. DOI: 10.1001/jama.2015.3058.

[23] G C Román et al. "Vascular dementia: diagnostic criteria for research studies. Report of the NINDS-AIREN International Workshop." In: *Neurology* 43.2 (Feb. 1993), pp. 250–260. DOI: 10.1212/WNL.43.2.250.

[24] Alistair Burns and Steve Iliffe. "Alzheimer's disease". In: *BMJ* 338 (2009). DOI: 10.1136/bmj.b158.

[25] C L Masters et al. "Amyloid plaque core protein in Alzheimer disease and Down syndrome." In: *Proceedings of the National Academy of Sciences* 82.12 (June 1985), pp. 4245–4249. DOI: 10.1073/pnas.82.12.4245.

[26] J. Hardy. "The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics". In: *Science* 297.5580 (July 2002), pp. 353–356. DOI: 10.1126/science.1072994.

[27] Virginia M Y Lee et al. "Developing therapeutic approaches to tau, selected kinases, and related neuronal protein targets." In: *Cold Spring Harbor perspectives in medicine* 1.1 (Sept. 2011), a006437–a006437. DOI: 10.1101/cshperspect.a006437.

[28] T. Kim et al. "Human LilrB2 Is a -Amyloid Receptor and Its Murine Homolog PirB Regulates Synaptic Plasticity in an Alzheimer's Model". In: *Science* 341.6152 (Sept. 2013), pp. 1399–1404. DOI: 10.1126/science.1242077.

[29] Shichun Tu et al. "Oligomeric Aβ-induced synaptic dysfunction in Alzheimer's disease". In: *Mol Neurodegener* 9.1 (Nov. 2014), pp. 1–12. DOI: 10.1186/1750-1326-9-48.

[30] Robert D Bell and Berislav V Zlokovic. "Neurovascular mechanisms and blood-brain barrier disorder in Alzheimer's disease." In: *Acta neuropathologica* 118.1 (July 2009), pp. 103–13. DOI: 10.1007/s00401-009-0522-3.

[31] Bruce D. Trapp and Klaus-Armin Nave. "Multiple sclerosis: an immune or neurodegenerative disorder?" In: *Annual review of neuroscience* 31.1 (July 2008), pp. 247–69. DOI: 10.1146/annurev.neuro.30.051606.094313.

[32] Claudia Lucchinetti et al. "Heterogeneity of multiple sclerosis lesions: Implications for the pathogenesis of demyelination". In: *Annals of Neurology* 47.6 (June 2000), pp. 707–717. DOI: 10.1002/1531-8249(200006)47:6<707::AID-ANA3>3.0.CO;2-Q.

[33] Olga Ciccarelli et al. "Pathogenesis of multiple sclerosis: insights from molecular and metabolic imaging." In: *The Lancet. Neurology* 13.8 (Aug. 2014), pp. 807–22. DOI: 10.1016/S1474-4422(14)70101-2.

[34] Don H Mahad, Bruce D Trapp, and Hans Lassmann. "Pathological mechanisms in progressive multiple sclerosis". In: *The Lancet Neurology* 14.2 (2015), pp. 183–193. DOI: 10.1016/S1474-4422(14)70256-X.

[35] Rohit Bakshi et al. "Imaging of multiple sclerosis: role in neurotherapeutics." In: *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics* 2.2 (Apr. 2005), pp. 277–303. DOI: 10.1602/neurorx.2.2.277.

[36] Joanna M Wardlaw, Maria C Valdés Hernández, and Susana Muñoz-Maniega. "What are white matter hyperintensities made of? Relevance to vascular cognitive impairment." In: *Journal of the American Heart Association* 4.6 (June 2015), p. 001140. DOI: 10.1161/JAHA.114.001140.

[37] F-E de Leeuw et al. "Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. The Rotterdam Scan Study". In: *Journal of Neurology, Neurosurgery & Psychiatry* 70.1 (Jan. 2001), 9 LP –14.

[38] Yulin Ge et al. "Dirty-Appearing White Matter in Multiple Sclerosis: Volumetric MR Imaging and Magnetization Transfer Ratio Histogram Analysis". In: *American Journal of Neuroradiology* 24.10 (2003), pp. 1935–1940.

[39] J.H. Simon et al. "Standardized MR Imaging Protocol for Multiple Sclerosis: Consortium of MS Centers Consensus Guidelines". In: *AJNR Am J Neuroradiol* 27.2 (Feb. 2006), pp. 455–461.

[40] Joanna M Wardlaw et al. *Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration.* Aug. 2013. DOI: 10.1016/S1474-4422(13)70124-8.

[41] Chris H Polman et al. "Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria." In: *Annals of neurology* 69.2 (Feb. 2011), pp. 292–302. DOI: 10.1002/ana.22366.

[42] F Fazekas, J B Chawluk, and A Alavi. "MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging". In: *American Journal of Neuroradiology* 8.3 (Aug. 1987), pp. 421–426. DOI: 10.2214/ajr.149.2.351.

[43] Christine Egger et al. "MRI FLAIR lesion segmentation in Multiple Sclerosis: Does automated segmentation hold up with manual annotation?" In: *NeuroImage: Clinical* 13 (2016), pp. 264–270. DOI: 10.1016/j.nicl.2016.11.020.

[44] F D Lublin et al. "Defining the clinical course of multiple sclerosis: the 2013 revisions". In: *Neurology* 83 (2014). DOI: 10.1212/WNL.0000000000000560.

[45] Anthony Traboulsee et al. "Canadian Expert Panel Recommendations for MRI Use in MS Diagnosis and Monitoring." In: *The Canadian journal of neurological sciences. Le journal canadien des sciences neurologiques* 42.3 (May 2015), pp. 159–67. DOI: 10.1017/cjn.2015.24.

[46] Maria Pia Sormani and Paolo Bruzzi. "MRI lesions as a surrogate for relapses in multiple sclerosis: A meta-analysis of randomised trials". In: *The Lancet Neurology* 12.7 (2013), pp. 669–676. DOI: 10.1016/S1474-4422(13)70103-0.

[47] Kyle Fahrbach et al. "Relating Relapse and T2 Lesion Changes to Disability Progression in MS: A Systematic Literature Review and Regression Analysis (P05.090)". In: *Neurology* 78.Meeting Abstracts 1 (Nov. 2012), P05.090–P05.090. DOI: 10.1212/WNL.78.1_MeetingAbstracts.P05.090.

[48] Tjalf Ziemssen et al. *Optimizing therapy early in multiple sclerosis: An evidence-based view*. Sept. 2015. DOI: 10.1016/j.msard.2015.07.007.

[49] J I O'Riordan et al. "The prognostic value of brain MRI in clinically isolated syndromes of the CNS. A 10-year follow-up". In: *Brain* 121.3 (Mar. 1998), pp. 495–503. DOI: 10.1093/brain/121.3.495.

[50] Daisy Mollison et al. "The clinico-radiological paradox of cognitive function and MRI burden of white matter lesions in people with multiple sclerosis: A systematic review and meta-analysis". In: *PLoS ONE* 12.5 (May 2017). Ed. by Orhan Aktas, e0177727. DOI: 10.1371/journal.pone.0177727.

[51] Bruno Dubois et al. "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria". In: *Lancet Neurology* 6.8 (Aug. 2007), pp. 734–746. DOI: 10.1016/S1474-4422(07)70178-3.

[52] S.a Sorbi et al. "EFNS-ENS Guidelines on the diagnosis and management of disorders associated with dementia". In: *European Journal of Neurology* 19.9 (2012), pp. 1159–1179. DOI: 10.1111/j.1468-1331.2012.03784.x.

[53] Martijn V. Verhagen et al. "The impact of MRI combined with visual rating scales on the clinical diagnosis of dementia: a prospective study". In: *European Radiology* 26.6 (June 2016), pp. 1716–1722. DOI: 10.1007/s00330-015-3957-z.

[54] Guy M. McKhann et al. "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". In: *Alzheimer's and Dementia* 7.3 (May 2011), pp. 263–269. DOI: 10.1016/j.jalz.2011.03.005.

[55] Danielle van Westen et al. "Cerebral white matter lesions - associations with Aβ isoforms and amyloid PET." In: *Scientific reports* 6 (2016), p. 20709. DOI: 10.1038/srep20709.

[56] Ki Woong Kim, James R. MacFall, and Martha E. Payne. "Classification of White Matter Lesions on Magnetic Resonance Imaging in Elderly Persons". In: *Biological Psychiatry* 64.4 (Aug. 2008), pp. 273–280. DOI: 10.1016/j.biopsych.2008.03.024.

[57] Aaron Carass et al. "Longitudinal multiple sclerosis lesion segmentation data resource". In: *Data in Brief* 12 (June 2017), pp. 346–350. DOI: 10.1016/j.dib.2017.04.004.

[58] Tomoko Okuda et al. "Brain Lesions: When Should Fluid-attenuated Inversion-Recovery Sequences Be Used in MR Evaluation?" In: *Radiology* 212.3 (Sept. 1999), pp. 793–798. DOI: 10.1148/radiology.212.3.r99se07793.

[59] M. Rovaris et al. "The contribution of fast-FLAIR MRI for lesion detection in the brain of patients with systemic autoimmune diseases." In: *Journal of neurology* 247.1 (Jan. 2000), pp. 29–33. DOI: 10.1007/s004150050006.

[60] Rohit Bakshi et al. "Fluid-attenuated inversion recovery magnetic resonance imaging detects cortical and juxtacortical multiple sclerosis lesions". In: *Archives of neurology* 58.5 (May 2001), pp. 742–748. DOI: 10.1001/archneur.58.5.742.

[61] Frederik Barkhof and Philip Scheltens. "Imaging of white matter lesions." In: *Cerebrovascular diseases (Basel, Switzerland)* 13 Suppl 2.suppl 2 (2002), pp. 21–30. DOI: 49146.

[62] Rola Harmouche et al. "Bayesian MS lesion classification modeling regional and local spatial information". In: *Proceedings - International Conference on Pattern Recognition*. Vol. 3. IEEE, 2006, pp. 984–987. DOI: 10.1109/ICPR.2006.318.

[63] Renske de Boer et al. "White matter lesion extension to automatic brain tissue segmentation on MRI." In: *NeuroImage* 45.4 (May 2009), pp. 1151–61. DOI: 10.1016/j.neuroimage.2009.01.011.

[64] Martijn D Steenwijk et al. "Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs)." In: *NeuroImage. Clinical* 3 (Jan. 2013), pp. 462–9. DOI: 10.1016/j.nicl.2013.10.003.

[65] Martha E Payne et al. "Development of a semi-automated method for quantification of MRI gray and white matter lesions in geriatric subjects". In: *Psychiatry Research - Neuroimaging* 115.1-2 (Aug. 2002), pp. 63–77. DOI: 10.1016/S0925-4927(02)00009-4.

[66] M Ghazel, A Traboulsee, and R K Ward. "Semi-Automated Segmentation of Multiple Sclerosis Lesions in Brain MRI using Texture Analysis". In: *2006 IEEE International Symposium on Signal Processing and Information Technology*. Aug. 2006, pp. 6–10. DOI: 10.1109/ISSPIT.2006.270760.

[67] Yasuo Kawata et al. "Computer-aided evaluation method of white matter hyperintensities related to subcortical vascular dementia based on magnetic resonance imaging". In: *Computerized Medical Imaging and Graphics* 34.5 (July 2010), pp. 370–376. DOI: 10.1016/j.compmedimag.2009.12.014.

[68] Mariangela Iorio et al. "White matter hyperintensities segmentation: a new semi-automated method." In: *Frontiers in aging neuroscience* 5 (Jan. 2013), p. 76. DOI: 10.3389/fnagi.2013.00076.

[69] Olaf Dietrich et al. "Influence of multichannel combination, parallel imaging and other reconstruction techniques on MRI noise characteristics." eng. In: *Magnetic resonance imaging* 26.6 (July 2008), pp. 754–762. DOI: 10.1016/j.mri.2008.02.001.

[70] P. Santago. "Statistical models of partial volume effect". In: *Image Processing, IEEE Transactions on* 4.11 (1995), pp. 1531–1540.

[71] April Khademi, Anastasios Venetsanopoulos, and Alan R Moody. "Generalized method for partial volume estimation and tissue segmentation in cerebral magnetic resonance images". In: *Journal of Medical Imaging* 1.1 (Apr. 2014), p. 14002. DOI: 10.1117/1.JMI.1.1.014002.

[72] W.J. Niessen et al. "Multiscale Segmentation of Three-Dimensional MR Brain Images". In: *International Journal of Computer Vision* 31.2/3 (1999), pp. 185–202. DOI: 10.1023/A:1008070000018.

[73] Jaber Juntu et al. "Bias field correction for mri images". In: *Computer Recognition Systems* 30 (2005), pp. 543–551. DOI: 10.1007/3-540-32390-2_64.

[74] Uroš Vovk, Franjo Pernuš, and Boštjan Likar. "A review of methods for correction of intensity inhomogeneity in MRI". In: *IEEE Transactions on Medical Imaging* 26.3 (Mar. 2007), pp. 405–421. DOI: 10.1109/TMI.2006.891486.

[75] Rohit Bakshi et al. "Intraventricular CSF pulsation artifact on fast fluid-attenuated inversion-recovery MR images: Analysis of 100 consecutive normal studies". In: *American Journal of Neuroradiology* 21.3 (2000), pp. 503–508.

[76] Maxim Zaitsev, Julian Maclaren, and Michael Herbst. *Motion artifacts in MRI: A complex problem with many partial solutions*. Oct. 2015. DOI: 10.1002/jmri.24850.

[77] Ioannis Kapouleas. "Automatic detection of white matter lesions in magnetic resonance brain images". In: *Computer Methods and Programs in Biomedicine* 32.1 (May 1990), pp. 17–35. DOI: 10.1016/0169-2607(90)90082-K.

[78] Xavier Lladó et al. "Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches". In: *Information Sciences* 186.1 (2012), pp. 164–185. DOI: http://dx.doi.org/10.1016/j.ins.2011.10.011.

[79] D Mortazavi, A Z Kouzani, and H Soltanian-Zadeh. "Segmentation of multiple sclerosis lesions in MR images: a review". In: *Neuroradiology* 54 (2012). DOI: 10.1007/s00234-011-0886-7.

[80] Daniel Garcia-Lorenzo et al. "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging." eng. In: *Medical image analysis* 17.1 (Jan. 2013), pp. 1–18. DOI: 10.1016/j.media.2012.09.004.

[81] Maria Eugenia Caligiuri et al. "Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review". In: *Neuroinformatics* 13.3 (July 2015), pp. 261–276. DOI: 10.1007/s12021-015-9260-y.

[82] *MS Lesion Segmentation Challenge 2008*. 2008.

[83] *MS Lesion Segmentation Challenge 2016*. 2016.

[84] John Mazziotta et al. "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)." In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 356.1412 (2001), pp. 1293–322. DOI: 10.1098/rstb.2001.0915.

[85] Jesse Knight and April Khademi. "Disease-Inspired Feature Design for Computer-Aided Diagnosis of Breast Cancer Digital Pathology Images". In: *Medical Image Analysis and Informatics: Computer-aided Diagnosis and Therapy 2*. Ed. by Paulo Mazzoncini de Azevedo Marques et al. CRC Press, 2017.

[86] K. Van Leemput et al. "Automated segmentation of multiple sclerosis lesions by model outlier detection". In: *IEEE Transactions on Medical Imaging* 20.8 (2001), pp. 677–688. DOI: 10.1109/42.938237.

[87] Clifford R Jack et al. "FLAIR histogram segmentation for measurement of leukoaraiosis volume". In: *Journal of Magnetic Resonance Imaging* 14.6 (Dec. 2001), pp. 668–676. DOI: 10.1002/jmri.10011.

[88] Alex P Zijdenbos, Reza Forghani, and Alan C Evans. "Automatic pipeline analysis of 3-D MRI data for clinical trials: application to multiple sclerosis." In: *IEEE transactions on medical imaging* 21.10 (Oct. 2002), pp. 1280–91. DOI: 10.1109/TMI.2002.806283.

[89] Petronella Anbeek et al. "Probabilistic segmentation of white matter lesions in MR imaging." In: *NeuroImage* 21.3 (Mar. 2004), pp. 1037–44. DOI: 10.1016/j.neuroimage.2003.10.012.

[90] Petronella Anbeek et al. "Probabilistic segmentation of brain tissue in MR imaging." In: *NeuroImage* 27.4 (Oct. 2005), pp. 795–804. DOI: 10.1016/j.neuroimage.2005.05.046.

[91] F. Admiraal-Behloul et al. "Fully automatic segmentation of white matter hyperintensities in MR images of the elderly". In: *NeuroImage* 28.3 (2005), pp. 607–617. DOI: 10.1016/j.neuroimage.2005.06.061.

[92] Z. Lao et al. "Automated Segmentation of White Matter Lesions in 3D Brain MR Images, using Multivariate Pattern Classification". In: *3rd IEEE International Symposium on Biomedical Imaging: Macro to Nano, 2006.* IEEE, 2006, pp. 307–310. DOI: 10.1109/ISBI.2006.1624914.

[93] Ying Wu et al. "Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI". In: *NeuroImage* 32.3 (Sept. 2006), pp. 1205–1215. DOI: 10.1016/j.neuroimage.2006.04.211.

[94] Balasrinivasa Rao Sajja et al. "Unified approach for multiple sclerosis lesion segmentation on brain MRI." In: *Annals of biomedical engineering* 34.1 (Jan. 2006), pp. 142–51. DOI: 10.1007/s10439-005-9009-0.

[95] Rasoul Khayati et al. "Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model." eng. In: *Computers in biology and medicine* 38.3 (Mar. 2008), pp. 379–390. DOI: 10.1016/j.compbiomed.2007.12.005.

[96] M. Wels, M. Huber, and J. Hornegger. "Fully automated segmentation of multiple sclerosis lesions in multispectral MRI". In: *Pattern Recognition and Image Analysis* 18.2 (June 2008), pp. 347–350. DOI: 10.1134/S1054661808020235.

[97] E H Herskovits, R N Bryan, and F Yang. "Automated Bayesian segmentation of microvascular white-matter lesions in the ACCORD-MIND study." eng. In: *Advances in medical sciences* 53.2 (2008), pp. 182–190. DOI: 10.2478/v10039-008-0039-3.

[98] S Bricq, Christophe Collet, and Jean-Paul Armspach. "MS Lesion Segmentation based on Hidden Markov Chains". In: *The MIDAS Journal - MS Lesion Segmentation (MICCAI 2008 Workshop)*. 2008.

[99] Tim B Dyrby et al. "Segmentation of age-related white matter changes in a clinical multi-center study." eng. In: *NeuroImage* 41.2 (June 2008), pp. 335–345. DOI: 10.1016/j.neuroimage.2008.02.024.

[100] Jc Souplet et al. "An Automatic Segmentation of T2-FLAIR Multiple Sclerosis Lesions". In: *MICCAI Grand Challenge Workshop: Multiple Sclerosis Lesion Segmentation Challenge* (2008), pp. 1–11.

[101] Daniel García-Lorenzo et al. "Multiple sclerosis lesion segmentation using an automatic multimodal graph cuts". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 5762 LNCS. PART 2. 2009, pp. 584–591. DOI: 10.1007/978-3-642-04271-3_71.

[102] Ayelet Akselrod-Ballin et al. "Automatic Segmentation and Classification of Multiple Sclerosis in Multichannel MRI". In: *IEEE Transactions on Biomedical Engineering* 56.10 (2009), pp. 2461–2469.

[103] Christopher Schwarz et al. "Fully-automated white matter hyperintensity detection with anatomical prior knowledge and without FLAIR". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 5636 LNCS. NIH Public Access, 2009, pp. 239–251. DOI: 10.1007/978-3-642-02498-6_20.

[104] Erin Gibson et al. "Automatic segmentation of white matter hyperintensities in the elderly using FLAIR images at 3T". In: *Journal of Magnetic Resonance Imaging* 31.6 (June 2010), pp. 1311–1322. DOI: 10.1002/jmri.22004.

[105] Navid Shiee et al. "A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions." In: *NeuroImage* 49.2 (Jan. 2010), pp. 1524–1535. DOI: 10.1016/j.neuroimage.2009.09.005.

[106] Mark Scully et al. "An Automated Method for Segmenting White Matter Lesions through Multi-Level Morphometric Feature Classification with Application to Lupus." In: *Frontiers in human neuroscience* 4.April (2010), p. 27. DOI: 10.3389/fnhum.2010.00027.

[107] Daniel García-Lorenzo et al. "Trimmed-likelihood estimation for focal lesions and tissue segmentation in multisequence MRI for multiple sclerosis." In: *IEEE Transactions on Medical Imaging* 30.8 (Aug. 2011), pp. 1455–67. DOI: 10.1109/TMI.2011.2114671.

[108] Ezequiel Geremia et al. "Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images." eng. In: *NeuroImage* 57.2 (July 2011), pp. 378–390. DOI: 10.1016/j.neuroimage.2011.03.080.

[109] Sean D Smart, Michael J Firbank, and John T O'Brien. "Validation of automated white matter hyperintensity segmentation." In: *Journal of aging research* 2011 (Jan. 2011), p. 391783. DOI: 10.4061/2011/391783.

[110] Thomas Samaille et al. "Contrast-based fully automatic segmentation of white matter hyperintensities: method and validation." eng. In: *PloS one* 7.11 (2012), e48953. DOI: 10.1371/journal.pone.0048953.

[111] April Khademi, Anastasios Venetsanopoulos, and Alan R Moody. "Robust white matter lesion segmentation in FLAIR MRI." eng. In: *IEEE transactions on bio-medical engineering* 59.3 (Mar. 2012), pp. 860–871. DOI: 10.1109/TBME.2011.2181167.

[112] Paul Schmidt et al. “An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis”. eng. In: *Neuroimage* 59.4 (Feb. 2012), pp. 3774–3783. DOI: 10.1016/j.neuroimage.2011.11.032.

[113] Bassem a Abdullah, Akmal a Younis, and Nigel M John. “Multi-Sectional Views Textural Based SVM for MS Lesion Segmentation in Multi-Channels MRIs.” eng. In: *The open biomedical engineering journal* 6 (2012), pp. 56–72. DOI: 10.2174/1874230001206010056.

[114] Elizabeth M. Sweeney et al. “OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI”. In: *NeuroImage: Clinical* 2.1 (2013), pp. 402–413. DOI: 10.1016/j.nicl.2013.03.002.

[115] Sushmita Datta and Ponnada A. Narayana. “A comprehensive approach to the segmentation of multichannel three-dimensional MR brain images in multiple sclerosis”. In: *NeuroImage: Clinical* 2 (2013), pp. 184–196. DOI: 10.1016/j.nicl.2012.12.007.

[116] Vamsi Ithapu et al. “Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer’s disease risk and aging studies”. In: *Human Brain Mapping* 35.8 (2014), pp. 4219–4235. DOI: 10.1002/hbm.22472.

[117] Byung Il Yoo et al. “Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images”. In: *Neuroradiology* 56.4 (2014), pp. 265–281. DOI: 10.1007/s00234-014-1322-6.

[118] Rola Harmouche et al. “Probabilistic multiple sclerosis lesion classification based on modeling regional intensity variability and local neighborhood Information”. In: *IEEE Transactions on Biomedical Engineering* 62.5 (2015), pp. 1281–1292. DOI: 10.1109/TBME.2014.2385635.

[119] Nicolas Guizard et al. “Rotation-invariant multi-contrast non-local means for MS lesion segmentation”. In: *NeuroImage: Clinical* 8 (2015), pp. 376–389. DOI: 10.1016/j.nicl.2015.05.001.

[120] Saurabh Jain et al. “Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images.” In: *NeuroImage: Clinical* 8 (2015), pp. 367–75. DOI: 10.1016/j.nicl.2015.05.003.

[121] Xavier Tomas-Fernandez and Simon K Warfield. “A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion segmentation”. In: *IEEE Transactions on Medical Imaging* 34.6 (June 2015), pp. 1349–1361. DOI: 10.1109/TMI.2015.2393853.

[122] Rui Wang et al. “Automatic segmentation and volumetric quantification of white matter hyperintensities on fluid-attenuated inversion recovery images using the extreme value distribution.” eng. In: *Neuroradiology* 57.3 (Mar. 2015), pp. 307–320. DOI: 10.1007/s00234-014-1466-4.

[123] Pallab Kanti Roy et al. “Automatic white matter lesion segmentation using contrast enhanced FLAIR intensity and Markov Random Field.” eng. In: *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 45 (Oct. 2015), pp. 102–111. DOI: 10.1016/j.compmedimag.2015.08.005.

[124] Tom Brosch et al. *Deep Convolutional Encoder Networks for Multiple Sclerosis Lesion Segmentation*. Vol. 9556. Springer International Publishing, 2015, pp. 144–155. DOI: 10.1007/978-3-319-30858-6.

[125] Mário João Fartaria et al. “Automated detection of white matter and cortical lesions in early stages of multiple sclerosis.” In: *Journal of magnetic resonance imaging : JMRI* (Nov. 2015). DOI: 10.1002/jmri.25095.

[126] Hrishikesh Deshpande, Pierre Maurel, and Christian Barillot. “Classification of multiple sclerosis lesions using adaptive dictionary learning”. In: *Computerized Medical Imaging and Graphics* 46 (2015), pp. 2–10. DOI: 10.1016/j.compmedimag.2015.05.003.

[127] Eloy Roura et al. “A toolbox for multiple sclerosis lesion segmentation”. In: *Neuroradiology* 57.10 (Oct. 2015), pp. 1031–1043. DOI: 10.1007/s00234-015-1552-2.

[128] Jesse Knight and April Khademi. “MS Lesion Segmentation Using FLAIR MRI Only”. In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI*. Athens, Greece, 2016, TBD.

[129] Roey Mechrez, Jacob Goldberger, and Hayit Greenspan. “Patch-Based Segmentation with Spatial Consistency: Application to MS Lesions in Brain MRI”. In: *International Journal of Biomedical Imaging* 2016 (2016), pp. 1–13. DOI: 10.1155/2016/7952541.

[130] Maddalena Strumia et al. “White Matter MS-Lesion Segmentation Using a Geometric Brain Model.” In: *IEEE transactions on medical imaging* PP.99 (2016), p. 1. DOI: 10.1109/TMI.2016.2522178.

[131] Ludovica Griffanti et al. “BIANCA (Brain Intensity AbNormality Alassification Algorithm): A new tool for automated segmentation of white matter hyperintensities”. In: *NeuroImage* 141 (2016), pp. 191–205. DOI: 10.1016/j.neuroimage.2016.07.018.

[132] Sergi Valverde et al. “Automated tissue segmentation of MR brain images in the presence of white matter lesions”. In: *Medical Image Analysis* 35 (Aug. 2017), pp. 446–457. DOI: 10.1016/j.media.2016.08.014.

[133] Tom Brosch et al. “Deep 3D Convolutional Encoder Networks With Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation”. In: *IEEE Transactions on Medical Imaging* 35.5 (May 2016), pp. 1229–1239. DOI: 10.1109/TMI.2016.2528821.

[134] Mahsa Dadar et al. “Validation of a Regression Technique for Segmentation of White Matter Hyperintensities in Alzheimer’s Disease”. In: *IEEE Transactions on Medical Imaging* 99.PP (2017), pp. 1–1. DOI: 10.1109/TMI.2017.2693978.

[135] Tianming Zhan et al. “Multimodal spatial-based segmentation framework for white matter lesions in multi-sequence magnetic resonance images”. In: *Biomedical Signal Processing and Control* 31 (2017), pp. 52–62. DOI: 10.1016/j.bspc.2016.06.016.

[136] April Khademi and Alan R. Moody. “Multiscale partial volume estimation for segmentation of white matter lesions using flair MRI”. In: *Proceedings - International Symposium on Biomedical Imaging*. Vol. 2015-July. IEEE, Apr. 2015, pp. 568–571. DOI: 10.1109/ISBI.2015.7163937.

[137] John Ashburner and Karl J Friston. “Multimodal Image Coregistration and Partitioning - a Unified Framework”. In: *Neuroimage* 6.3 (Oct. 1997), pp. 209–217. DOI: 10.1006/nimg.1997.0290.

[138] John Ashburner and Karl J Friston. “Unified segmentation.” eng. In: *NeuroImage* 26.3 (July 2005), pp. 839–851. DOI: 10.1016/j.neuroimage.2005.02.018.

[139] N K Subbanna et al. “MS Lesion Segmentation using Markov Random Fields”. In: *MICCAI Workshop on Medical Image Analysis on Multiple Sclerosis*. 2009, pp. 37–48.

[140] Pierre Louis Bazin and Dzung L. Pham. “Homeomorphic brain image segmentation with topological and statistical atlases”. In: *Medical Image Analysis* 12.5 (Oct. 2008), pp. 616–625. DOI: 10.1016/j.media.2008.06.008.

[141] Nicholas J Tustison et al. “N4ITK: improved N3 bias correction.” In: *IEEE Transactions on Medical Imaging* 29.6 (June 2010), pp. 1310–20. DOI: 10.1109/TMI.2010.2046908.

[142] Stephen M Smith. “Fast robust automated brain extraction.” In: *Human brain mapping* 17.3 (Nov. 2002), pp. 143–55. DOI: 10.1002/hbm.10062.

[143] Y. Zhang, M. Brady, and S. Smith. “Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm”. In: *IEEE Transactions on Medical Imaging* 20.1 (2001), pp. 45–57. DOI: 10.1109/42.906424.

[144] Jesper L R Andersson, Mark Jenkinson, and Stephen Smith. *Non-linear registration AKA Spatial normalisation. FMRIB Technial Report TR07JA2*. Tech. rep. Oxford, UK: FMRIB Centre, 2007.

[145] L G Nyul and J K Udupa. “On standardizing the MR image intensity scale.” eng. In: *Magnetic resonance in medicine* 42.6 (Dec. 1999), pp. 1072–1081.

[146] Jesse Knight, Alan R Moody, and April Khademi. “Noise in parallel MRI: how to determine whether single-coil assumptions still hold (they don’t) (Poster)”. In: *Imaging Network Ontario Symposium*. Toronto, 2016. DOI: 10.13140/RG.2.2.11028.91527.

[147] John G. Sled et al. “Regional Variations in Normal Brain Shown by Quantitative Magnetization Transfer Imaging”. In: *Magnetic Resonance in Medicine* 51.2 (Feb. 2004), pp. 299–303. DOI: 10.1002/mrm.10701.

[148] E M Sweeney et al. “Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI”. In: *American Journal of Neuroradiology* 34.1 (Jan. 2013), pp. 68–73. DOI: 10.3174/ajnr.A3172.

[149] Paul Schmidt. “Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging”. PhD thesis. Ludwig-Maximilians-Universität München, 2017.

[150] Paul Schmidt, Mark Mühlau, and Volker Schmid. “Fitting large-scale structured additive regression models using Krylov subspace methods”. In: *Computational Statistics and Data Analysis* 105 (2017), pp. 59–75. DOI: 10.1016/j.csda.2016.07.006.

[151] Paul Schmidt. *LST: A lesion segmentation tool for SPM*. 2015.

[152] John Ashburner and Karl J. Friston. “Voxel-Based Morphometry - The Methods”. In: *NeuroImage* 11.6 (2000), pp. 805–821. DOI: 10.1006/nimg.2000.0582.

[153] Stephen M. Smith et al. “Advances in functional and structural MR image analysis and implementation as FSL”. In: *NeuroImage*. Vol. 23. SUPPL. 1. Jan. 2004, S208–S219. DOI: 10.1016/j.neuroimage.2004.07.051.

[154] Jean Talairach and Pierre Tournoux. *Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging*. 1988.

[155] A.C. Evans et al. “3D statistical neuroanatomical models from 305 MRI volumes”. In: *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference* January 1993 (1993), pp. 1813–1817. DOI: 10.1109/NSSMIC.1993.373602.

[156] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. "Deformable medical image registration: a survey." In: *IEEE transactions on medical imaging* 32.7 (July 2013), pp. 1153–90. DOI: 10.1109/TMI.2013.2265603.

[157] Mark Jenkinson et al. "Improved optimization for the robust and accurate linear registration and motion correction of brain images". In: *NeuroImage* 17.2 (Oct. 2002), pp. 825–841. DOI: 10.1016/S1053-8119(02)91132-8.

[158] B. B. Avants et al. "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain". In: *Medical Image Analysis* 12.1 (Feb. 2008), pp. 26–41. DOI: 10.1016/j.media.2007.06.004.

[159] Arno Klein et al. "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration." In: *NeuroImage* 46.3 (July 2009), pp. 786–802. DOI: 10.1016/j.neuroimage.2008.12.037.

[160] K Kazemi and N Noorizadeh. "Quantitative Comparison of SPM, FSL, and Brainsuite for Brain MR Image Segmentation." In: *Journal of biomedical physics & engineering* 4.1 (Mar. 2014), pp. 13–26.

[161] Massachusetts General Hospital Center for Morphometric Analysis. *Internet Brain Segmentation Repository (v2.0)*. Massachusetts, 2003.

[162] Anderson M Winkler, Peter Kochunov, and David C Glahn. *FLAIR Templates*. 2012.

[163] J G Sled, a P Zijdenbos, and a C Evans. "A nonparametric method for automatic correction of intensity nonuniformity in MRI data." In: *IEEE transactions on medical imaging* 17.1 (1998), pp. 87–97. DOI: 10.1109/42.668698.

[164] John Ashburner and Karl J Friston. "Nonlinear spatial normalization using basis functions". In: *Human Brain Mapping* 7.4 (1999), pp. 254–266. DOI: 10.1002/(SICI)1097-0193(1999)7:4<254::AID-HBM4>3.0.CO;2-G.

[165] Boubakeur Belaroussi et al. *Intensity non-uniformity correction in MRI: Existing methods and their validation*. Apr. 2006. DOI: 10.1016/j.media.2005.09.004.

[166] Marco Ganzetti, Nicole Wenderoth, and Dante Mantini. "Quantitative Evaluation of Intensity Inhomogeneity Correction Methods for Structural MR Brain Images". In: *Neuroinformatics* 14.1 (Jan. 2016), pp. 5–21. DOI: 10.1007/s12021-015-9277-2.

[167] Benoit M. Dawant, Alex P. Zijdenbos, and Richard A. Margolin. "Correction of Intensity Variations in MR Images for Computer-Aided Tissue Classification". In: *IEEE Transactions on Medical Imaging* 12.4 (1993), pp. 770–781. DOI: 10.1109/42.251128.

[168] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. 3rd ed. Upper Saddle River, New Jersey: Prentice Hall, Inc., Feb. 2006.

[169] László G Nyúl, Jayaram K. Udupa, and Xuan Zhang. "New variants of a method of MRI scale standardization". In: *IEEE Transactions on Medical Imaging* 19.2 (2000), pp. 143–150. DOI: 10.1109/42.836373.

[170] Jesse Knight, Graham Taylor, and April Khademi. "Equivalence of histogram equalization, histogram matching and the Nyul algorithm for intensity standardization in MRI". In: *3rd Annual Conference on Vision and Imaging Systems*. Waterloo, 2017.

[171] Russell T. Shinohara et al. "Statistical normalization techniques for magnetic resonance imaging". In: *NeuroImage: Clinical* 6 (2014), pp. 9–19. DOI: 10.1016/j.nicl.2014.08.008.

[172] Benjamin Kedem. "Spectral Analysis and Discrimination by Zero-Crossings". In: *Proceedings of the IEEE* 74.11 (1986), pp. 1477–1493. DOI: 10.1109/PROC.1986.13663.

[173] Thomas P Minka. "A comparison of numerical optimizers for logistic regression". 2003.

[174] Martin A. Tanner et al. "The calculation of posterior distributions by data augmentation". In: *Journal of the American Statistical Association* 82.398 (June 1987), pp. 528–540. DOI: 10.2307/2289457.

[175] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z Ghahramani et al. Curran Associates, Inc., 2014, pp. 2672–2680.

[176] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105.

[177] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67.2 (Apr. 2005), pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

[178] Robert Tibshirani. "Regression selection and shrinkage via the lasso". In: *Journal of the Royal Statistical Society B* 58.1 (1996), pp. 267–288. DOI: 10.2307/2346178.

[179] Su-in Lee et al. "Efficient L1 Regularized Logistic Regression". In: *AAAI* 6.1 (2006), pp. 401–408. DOI: 10.1.1.64.1993.

[180] C. Tomasi and R. Manduchi. "Bilateral filtering for gray and color images". In: *Sixth International Conference on Computer Vision*. 1998, pp. 839–846. DOI: 10.1109/ICCV.1998.710815.

[181] Pablo Arbeláez et al. "Contour detection and hierarchical image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.5 (May 2011), pp. 898–916. DOI: 10.1109/TPAMI.2010.161.

[182] Colm Elliott et al. "Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI". In: *IEEE Transactions on Medical Imaging* 32.8 (Aug. 2013), pp. 1490–1503. DOI: 10.1109/TMI.2013.2258403.

[183] Alireza Akhondi-Asl et al. "A logarithmic opinion pool based STAPLE algorithm for the fusion of segmentations with associated reliability weights." In: *IEEE Transactions on Medical Imaging* 33.10 (Oct. 2014), pp. 1997–2009. DOI: 10.1109/TMI.2014.2329603.

[184] D.L. Collins. "Design and construction of a realistic digital brain phantom". In: *IEEE Transactions on Medical Imaging* 17.3 (1998), pp. 463–468.

[185] Frederick Klauschen et al. "Evaluation of automated brain MR image segmentation and volumetry methods". In: *Human Brain Mapping* 30.4 (Apr. 2009), pp. 1310–1327. DOI: 10.1002/hbm.20599.

[186] Lucas D. Eggert et al. "Accuracy and Reliability of Automated Gray Matter Segmentation Pathways on Real and Simulated Structural Magnetic Resonance Images of the Human Brain". In: *PLoS ONE* 7.9 (Sept. 2012). Ed. by Yong Fan, e45081. DOI: 10.1371/journal.pone.0045081.

[187] Terry K Koo and Mae Y Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research". In: *Journal of Chiropractic Medicine* 15.2 (June 2016), pp. 155–163. DOI: 10.1016/j.jcm.2016.02.012.

[188] D. G. Altman and J. M. Bland. "Measurement in Medicine: the Analysis of Method Comparison Studies". In: *Statistician* 32.July 1981 (Sept. 1983), pp. 307–317. DOI: 10.2307/2987937.

[189] Douglas M Hawkins. "The Problem of Overfitting". In: *Journal of Chemical Information and Computer Sciences* 44.1 (2004), pp. 1–12. DOI: 10.1021/ci0342472.

[190] Sylvain Arlot and Alain Celisse. "A survey of cross-validation procedures for model selection". In: *Statistics Surveys* 4 (2009), pp. 40–79. DOI: 10.1214/09-SS054.

[191] Krzysztof Geras and Charles Sutton. "Multiple-source cross-validation". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, 2013, pp. 1292–1300.

[192] M.-P. Dubuisson and A.K. Jain. "A modified Hausdorff distance for object matching". In: *Proceedings of 12th International Conference on Pattern Recognition*. Vol. 1. IEEE Comput. Soc. Press, pp. 566–568. DOI: 10.1109/ICPR.1994.576361.

[193] Carole H. Sudre et al. "Bayesian Model Selection for Pathological Neuroimaging Data Applied to White Matter Lesion Segmentation". In: *IEEE Transactions on Medical Imaging* 34.10 (Oct. 2015), pp. 2079–2102. DOI: 10.1109/TMI.2015.2419072.

[194] Andriy Fedorov et al. "3D Slicer as an image computing platform for the Quantitative Imaging Network". In: *Magnetic Resonance Imaging* 30.9 (Nov. 2012), pp. 1323–1341. DOI: 10.1016/j.mri.2012.05.001.

Appendix A

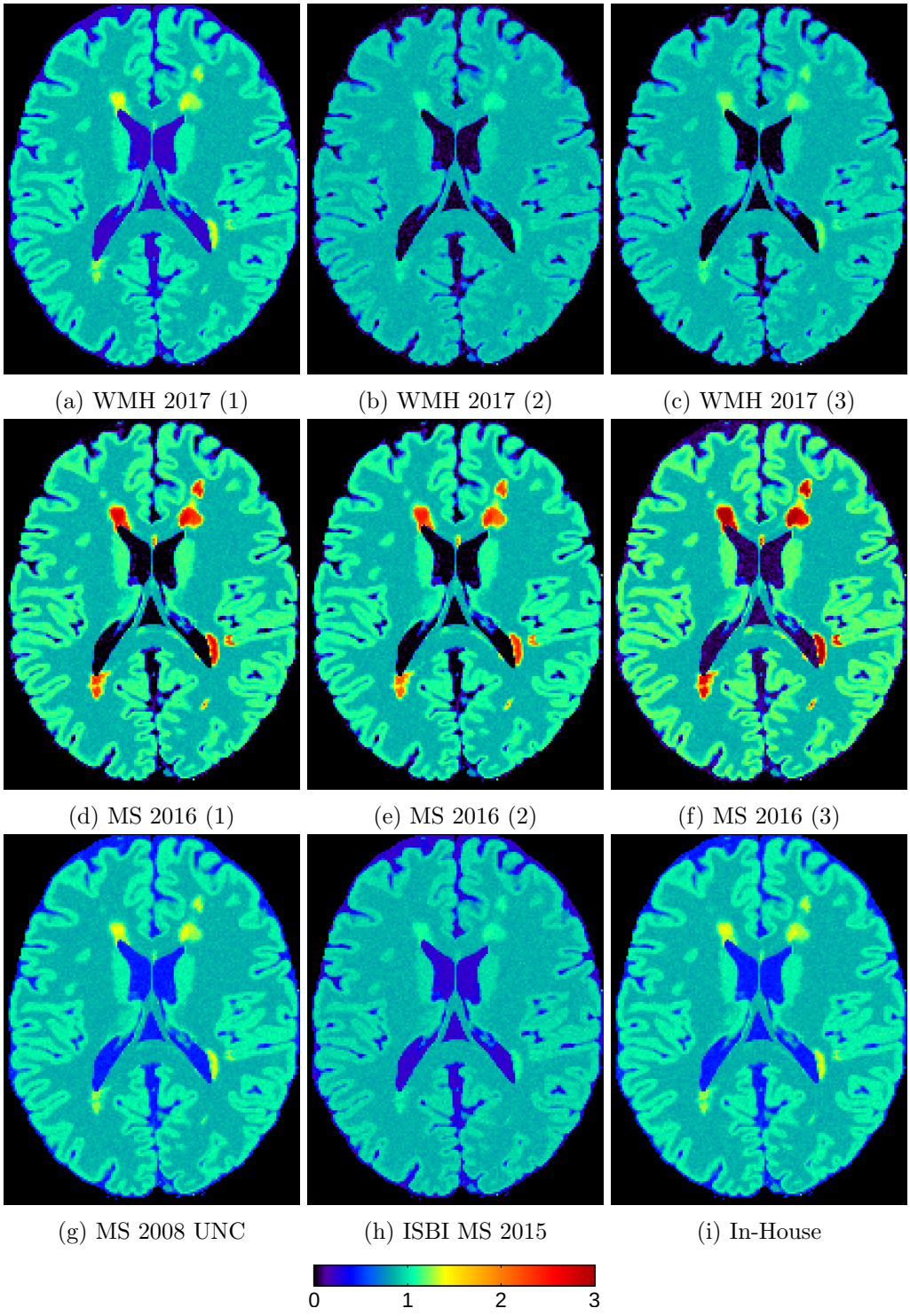
Maths

This section presents various mathematical 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 ms) and one simulated T2 image (TE/TR = 100/5500 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 [184],[1] while in Figure A.3, the PMF for each tissue class from the same data are given.

[1] http://brainweb.bic.mni.mcgill.ca/



(a) WMH 2017 (1) (b) WMH 2017 (2) (c) WMH 2017 (3)

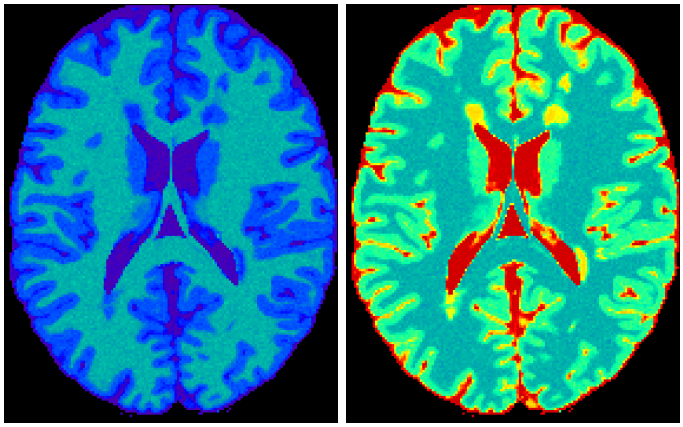
(d) MS 2016 (1) (e) MS 2016 (2) (f) MS 2016 (3)

(g) MS 2008 UNC (h) ISBI MS 2015 (i) In-House

Figure A.1: Simulated FLAIR images using scan parameters from the experimental database. Colourmap is arbitrary but consistent.

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.

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



(a) T1 (b) T2

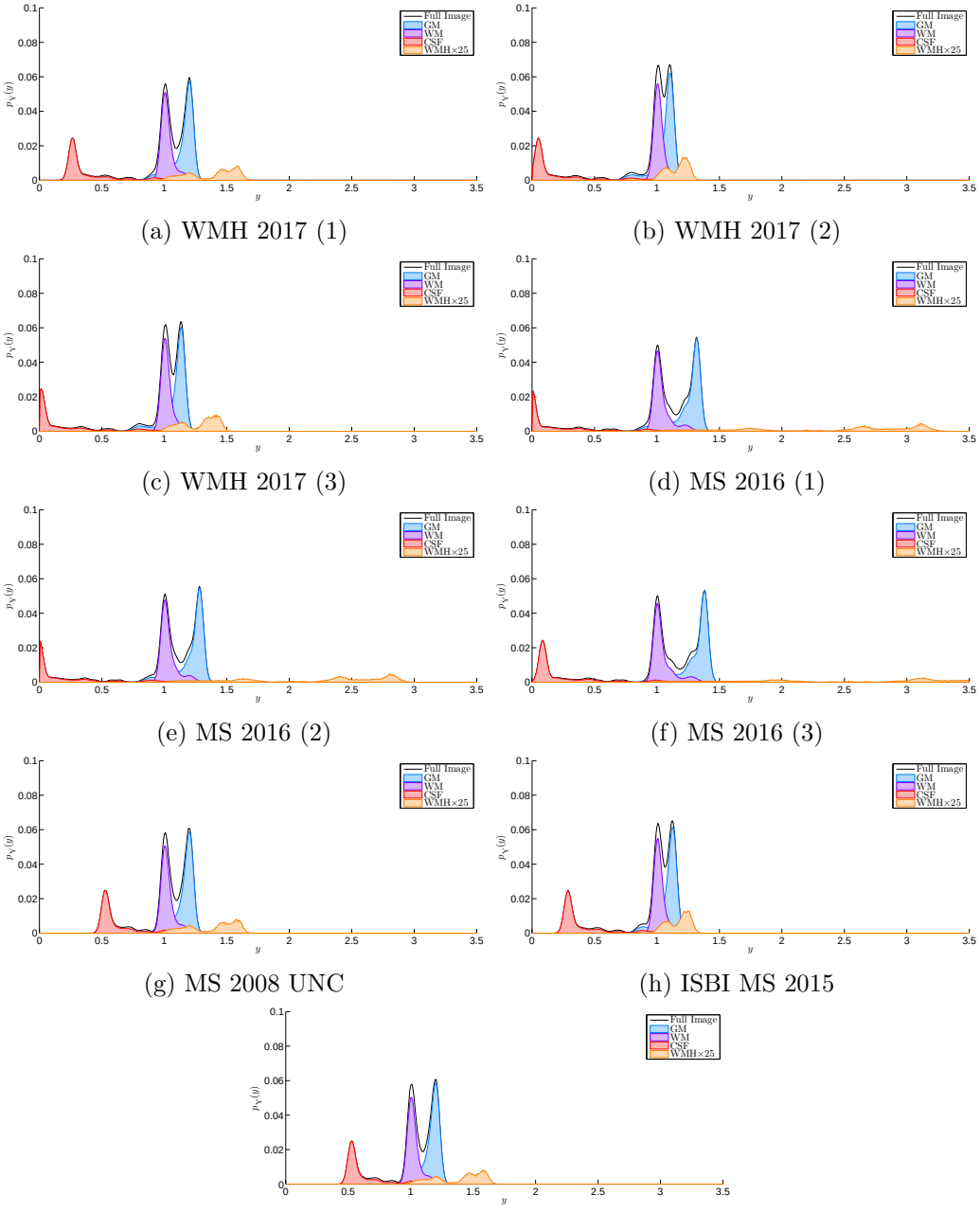
Figure A.2: Simulated T1 (TE/TR = 5/15 ms) and T2 (TE/TR = 100/5500 ms) images.

A.2 Graylevel Standardization

A.2.1 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, ${F_{\tilde{\text{Y}}}}^{-1}$,

$$\tau(y) = {F_{\tilde{\text{Y}}}}^{-1}\big(F_{\text{Y}}(y)\big) \tag{A.1}$$



(a) WMH 2017 (1)

(b) WMH 2017 (2)

(c) WMH 2017 (3)

(d) MS 2016 (1)

(e) MS 2016 (2)

(f) MS 2016 (3)

(g) MS 2008 UNC

(h) ISBI MS 2015

(i) In-House

(j) T1 e.g.

(k) T2 e.g.

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.

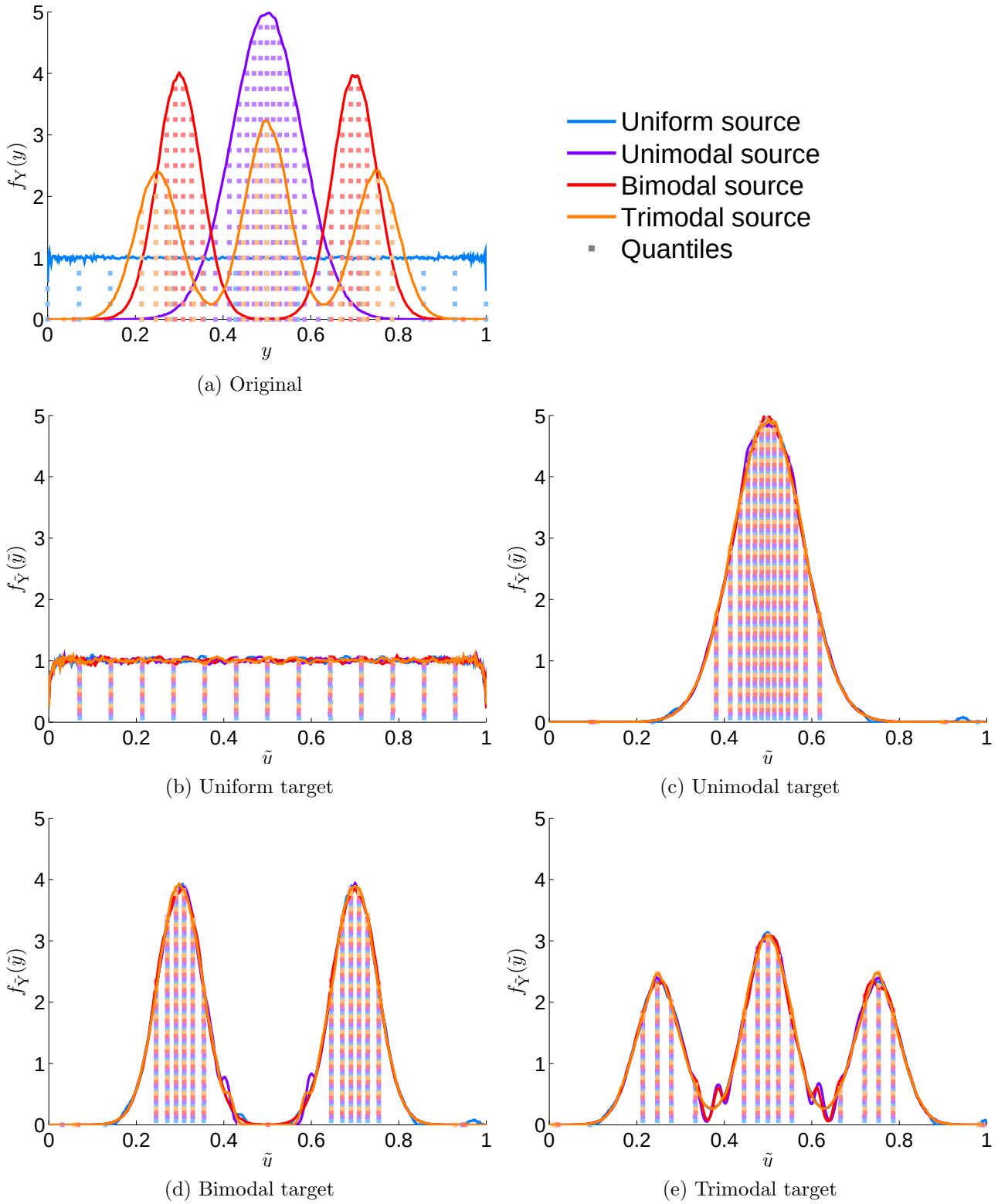
The second transformation in the cascade does not depend on Y, and so it is applied equally to images. Therefore the choice of target PMF is not important for the objective of graylevel standardization. This result is verified experimentally using synthetic images. Four $100 \times 100 \times 100$ images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

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

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.

A.2.2 Nyul Approximation of Histogram Matching

This topic is treated in [170].



(a) Original

(b) Uniform target

(c) Unimodal target

(d) Bimodal target

(e) Trimodal target

Figure A.4: Histogram matching of synthetic data to different target histograms. Quantiles show high agreement regardless of the target histogram.

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 [194] 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.

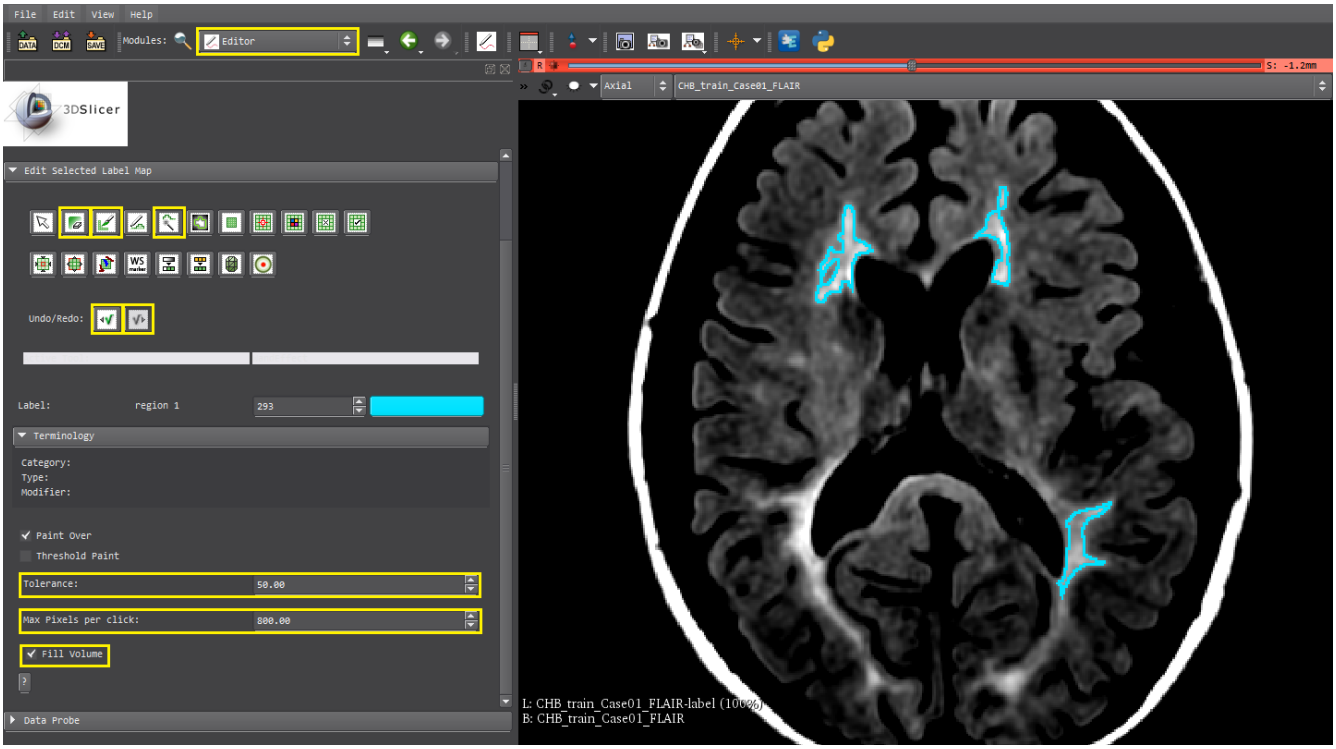


Figure B.1: 3D Slicer user interface for performing in-house manual segmentations and revisions. The tools used are highlighted in yellow, while the in-progress segmentation is shown in blue.

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 [84] 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.

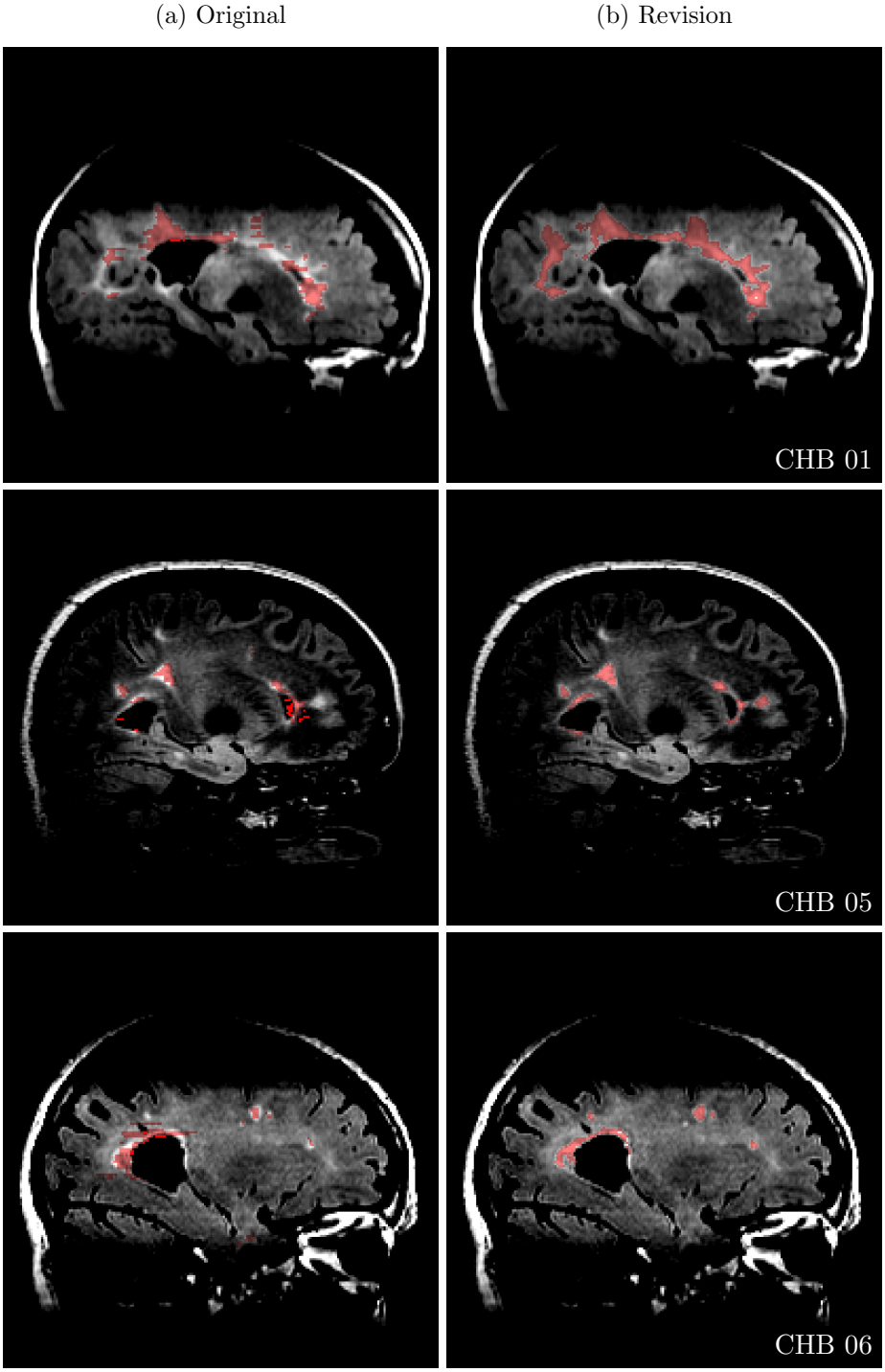


Figure B.2: Example revisions to the manual segmentations for the MS 2008 challenge dataset.

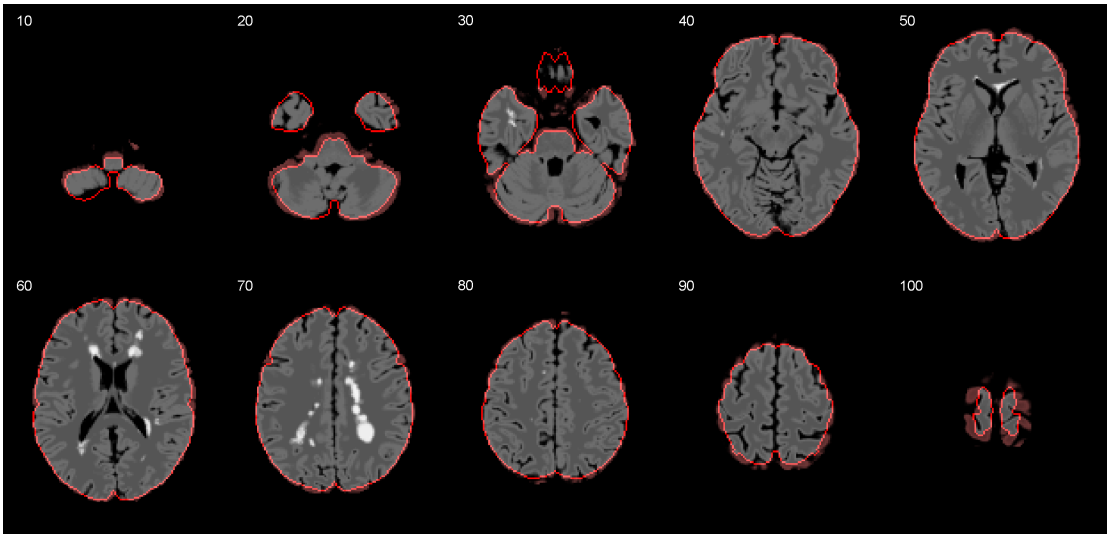


Figure B.3: Manually refined brain mask in MNI space, overlaid on a simulated BrainWeb FLAIR image. Mask outline is highlighted in red; inclusions are shown in grayscale; exclusions tinted red.

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 $\tilde{\boldsymbol{\mathcal{Y}}}_\gamma(x)$, with $\tilde{\mathcal{Y}}_\gamma^0 = 1$, and labels $\mathcal{C}_\gamma(x)$ – 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 $\boldsymbol{\beta}^{(0)}(x)$, 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^{th} feature from the n^{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

$$\mathbb{Y}_n^k = \begin{cases} +\mathbb{Y}_n^k, & \mathbb{C}_n \geq 0.5 \\ -\mathbb{Y}_n^k, & \mathbb{C}_n < 0.5 \end{cases}, \qquad \forall\, k \in \{1, \ldots, K\}. \tag{B.1}$$

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 $\Delta\mathbb{B}^{(t)}$. These calculations are performed in **??**. 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.

$$\mathbb{S} = \frac{1}{1 + e^{+\eta}}, \qquad \eta = \mathbb{B}^0 + \left(\mathbb{B}^1 \circ \mathbb{Y}^1\right) \tag{B.2}$$

$$\mathbb{A} = \mathbb{S} \circ (1 - \mathbb{S}) \tag{B.3}$$

$$\begin{aligned} \mathbb{G} &= \nabla_{\mathbb{B}}\mathcal{L} - \lambda\mathbb{B} \\ &= \begin{bmatrix} \mathbb{G}^0 \\ \mathbb{G}^1 \end{bmatrix} \\ &= \begin{bmatrix} \sum_{n=1}^{N} \left(\mathbb{Y}_n^0 \circ \mathbb{S}\right) \\ \sum_{n=1}^{N} \left(\mathbb{Y}_n^1 \circ \mathbb{S}\right) \end{bmatrix} - \lambda \begin{bmatrix} \mathbb{B}^0 \\ \mathbb{B}^1 \end{bmatrix} \end{aligned} \tag{B.4}$$

$$\begin{aligned} \mathbb{H} &= \nabla_{\mathbb{B}}^1\mathcal{L} - \lambda\mathbb{I} \\ &= \begin{bmatrix} \mathbb{H}^{0,0} & \mathbb{H}^{0,1} \\ \mathbb{H}^{1,0} & \mathbb{H}^{1,1} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{n=1}^{N} \left(\mathbb{A} \circ \mathbb{Y}_n^0 \circ \mathbb{Y}_n^0\right) & \sum_{n=1}^{N} \left(\mathbb{A} \circ \mathbb{Y}_n^1 \circ \mathbb{Y}_n^0\right) \\ \sum_{n=1}^{N} \left(\mathbb{A} \circ \mathbb{Y}_n^0 \circ \mathbb{Y}_n^1\right) & \sum_{n=1}^{N} \left(\mathbb{A} \circ \mathbb{Y}_n^1 \circ \mathbb{Y}_n^1\right) \end{bmatrix} - \lambda \begin{bmatrix} 1 & \\ & 1 \end{bmatrix} \end{aligned} \tag{B.5}$$

$$\begin{aligned} \mathbb{D} &= \det \mathbb{H} \\ &= \left(\mathbb{H}^{0,0} \circ \mathbb{H}^{1,1}\right) - \left(\mathbb{H}^{0,1} \circ \mathbb{H}^{1,0}\right) \end{aligned} \tag{B.6}$$

$$\begin{aligned} \Delta\mathbb{B} &= -\mathbb{H}^{-1}\mathbb{G} \\ &= \frac{1}{\mathbb{D}} \begin{bmatrix} \left(\mathbb{H}^{1,1} \circ \mathbb{G}^0 - \mathbb{H}^{1,0} \circ \mathbb{G}^1\right) & \left(\mathbb{H}^{0,1} \circ \mathbb{G}^1 - \mathbb{H}^{0,0} \circ \mathbb{G}^0\right) \end{bmatrix}^T \end{aligned} \tag{B.7}$$

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 $\boldsymbol{\beta}(x)$ 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 `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 `spm_diffeo` uses only `.nii` files for all input and output data flows, so the estimated $\boldsymbol{\beta}(x)$ images must be written from Matlab to file before transformation, and loaded from file into Matlab afterwards.

B.3.3 Half Resolution Model Estimation

Following the results from § 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 $\boldsymbol{\beta}(x)$ at a lower resolution, followed by interpolative upsampling. This has the additional advantage of requiring significantly less time for model estimation at($\mathcal{O}(n^3)$ 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 § B.3.1, yielding low resolution parameter images. Next, these images, $\boldsymbol{\beta}_{\mathrm{r}}(x)$, were linearly interpolated to the original resolution ($\mathrm{r} = 1$), before application of the smoothing filter to yield the final $\boldsymbol{\beta}(x)$. For reference, an example parameter image at each resolution is shown in Figure B.4.

All results were computed with this adaptation (specifically $\mathrm{r} = 0.5$) except for the parameter image smoothing experiments described in § 4.7.3, which justify this modification.

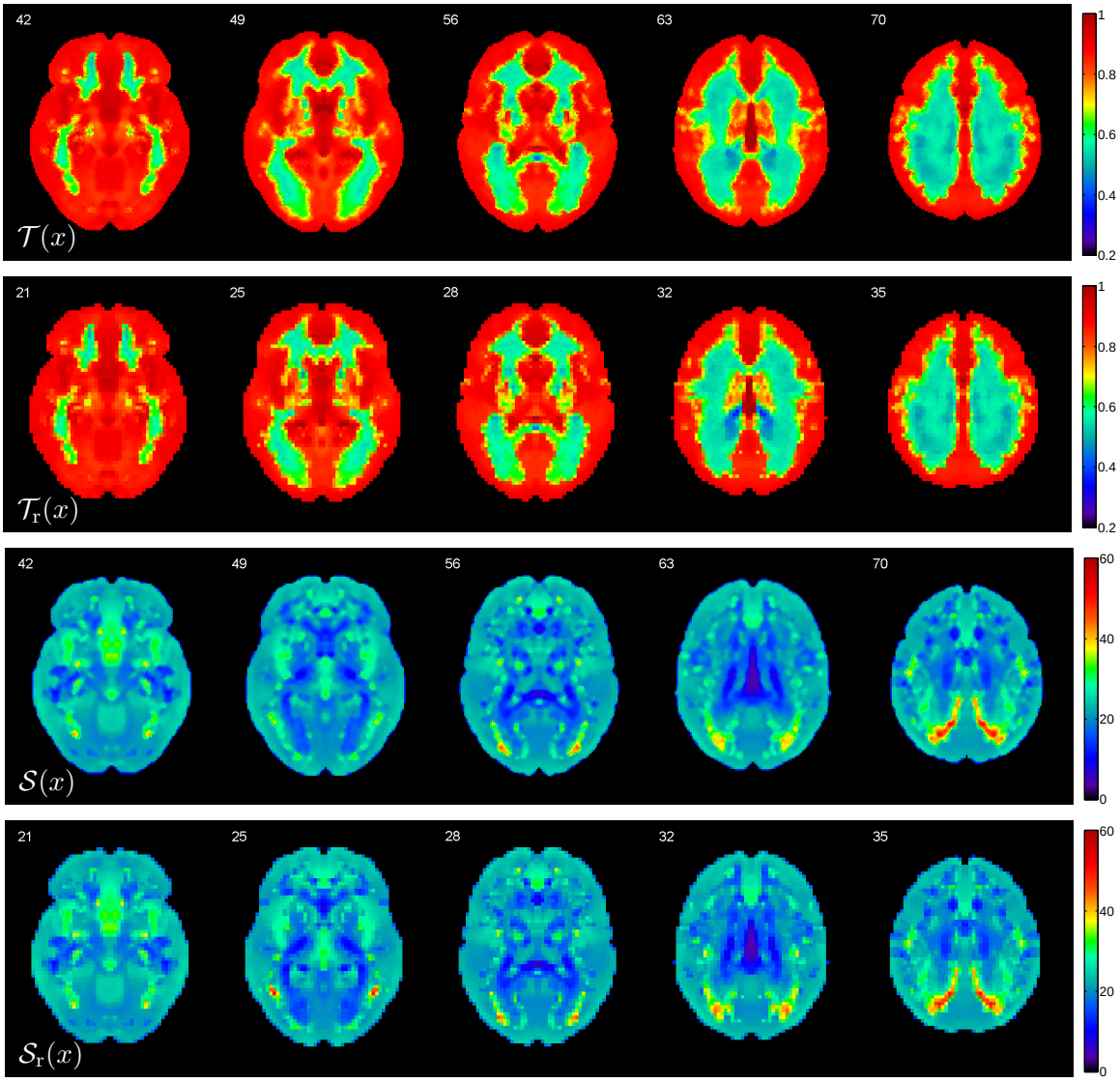


Figure B.4: Comparison of parameter images estimated at full and half-resolution.

Appendix C

Code