The method used is slightly adapted from the framework described in [2]. It consists in two steps: model selection and post-processing.

I. Pre-processing

As part of the preprocessing, the FLAIR image is registered to the T1 image. A label fusion technique, known as GIF (Geodesic Information Flows) [1], is used for three purposes:

a skull-stripping of the image
b patient-specific priors obtained as the average of the registered segmentations
c brain parcellation.

In order to model the bias field as additive, multivariate image intensities $\mathbf{y}_n$ with n representing an individual voxel are log-transformed and normalised using the skull-stripped mask.

II. Model selection

A. Gaussian mixture model

The three-level hierarchy of the Gaussian mixture model can be described as follows

Level 1 At the first level (Level 1), indexed by l, the model is robustly divided into two density functions I and O, that correspond respectively to the inlier part (I), modelling the healthy tissues, and to the outlier part (O), related to the unexpected observations, such that

$$f(\mathbf{y}_n|\Xi_\mathbf{K}) = b_{nI} \cdot I(\mathbf{y}_n|\Xi_\mathbf{K}) + b_{nO} \cdot O(\mathbf{y}_n|\Xi_\mathbf{K}),$$

with $b_{nI} + b_{nO} = 1$ and $b_{nl} \geq 0$, and introducing $\mathbf{b}$ the vector formed with these parameters. Initially, flat priors are used to differentiate inliers and outliers.

Level 2 The second level (Level 2), indexed by j characterizes the anatomical tissue classes (*i.e.* if an inlier or outlier voxel belongs to WM, GM, CSF or other non-brain (NB) tissues). The number of anatomical classes J_l is considered the same for both the inlier and outlier classes since the model is built under an assumption of symmetry, simplifying $J_I = J_O = J$. The distribution is thus:

$$f(\mathbf{y}_n|\Xi_\mathbf{K}) = \sum_{l\in I,O} b_{nl} \sum_{j=1}^{J} a_{nl_j} \Phi\left(\mathbf{y}_n \middle| \Theta_{l_j}\right),$$

where b_{nl}, a_{nl_j} and $\Phi\left(\mathbf{y}_n \middle| \Theta_{l_j}\right)$ are respectively the local mixing probabilities of l, the local *a priori* probability for l_j and the likelihood of the data at voxel n for the tissue class l_j. The local *a priori* probabilities for tissue classes j are obtained from the statistical atlases such that a_{nl_j} satisfies $\sum_{j=1}^{J} a_{nl_j} = 1$ and $a_{nl_j} \geq 0$, $\forall l \in \{I,O\}$.

Level 3 The third level (Level 3), indexed by k, characterizes the multiple intensity clusters of each inlier or outlier tissue class and models the acquisition noise in the observations from the expected biological mean signal. Each anatomical class density distribution is modelled by a mixture of multiple components with distribution $\mathscr{M}$, that can be Gaussian ($\mathscr{G}$) and/or uniform ($\mathscr{U}$) such that

$$\begin{aligned}\Phi\left(\mathbf{y}_n \middle| \Theta_{l_j}\right) &= \sum_{k=1}^{K_{l_j}+1} w_{l_{j_k}} \mathscr{M}\left(\mathbf{y}_n \middle| \theta_{l_{j_k}}\right) \\ &= \sum_{k=1}^{K_{l_j}} w_{l_{j_k}} \mathscr{G}\left(\mathbf{y}_n \middle| \theta_{l_{j_k}}\right) + w_{l_{j_{K_{l_j}+1}}} \mathscr{U}_{l_j}\end{aligned}$$

where K_{l_j} is the number of Gaussian components in class l_j, $w_{l_{j_k}}$ is the mixing proportion (≥ 0) of class l_{j_k} and $\theta_{l_{j_k}}$ are the corresponding Gaussian parameters. The uniform distribution in each class l_j is only parameterised by the mixing coefficient $w_{l_{j_{K_{l_j}+1}}}$. The mixing coefficients for class l_j are gathered in the vector $\mathbf{w}_{l_j}$, with $\mathbf{W}$ being the set of all such vectors that satisfy $\sum_{k=1}^{K_{l_j}+1} w_{l_{j_k}} = 1$, $\forall l \in \{I,O\}$ and $\forall j \in \{1,\cdots,J\}$.

Adopting the notation $\omega_{nl_{j_k}} = b_{nl} a_{nl_j} w_{l_{j_k}}$ with $\boldsymbol{\pi} = \{\mathbf{b}, \mathbf{a}, \mathbf{W}\}$ the set of *a priori* mixing weights at the different hierarchical levels and considering the observations as independent and identically distributed (*iid*), the multi-layered mixture model can finally be expressed as follows:

$$f(\mathbf{Y}|\Xi_\mathbf{K}) = \prod_{n=1}^{N} \sum_{l\in I,O} \sum_{j=1}^{J} \left[\sum_{k=1}^{K_{l_j}+1} \omega_{nl_{j_k}} \mathscr{M}\left(\mathbf{y}_n \middle| \theta_{l_{j_k}}\right) \right].$$

B. Model evolution

The number of Gaussian components needed is obtained using a split and merge strategy. Possible model modifying operations (split of a component at the third level or merge of two component) are ordered based on their likelihood to provide a successful improvement after convergence. Each operation is tested by initialising the change and optimising an expectation-maximisation (EM) algorithm on the new model. Priors over the covariance of the Gaussian components are modelled according to an Inverse Wishart distribution. After convergence, the Bayesian Inference criterion is used to check if the newly optimised model should be accepted. This step is used to enforce a balance between model fit and model complexity. The final model is obtained once no further modifying operation successfully satisfies this criterion.

C. Initialisation / Outlier atlas

Considering a mixture with eight classes separated in inliers (I) and outliers (O), an initial expectation-maximisation algorithm is run performing bias field correction and providing initial separation between inliers and outliers. Atlases of outlierness are obtained as smoothed version of the typicality

map defined by Van Leemput et al. [3]. The typicality value for a given voxel n follows the expression:

$$t_n = \sum_{j=1}^{J} p_{nj} \frac{\mathscr{G}\left(\mathbf{y}_n \middle| \theta_{I_j}\right)}{\mathscr{G}\left(\mathbf{y}_n \middle| \theta_{I_j}\right) + \dfrac{1}{\sqrt{(2\pi)^D \left|\Lambda_{I_j}\right|} \exp\left(-\dfrac{1}{2}\kappa^2\right)}}$$

where $p_{nj} = p_{nI_j} + p_{nO_j}$ and $\kappa = 3$.

III. Post-Processing

A. Candidate voxels selection

After convergence of the model selection process, a three-level Gaussian mixture model is obtained. The classes corresponding to the inlier of the WM are gathered in order to produce a distribution for intensity reference with parameters $\boldsymbol{\mu}_{IWM}$, $\boldsymbol{\Lambda}_{IWM}$. For the voxels classified as outliers, intensity comparisons are performed with respect to this reference class and a probability map of candidate lesion is obtained as such that

$$L_n = \begin{cases} p_{nO} \cdot \max\left(1, \dfrac{d_{Mahal}(y_n^{FLAIR})}{3}\right) & \text{if } y_n^{FLAIR} > \mu_{IWM}^{FLAIR} \\ 0 & otherwise \end{cases}$$

where L_n represents the lesion probability, p_{nO} the probability to be classified as outlier and d_{Mahal} is the Mahalanobis distance to the inlier white matter reference distribution.

B. False positive correction

After thresholding the candidate probability map at 0.5, connected components are extracted and their location with respect to the ventricles, the choroid plexus, septum pellucidum and cortical sheet is assessed using the brain parcellation obtained from GIF. This information is used to separate true lesion elements from spurious false positives resulting in the final segmentation.

IV. Results on training set

Evaluation of the automated segmentation results was obtained for the training set using the open source package NiftyNet (http://www.niftynet.io) and Table I summarises the results obtained for Dice score coefficient (DSC), sensitivity (Sens), average pairwise distance (AveDist) and positive predictive value (PPV) for the three scanner types. Linear relationships between reference and segmented volumesare plotted in Figure 1. The values of R^2 varied between 0.95 and 0.98.

Lastly Figure 2 present a segmentation example compared to the gold standard in one case per scanner type.

	GE	Singapore	Utrecht	Overall
DSC	56.4 [52.0 ; 72.1]	75.3 [59.3 ; 77.2]	72.6 [45.4 ; 80.1]	69.0 [51.2 ; 76.9]
Sens	68.5 [56.9 ; 76.3]	64.4 [59.3 ; 69.7]	81.7 [69.0 ; 88.5]	68.9 [58.4 ; 78.8]
AveDist	1.42 [0.57 ; 2.58]	0.68 [0.47 ; 2.02]	0.67 [0.35 ; 2.67]	0.83 [0.47 ; 2.66]
PPV	56.1 [46.2 ; 66.7]	85.4 [57.9 ; 90.7]	64.6 [37.8 ; 76]	64.8 [47.0 ; 82.4]

TABLE I: Evaluation of segmentation results for each scanner type presented under the form median [1st Quartile ; 3rd Quartile]

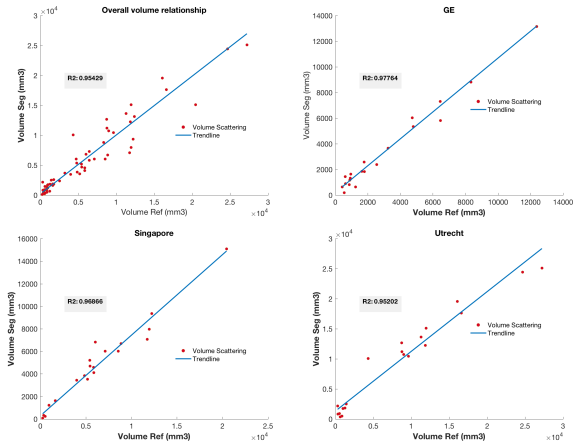

Fig. 1: Plots of the relationship between segmented and reference volumes for the different scanner types

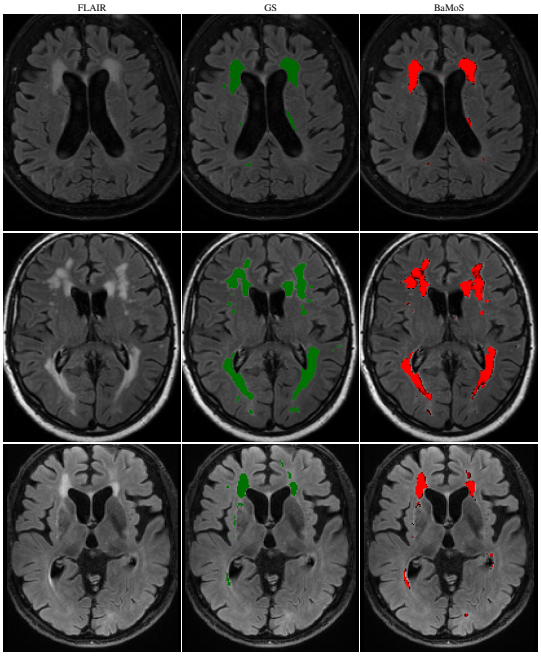

Fig. 2: Example of segmentation results (3rd column) compared to gold standard segmentations (2nd column)

References

[1] M. Jorge Cardoso, Marc Modat, Robin Wolz, Andrew Melbourne, David Cash, Daniel Rueckert, and Sébastien Ourselin. Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion. *IEEE Transactions on Medical Imaging*, 34(9):1976–1988, apr 2015.

[2] Carole Sudre, M Jorge Cardoso, Willem Bouvy, Geert Biessels, Josephine Barnes, and Sébastien Ourselin. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. *IEEE Transactions on Medical Imaging*, 34(10):2079–2102, apr 2015.

[3] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, Alan Colchester, and Paul Suetens. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Transactions on Medical Imaging*, 20(8):677–688, aug 2001.