Magnetic Resonance Imaging xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

journal homepage: www.elsevier.com/locate/mri

Original contribution

Whole volume brain extraction for multi-centre, multi-disease FLAIR MRI datasets☆

April Khademi[a,*], Brittany Reiche[b], Justin DiGregorio[a], Giordano Arezza[a], Alan R. Moody[c]

[a] *Image Analysis in Medicine Lab (IAMLAB), Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada*
[b] *University of Guelph, Guelph, ON N1G 2W1, Canada*
[c] *Department of Medical Imaging, University of Toronto, Toronto M5S 1A1, Canada*

ARTICLE INFO

Keywords:
Magnetic Resonance Imaging (MRI)
Fluid Attenuation Inversion Recovery (FLAIR) MRI
Brain extraction
Segmentation
Multicentre studies

ABSTRACT

Automatic segmentation of the brain from magnetic resonance images (MRI) is a fundamental step in many neuroimaging processing frameworks. There are mature technologies for this task for T1- and T2-weighted MRI; however, a widely-accepted brain extraction method for Fluid-Attenuated Inversion Recovery (FLAIR) MRI has yet to be established. FLAIR MRI are becoming increasingly important for the analysis of neurodegenerative diseases and tools developed for this sequence would have clinical value. To maximize translation opportunities and for large scale research studies, algorithms for brain extraction in FLAIR MRI should generalize to multi-centre (MC) data. To this end, this work proposes a fully automated, whole volume brain extraction methodology for MC FLAIR MRI datasets. The framework is built using a novel standardization framework which reduces acquisition artifacts, standardizes the intensities of tissues and normalizes the spatial coordinates of brain tissue across MC datasets. Using the standardized datasets, an intuitive set of features based on intensity, spatial location and gradients are extracted and classified using a random forest (RF) classifier to segment the brain tissue class. A series of experiments were conducted to optimize classifier parameters, and to determine segmentation accuracy for standardized and unstandardized (original) data, as a function of scanner vendor, feature type and disease type. The models are trained, tested and validated on 156 image volumes (~8000 image slices) from two multi-centre, multi-disease datasets, acquired with varying imaging parameters from 30 centres and three scanner vendors. The image datasets, denoted as CAIN and ADNI for vascular and dementia disease, respectively, represent a diverse collection of MC data to test the generalization capabilities of the proposed design. Results demonstrate the importance of standardization for segmentation of MC data, as models trained on standardized data yielded a drastic improvement in brain extraction accuracy compared to the original, unstandardized data (CAIN: $DSC = 91\%$ and ADNI: $DSC = 86\%$ vs. CAIN: 78% and ADNI: 65%). It was also found that models created from one scanner vendor based on unstandardized data yielded poor segmentation results in data acquired from other scanner vendors, which was improved through standardization. These results demonstrate that to create consistency in segmentations from multi-institutional datasets it is paramount that MC variability be mitigated to improve stability and to ensure generalization of machine learning algorithms for MRI.

1. Introduction

Neurodegenerative diseases impact the well-being of those affected, while presenting a large economic burden on healthcare systems. To reduce this burden, the etiology and progression of these diseases must be understood so that treatment can be applied early, before irreversible brain damage has occurred. For this, researchers are investigating magnetic resonance images (MRI) of the brain to identify precursors and to further characterize neurodegenerative disease. White matter lesions (WML) are one such pathological feature identified on MRI that are associated with ischemic [1], vascular [2, 3], dementia [4] and demyelinating [5] diseases.

To better understand the relationship between neurodegenerative diseases and WML, images from large patient cohorts need to be

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

* Corresponding author.

E-mail address: akhademi@ryerson.ca (A. Khademi).

https://doi.org/10.1016/j.mri.2019.08.022
Received 22 January 2019; Received in revised form 1 May 2019; Accepted 15 August 2019

analyzed and correlated with patient outcomes. Accurate and quantitative measurement of WML volume, as well as other biomarkers such as brain volume, can be extracted from large databases to model disease progression, and to identify new risk factors [1, 6-10]. However, manual biomarker measurement is subjective and extremely laborious, especially for large-scale studies [11, 12]. Automated algorithms that measure biomarkers are an ideal alternative, since they can compute lesion volume and other quantitative metrics for thousands of patients in an objective, accurate, and efficient manner.

Fluid-Attenuated Inversion Recovery (FLAIR) MRI is becoming increasingly important for diagnosis and treatment of neurodegenerative disease [1-5, 13]. This is because the cerebral spinal fluid (CSF) signal is nulled, which emphasizes the appearance of WML, increasing the ease of analysis [2, 14-16]. Many works have analyzed FLAIR via multimodality approaches that co-register FLAIR to T1- and T2-weighted MRI. However, such methods increase image acquisition costs (multiple scans) and error introduced during registration due to the differing contrast appearances in the three sequences [17]. As FLAIR is routinely acquired, methods dedicated solely to this sequence would have clinical value.

Automatic WML segmentation frameworks have already been developed to analyze WML, for example [18-20]. Critical for robust application of these algorithms is the preprocessing step of brain extraction. Brain extraction, or skull stripping, removes any non-brain tissues from the image (i.e., skull and eyes), as they can interfere with WML segmentation schemes. Moreover, brain extraction methods allow for automated brain volume measurement.

There are existing brain extraction algorithms for other MR sequences, such as T1- and T2-weighted MRI [21-23]. In [21], the ROBEX algorithm uses a machine learning approach that was trained on T1-weighted data, which likely cannot generalize to FLAIR images due to differences in intensities for tissues between sequences. The other popular method, brain extraction tool (BET) [22], uses a deformable model that is initialized at the centre of the brain volume, and expands until it reaches a significant threshold. The threshold is often met as the model passes through WML, and the resultant images tend to be undersegmented. In [24], we explored these tools on FLAIR MRI, and found that the segmentations were not optimal (average dice similarity of 56.58% for BET and 60.6% for ROBEX). In [23], a convolutional neural network was used to perform brain extraction, but these models are known to be computationally intensive, and the authors noted that results were sensitive to cases where datasets had vastly different acquisition parameters. Moreover, specialized hardware (GPUs) are needed to run these algorithms, which are usually not available in hospital systems, reducing clinical utility and translation of such methods.

Few works have been developed to handle brain extraction in FLAIR MRI. Some approaches are semi-automated requiring user input [19], which is laborious and subjective. Others require the use of multiple modalities (i.e. T1, T2, etc.) [25-27], but multi-sequence registration can create segmentation errors [17]. In [25], the authors propose a FLAIR-only MRI brain extraction method based on edge detection, the local moment of inertia and morphological processing. The method was validated on 30 cases, but was developed specifically for images from a single centre, acquired with the same device and imaging parameters. Therefore, it is unknown if the method will generalize to multicentre datasets and the authors acknowledge this limitation in the conclusion of their manuscript.

To combat these downfalls, this work presents a whole volume, fully-automatic brain extraction approach designed using solely the FLAIR modality, that is efficient and robust to multi-centre (MC) image variability. To develop a method that is robust to MC variability can be challenging, as differences between acquisition parameters, scanner vendor hardware, reconstruction algorithms, and patient-dependent artifacts creates significant variability in image properties within large datasets. For example, there can be large differences in the distributions of acquisition noise, intensity inhomogeneity, intensity non-standardness, differences in voxel resolution, and patient orientation [28]. All of these sources of variability affect the results of automated segmentation algorithms [29].

The proposed work handles all of the challenges of MC analysis in one framework. Images are standardized [30, 31], which reduces variability in MC datasets by normalizing the intensity scale, suppressing acquisition artifacts, and normalizing voxel resolution and patient orientation. Using the standardized image volumes, brain extraction is performed using a Random Forest (RF) classifier [32], based on an intuitive feature set. After classification, mathematical morphology is applied as a post-processing step to suppress false positives, which are a common issue in machine learning-based approaches to brain extraction [21, 23]. This simple step deeply contrasts other approaches, which require complex models to perform post-processing, such as generative models or graph cuts [21].

In total, 156 image volumes with ground truths (approximately 8000 image slices) are used to train, test and validate the proposed brain segmentation scheme. Experiments include optimization of classifier parameters, followed by the analysis of the effects of standardization, scanner vendor, and feature type on segmentation accuracy. The FLAIR MRI dataset used in this work, is collected from two multi-centre datasets from the Canadian Atherosclerosis Imaging Network (CAIN) (vascular disease) [33] and the Alzheimer's Disease Neuroimaging Initiative (ADNI) (dementia disease) [34]. The cases selected for this study were acquired at thirty centres with varying acquisition parameters and scanner vendors, and represent two different diseases. Vascular disease can be associated with strokes, and large lesion burdens, whereas dementia may be characterized by brain atrophy with varying lesion loads. Validation using these images will demonstrate the robustness of the framework across a diverse collection of images, scanning hardware, imaging centres and diseases, which increases its potential for clinical translation.

Because of image standardization, we hypothesis that a simple feature set will generate accurate and robust automatic whole volume brain segmentations on MC data. We also strongly believe that part of the elegance of the solution, is in fact, in its simplicity. Variability from patients and scanners have been systematically and deliberately reduced, which allows such a tool to generalize multi-institutional FLAIR MRI from multiple diseases. Also, feature sets and classification sampling strategies are chosen to maximize robustness. As will be shown, the framework is robust to MC variability and has been designed with simplified models and intuitive feature sets which are feasible due to the novel image standardization framework.

This work will represent one of the first approaches to automatic whole volume brain extraction in multicentre, multidisease FLAIR MRI, validated on two MC and multi-disease datasets. Practical implementation and clinical use is a major consideration of this design, and it can be easily translated into routine clinical workflow. The interpretable feature set allows for analysis of failures in a robust manner, the minimal processing allows for real-time implementation that does not require specialized hardware and dependence on only the FLAIR sequence eliminates the reliance on other sequences.

The remainder to the paper is detailed as follows: Section 2 will outline the methods and materials used for this work, which includes description of the standardization framework, feature extraction, classifier sampling and construction, validation metrics, as well as the data that is used. Section 3 outlines the optimization of the classifier parameters, followed by experiments conducted to analyze the effects of image standardization, scanner vendor, and disease classification on segmentation accuracy. Section 4 contains the discussions and Section 5.

2. Methods and materials

Given an axial image volume, $I(x,y,z)$, where x,y are the spatial

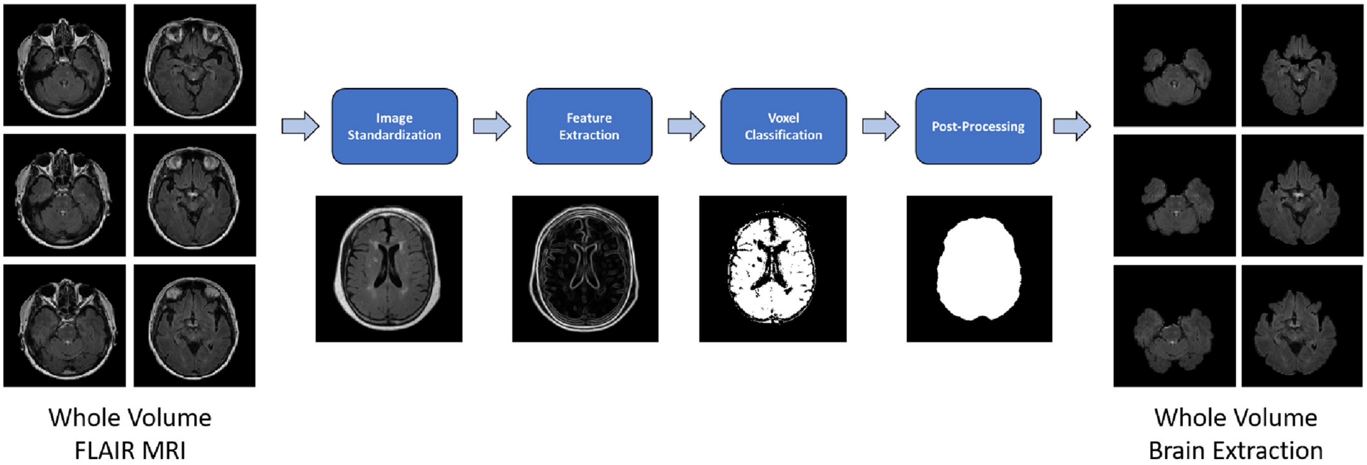


Fig. 1. Overview of proposed whole volume FLAIR MRI brain extraction technique. Note: only seven image slices are shown for demonstrative purposes.

coordinates of each slice, and z is the slice number, the goal of the brain extraction algorithm is to find a binary segmentation mask $b(x,y,z)$ that identifies voxels from the brain tissue class. This mask may be multiplied (voxel-wise) by the original image in order to extract the brain region from the whole volume. In this work, as shown in Fig. 1, several steps are utilized to find the brain mask $b(x,y,z)$. First, the data is preprocessed using a novel standardization framework, that reduces noise, standardizes the intensity scale, and normalizes spatial dimensions and orientation. These steps are designed to systematically and deliberately reduce each of the sources of variability in multi-centre FLAIR MRI. Based on the standardized data, an intuitive feature set is extracted from each voxel. Using an optimized random forest classifier, with carefully designed sampling strategies, the segmentation model is constructed to extract the brain for whole volume FLAIR MRI. A simple morphological post-processing step is used to tidy the binary segmentation mask $b(x,y,z)$. This section briefly describes the standardization framework presented in [30] and [31], followed by details of the process of brain extraction, which is achieved through feature extraction, classification and post-processing.

2.1. Standardization framework

Although access to data is growing, the ability to apply algorithms on multicentre data has been limited due the multicentre effect (MCE): different acquisition systems create differences in noise, intensity, contrast and resolution in MC datasets. As algorithms are quantitative, small changes in the images (i.e. intensity values) can have a large negative impact on the reliability of the results. To manage variability in MC FLAIR MRI data, a first-of-its kind standardization pipeline was developed in [30] and [31] that reduces variability to improve performance of image analysis and machine learning tools on MC data. Most notably, it reduces noise and bias field, makes the intensity distributions of images consistent across and within imaging centres and normalizes brain orientation and voxel resolutions. The standardization framework is briefly outlined in this section.

2.1.1. Denoising and background suppression

To remove acquisition noise, a median filter was employed to remove spurious pixels while maintaining edge information. For background subtraction, the upper and lower 2% of the image histogram were cropped to remove outlier intensities and a K-means classifier (k = 2) was used to segment the image into its foreground and background components. Morphological processing and filling was completed to ensure that foreground regions included the ventricles. The background mask was used to zero-out all non-tissue pixels and background noise.

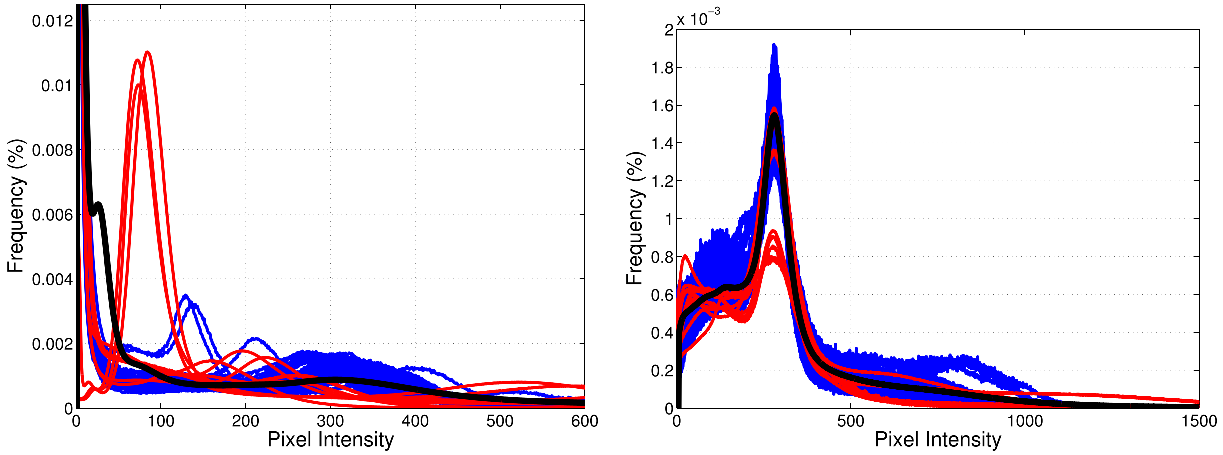
2.1.2. Bias field correction

Bias field correction was performed in a way similar to [35]. Each image slice is divided by a low-pass filtered version of itself, which represents the low frequency bias field artifact and therefore, suppresses the modulation of intensities from within the same tissue class.

2.1.3. Intensity standardization

A novel intensity standardization for MC FLAIR MRI was developed in [30] and [31] and is used to align the histograms of all images in a database to an atlas, yielding a consistent intensity interval for the same tissues between images. This is accomplished by aligning the largest mode of the histogram between volumes, which corresponds to the gray matter (GM) and white matter (WM) intensities (broadly: the brain), yielding similar histograms between images. More specifically, the intensity of the GM/WM (brain) peak of the volume is determined, and a scaling factor is computed by dividing the intensity of the atlas' brain peak by the volume's brain peak. This factor is multiplied by the original volume such that the brain peak is now aligned with that of the atlas. As an optional post-processing step, we perform what is called slice refinement. In this stage, the peak of each slice is detected and shifted to be aligned with the peak of the volume. This step is completed to ensure that the brain peaks of each slice are optimally aligned with that of the volume. This work has been applied and validated on 350000 FLAIR MRI from more than 60 international imaging centres, for 3 T and 1.5 T data, scanned by GE, Siemens and Philips scanners for patients with dementia (ADNI) and vascular disease (CAIN), as outlined in [30] and [31]. The Kullback Leibler (KL) distance [36] was used to measure the similarity of the volume histograms in a set, before and after standardization where a small distance indicates a high degree of similarity. The KL divergence between the volume histograms and the mean volume histogram per scanner/disease was reduced from 1.013 ± 1.635 to 0.094 ± 0.057 on average, for the original and standardized data, respectively, indicating a high degree of similarity in the intensity distributions in the standardized dataset. In addition, the standard deviation of the KL distance was greatly reduced indicating higher consistency in the standardized data.

Fig. 2 contains the volume histogram of the CAIN and ADNI datasets used for brain extraction in this work, which were acquired by GE, Siemens and Philips scanners, before and after denoising and intensity standardization. Intensity standardization aligns histograms and makes the intensity ranges and distributions more similar across datasets, indicating that the same tissues are mapped to the same intensity ranges regardless of the institution or scanner vendor used to acquire the data. Some example images before and after standardization are shown in Fig. 3 for various scanner vendors (note the same display range is used to view all images). Prior to standardization, images have varying



(a) Original Volume Histograms

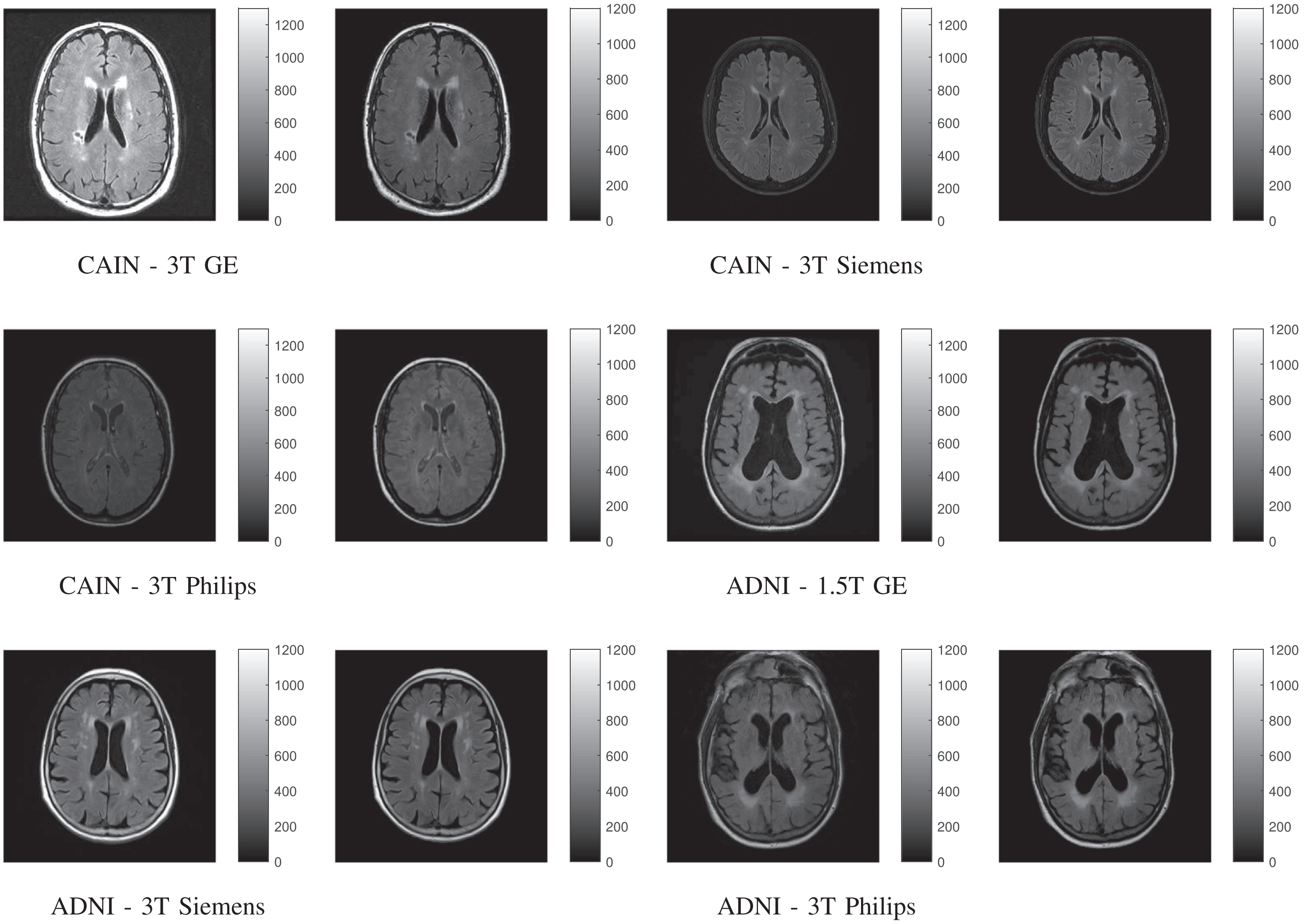
(b) Standardized Volume Histograms

Fig. 2. Intensity histograms of standardized FLAIR MRI volumes. Red: ADNI data, blue: CAIN data, black: average. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

intensities and contrasts, for the same tissue classes. After standardization, the intensities from each tissue classes in all the images appear to be more uniform over both datasets, and for all three scanner vendors. These characteristics should allow for the robust application of image analysis and machine learning algorithms to large multicentre FLAIR MRI datasets.

2.1.4. *Voxel resolution and patient orientation normalization*

To ensure that voxel resolution and patient orientation was normalized in the dataset, images were registered to the atlas using affine registration based on the demons algorithm [37]. Squared differences was used as the error metric, and optimization was achieved using gradient descent. The number of iterations was limited to 100, and



CAIN - 3T GE

CAIN - 3T Siemens

CAIN - 3T Philips

ADNI - 1.5T GE

ADNI - 3T Siemens

ADNI - 3T Philips

Fig. 3. Original images and results from standardization steps. Each original image in shown, followed by its standardized version.

images were transformed using cubic interpolation, as this method minimizes transformation artifacts. Non-linear registration methods were also investigated; however, it has been noted that with varying levels of WML, strong distortions can occur in intensity-based registration. The output from this phase of the standardization pipeline is normalized voxel sizes between images, and a similar position of the brain in all images, therefore maximally aligning anatomy between subjects.

2.2. Whole volume brain extraction

In this work, we focus on the design of a brain extraction tool for FLAIR MRI that is robust to MC and multi-disease datasets of patients with varying lesion loads based on an intuitive feature set. Because of image standardization, we hypothesis that a simple feature set will generate accurate and robust automatic whole volume brain segmentations on MC data. Three groups of features are investigated that are based on intensity, spatial or gradient based features. To increase computational efficient and reduce memory consumption, the RF model is constructed using a novel training voxel sample selection strategy. This section will outline the feature extraction, parameters of the RF used for voxel classification, the sampling strategy and post-processing.

2.2.1. Feature extraction

In this work, we focus on an intuitive feature set that is robust to MC and multi-disease variability. Intuitive features are important for understanding system failures and analysis of the results on these features can be used to improve performance. Also, simplified models and features lend to clinical translation since implementation is real-time and does not require specialized hardware systems to run. For classification, 28 features were extracted on a per-voxel basis for each volume. These features can be grouped into three categories: Intensity, Spatial and Gradient-based features. These features are described here.

2.2.1.1. Intensity-based features. Intensity standardization has normalized the intensity of tissues across images. Tissues have been mapped to similar intensity ranges yielding a consistent intensity interval for tissues across volumes, as shown in Figs. 2 and 3. Visually, since brain tissue is often darker than non-brain tissues (i.e. skull, ears, eyes), intensity should be a highly discriminatory feature. Therefore, the first feature considered is the voxel intensity and is extracted as a feature, as in $F_1(x,y,z) = I(x,y,z)$. To further exploit intensity information while minimizing noise, a smoothed version of the image was also computed, and mean neighbourhood values of pixel intensities were calculated for kernels of 5 mm and 7 mm in size, resulting in $F_2(x,y,z)$ and $F_3(x,y,z)$. These kernel sizes were selected to reduce noise significantly, while maintaining global intensity characteristics.

2.2.1.2. Spatial location-based features. Spatial registration in the standardization framework normalizes spatial coordinates across images, resulting brain tissue locations being approximately consistent across volumes. Therefore, as a feature, spatial location is a natural choice. The positional features included the (x,y,z) coordinates of each voxel, with $F_4(x,y,z) = x$, $F_5(x,y,z) = y$ and $F_6(x,y,z) = z$.

2.2.1.3. Gradient-based features. A key defining characteristic of brain tissue is its smoothness relative to the surrounding tissues, which have sharp edges. For example, the interior of brain regions have approximately uniform intensity, while the skull and exterior tissues have rapidly changing intensity characteristics. Additionally, although the intensities of WML are typically similar to the skull, and non-brain tissues, the edge content of WML are highly variable. Due to partial volume averaging, the edge strength of WML boundaries are often low, and dispersed [38]. These visual clues led to the investigation of edge-based features. Ultimately, a set of features that can robustly discriminate between brain and non-brain tissues, while identifying WML as “brain” tissues is desired.

In order to capture edges of different sizes and scales, such as the sharp brain boundary, while being robust to the softer boundary between WML and normal brain tissues, image gradients are investigated for different sized neighbourhoods and scales. Larger neighbourhoods will mostly likely be the most robust to WML edges which are more diffuse, but smaller windows should capture the strong edge features such as the brain-background boundary.

There are two types of gradient features calculated. The first is based on the first order gradient magnitudes (gradient magnitude features), computed from the average of three different-sized regions. The second set of gradient features are computed from Gaussian scale space, where the first order and second order gradients are computed for each of the x, y and z directions separately, for three different scales. These features are called the first order and second order Gaussian scale space features, respectively.

For the gradient magnitude features, the magnitude of the image gradient is calculated in 3D:

$$||\nabla I|| = \sqrt{\frac{\partial I}{\partial x}^2 + \frac{\partial I}{\partial y}^2 + \frac{\partial I}{\partial z}^2}, \tag{1}$$

where the digital gradient is approximated using the Sobel operator. To capture edge information at different scales, the average gradient magnitude of the neighbourhood surrounding each pixel was computed for four different-sized regions to obtain the dominating edge strength value for that region, while reducing noise. The four regions that were considered had widths of 4, 5, 8 and 16 mm, resulting in four gradient magnitude features $F_7(x,y,z)$, $F_8(x,y,z)$, $F_9(x,y,z)$ and $F_{10}(x,y,z)$. These scales were selected to capture both small and large-scale edge characteristics.

As gradients are sensitive to noise, Gaussian scale space features were also considered. Gaussian scale space edge detection involves smoothing (convolving) the image with a Gaussian kernel of some width σ, and then taking the gradient of the smoothed image. The benefit of using Gaussian scale space to compute the gradient is the ability to investigate edge content at different scales. Dependent on the size of the smoothing kernel, it is possible to isolate larger or smaller edge features. For example, a large σ smooths out small objects, while retaining large objects that are of the same size approximately as σ. Then, taking the gradient of the smoothed image examines the edge content of these large objects. Therefore, Gaussian scale space allows us to robustly investigate the edge content of different sized objects, and edge strengths, which could be robust in differentiating between the brain and non-brain tissues.

To extract Gaussian scale space gradient features, first, the 3D Gaussian smoothing kernel is used:

$$G(x, y, z, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2 + y^2 + z^2)}{2\sigma^2}}, \tag{2}$$

where σ is the standard deviation of the Gaussian function, which is proportional to the scale of the objects being detected. Scales of 1, 2, and 8 mm were used because they present valuable local information of each image at different scales. To compute the features at a specific scale, the original image is convolved with the 3D Gaussian smoothing kernel, and the edges are detected, as in

$$H(x, y, z, \sigma) = \nabla(I(x, y, x) * G(x, y, z, \sigma)). \tag{3}$$

Instead of looking at gradient magnitudes which combines the edge information from all directions, for the Gaussian scale space features, the individual gradient directions are separately investigated since the images have been smoothed and these estimates should be less noisy. The motivation behind this is to determine whether individual edge directions are discriminating features. The individual gradients along each direction were computed as features, where $F_{11}(x,y,z)$, $F_{12}(x,y,z)$,

$F_{13}(x,y,z)$ are the gradient along the x-direction for the three different scales, $F_{14}(x,y,z)$, $F_{15}(x,y,z)$, $F_{16}(x,y,z)$, are the gradients along the y-direction for the three different scales and $F_{17}(x,y,z)$, $F_{18}(x,y,z)$, $F_{19}(x,y,z)$ are the gradients along the z-direction for the three different scales as well. In addition to these, the second order gradient was computed, and the gradient along the respective directions, for each scale was taken, resulting in $F_{20}(x,y,z)$, $F_{21}(x,y,z)$, $F_{22}(x,y,z)$, $F_{23}(x,y,z)$, $F_{24}(x,y,z)$, $F_{25}(x,y,z)$, $F_{26}(x,y,z)$, $F_{27}(x,y,z)$, $F_{28}(x,y,z)$.

2.2.2. Classifier construction and training

The RF was constructed with optimized parameters for the number of features, trees, and examples; see Section 3 for the experimental design and results. The number of features analyzed at each node was set to 2, since with a large number of trees, the strength of individual trees is less relevant, and a higher value for this parameter would likely increase the correlation between trees, which is known to increase error [39]. The minimum number of training voxels present at a node and leaf was set to 20 [21]. Generalization of the model was ensured by implementing pruning after training, which is the process of randomly removing some branches of each tree after training. This is known to suppress any effects of over-fitting of the training set, therefore increasing generalization to new data [39].

Due to the high correlation of voxels in an image volume, not all voxels from the training volumes are needed to construct the classifier, which reduces classifier complexity and computational load. Instead, a selective training sampling strategy is explored which randomly samples voxels from the training set to build the model. Each voxel from volumes within the training set is classified as positive, negative, or restricted negative using the corresponding ground truth brain masks. Positive voxels represented brain tissue and negative voxels represented non-brain tissue. Restricted negative voxels represent negative cases that lay on the brain-skull boundary which were identified early on as the most difficult cases to correctly classify. The labels for all voxels are provided by expert generated binary brain masks.

The training voxels are then randomly selected with an even number of positive and negative cases which avoids biases associated with class imbalances during training [39]. Within the negative class, 75% of negative examples are restricted to lay within 10 mm of the brain boundary. This is done since through preliminary experiments, it was found that the most difficult negative examples to correctly classify lay near the brain boundary where there is some intensity overlap between brain tissue and surrounding cerebrospinal fluid. Therefore, it is desirable to expose the classifier to more of these challenging cases, which should increase accuracy and make the classifier more robust [21]. To find the “restricted negative” training voxels, a combination of edge detection and morphology were used. This edge provides a pool of voxel indices from which the difficult negative training instances are sampled.

The number of training voxels per volume is evenly distributed (i.e. for 150,000 training voxels with 76 training volumes, ~1974 voxels would be randomly selected from each volume). From each of the selectively sampled voxels, the 28 features were computed. Compared to entire volumes, which can have 512 × 512 × 45 voxels for example, the proposed sampling strategy can be used to build models more computationally efficiently and less memory intensive.

2.2.3. Post-processing

A simple post-processing scheme was implemented to reduce false positives which are a common issue brain extraction approaches using machine learning [21,23]. Initially, the brain masks $b(x,y,z)$ were eroded using a kernel size of 4 to remove small structures that were connecting brain to non-brain tissues. Any small clusters of voxels not connected to the central brain mass were then removed using connectivity analysis. The remaining mask was dilated by a kernel size of 6 to reduce the impact of the initial erosion step, and this was followed by hole filling. In contrast to similar approaches to brain extraction [21], this is a very simple approach to false positive reduction.

2.3. Validation metrics

Segmentation accuracy was objectively compared to the groundtruth using multiple metrics. To measure the amount of intersection between a segmented object and the groundtruth, the Dice Similarity Coefficient (*DSC*) [40] was calculated:

$$DSC = \frac{2|A \cap B|}{|A| + |B|}, \tag{4}$$

where A and B are the binary masks of the brain for the groundtruth reference and automatic segmentation, respectively. The Hausdorff Distance (*HD*) was also calculated, which is a measure of maximum surface-to-surface distance [41]. It is calculated as the sum of distances between boundary points of the automatic segmentation and their closest neighbours in the groundtruth mask. In contrast to the *DSC*, this metric penalizes cases in which two overlapping objects still have different boundaries.

In addition to these metrics, classification accuracy was further quantified using sensitivity (*sens*), also known as Overlap Fraction, and is a measure of the true positive (*TP*) rate:

$$sens = \frac{TP}{TP + FN}, \tag{5}$$

where *FN* are false negatives. In addition, the specificity (*spec*) was calculated as a measure of the true negative (*TN*) rate:

$$spec = \frac{TN}{TN + FP}, \tag{6}$$

where *FP* are false positives. Extra Fraction [42] was also calculated, which is a measure of the false positive rate:

$$EF = \frac{FP}{TP + FN}. \tag{7}$$

In an ideal automatic segmentation, the *DSC*, specificity, and sensitivity measures should be close to one, while *HD* and *EF* should be close to zero.

2.4. Datasets

In this work, two datasets are used for training and testing. The two FLAIR MRI datasets represent a highly diverse set of data from two studies that focus on research of vascular and dementia disease, collected from over 30 international imaging centres. The FLAIR MRI data is collected from GE, Philips and Siemens scanners, with widely varying imaging parameters, with variable characteristics that adequately represent multi-institutional data. From these datasets, 156 volumes have been manually annotated.

The first dataset is from the Canadian Atherosclerosis Imaging Network (CAIN), a pan-Canadian study to study vascular disease [33]. Neuroimaging, neuropsychological assessments, and clinical data were collected from patients across nine imaging centres in Canada, who presented with symptomatic vascular disease. There is approximately 400 patients, with longitudinal follow up. Subjects have a mean age of 73.2 ± 8.25 years old, and roughly 56% are male. FLAIR MRI was acquired for each subject in the study on a 3 T GE, Siemens or Philips scanner in the axial plane, and with varying imaging parameters. In this work, 135 image volumes (roughly 7000 image slices) were used for training and testing, and the volumes selected were randomly chosen from the entire dataset constrained to contain an equal distribution of images from each of the scanner vendors and institutions. Acquisition parameters for the CAIN dataset can be found in Table I. Note that there are multiple values for TR/TE/TI and pixel spacing as represented by the range found in the data used.

The second dataset used in this work is from the Alzheimer's Disease

Table I
Summary of datasets and acquisition parameters. All images were acquired at 3 T.

Database	Disease	No. volumes	Centres	Scanner vendors	TR (ms)	TE (ms)	TI (ms)	Pixel spacing (mm)	Slice thickness (mm)
CAIN	Vascular	135 (~7000 slices)	9	GE, Philips, Siemens	8000–11,000	117–150	2200–2800	0.4286–1	3
ADNI	Dementia	21 (~1000 slices)	21	GE, Philips, Siemens	650–11,900	20–193	2000–2800	0.7813–1.0156	5

Neuroimaging Initiative (ADNI), which is an open source dataset for researchers to analyze large amounts of patient data related to dementia. Many patients have undergone a 3 T FLAIR MRI sequence on a GE, Philips or Siemens scanner, and there are roughly 60 international imaging centres where images have been collected. From these subjects, 21 subjects (roughly 1000 image slices) were randomly selected from different centres, with 7 subjects selected evenly for each scanner vendor (i.e. GE, Siemens, Philips). Random sampling resulted in at least three volumes from each of the disease classifications in ADNI (i.e. Normal, Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), Subjective Memory Concerns (SMC), and AD). In total, 3 Normal, 3 EMCI, 6 AD, 3 SMC and 6 LMCI FLAIR MRI volumes were selected. The progression of the disease can affect the prevalence of WML, as well as the morphological characteristics of the brain, making this a diverse dataset that will be optimal for validating the robustness of the framework. Acquisition parameters for the ADNI dataset used in this work are listed in Table I, where there were multiple values for TR/TE/TI and pixel spacing as represented by the respective ranges.

A single biomedical summer student was trained by a radiologist to segment the brain tissue through two training sessions. The student completed the manual outlines and the radiologist verified the manual segmentations. Ground truth annotations were generated using the Pathcore Sedeen Viewer[1]. Manual segmentations were generated as XML coordinate files and converted to binary masks. This was completed for all 135 CAIN image volumes and 21 ADNI image volumes.

For intensity and spatial standardization, a FLAIR MRI atlas was used[2]. The FLAIR template was created from 853 subjects, with a mean age of 43.85 years ± 14.81, with 515 females. The subjects had varying degrees of WML disease. The FLAIR sequence used echo time (TE): 353 ms, inversion time (TI): 1800 ms, repetition time (TR): 5000 ms, flip angle 180, echo train length: 221, voxel size: 0.7 mm^3. Images were acquired in the axial plane on a Siemens 3 T scanner.

3. Results

In this section, the experimental design and results will be detailed. Experiments are split into two different stages. The first phase focuses on optimization of the classifier, and there are three experiments that optimize the number of features, trees, and training examples. The optimized classifier is then used for the second phase of experiments, focused on performance evaluation of the brain segmentation results. For this phase, there are four different experiments that evaluate the effects of dataset (disease type) used to generate the model, standardization, scanner vendor, and feature type on the segmentation accuracy.

The databases from where the FLAIR MRI data was sampled from, and details on how the manual segmentations were generated are described in the previous section. In total, there are 135 CAIN volumes and 21 ADNI volumes with manual segmentations (roughly 8000 image slices), which provide a strong representation of the multi-centre variability of FLAIR MRI created by diverse acquisition parameters and pathology. For all optimization experiments, classifiers were trained and tested with 108 CAIN image volumes, while 27 volumes were reserved for validation in the segmentation experiments. For the segmentation experiments, a variety of training and testing datasets are used, and are specifically detailed in the experiments. A summary of the experiments, data organization, training and testing data splits, and whether original or standardized data was used are detailed for the optimization and segmentation experiments in Tables II and III, respectively.

3.1. *Classifier optimization experiments*

In this subsection, the design and results for the classifier optimization experiments are detailed. In the optimization phase, three parameters are optimized for the classifier, including the number of: extracted features per voxel, trees, and training examples. Classifiers were trained and tested with 108 CAIN image volumes, while 27 volumes were reserved for validation in the segmentation experiments. For the optimization experiments, the remaining set of 108 CAIN volumes were split into 76 volumes for training, and 32 for testing (this is an approximation of the 70/30 split common in the machine learning approaches [39]). The training and testing volumes were randomly shuffled to ensure that parameters were not dependent on any specific training set, and all results are reported as an average of five runs for each experiment. Additionally, where applicable, a separate model was generated using the standardized data, and a separate model was generated using the original data in order to examine the utility of the standardization approach. To summarize, the final, optimized classifier for standardized data was constructed using 15 features, 200 trees and trained with 150,000 training voxels. This section will show these results.

3.1.1. *Feature selection*

A set of 28 features are extracted from every voxel in the image volumes. It is useful to reduce the dimensionality of this set to reduce computational costs, as well as to increase accuracy by reducing noise. In addition, the authors hypothesize that due to standardization, a small feature set should yield good performance due to the reduced variability. To determine the most discriminatory features, the Minimum-Redundancy Maximum-Relevancy (mRmR) algorithm was used [43]. This approach calculates mutual information between the features and ranks the features in order of importance. Although feature importance is measured, the number of these features that should be used to yield optimal performance of the classifier is not reported by the mRmR algorithm. To overcome this and determine the optimal number of features to extract per voxel, the classifier was trained with the top N features from mRmR, and classification accuracy (ACC) was calculated by:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{8}$$

The top N features from the mRmR algorithm that results in the greatest accuracy will be retained and used in the final, optimized classifier. To determine the optimal feature subset, a model was trained with 150 trees and 80,000 examples, for original and standardized data separately. The classification accuracy, as a function of the top N features determined by mRmR, for both standardized and original data is shown in Fig. 4. According to this graph, two things are deduced. First, it can be seen that standardization yields roughly 10% increase in classification accuracy over all feature combinations. Second, the optimal feature set for the standardized data (highest classification

[1] https://pathcore.com/sedeen/

[2] https://brainder.org/download/flair/

Table II
Summary of data organization for classifier optimization experiments.

Experiment	Analysis	Original/standardized	Training/testing (# volumes)
Feature selection	Optimization	Original, standardized	CAIN 108 - with 76/32 split
No. trees	Optimization	Original, standardized	CAIN 108 - with 76/32 split
No. examples	Optimization	Original, standardized	CAIN 108 - with 76/32 split

Table III
Summary of data organization for segmentation experiments.

Experiment	Analysis	Original/standardized	Training data (# volumes)	Testing data (# volumes)
CAIN (vascular disease) model	Segmentation	Original, standardized	CAIN (108)	CAIN (27), ADNI (21)
ADNI (dementia disease) model	Segmentation	Original, standardized	ADNI (21)	CAIN (27)
Effect of scanner vendor [×3 vendors]	Segmentation	Original, standardized	CAIN (36)	CAIN (72)
Feature effects	Segmentation	Original, standardized	CAIN (108)	CAIN (27)

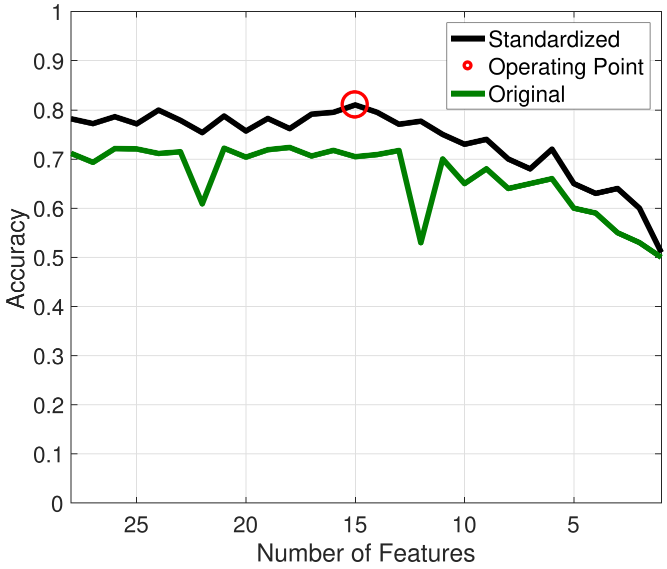


Fig. 4. Optimization of feature set, as measured by classification accuracy.

accuracy) was found to be 15 (classification accuracy of 80.15% ± 8.97). The top 15 selected features were (in order): smoothed intensity with a kernel of 7 mm - $F_3(x,y,z)$, z coordinate - $F_6(x,y,z)$, x coordinate - $F_4(x,y,z)$, y coordinate - $F_5(x,y,z)$, pixel intensity - $F_1(x,y,z)$, gradient magnitude at a scale of 16 mm - $F_{10}(x,y,z)$, gradient magnitude at a scale of 5 mm - $F_8(x,y,z)$, smoothed intensity with a kernel of 5 mm - $F_2(x,y,z)$, second order Gaussian scale features in the y direction at 1 mm - $F_{23}(x,y,z)$, gradient magnitude at a scale of 8 mm - $F_9(x,y,z)$, second order Gaussian scale features in the x direction at 1 mm - $F_{20}(x,y,z)$, second order Gaussian scale features in the y direction at 8 mm - $F_{25}(x,y,z)$, second order Gaussian scale features in the z direction at 8 mm - $F_{28}(x,y,z)$, gradient magnitude at a scale of 4 mm - $F_{28}(x,y,z)$, and second order Gaussian scale features in the y direction at 2 mm - $F_{24}(x,y,z)$.

These features demonstrate several important characteristics of the images: (*i*) that intensity and position of the brain are important to classification, which can be attributed to the normalization of these features via standardization; (*ii*) that the Gaussian derivatives at varying scales capture diverse local representations edge content which assists in classification; and (*iii*) that the gradient edge magnitude feature indicates that texture of the brain (smoothness) is a highly discriminatory feature, but also that the large neighbourhood kernel is required to be robust.

These results are compared to those obtained by the classifier trained on original images. As can be seen in Fig. 4, the original image classifier achieved a lower accuracy than those trained on standardized images, which likely can be attributed to the variability in the images. To optimize the next parameter (number of trees) for the original image classifier, 21 features were used, based on the performance reported in Fig. 4.

3.1.2. Number of trees

At its inception, it was claimed that a RF could be constructed with any number of trees, as they do not overfit [32]; but recent work has shown that they can overfit noisy datasets [44]. For this reason, experiments were conducted with varying numbers of trees (from one to 500), and the setting that yielded the best performance was selected for the final classifier.

For this experiment, the models were trained with the optimized feature set and 80,000 examples. The number of trees did not have a significant impact on the accuracy of the classifier (average accuracy was 94.76% ± 1.26). This is likely due to standardization, as a low amount of variability in image properties reduces noise in the feature set, making individual trees more robust. Because of this, 200 trees were selected for experiments that used standardized data, as this yielded optimal results in similar work [21].

The model trained on original, unstandardized images achieved a much lower classification accuracy on average (77.66% ± 1.37), with a maximum accuracy of 79.56% using 300 trees. Similar to the previous experiment, this is likely due to the fact that the classifier cannot generalize to the large range of variability in MC datasets that is present in the original, unstandardized data. For the next parameter optimization experiment (number of training examples), 300 trees were used for the experiments on original data.

3.1.3. Number of training examples

The number of training examples was also analyzed, as there is trade-off between classification accuracy and overfitting, as well as computational complexity. It is known that increased complexity of a classifier can eventually lead to an increase in error as well [39]. A range of values were used, from 100 to 500,000. The number of examples with the greatest accuracy was implemented in the final classifier. As this was the last classifier parameter to be optimized, sensitivity and specificity were also analyzed. In Fig. 5, the optimal sensitivity and specificity were achieved with 150,000 examples (while accuracy was saturated for all other values), and so this parameter was selected for the final classifier. This number will likely yield good generalization across datasets without overfitting.

As seen in Fig. 5, the performance of the original image-trained classifier was compared the performance achieved using standardized images. Although the accuracy and specificity are comparable, the classifier trained on original images yielded a much lower sensitivity rate, which indicates a low rate of true positives. The superior

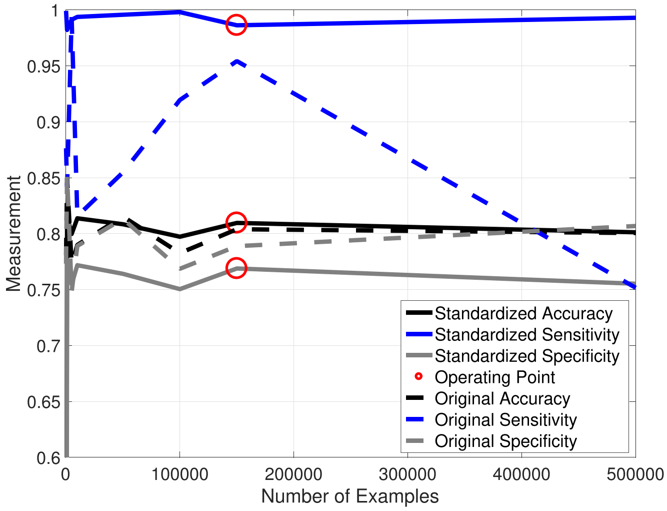


Fig. 5. Optimization of the number of training examples for the RF classifier.

performance of classifiers trained on standardized data for all optimization steps demonstrates that thorough standardization of image datasets can greatly improve the accuracy and robustness of subsequent analysis.

3.2. Segmentation experiments

In this section, the segmentation results for standardized and original data, as a function of feature type, scanner vendor and disease will be explored. Example brain extractions across datasets are shown in Fig. 6, where the top row shows the original images with the manual segmentation outlines in red, and the bottom row has the resultant extracted brains for the corresponding image slice. The example images have varying disease levels (WML burdens, enlarged ventricles, atrophy), intensity ranges and sizes, acquired across three different scanner vendors. Despite the large variability in the data, the brain has been robustly extracted demonstrating the ability of the method to adapt to a wide variety, diverse set of images.

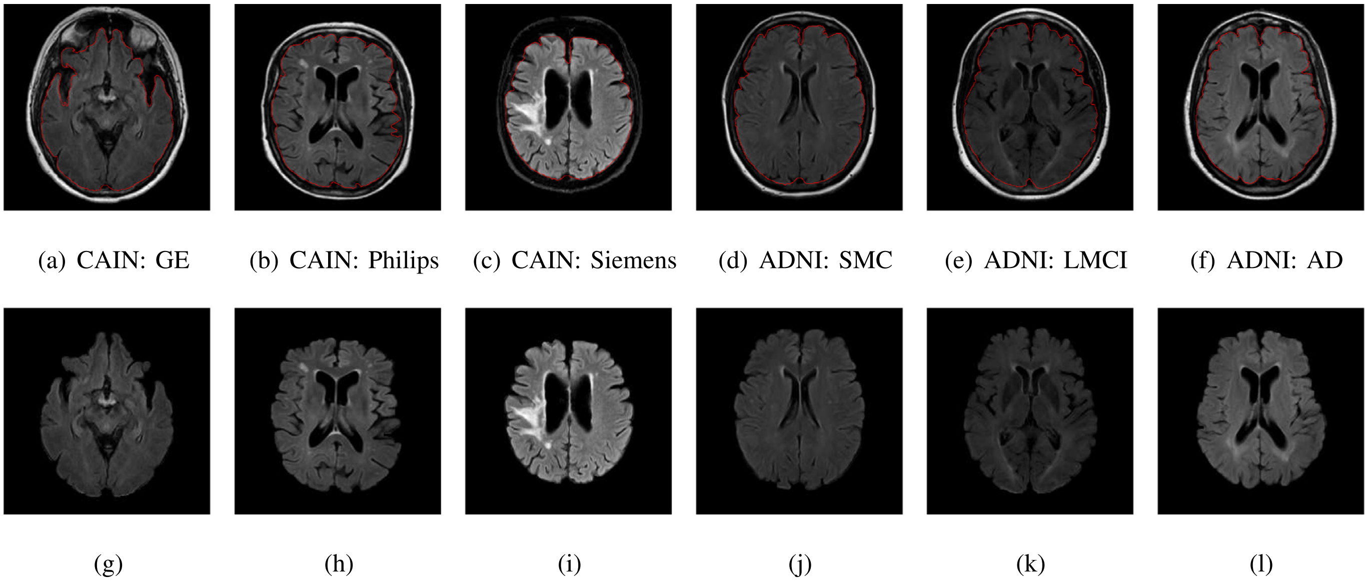
Several experiments were conducted to validate performance in terms of segmentation accuracy. Data splits can be found in Table III.

For all experiments the data for training and validation were carefully selected to ensure that the optimized classifier was validated using the unseen, validation sets. In the first experiment "CAIN (Vascular Disease) Model", the 108 CAIN image volumes that were used to optimize classification parameters were used for training, while the reserved 27 CAIN and 21 ADNI volumes were utilized for segmentation validation experiments for original and standardized data separately. These tests examine the model's performance trained using a multicentre, vascular disease dataset. A similar experiment was conducted for the "ADNI (Dementia Disease) Model", where the ADNI dataset was used to generate the model, and the reserved CAIN data was used for validation to investigate the effects of training solely on multicentre, dementia disease cases. The "Effect of Scanner Vendor" experiment was based on the training dataset (108 CAIN image volumes), where 36 CAIN volumes from a single scanner type was used to develop the model, and the remaining 72 image volumes from the two other (unseen) scanner vendors (36 per scanner type) were used to test the generalization capability to other scanners. In these cases, the segmentation performance as a function of scanner type (GE, Philips, Siemens) was observed. The last experiment, "Feature Effects", examines the effect of each of the feature groups, using the original 108 CAIN volumes for training, and the held out 27 CAIN volumes for validation. The results are shown and discussed for each experiment and effect separately below.

3.2.1. Generalization and the effects of standardization

Two classifiers were generated to analyze the effect of standardization on segmentation performance: one trained using original data and one trained with standardized data. Both classifiers were constructed using the optimized parameters determined by the Classifier Optimization Experiments in Section 3.1. The experiments were conducted using both CAIN and ADNI training data separately (108 and 21 training volumes, respectively). Classifiers were tested on unseen datasets (CAIN and ADNI, with 27 and 21 testing image volumes, respectively), providing insight into the framework's ability to generalize to different diseases in multicentre data. Table IV contains a summary of performance metrics for these segmentation experiments, and graphically as boxplots for each metric and method in Fig. 8.

First, the results of standardized versus original images are compared. Over all metrics and models, the standardized data yielded better results as compared to the unstandardized data. The *DSC*,


(a) CAIN: GE (b) CAIN: Philips (c) CAIN: Siemens (d) ADNI: SMC (e) ADNI: LMCI (f) ADNI: AD

(g) (h) (i) (j) (k) (l)

Fig. 6. Sample segmentation results across datasets for different scanners and disease classifications. For each example, the original image and groundtruth outline is shown, followed by the automated segmentation.

Table IV
Summary of segmentation metrics across MC and multi-disease datasets.

Data type	Exp.	Training data	Testing data	DSC	HD	Sensitivity	Specificity	EF
	CAIN Model	CAIN (108)	CAIN (27)	77.68 ± 22.5	5.26 ± 8.38	84.1 ± 24.5	98.6 ± 0.77	7.8 ± 4.4
Original	CAIN Model	CAIN (108)	ADNI (21)	75.1 ± 17.4	4.62 ± 4.44	90.0 ± 20.1	96.6 ± 1.9	19.6 ± 11.6
	ADNI Model	ADNI (21)	CAIN (27)	65.2 ± 29.9	8.0 ± 8.9	68.8 ± 32.2	99.3 ± 0.51	4.4 ± 3.43
	CAIN Model	CAIN (108)	CAIN (27)	91 ± 1.52	1.11 ± 0.8	96.9 ± 2.1	98.4 ± 0.62	6.55 ± 3.1
Standardized	CAIN Model	CAIN (108)	ADNI (21)	86.2 ± 5.4	3.71 ± 2.83	95.5 ± 3.1	97.5 ± 1.6	11.1 ± 6.46
	ADNI Model	ADNI (21)	CAIN (27)	86.2 ± 5	2.07 ± 1.29	89.5 ± 6.7	99.1 ± 0.53	3.7 ± 2.5

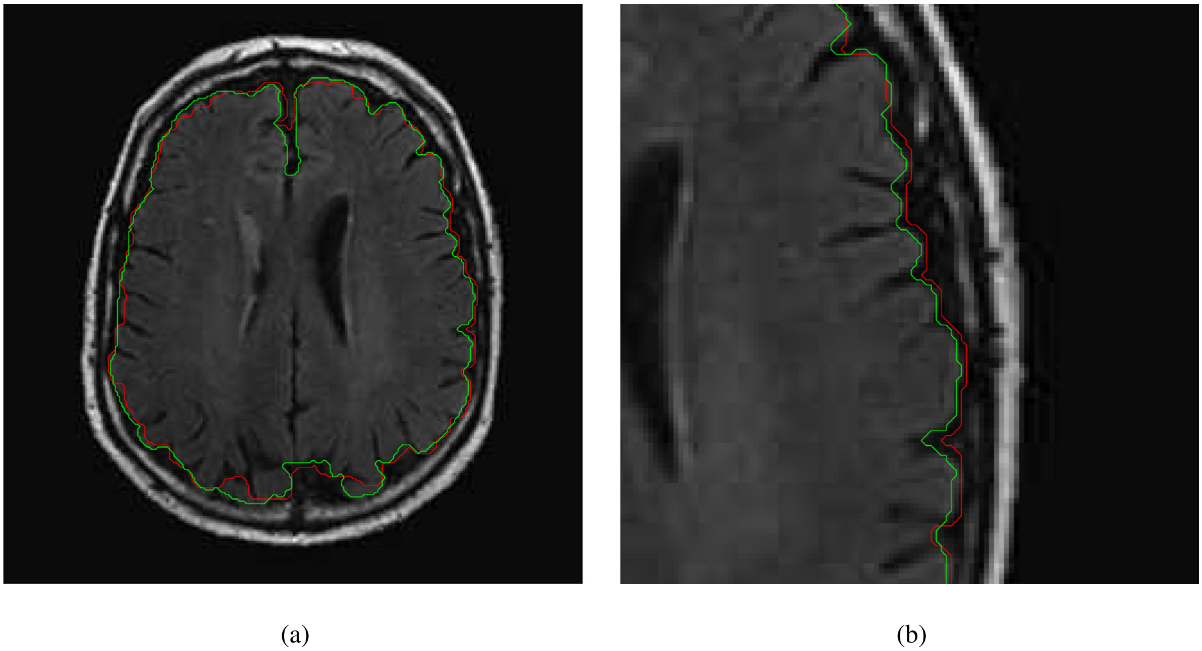


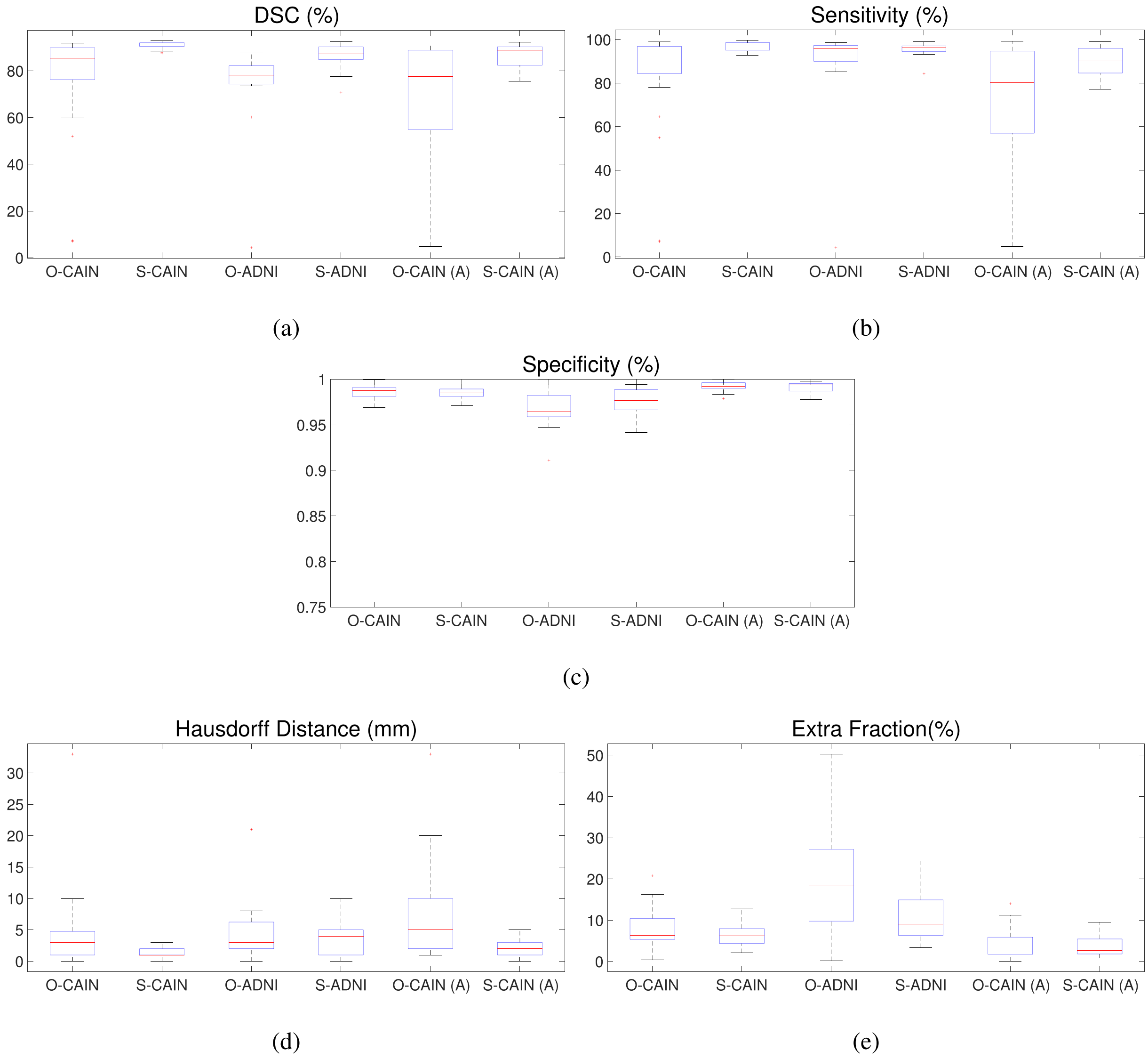
Fig. 7. Example segmentations from the CAIN dataset, where the green line represents the manual groundtruth delineation, and the red line is the automated segmentation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sensitivity, and specificity were all increased using standardized data, indicating better agreement with the groundtruth segmentations. In the case of *HD*, the smaller distance achieved by the standardized data demonstrates that the fine details of the sulci were better localized than with the original data, and a lowered *EF* measurement indicates that the *FP* rate was reduced. Fig. 7 demonstrates this point further, as it shows the differences between outlines of the groundtruth and automated segmentations. As can be seen, the automated outline follows the sulci closely. Additionally, the standard deviation over all tests is much lower with the standardized data, indicating more reliable and consistent segmentations. This is important since this indicates that the performance on new data will be reproducible. Due to the variability in image properties present in the original data, the classifier trained on original images were not as robust and the RF classifier was not able to generalize as well. The improvement in segmentation accuracy following standardization implies that standardization reduced the variability in images that previously had a negative effect on the RF classifier. This is an indication that standardization is beneficial to automated algorithms for analysis; as MC variability is mitigated, the algorithms more robust and accurate.

Considering the standardized results, for the different diseases and models, the CAIN model, tested on the CAIN validation set yielded an impressive *DSC* of 91% ± 1.52 across different scanner types and imaging acquisition parameters. When using this same model and testing with the ADNI data, the results are slightly reduced to *DSC* of 86.2% ± 5.4. Although the results are still high, the slight drop in performance can be attributed to perhaps differences in brain morphology of patients with dementia disease (compared to those with vascular disease). Vascular disease can be associated with strokes, and large lesion burdens, whereas dementia may be characterized by brain atrophy with varying lesion loads. Classifiers trained on the ADNI dataset generated similar results with a *DSC* of 86.2% ± 5, and the slight difference in performance could be attributed to the same reasoning as above. Also, due to the size of the ADNI training set being 20% of the size of the CAIN training set (21 images versus 108 images), there were less examples for training, which could have also affected results. Overall, for a variety of scanner types, centres and disease, the results on standardized data are extremely promising for multicentre, multi-disease FLAIR MRI data.

3.2.2. Effect of scanner vendor

In these sets of experiments, the effects of training using one scanner vendor's data, and testing on the other two vendor's data is analyzed. The classifier was trained on a dataset of 36 image volumes from each vendor, and tested on the remaining two (36 image volumes each). The three vendors under consideration were GE, Siemens and Philips. This experiment was conducted using both original and standardized data. Due to standardization, the classifier should yield better segmentation



(a) (b)

(c)

(d) (e)

Fig. 8. Overall segmentation metrics across multiple datasets for Experiments 4 and 5. *O* stands for original, non-standardized data, and *S* indicates standardized data; *CAIN* and *ADNI* represent the validation datasets, respectively, and *(A)* represents the experiment where ADNI was used as the training set.

accuracy using standardized images, as there should now be little variability in characteristics between images acquired on different scanners. The effect of scanner vendor on segmentation accuracy is graphically shown in Fig. 9. For each of the scanner vendor-dependent trained classifiers, several observations can be made and these are highlighted here for each model developed.

GE-trained classifier:

- On the original, unstandardized data model, the *DSC* and sensitivity values were low and/or highly variable, and the HD was lower and highly variable also when tested on original data from Philips and Siemens scanners indicating poor generalization;
- For the standardized data, the *DSC* and sensitivity were increased and HD was reduced for both Philips and Siemens data and the variability in these metrics were significantly reduced, indicating better segmentations and generalization, as well as consistency and reliability;
- The specificity decreased slightly for the standardized Siemens data (TN), which likely came at the cost of increased sensitivity (TP); and
- *HD* and *EF* both showed improvements using standardized images, indicating more accurate segmentations around the brain boundary, as the distance between the segmentation borders and FP rate were reduced.

Philips-trained classifier:

- The classifier could not generalize at all to the original GE data, and yielded dismal results due to over-segmentation (*DSC* and sensitivity were less than 10%, and specificity was 100%) indicating generalization failure;
- The classifier struggled with original Siemens data (*DSC* and sensitivity less than 80%) but this was improved with standardization as the values were both increased, and the variability was reduced thereby showing more reliability on standardized data;
- Using the standardized data, the *EF* increased slightly in both Siemens and GE data due to an increase in sensitivity (TP), which introduced some false positives (as measured by *EF*); however at a rate of 3%, this increase is appropriate given a large increase in sensitivity (TP)
- On standardized data, the *HD* metric was improved as compared to the model generated with original data in both other scanners indicating more detailed and refined segmentations using the standardized data.

Siemens-trained classifier:

- Standardization increased classifier performance on both scanner

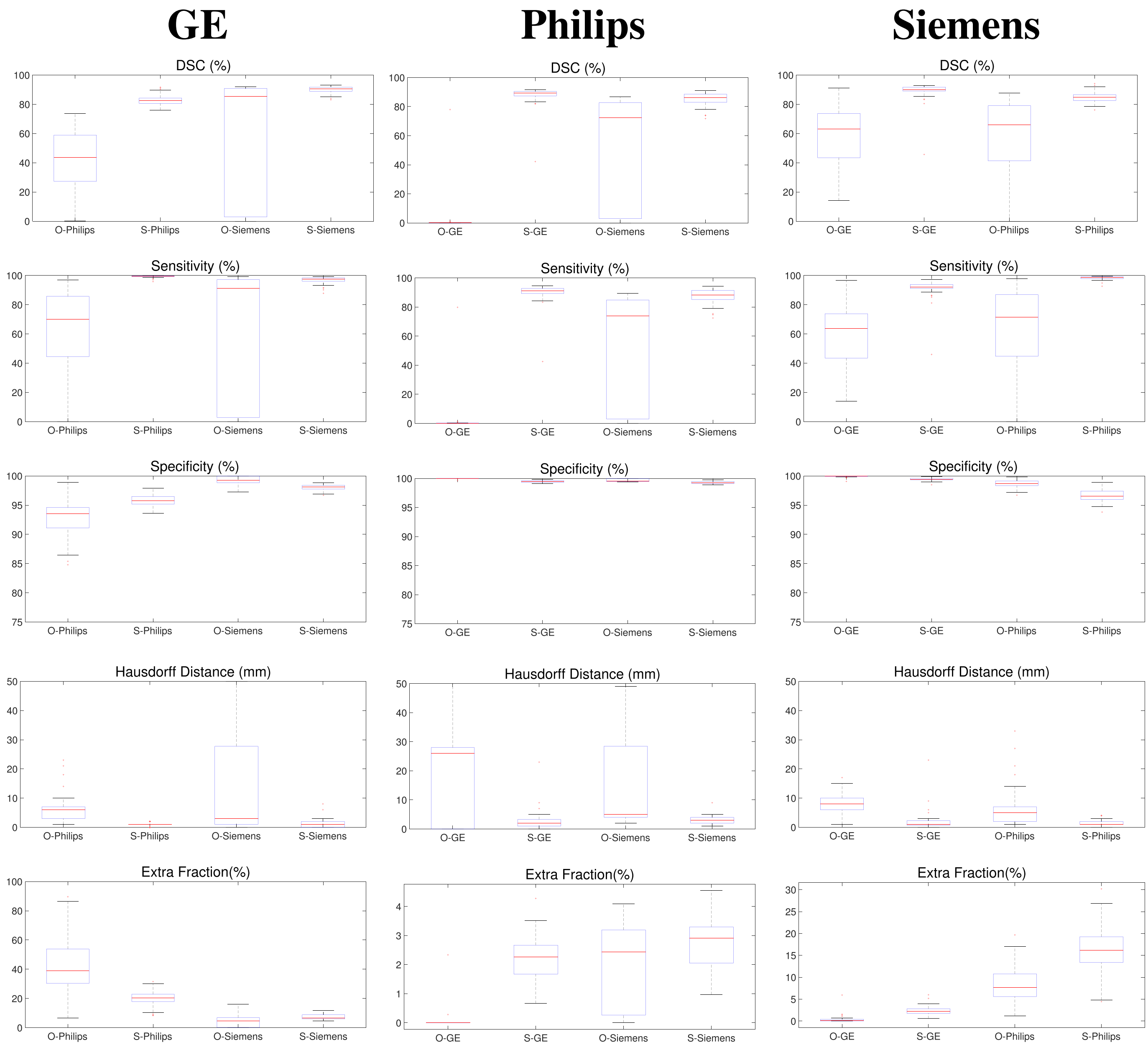


Fig. 9. Effect of scanner vendor on segmentation accuracy. The first column was trained on GE images; the second on Philips, and the third, Siemens. (O) indicates original data, and (S) indicates standardized data.

vendors (*DSC*, sensitivity, and *HD* improved), and the variability of each of these metrics was greatly reduced indicating more consistent and reproducible segmentation results; and

- Even with standardized data, the Siemens classifier could not generalize completely to the Philips data, as shown by the increase of FP rate (*EF* of 17%).

Over all models, testing on data generated from unseen scanner vendors yielded poorer performance on the original images as compared to the standardized images, demonstrating the effectiveness of the standardization methodology. A common result across these experiments is that there was generally an improvement in all metrics with standardization indicating higher quality brain segmentations, except for *EF* in some cases. Because standardization increased sensitivity, which is a measure of the TP rate, this likely increased the rate of positive cases in general, which lead to the small increase in *EF*. This small increase in FP is a trade-off for the significant increases to the TP rate. However, the variability over all metrics were greatly reduced, thereby indicating more reliable and consistent results.

These experiments highlight how segmentation using machine learning models can greatly be effected by the type of scanning device used to acquire FLAIR MRI in multi-institutional datasets. Despite the varying imaging parameters that can create intensity and contrast differences for the same sequence, this scanner-vendor dependent effect is likely due to more than this. It is probably caused by the differences in hardware, software and reconstruction algorithms used to create the images, which are used by companies as a competitive advantage, and therefore are proprietary. Depending on the way the images are reconstructed, there can be highly variable image properties and characteristics, including contrast, spatial correlation in noise, imaging artifacts, and more [28]. These results demonstrate that to create consistency in the multi-institutional FLAIR MRI datasets acquired by different vending hardware and software solutions, it is paramount that MC variability be reduced to improve stability and generalizable of

Table V
Segmentation results on original and standardized CAIN images using intensity-, spatial-, and gradient-based features.

		DSC (%)	HD (mm)	Sensitivity (%)	Specificity (%)	Extra fraction (%)
Original	Intensity	74.6 ± 21.3	5.44 ± 8.40	78.6 ± 22.8	99.1 ± 0.8	5.0 ± 5.1
	Spatial	64.4 ± 19.6	10.9 ± 4.9	70.0 ± 20.4	99.3 ± 0.52	3.83 ± 3.0
	Gradient	77.9 ± 22.3	4.78 ± 8.40	84.3 ± 24.6	98.6 ± 0.80	7.67 ± 4.3
Standardized	Intensity	84.5 ± 5.7	**1.37 ± 2.5**	86.9 ± 7.3	99.4 ± 0.53	2.7 ± 2.5
	Spatial	58.6 ± 5.5	7.0 ± 1.86	59.8 ± 5.9	**99.6 ± 0.40**	**2.0 ± 2.3**
	Gradient	**90.3 ± 2.4**	1.44 ± 2.5	**96.3 ± 3.0**	98.4 ± 0.7	6.7 ± 3.2

brain segmentation classifiers.

3.2.3. *Effect of feature type*

There were three types of features investigated in this work, including intensity-based, spatial-based and gradient-based features. To determine the importance of each of these three feature groups, three classifiers were trained separately with each of the feature types, for standardized and unstandardized data. Analysis of the effects of these features will shed light into which features are the most important for classification, as well as how much image standardization played a role in automated segmentation across diverse datasets.

In this experiment, the effects of different feature types used for classification was investigated on both original and standardized images in the CAIN dataset, and results can be found in Table V. As can been seen, standardized features yielded better results than those generated from the original images. Additionally, in both sets of experiments, the spatial features yielded the worst results, with results improving with intensity- and gradient-based features. In the standardized data, the poor performance of the spatial features can likely be attributed to suboptimal registration performance. FLAIR MRI are difficult images to register, since there is no contrast between WM and GM, which are usually the defining features in registration techniques focused on T1 or T2 MRI. However, it should also be noted that spatial-based features yielded the best specificity and *EF* scores for standardized data, while achieving the lowest scores on all other metrics. This indicates that the model tended to over-segment, which would give the best specificity and *EF* scores (which are measurements of *TN* and *FP*, respectively), while giving low scores in sensitivity (*TP* measurement), *DSC* (similarity), and *HD* (distance between the segmentations). Interestingly, on average, the gradient-based features performed the best, and slightly better than the intensity-based features, despite the standardized intensity scale. We hypothesis that the texture differences between classes (brain and non-brain) must be much more of a discriminating and consistent feature across the datasets. In some of the Siemens data, the skull can be noted to be darker (see Fig. 6 (c)), which may be due to a fat suppression sequence, and could have contributed to a reduction in performance of just the intensity-based features. Gradients, on the other hand, consider relative differences in intensity, or contrasts between structures, which must be more a consistent feature across datasets, for standardized data.

4. Discussion

FLAIR MRI has been gaining popularity for the analysis of neurodegenerative disease, which is primarily due to the superior visualization of WML with this sequence. However, there is little work on the development of algorithms solely for this modality. To address this lapse, this work presents a novel approach to brain extraction for multicentre (MC) FLAIR MRI. The framework was validated on MC and multi-disease datasets, which included images from CAIN (vascular disease) and ADNI (dementia disease). By creating a framework solely for FLAIR, the need for co-registration to T1- and T2-weighted images has been eliminated, therefore reducing image acquisition costs and errors introduced due to registration. This framework also provides a gateway for the robust application of existing WML analysis algorithms to large, MC datasets, from which neuroimaging biomarkers can be correlated with clinical outcomes. This work represents one of the first approaches to segmentation of MC and multi-disease neurological FLAIR MRI using solely the FLAIR modality.

One of the major novelties of this approach surrounds the pre-processing of images using standardization that reduces variability in MC datasets allowing for intuitive, simple features sets to produce robust and accurate segmentations. Results demonstrate that the standardization of images is extremely beneficial for the segmentation of these diverse datasets, and also that the framework generalizes well to other diseases from which it was trained.

The use of standardization to improve segmentation accuracy has been previously mentioned, but not addressed in machine learning approaches to brain segmentation [21]. Other works address the fact that scanner manufacturers have an effect on accuracy [23], and proposed that models should be trained on MC data for increased robustness. This work addressed both of these concerns. By standardizing the images, segmentation accuracy is increased and the results were much more consistent and reliable. In [21], a RF was used to assign probabilities to pixels along the brain boundary, followed by the application of a probabilistic generative model has used to create the final classification. In [23], a convolutional neural network was used, and this model is generally regarded as being very complex [45]. In this work, an RF was constructed to be minimally complex to increase generalization accuracy, and simple post-processing was conducted using mathematical morphology. This is a testament to the fact that standardization of images can greatly simplify segmentation algorithms, while maintaining, or increasing, segmentation accuracy and robustness.

The framework was tested across diverse datasets in order to validate the algorithm's robustness and accuracy when faced with a large amount of variability in image properties. Results showed that image standardization is extremely beneficial to processing of MC images, as demonstrated by the segmentation accuracy achieved using original versus standardized data. Using the CAIN dataset, standardization had a substantial effect on *DSC* (77.7% versus 91% for original and standardized data, respectively), sensitivity (84.1% versus 96.9%), *HD* (5.26 versus 1.11), and *EF* (7.8% versus 6.5%). This indicates that the false negative rate was greatly reduced using standardized data, therefore increasing the accuracy of the overall segmentation. The specificity was also increased (84.1% versus 96.9%), which indicates that the standardized model increased the *TN* rate. When applied to databases containing other diseases, segmentation accuracy was maintained (*DSC* of 91% and 86.2% for ADNI), indicating that the framework was robust to the morphological differences in these images due to varying disease manifestations.

It has been noted that scanner vendors may have an effect on segmentation accuracy due to differences in acquisition artifacts and reconstruction algorithms [29]. In this work, the classifier was trained on images acquired by one scanner manufacturer, and applied to others in order to analyze these differences. In general, it was found that prior to standardization, all vendors struggled to generalize to the others; but that they were successful following standardization. In particular, it was found that Philips data could not generalize to GE data, which indicates that there are subtle characteristics in these images that may not be apparent to a human observer, but can greatly confound automated algorithms. Across all cases, standardization increased segmentation accuracy to acceptable levels; for example, *DSC* was increased from 44% to 84% when applying GE data to Philips. In a more radical case, standardization of Philips data increased the *DSC* from 2% to 88% for GE data. This truly highlights the necessity and utility of standardization for the analysis of MC data.

Three intuitive feature groups were used, based on the intensity, spatial locations and gradient of positive (brain) and negative (nonbrain) voxels. From these experiments, it was clear that standardization is beneficial to the segmentation task, as the performance metrics improved for all the feature groups for standardized data, in comparison to the original. Interestingly, using the standardized data, the strongest feature group that yielded the highest segmentation performance on their own were the gradient features. This could indicate that relative differences in intensities are more discriminating than absolute intensity, since intensity-based features yielded lower segmentation results than that achieved by the gradient features.

In all, the framework requires approximately 7 min of single-threaded run time on an i7 intel processor with 16 GB of RAM for standardization, feature extraction, and classification; however, nearly half of this time is spent performing image registration. Future work will include optimization of image registration to streamline the algorithm or the investigation of brain extraction that excludes the spatial coordinates as a feature. For reference, in [23], a convolutional neural network was implemented using Graphics Processing Units (GPUs), and computation took an average of 40 to 60 s. In a clinical setting, where GPU computation may not be available, these same calculations could take up to 12 times longer on conventional CPUs [46]. In addition, the authors plan to apply this framework to large databases of FLAIR MRI; to the knowledge of the authors, little work has been done in quantifying disease in FLAIR images exclusively, with the exception of segmentation of WML. With this framework, novel imaging biomarkers from FLAIR MRI can be discovered hopefully yielding new insights into neurodegenerative disease.

5. Conclusion

This work proposes a fully-automated, whole volume brain extraction methodology for multicentre (MC) FLAIR MRI. The pipeline consists of image standardization, which reduces acquisition noise and artifacts, standardizes the intensity scale and spatially normalizes voxel coordinates of brain tissue across datasets, followed by feature extraction by intensity-, gradient- or spatial location-based features and classification with a random forest. Using a novel sampling strategy, the classifier is trained using selective voxels from the positive and negative classes, for both original and standardized data separately. Two multicentre, international datasets are used to optimize classifier parameters and analyze segmentation performance. The first dataset is from the Canadian Atherosclerotic Imaging Network (CAIN) and contains images from subjects with vascular disease and 135 volumes (roughly 7000 image slices) with ground truths were used for training, testing and validation. The second dataset used is from Alzheimer's Disease Neuroimaging Initiative (ADNI), and 21 volumes (approximately 1000 image slices) were used for training and validation. Experiments were conducted to optimize classifier parameters, and determine segmentation accuracy as a function of data type (standardization vs. original data), scanner type, feature types and disease. Across all experiments, the segmentation results demonstrate that standardization significantly improves the performance across instructions, features, scanner vendors and disease types, thereby proving the necessity and efficiency of the proposed standardization methodology for multicentre (MC) FLAIR MRI. With this work, the need for registration to T1- and T2-weighted images is eliminated, reducing acquisition costs and dependence on registration which introduces error. Future work will involve the application of this framework to large, MC datasets to investigate novel neuroimaging biomarkers from FLAIR MRI related to neurodegenerative diseases.

Acknowledgments

- We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through the NSERC Discovery Grant program.
- The Canadian Atherosclerosis Imaging Network (CAIN) was established through funding from a Canadian Institutes of Health Research Team Grant for Clinical Research Initiatives (CIHR-CRI 88057). Funding for the infrastructure was received from the Canada Foundation for Innovation (CFI-CAIN 20099), with matching funds provided by the governments of Alberta, Ontario, and Quebec.
- Data collection and sharing for this project was partially funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defence award number W81XWH-12-2-0012).

References

[1] Staals J, Makin S D, Doubal F N, Dennis M S, Wardlaw J M. Stroke subtype, vascular risk factors, and total MRI brain small-vessel disease burden. Neurology 2014;83(14):1228–34.

[2] Wardlaw J M, Valdes Hernandez M C, Munoz-Maniega S. What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. J Am Heart Assoc 2015;4(6):001140.

[3] Wardlaw J, Smith E, Biessels G, Cordonnier C, Fazekas F, Frayne R, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Nuerol 2013;12:822–38: https://www.ncbi.nlm.nih.gov/pubmed/23867200.

[4] Olsson E, Klasson N, Berge J, Eckerstrm C, Edman A, Malmgren H, et al. White matter lesion assessment in patients with cognitive impairment and healthy controls: reliability comparisons between visual rating, a manual, and an automatic volumetrical MRI method - the gothenburg MCI study. J Aging Res 2013;12:10: https://doi.org/10.1155/2013/198471.

[5] Eichinger P, Schön S, Pongratz V, Wiestler H, Zhang H, Bussas M, et al. Accuracy of unenhanced MRI in the detection of new brain lesions in multiple sclerosis. J Radiological Soc North Amer 2019:10: https://doi.org/10.1148/radiol.2019181568.

[6] Schmidt P, Gaser C, Arsic M, Buck D, Frschler A, Berthele A, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. NeuroImage Feb. 2012;59(4):3774–83.

[7] van Dijk E J, Breteler M M, Schmidt R, Berger K, Nilsson L -G, Oudkerk M, et al. The association between blood pressure, hypertension, and cerebral white matter lesions: cardiovascular determinants of dementia study. Hypertension Nov. 2004;44(5):625–30.

[8] Gons RAR, van Norden AGW, de Laat KF, van Oudheusden LJB, van Uden IWM, Zwiers MP, et al. Cigarette smoking is associated with reduced microstructural integrity of cerebral white matter. Brain Jul. 2011;134(7):2116–24.

[9] Wardlaw J M, Allerhand M, Doubal F N, Hernandez M V, Morris Z, Gow A J, et al. Vascular risk factors, large-artery atheroma, and brain white matter hyperintensities. Neurology 2014;82(15):1331–8.

[10] Debette S, Seshadri S, Beiser A, Au R, Himali J J, Palumbo C, et al. Midlife vascular risk factor exposure accelerates structural brain aging and cognitive decline. Neurology 2011;77(5):461–8.

[11] Bilello M, Suri N, Krejza J, Woo J H, Bagley L J, Mamourian A C, et al. An approach to comparing accuracies of two flair MR sequences in the detection of multiple sclerosis lesions in the brain in the absence of gold standard. Acad Radiol 2010;17(6):686–95.

[12] Wilke M, de Haan B, Juenger H, Karnath H -O. Manual, semi-automated, and automated delineation of chronic brain lesions: a comparison of methods. NeuroImage 2011;56(4):2038–46.

[13] Malloy P, Correia S, Stebbins G, Laidlaw D. Neuroimaging of white matter in aging and dementia. Clin Neuropsychol 2007;21:73–109.

[14] Altaf N, Morgan P S, Moody A, MacSweeney S T, Gladman J R, Auer D P. Brain white matter hyperintensities are associated with carotid intraplaque hemorrhage

1. Radiology 2008;248(1):202–9.

[15] Altaf N, Daniels L, Morgan P, Lower J, Gladman J, MacSweeney S, et al. Cerebral white matter hyperintense lesions are associated with unstable carotid plaques. Eur J Vacs and Endovasc Surg 2006;31:8–13.

[16] de Groot M, Verhaaren BF, de Boer R, Klein S, Hofman A, van der Lugt A, et al. Changes in normal-appearing white matter precede development of white matter lesions. Stroke 2013;44(4):1037.

[17] Soltanian-Zadeh H, Peck D. Feature space analysis: effects of MRI protocols. Med Phys 2001/11/;28(11):2344–51.

[18] Garcia-Lorenzo D, Francis S, Narayanan S, Arnold D L, Collins D L. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med Image Anal Jan. 2013;17(1):1–18.

[19] Khademi A, Venetsanopoulos A, Moody A R. Robust white matter lesion segmentation in FLAIR MRI. IEEE Trans Biomed Eng Mar. 2012;59(3):860–71.

[20] Khademi A, Moody A R. Multiscale partial volume averaging estimation for segmentation of WML in FLAIR MRI. 2015. p. 568–71.

[21] Iglesias J E, Cheng-Yi Liu, Thompson P M, Zhuowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans Med Imaging 2011;30(9):1617–34.

[22] Smith S M. Fast robust automated brain extraction. Hum Brain Mapp Nov. 2002;17(3):143–55.

[23] Kleesiek J, Urban G, Hubert A, Schwarz D, Maier-Hein K, Bendszus M, et al. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. NeuroImage Apr. 2016;129:460–9.

[24] DiGregorio J, Moody A, Khademi A. Brain extraction methods for neurological FLAIR MRI. The 16th Annual Imaging Network of Ontario IMNO Symposium. 2018.

[25] Zhong Y, Qi S, Kang Y, Feng W, Haacke E M. Automatic skull stripping in brain MRI based on local moment of inertia structure tensor. Information and Automation (ICIA), 2012 International Conference on. IEEE; 2012. p. 437–40.

[26] de Boer R, van der Lijn F, Vrooman H, Vernooij M, Ikram M, Breteler M, et al. Automatic segmentation of brain tissue and white matter lesions in MRI. International Symposium on Biomedical Imaging (ISBI. 2007. p. 652–5.

[27] Hah T T T, Kim J Y, Choi S H. White matter hyperintensities extraction based T2-FLAIR MRI using non-local means filter and nearest neighbor algorithm. IT Convergence and Security (ICITCS), 2014 International Conference on. IEEE; 2014. p. 1–4.

[28] Khademi A, Hosseinzadeh D, Venetsanopoulos A, Moody A R. Nonparametric statistical tests for exploration of correlation and nonstationarity in images. International Conference on Digital Signal Processing (DSP). 2009. p. 1–6.

[29] Fennema-Notestine C, Ozyurt I B, Clark C P, Morris S, Bischoff-Grethe A, Bondi M W, et al. Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. Hum Brain Mapp 2006;27(2):99–113.

[30] Reiche B, Moody A R, Khademi A. Pathology-preserving intensity standardization framework for multi-institutional FLAIR MRI datasets. Magn Reson Imaging 2019. Minor Revision, April.

[31] A Khademi, B Reiche, G Arezza. Method and system for standardized processing of MR images; 2018. PCT Patent Application No. PCT/CA2018/051606, Filed:. Dec. 14, 2018

[32] Breiman L. Random forests. Mach Learn 2001;4(1):5–32.

[33] Tardif J, Spence J, Heinonen T, Moody A, Pressacco J, Frayne R, et al. Atherosclerosis imaging and the Canadian Atherosclerosis Imaging Network. Can J Cardiol 2013:297–303.

[34] Aisen P S, Petersen R C, Donohue M, Weiner M W. Alzheimer's disease neuroimaging initiative 2 clinical core: progress and plans. Alzheimers Dement 2015;11(7):734–9: http://linkinghub.elsevier.com/retrieve/pii/S1552526015001727.

[35] Zhong Y, Utriainen D, Wang Y, Kang Y, Haacke E M. Automated white matter hyperintensity detection in multiple sclerosis using 3D T2 FLAIR. Int J Biomed Imaging 2014;2014:1–7.

[36] Kullback S, Leibler R. On information and sufficiency. Ann Math Stat 1951;22:79–86.

[37] Thirion J -P. Image matching as a diffusion process: an analogy with Maxwell's demons. Med Image Anal 1998;2:243–60.

[38] Khademi A, Moody A R, Venetsanopoulos A. Generalized partial volume averaging estimation for cerebral MRI. J Med Image Anal 2014;1(1).

[39] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Springer series in statistics. 1. Berlin: Springer; 2001.

[40] Zou K H, Warfield S K, Bharatha A, Tempany C M, Kaus M R, Haker S J, et al. Statistical validation of image segmentation quality based on a spatial overlap index 1: scientific reports. Acad. Radiol. 2004;11(2):178–89.

[41] Beauchemin M, Thomson K P, Edwards G. On the Hausdorff distance used for the evaluation of segmentation results. Can J Remote Sens 1998;24(1):3–8.

[42] Stokking R, Vincken K L, Viergever M A. Automatic morphology-based brain segmentation (MBRASE) from MRI-T1 data. NeuroImage Dec. 2000;12(6):726–38.

[43] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1226–38.

[44] Segal M R. Machine learning benchmarks and random forest regression. Center for Bioinformatics & Molecular Biostatistics 2004.

[45] He K, Sun J. Convolutional neural networks at constrained time cost. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. p. 5353–60.

[46] Jia Y. Learning semantic image representations at a large scale Ph.D. dissertation Berkeley: University of California; 2014: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-93.html.