The method proposed is using a deep-learning network architecture implemented in the NiftyNet open-source package http://www.niftynet.io and the network is trained using a generative training database using automatically segmented WMH cases Note that the challenge training set was only used for the final evaluation of the method but not at any point during the network training.

I. Network

A. Architecture

The network architecture chosen in this case is the HighResNet architecture proposed by Li et al [1]. This network is designed for 3D image segmentation using dilated convolutions and residual connections. Dilated convolutions allow for the consideration of a larger receptive field while requiring the same number of parameters as a classical convolution. Using dilated convolutions is thus beneficial to avoid overfitting. Residual connections are added to group every two convolution layers. In total 20 convolution layers are included with progressive kernel dilation (1, 2 4) every 6 convolutions. Each convolution block is composed of feature batch normalisations, ReLU activation function and the considered convolution layer. A final fully convolutional layer is applied before a softmax layer at the end of the network. The resolution of the input is maintained throughout the network. Figure 1 summarises the network architecture.

B. Loss function

In order to account for the high level of imbalance between lesion and background voxels, the Generalised Dice Loss function was used as cost function for the training [3]. For the binary segmentation problem presented here, it can be expressed as:

$$\mathrm{GDL} = 1 - 2\frac{\sum_{l=1}^{2} w_l \sum_n r_{ln} p_{ln}}{\sum_{l=1}^{2} w_l \sum_n r_{ln} + p_{ln}},$$

where r_{ln} (resp. p_{ln}) denotes the value of the reference (resp. prediction) for class l at voxel n and p_{ln}. The weight w_l is assigned the value $\frac{1}{(\sum r_{ln})^2}$, that is the squared inverse of the reference volume over the considered patch for class l.

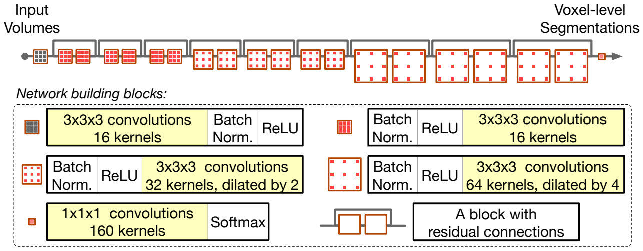

Fig. 1: Illustration of the network architecture

C. Training strategy / Hyperparameters

Intensities of coregistered images corrected for bias field were first standardised using the piecewise linear standardisation described by Nyul et al. [2] For the first step of the training, an adaptive estimation moment optimiser was used with an initial learning rate of 10^{-3}. For the second stage of the training, only the weights belonging to the fully connected layers were refined using an oscillating learning rate. Sampling windows of size 64^3 were taken as input and considered for training if they contained lesion at a ratio of at least 0.00001.

II. Database

In order to build a suitable and large database for training of the network, 2660 cases of automatically segmented WMH were gathered. Automated segmentation was performed based on the method described in [4] modelling the data as a Gaussian mixture model identifying dynamically the number of components to consider to model simultaneously healthy and abnormal observations. Imaging data comprised multiple scanner types and acquisition protocols at various field strengths in order to represent the variability of cases and limit overfitting. This generative pre-training enables the use of a large labelled database without requiring the need for manual delineations. In this still preliminary attempt at using deep-learning techniques for lesion segmentation purposes, only the final results were used for training but one may envision using the other outputs (brain parcellation, tissue segmentation...) to further inform the network at the training stage and/or adapt to different tasks.

For further refinement, 80 cases from this database were manually segmented and used after the first training stage.

III. Results on training set

The segmentation obtained on the challenge data was evaluated using the NiftyNet package and the results obtained for Dice score coefficient, sensitivity, average pairwise difference and positive predictive values are summarised in Table I.

The segmentations obtained with this method were further compared to segmentations obtained using the generative model used to produce the training database. Table II gathers the corresponding evaluation measures. Figure 3 present a segmentation example compared to the gold standard in one case per scanner type.

	GE	**Singapore**	**Utrecht**	**Overall**
DSC	60.3 [45.3 ; 67.7]	76.8 [58.9 ; 83.3]	60.3 [44.9 ; 75.2]	63.7 [46.4 ; 76.8]
Sens	85.0 [72.2 ; 90.7]	81.9 [67.8 ; 86.8]	55.9 [43.5 ; 72.5]	75.5 [55.9 ; 86.7]
AveDist	1.45 [0.72 ; 3.07]	0.58 [0.34 ; 1.78]	1.28 [0.81 ; 2.57]	1.15 [0.58 ; 2.77]
PPV	47.8 [34.9 ; 55.7]	75.9 [51.6 ; 84.2]	75.4 [49.2 ; 84.3]	65.0 [40.1 ; 82.3]

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

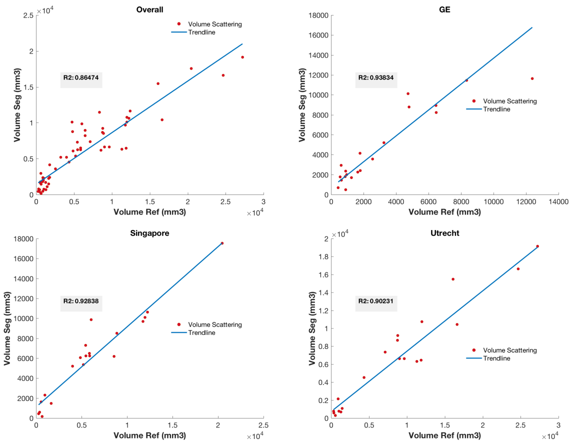

Fig. 2: Plots of the relationship between segmented and reference volumes for the different scanner types

	GE	Singapore	Utrecht	Overall
DSC	65.5 [54.4 ; 71.6]	75.5 [69.6 ; 81.1]	59.6 [44.9 ; 74.2]	68.2 [53.6 ; 77.3]
Sens	55.7 [41.4 ; 61.8]	65.0 [60.3 ; 72.2]	87.2 [75.4 ; 92.5]	65.0 [53.6 ; 81.7]
AveDist	0.84 [0.54 ; 1.65]	0.49 [0.33 ; 1.57]	1.41 [0.76 ; 3.13]	0.81 [0.49 ; 2.25]
PPV	81.7 [75.8 ; 87.9]	90.4 [81.6 ; 93.9]	44.6 [34.1 ; 64.3]	77.9 [59.9 ; 90.0]

TABLE II: Evaluation of segmentation results for each scanner type presented under the form median [1st Quartile ; 3rd Quartile] when compared to the segmentation obtained with the generative model used to create the database.

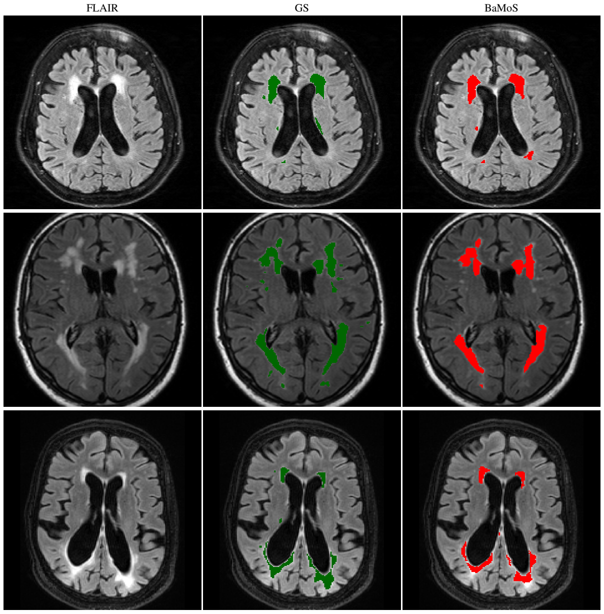

Fig. 3: Example of segmentation results (3rd column) compared to gold standard segmentations (2nd column)

References

[1] Wenqi Li, Guotai Wang, Lucas Fidon, Sebastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. *On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task*, pages 348–360. Springer International Publishing, Cham, 2017.

[2] László G Nyúl, Jayaram K Udupa, et al. On standardizing the mr image intensity scale. *image*, 1081, 1999.

[3] C. H Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. *ArXiv e-prints*, July 2017.

[4] 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.