Deep Bayesian networks for uncertainty estimation and adversarial resistance of white matter hyperintensity segmentation

Parisa Mojiri Forooshani[1,2], Mahdi Biparva[1,2], Emmanuel E. Ntiri[1,2], Joel Ramirez[1,2], Lyndon Boone[1,17], Melissa F. Holmes[1,2], Sabrina Adamo[1,2], Fuqiang Gao[1,2], Miracle Ozzoude[1,2], Christopher J.M. Scott[1,2], Dar Dowlatshahi[3], Jane M. Lawrence-Dewar[4], Donna Kwan[5], Anthony E. Lang[6], Karine Marcotte[7,8], Carol Leonard[9,10], Elizabeth Rochon[10,11], Chris Heyn[1,12], Robert Bartha[13], Stephen Strother[14], Jean-Claude Tardif[15], Sean Symons[12], Mario Masellis[16], Richard H. Swartz[16], Alan Moody[12], Sandra E. Black[1,2,16,*], Maged Goubran[1,2,17,*]

[1] Hurvitz Brain Sciences Research Program, Sunnybrook Research Institute, University of Toronto, Toronto, Canada
[2] Heart and Stroke Foundation, Canadian Partnership for Stroke Recovery, Toronto, Canada
[3] Department of Medicine, University of Ottawa Brain and Mind Institute, the Ottawa Hospital Research Institute, Ottawa, Canada
[4] Thunder Bay Regional Health Research Institute, Thunder Bay, Canada
[5] Department of Psychology, Faculty of Health, York University, Toronto, Canada
[6] The Edmond J. Safra Program in Parkinson's Disease, Toronto Western Hospital, Toronto, Canada
[7] School of Speech Pathology and Audiology, University of Montreal, Montreal, Canada
[8] Research Center of Centre intégré universitaire de santé et de services sociaux du Nord-de-l'île-de-Montréal, Montréal, Canada
[9] Audiology and Speech-Language Pathology program, School of Rehabilitation Sciences, University of Ottawa, Ottawa, Canada
[10] Department of Speech-Language Pathology and the Rehabilitation Sciences Institute, University of Toronto, Toronto, Canada
[11] KITE Research Institute, Toronto Rehab, University Health Network, Toronto, Canada
[12] Department of Medical Imaging, University of Toronto, Toronto, Canada
[13] Department of Medical Biophysics, Schulich School of Medicine and Dentistry, Robarts Research Institute, Western University, London, Canada
[14] Department of Medical Biophysics, Rotman Research Institute, Baycrest, University of Toronto, Canada
[15] Montreal Heart Institute, Universite de Montreal, Montreal, Canada
[16] Department of Medicine (Neurology division), Sunnybrook HSC and University of Toronto, Toronto, Canada
[17] Department of Medical Biophysics, University of Toronto, Toronto, Canada

*co-senior authors

Corresponding author:
Maged Goubran, PhD
maged.goubran@utoronto.ca
416-480-6100 Ext. 85410
2075 Bayview Avenue, M6 West RM 176
Toronto, Canada, M4N 3M5

Running title:
Deep Bayesian networks for WMH segmentation

Abstract

White matter hyperintensities (WMH) are frequently observed on structural neuroimaging of elderly populations and are associated with cognitive decline and increased risk of dementia. Many existing WMH segmentation algorithms produce suboptimal results in populations with vascular lesions or brain atrophy, or require parameter tuning and are computationally expensive. Additionally, most algorithms do not generate a confidence estimate of segmentation quality, limiting their interpretation. MRI-based segmentation methods are often sensitive to acquisition protocols, scanners, noise-level, and image contrast, failing to generalize to other populations and out-of-distribution datasets. Given these concerns, we propose a novel Bayesian 3D Convolutional Neural Network (CNN) with a U-Net architecture that automatically segments WMH, provides uncertainty estimates of the segmentation output for quality control and is robust to changes in acquisition protocols. We also provide a second model to differentiate deep and periventricular WMH. 432 subjects were recruited to train the CNNs from four multi-site imaging studies. A separate test set of 158 subjects was used for evaluation, including an unseen multi-site study. We compared our model to two established state-of-the-art techniques (BIANCA and DeepMedic), highlighting its accuracy and efficiency. Our Bayesian 3D U-Net achieved the highest Dice similarity coefficient of 0.89 ± 0.08 and the lowest modified Hausdorff distance of 2.98 ± 4.40 mm. We further validated our models highlighting their robustness on 'clinical adversarial cases' simulating data with low signal-to-noise ratio, low resolution, and different contrast (stemming from MRI sequences with different parameters). Our pipeline and models are available at: https://hypermapp3r.readthedocs.io

Keywords:

White Matter Hyperintensity (WMH), Image segmentation, Bayesian neural networks, Uncertainty estimation, Adversarial attacks, Deep learning, Vascular lesions

Introduction

Clinical Motivation

White matter hyperintensities (WMH) are commonly observed MRI-based biomarkers of cerebral small vessel disease and have been associated with aging and neurodegenerative diseases such as Parkinson's disease (PD) and Alzheimer's disease (AD) (S. Lee et al. 2016; Prins and Scheltens 2015; S.-J. Lee et al. 2010). WMH have been linked to global cognitive impairment (Jokinen et al. 2020), a decline in mental processing speed (van den Heuvel 2006), increased risk of late onset depression (Herrmann, Le Masurier, and Ebmeier 2008), and increased risk of stroke, dementia, and mortality (Debette and Markus 2010). Since WMH often occurs in the preclinical stage of dementia, it may represent one of the pathological processes indicating progression from mild cognitive impairment (MCI) to dementia (Smith et al. 2008) and should be considered as a covariant of interest at baseline and longitudinally in future AD treatments (Carmichael et al. 2010). Furthermore, periventricular venous collagenosis is associated with WMH in both AD and non-AD patients (Keith et al. 2017).

WMH appear as hyperintense (bright) on both T2-weighted (T2w) and fluid attenuated inversion recovery (FLAIR) MRI images and hypointense (dark) on T1-weighted (T1w) images. While T1w-based WMH estimates have been shown to have some degree of correlation with estimates based on T2w and FLAIR image (Dadar et al. 2018), the inclusion of the two sequences provides improved contrast and visualization of WMH borders, and thus a more accurate segmentation and volume quantification. This is especially important for the classification of both deep WMH (dWMH, found in deep white matter) and periventricular WMH (pvWMH, which extend from the ventricular wall). dWMH presence has been linked to the incidence of migraine (Hong et al. 2020), while increases in pvWMH have been associated with AD progression (Kilgore et al. 2020), as well as stroke (Hernández, Piper, and Bastin 2014). There has been debate in the literature regarding whether or not WMH should be differentiated based on their location, or if they should be considered to be the same based on continuous spectrum (Wardlaw, Valdés Hernández, and Muñoz-Maniega 2015). The exploration of WMH involvement in neurodegenerative disease warrants an accurate segmentation, quantification and classification of WMH and its subtypes.

Related Works

Automated segmentation

There has been a large amount of work revolving around WMH segmentation in clinical and neuroscience research. Traditionally, manual tracings are employed as the gold standard. Individuals trained to segment WMH are aware of the features that distinguish WMH from normal-appearing white matter and other tissue types. The large amount of time and training required to manually segment WMH necessitate the need for automatic methods to segment WMH. Despite the bevy of available segmentation methods (Griffanti et al. 2016; Zhang et al. 2020; Y. Wang et al. 2012), several challenges still exist, such as the lack of uniformity across different image types and differences in imaging protocols/scanners, as well as the lack of methods for quality control. For example, segmentation methods that rely on image intensity (Yoo et al. 2014; Griffanti et al. 2016) had varying degrees of success. While methods that focus on thresholding based on intensity often fail to detect smaller, lower intensity dWMH, thereby increasing the rate of false positives (Jeon et al. 2011). In addition, datasets used for training and testing are often small, and may not always be tested across different scanner types.

Supervised methods utilize previously acquired ground truth, capitalizing on features from the data to produce accurate segmentations on other datasets. Many state-of-the-art (SOTA) WMH segmentation methods are based on a supervised learning formulation. BIANCA, short for "Brain intensity abnormality classification algorithm" (Griffanti et al. 2016) is a fully automated supervised method based on the k-nearest neighbor (kNN) algorithm. The LOCally Adaptive Threshold Estimation (LOCATE) method was

later proposed (Sundaresan et al. 2019) to determine local thresholds that are sensitive to spatial differences in the lesion probabilities to improve BIANCA's binary lesion masks.

In recent years, deep learning with convolutional neural networks (CNNs) has achieved SOTA performance for many medical image segmentation tasks, including WMH and stroke lesions (Guerrero et al. 2018). Kamnitsas et al. (Kamnitsas et al. 2017) proposed a network architecture consisting of two parallel convolutional pathways that processes the 3D input patches at multiple scales, followed by post-processing using a 3D densely connected conditional random field (CRF) to remove false positives. Their method was originally proposed for ischemic stroke and tumor segmentation, but it can be adopted for different lesion segmentations. The U-Net model architecture (Ronneberger, Fischer, and Brox 2015) has been widely used in segmenting biomedical images due to its performance and efficiency of using GPU memory. A CNN model for WMH segmentation that distinguishes between WMH and stroke was presented by Guerrero et al. (Guerrero et al. 2018). Li et al. (Li et al. 2018) presented a U-Net and an ensemble of models trained with random weight initializations to reduce overfitting and boost segmentation results. A skip connection U-Net model was proposed by Wu et al. (Wu et al. 2019) to capture more features and improve the model's receptive field and was evaluated on WMH segmentation.

Uncertainty and Bayesian networks

Uncertainty estimation is critical for understanding the reliability of segmentation networks, and for providing a quantitative assessment of confidence in their outputs. This is specifically important for medical imaging in a clinical setting. Current approaches do not provide uncertainty estimates for their segmentation results. Several studies have investigated uncertainty estimation for deep neural networks (Kendall and Gal 2017; Kendall, Badrinarayanan, and Cipolla 2017; G. Wang et al. 2019). As suggested by Kendall and Gal, there are two major types of predictive uncertainties for deep CNNs: epistemic (model) uncertainty and aleatoric (image-based) uncertainty. Epistemic uncertainty describes limitations in the learning procedure due to limited training data. Aleatoric uncertainty depends on noise or randomness in the input image.

Bayesian Neural Networks (BNNs) have been used to estimate model uncertainty; however, they are hard to implement and are computationally expensive. Previous works have used Stochastic Variational Gradient Descent (SVGD) to perform approximate Bayesian inference on uncertain CNN parameters (Zhu and Zabaras 2018). Other approximation methods have been developed such as Markov Chain Monte Carlo (Neal 2012) that are not scalable for large neural networks with millions of parameters, and variational methods (Graves 2011) to provide an analytical approximation to the posterior probability of unobserved variables, to perform statistical inference over these variables. Bayes by Backprop (BBB) is another approach that combines variational inference with traditional backpropagation to efficiently find the best approximation to the posterior (Blundell et al. 2015). Other studies have used ensembles of multiple models to generate uncertainty (Lakshminarayanan, Pritzel, and Blundell 2016). Alternatively, Monte Carlo (MC) Dropout has been used to demonstrate that dropout at test time can be cast as approximate Bernoulli variational inference to allow an efficient approximation of the model's posterior distribution without additional parameters (Kendall and Gal 2017).

Generalization and Robustness

Machine learning algorithms are usually evaluated by the model's generalization and robustness (Paschali et al. 2018). Generalization refers to the model performance on an unseen dataset. To build a deep learning model that generalizes well, a large and diverse amount of data is required to avoid overfitting (LeCun, Bengio, and Hinton 2015). This is a significant obstacle for using deep learning in the medical domain, where producing high-quality labeled data is time-consuming, expensive, and requires expert knowledge. Robustness refers to the ability of a model to correctly classify previously unseen examples with noise and slight perturbations, which are more challenging to classify and segment (Rozsa, Gunther, and Boult 2018). MR-based segmentation methods are mostly sensitive to acquisition protocols, scanners, noise-level and image contrast. Paschali et al. (Paschali et al. 2018) investigated the robustness of a variety of medical

imaging networks on adversarial cases and found that models that achieve comparable performance in generalizability may have significant differences in their relative exploration of the underlying data manifold, therefore resulting in varying robustness and model sensitivities. Data augmentation can be used as a regularization strategy to control overfitting and improve the model's robustness and performance (Nalepa, Marcinkiewicz, and Kawulok 2019). Common data augmentation schemes for segmentation of neuroimaging data include affine transformations, elastic deformation, random cropping, flipping about a spatial axis, adding noise and adjusting image contrast.

Contributions

In this work, we present several innovative methodological and experimental contributions. First, we developed and evaluated a 3D U-Net with MC Dropout as a Bayesian network to segment WMH and provide uncertainty quantification, generating an estimate of the model's confidence of the predicted segmentation. Second, a separate model was developed to distinguish and segment dWMH and pvWMH, respectively using initial WMH segmentation. Third, unlike previous work, we employed an unseen multi-site study for testing, in addition to the four large multi-site datasets with different diagnostic groups used for training. Fourth, we compared our methods against established SOTA techniques on a wide spectrum of white matter disease burden including very mild WMH cases with small volumes which are difficult to capture and segment. Finally, the network was trained using an augmentation scheme that included permutations to noise level, resolution, and contrast to achieve resistance to common changes due to different MRI acquisition protocols and scanners. This adversarial resistance was validated against simulated challenges in clinical datasets, referred to here as "clinical adversarial attacks". Our trained models are publicly available, and we developed an easy-to-use pipeline with a graphical user interface for making them accessible to users without programming knowledge.

Materials and Methods

Participants

To train the WMH segmentation model, a total of 432 subjects were recruited from four multicenter studies: 160 subjects with cerebrovascular disease +/- vascular cognitive impairment (CVD +/- VCI) or PD (55-86, 75% male) through the Ontario Neurodegenerative Disease Research Initiative (ONDRI) (Farhan et al. 2017), 203 individuals with non-surgical carotid stenosis (47-92, 61% male) through the Canadian Atherosclerosis Imaging Network (CAIN) study (ClinicalTrials.gov: NCT01440296), 37 subjects with nonfluent progressive aphasia, semantic dementia (SD) and normal controls through the Language Impairment in Progressive Aphasia (LIPA) study (Marcotte et al. 2017) (55-80), and 32 subjects with CVD, VCI or Alzheimer's disease (AD) through the Vascular Brain Health (VBH) study (Swardfager et al. 2017) (46-78, 50% male). Ground truth segmentations for WMH were generated using SABRE-Lesion Explorer semi-automated pipeline that generates intensity-based segmentations which are then manually edited by expert annotators trained by a neuroradiologist with an intraclass correlation of ≥ 0.9 (J. Ramirez et al. 2011; Joel Ramirez et al. 2014)). Sequences used for the generation of the ground truth included 3D T1w and T2w FLAIR to accurately delineate the WMH. Participant demographics, diagnosis, Montreal Cognitive Assessment (MoCA) scores, and volumes for the ground truth segmentations and vascular lesions are summarized in **Table 1**.

Table 1. Participant demographics, clinical diagnosis, MoCA scores, ICV and ventricle volumes, and WMH volumes in the training and test datasets.

	Training Dataset n = 432				**Test Dataset n = 158**				
	CAIN n=203	**ONDRI n= 160**	**LIPA n=37**	**VBH n=32**	**CAIN n=50**	**ONDRI n=37**	**LIPA n=10**	**VBH n=8**	**MITNEC n=53**
Demographics									
Diagnostic Group	CN	PD, CVD	CN, FTD	AD, CVD, CN	CN	PD, CVD	CN, FTD	AD, CVD, CN	MCI, AD, CVD
Age (years)	70.4 (7.8)	68.3 (6.9)	67.0 (7.3)	76.4 (9.0)	69 (7.5)	67.9 (5.6)	67.9 (12.1)	77.5 (11.7)	77 (8.7)
Sex (n, % female)	79 (39.0)	40 (25.0)	--	14 (43.8)	22 (44.0)	12 (31.5)	--	4 (50.0)	24 (45.2)
MOCA total score	--	25.9 (2.8)	--	--	--	25.3 (2.9)	--	--	22.2 (4.8)
Neuroimaging Metrics									
ICV (cc)	1241.5 (118.5)	1286.1 (135.8)	1255.7 (132.1)	1253.1 (171.0)	1234.9 (140.0)	1264.38 (151.0)	1254.5 (153.9)	1291.2 (149.6)	1225.8 (174.2)
vCSF (cc)	36.9 (19.7)	38.8 (19.9)	44.3 (20.0)	56.5 (37.2)	36.8 (19.7)	33.1 (18.1)	45.1 (18.2)	48.3 (21.3)	53.3 (21.6)
pvWMH (cc)	7.7 (9.3)	7.2 (10.1)	6.7 (9.9)	20.1 (22.6)	6.2 (4.9)	5.5 (7.9)	2.8 (1.6)	33.6 (35.2)	32.5 (22.0)
dWMH (cc)	1.1 (1.6)	0.7 (1.0)	0.6 (0.7)	2.1 (2.4)	0.8 (1.1)	0.6 (0.8)	0.5 (0.5)	2.4 (2.3)	2.2 (1.2)

Data is presented as mean ± standard deviation unless otherwise specified. ICV= intracranial volume; MoCA=Montreal Cognitive Assessment; vCSF= ventricular cerebrospinal fluid; pvWMH=periventricular white matter hyperintensities; dWMH=deep white matter hyperintensities; CN=cognitively normal; PD=Parkinson's Disease; VCI=vascular cognitive impairment; MCI=mild cognitive impairment; AD=Alzheimer's disease; CVD=cerebrovascular disease; FTD=Frontotemporal dementia.

The models were tested on a total of 158 subjects, including 53 with severe WMH burden (Fazekas 3/3) from the Medical Imaging Trials NEtwork of Canada (MITNEC) Project C6 (ClinicalTrials.gov: NCT02330510), a separate fifth multicenter (unseen) study not part of those used for training. The breakdown of subjects from the five studies is presented in **Table 1.**

Data preprocessing

The preprocessing techniques used in this study are described in our prior work (Goubran et al. 2019). Briefly, we conducted the following data preprocessing steps on all images prior to training: 1) bias-field correction for B1 inhomogeneities using the *N4* algorithm (Tustison et al. 2010), 2) skull-stripping to separate brain from non-brain tissues (Ntiri et al. 2021), 3) background-cropping using a bounding box such that all voxels outside the bounding box are zero-valued, and 4) Z-score normalization using the mean as the subtrahend and the standard deviation as the divisor for each patient volume.

Bayesian 3D CNN architecture

Our networks are based on the U-Net architecture (Ronneberger, Fischer, and Brox 2015; Çiçek et al. 2016), which consists of encoder and decoder pathways, with pooling and upsampling operations. The architecture of the network (**Figure 1(a)**) was based on the original U-Net (Çiçek et al. 2016; Ronneberger, Fischer, and Brox 2015), with some modifications in our 3D implementation. Similar to our prior work (Goubran et al.

2019), residual blocks were added to each encoding layer. Residual blocks resolve the gradient degradation problem that occurs with deeper networks with an increasing number of layers. Also similar to our previous work (Ntiri et al. 2021), dilated convolutions were used which help to enlarge the field of view (FOV) of convolutional filters without losing resolution or coverage (Yu and Koltun 2015). Instance normalization was used instead of batch normalization (Ioffe and Szegedy 2015) to avoid instability associated with batch normalization due to the stochasticity generated by small batch sizes. The network has a depth of 5 layers and 16 initial filters. Here, motivated by approximate Bayesian formulations in deep learning (Gal and Ghahramani 2016), Monte Carlo (MC) dropout layers were added to the network as one of the main contributions of the proposed network. To turn our baseline CNN into a Bayesian CNN, we added MC dropout layers in residual blocks after the first dilated convolution layer which is equivalent to placing a Bernoulli distribution over the weights.

To optimize the location and rate of MC dropout, validation experiments were performed, testing MC dropout on 1) all layers (both encoder and decoder), 2) encoder layers, 3) three central layers both encoder and decoder, and using dropout rates of 0.3 and 0.5. Based on these experiments, we observed that adding dropout on the encoder layers (the residual blocks) with a dropout rate of 0.3 produced the best results on the validation data.

Each encoder layer consists of a convolutional layer and a residual block (He et al. 2016). At each residual block, the input is split into two paths. The first path consisted of two dilated convolutional layers with kernel sizes 7x7x7 with a MC dropout layer with a rate of 0.3 between the two convolutions, while an identity map was applied to the input data in the second path. Element-wise addition was then applied to the results of the first and second paths, combining the outputs of the two.

At each decoder layer before the final layer, the upsampling block from the previous level is concatenated with the corresponding features on the encoder level. Upsampling modules consisted of an upsampling layer of size 3x3x3, followed by a convolutional layer. The output from the upsampling module was concatenated with the summation output from the respective later on the contracting side, before being passed to a feature block. The feature block consisted of two convolutional layers (one dilated convolution) with a stride of 1x1x1, and kernels of size 7x7x7 and 1x1x1, respectively. The number of filters was halved at each decoder step. At the last layer, the concatenated feature map is passed to a sigmoid function to generate a probability map for the class/voxels of interest.

Adversarial resistance and augmentation experiments

Data augmentation is an effective way to enlarge the size and quality of training data to improve model generalizability and avoid overfitting due to a limited dataset. Augmentation can equip a deep network with desired invariance and robustness properties. A large list of data augmentation strategies/schemes were evaluated to improve the adversarial resistance of our WMH model to common changes stemming from MRI sequences with different parameters, protocols, and scanners. The effects of three types of augmentations on model performance were evaluated: 1) Geometrical affine transformations, 2) Histogram-based transformations and 3) Pixel-level transformations. Affine transformations included flipping, changes in orientations, random cropping, and random scaling. Histogram-based transformations included histogram equalization, scaling and brightness modification, while pixel-level transformations included the addition of Rician noise (Gudbjartsson and Patz 1995), contrast shifting/scaling and gamma intensity transformations. Based on these experiments an optimal set of augmentations including four transformations per scan were performed on each of the subjects, including flipping along the horizontal axis (Left-Right axis), random rotating by an angle $\alpha = \pm 90$ along y and z axis, introducing Rician noise generated by applying the magnitude operation to images with added complex noise, where each channel of the noise is independently sampled from a Gaussian distribution with random standard deviation $\sigma = (0.01, 0.2)$, and changing image intensity by gamma $\gamma = (0.1, 0.5)$ such that each pixel/voxel intensity is updated as:

$$\left(\frac{x - min(x)}{intensity_range}\right)^{\gamma} * intensity_range + min(x)$$

where $intensity_range$ is $max(x) - min(x)$.

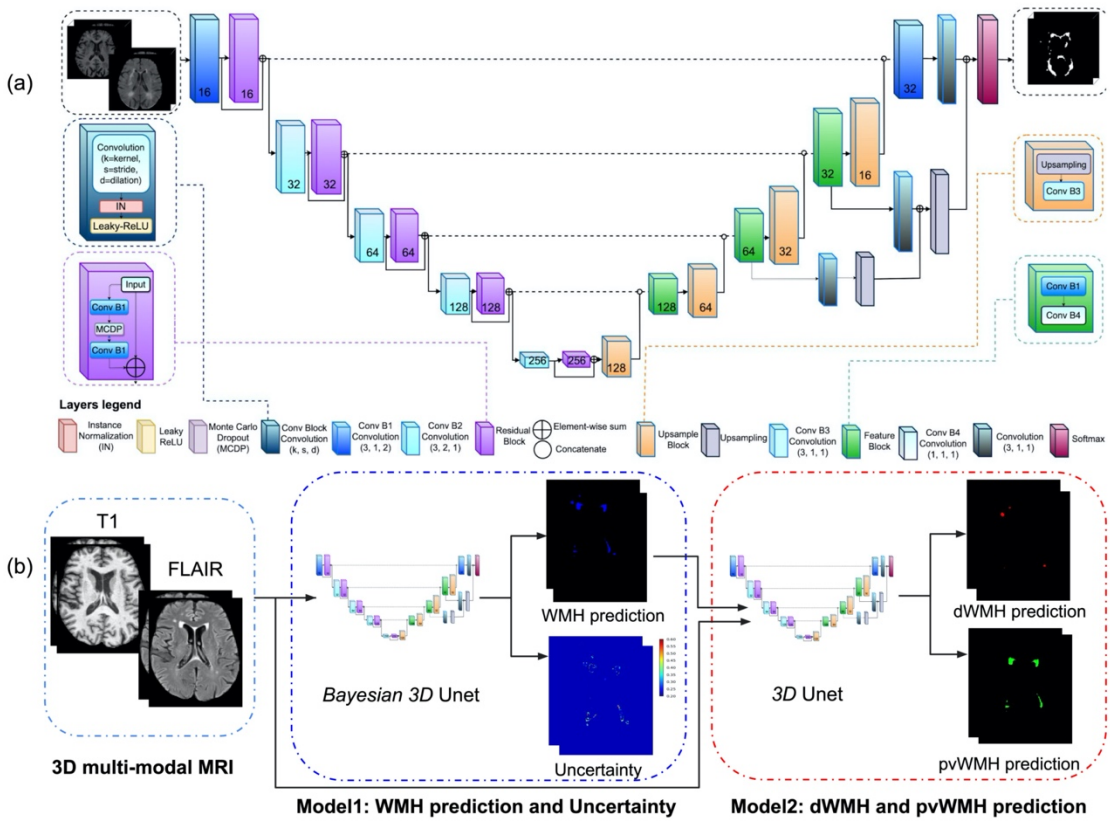

Figure 1. (a) Proposed architecture for the Bayesian 3D U-Net convolutional neural network with residual blocks and dilated convolutions. (b) Overall inference pipeline including a second network to differentiate dWMH and pvWMH.

Loss function

An "equally weighted" formulation of the Dice coefficient was used as the loss function to mitigate the class imbalance issue, in which the majority of voxels in the image do not represent the structure of interest (Milletari, Navab, and Ahmadi 2016). The Dice coefficient is a measure of similarity, determined by a calculation of the overlap between two binary images. Given a predicted binary volume P and the ground truth binary volume G, the Dice coefficient is defined as:

$$D = 2.\frac{\sum_i^N p_i g_i}{\sum_i^N {p_i}^2 + \sum_i^N {g_i}^2}$$

where the sums run over the N voxels, of the predicted binary volume $p_i \in P$ and the ground truth binary volume $g_i \in G$. When differentiated with respect to p_j (j-th voxel of the prediction), in order to calculate back-propagated gradients, we get:

$$\frac{\partial D}{\partial p_j} = 2 \cdot \left[\frac{g_j(\sum_i^N p_i^2 + \sum_i^N g_i^2) - 2p_j(\sum_i^N p_i g_i)}{(\sum_i^N p_i^2 + \sum_i^N g_i^2)^2}\right]$$

Model training
All models were trained for 200 epochs. Early stopping was set to 50 epochs where validation loss did not improve to avoid overfitting. The Adam optimizer (Kingma and Ba 2014) was used with an initial learning rate of $5x10^{-3}$, and a learning rate drop of 0.5 (after 10 epochs where validation loss did not improve). We have selected these training hyperparameters based on experiments.

A multi-contrast network that relies on T1w and FLAIR sequences as inputs were trained. The network was trained on images with various voxel sizes on the whole image, as opposed to patches. We tested different optimizers, learning and decay rates. The networks were trained on a V100-SXM2 graphics card with 32G of memory and a Volta architecture (NVIDIA, Santa Clara, CA).

Segmentation and uncertainty
The MC dropout sampling technique which places a Bernoulli distribution over the network's weights is implemented to estimate the uncertainty. At test time, by retrieving N stochastic outputs, the posterior distribution $p(Y|X)$ can then be approximated (Gal and Ghahramani 2016). With a set $y = \{y_1, y_2.., y_N\}$ from $p(Y|X)$, the final prediction $\hat{y}$ is obtained for X by maximum likelihood estimation:

$$\hat{y} = argmax_y \, p(y|X) \approx Mode(y)$$

where $Mode(y)$ is the most frequent element in y. This corresponds to the majority voting of multiple predictions.

The uncertainty is estimated by measuring how diverse the predictions are. Both variance and entropy of distribution $p(Y|X)$ can be used to estimate the uncertainty. However, the variance captures the spread among predictions. In this paper we use variance which provides a voxel-wise model uncertainty map:

$$var(y) \approx \frac{1}{N}\sum_{n=1}^{N} \hat{y}_n{}^2 - (\frac{1}{N}\sum_{n=1}^{N} \hat{y}_n)^2$$

dWMH and pvWMH segmentation
Distinguishing the dWMH and pvWMH is critical because of their different clinical implications. In this work, we trained a second network to distinguish and segment dWMH and pvWMH using the initial WMH segmentation from the first model. The inference pipeline is shown in **Figure 1(b)**.

Evaluation on clinical datasets
Our WMH model performance was compared against two other established SOTA segmentation tools: 1) BIANCA from the FSL suite (Griffanti et al. 2016), which is a fully automated and supervised method for WHM detection based on the k-nearest (k-NN) algorithm, and 2) DeepMedic (Kamnitsas et al. 2015, 2017) which is a 3D multi-scale CNN designed with parallel pathways that the second path operates on down-samples images. All segmentation tools used in the analysis were trained and evaluated on the same dataset. Volume and shape-based metrics were used to evaluate the segmentation performance of the segmentation methods including the Pearson correlation coefficient, the Dice similarity coefficient (DSC), the modified Hausdorff distance (HD95), the absolute volume difference (AVD), recall and F1-score for individual lesions. The Pearson correlation coefficient (Pearson and Galton 1895) was used as a measure of the correlation between the volumes from each segmentation prediction P, and volumes from ground truth manually segmented data G.

The DSC is a measure of the overlap between two datasets. Given a predicted binary mask, P and a binary ground truth volume G, the Dice similarity coefficient is defined as:

$$DSC(G,P) = 2.\frac{|G \cap P|}{|G| + |P|}$$

The Hausdorff distance measures how far two surfaces occupying the same space are. Given two sets of points representing objects occupying the same space, G and P, where $x \in G$ and $y \in P$, the Hausdorff distance from P to G is defined as the largest value in a derived set of closest distances between all points.

$$\check{\mathrm{H}}(G,P) = max_{x \in G}\{min_{y \in P}\{||x,y||\}\}$$

In the above function, $||x,y||$ is the Euclidean distance between points x and y. Because the Hausdorff distance between the two sets relative to G is not equal to the distance relative to P, the bidirectional Hausdorff distance is equal to the maximum value between the two directions. A smaller distance is indicative of a greater degree of similarity between the segmentation and the manual tracing. Here we used the 95th percentile instead of the maximum (100th percentile) distance to obtain a more robust distance estimate.

The AVD between the volumes of both ground truth (G) and predicted (P) images were also computed. An AVD of 0 signifies that the ground truth and the segmentation have the same number of voxels, though it is not indicative of a perfect segmentation. Let V_G and V_P be the volume of lesion regions in G and P respectively. Then the AVD as a percentage is defined as:

$$AVD = \frac{|V_G - V_P|}{V_G}$$

Each individual lesion is defined as a 3D connected component. Given this definition, let N_G be the number of individual lesions delineated in G, and N_{TP} be the number of correctly detected lesions after comparing P to G. Each individual lesion is defined as a 3D connected component. The Sensitivity for individual lesions (Recall) is defined as:

$$Recall(G,P) = \frac{N_{TP}}{N_G}$$

Let N_{FP} be the number of wrongly detected lesions in P. Then the F1-score for individual lesions is defined as:

$$F1(G,P) = \frac{N_{TP}}{N_{TP} + N_{FP}}$$

The non-parametric Mann-Whitney U test was employed with an α-level of 0.05 to assess the improvement between our models and all tested methods on these evaluation metrics.

Clinical adversarial attacks

In order to validate the robustness of our model on data with lower resolution or quality, we generated "clinical adversarial cases" to further test the model's robustness. These cases included: introduction of noise (to simulate data with lower signal-to-noise ratio 'SNR'), downsampling of image resolution (to simulate typically short clinical scans performed on low field strength magnets), and different contrasts (to simulate data with different scanners). Noise was introduced using a Rician distribution sampled from two channels of Gaussian noise with a standard deviation of σ. Input images were downsampled by a factor of 2 across all dimensions. The image intensities were changed with γ. All other SOTA methods were compared to our models on the most challenging adversarial cases, specifically: those with induced noise of $\sigma = 0.2$, those downsampled by a factor of 2 in each spatial axis, as well as changing image intensity with $\gamma = 0.5$.

Results

In this section, we first present our model segmentation results and the accompanying uncertainty maps to highlight the application of the uncertainty for quality control. We then compare the performance of our models to SOTA methods using volume and shape-based evaluation metrics. We also evaluate the second model's performance across dWMH and pvWMH segmentation. Furthermore, we focus on the evaluation of more challenging mild cases across tested methods. The models are finally tested against generated clinical adversarial cases to evaluate their robustness towards increased noise, lower resolution and different contrasts.

Uncertainty maps for quality control

Figure 2 highlights qualitative WMH segmentation results for an example subject based on the Bayesian model with Monte Carlo simulation ($N = 20$ models) to obtain epistemic uncertainty. This example demonstrates instances of both over-segmentation and under-segmentation. It can be observed that the uncertainty map shows high confidence (low uncertainty) in correctly segmented regions and lower confidence (high uncertainty) in mis-segmented regions as highlighted by red (under-segmentation) and green (over-segmentation) arrows. Uncertainty map can thus represent an estimate of segmentation confidence for quality control in both clinical and research settings.

Evaluation of clinical datasets

Average DSC, modified HD95, AVD, recall and F-1 score for HyperMapper Baseline, HyperMapper Bayesian, BIANCA, and DeepMedic for WMH segmentations are summarized in **Table 2** and **Figure 3**. Our Bayesian model had the highest DSC (0.89 ± 0.08) and the lowest HD95 (2.98 ± 4.40 mm) across tested methods. The Bayesian model outperformed the Baseline model, which demonstrates that test-time dropout helps improve segmentation accuracy besides providing an uncertainty map, with the caveat of increased computation time. BIANCA achieved the highest recall, which indicates a high proportion of true positive voxels; however, it produced the lowest F-1 score, which represents low precision or a high number of false positive voxels. Notably, our models had the highest F-1 score which represents lower false positives. DeepMedic and BIANCA had 1.5 to 10 times higher Hausdorff values than our Bayesian model. In addition, our models were faster than BIANCA and DeepMedic. Pearson correlations between manual segmentation volume and volume quantified by the three tested methods are summarized in **Suppl**. **Figure 1**. Our models and DeepMedic had the highest volume correlations with manual WMH compared to the other tested techniques ($r = 0.99,\ p < 0.0001$), while BIANCA had the lowest volume correlations ($r = 0.91,\ p < 0.0001$). A qualitative comparison between all WMH segmentation methods on one subject is shown in **Figure 4**.

The cases with the highest and lowest DSC between the manual segmentations and the Bayesian model's predictions, along with their uncertainties are displayed in **Figure 5** (A and B, respectively). In both cases, the model was able to accurately segment a wide spectrum of white matter disease burden including both very mild and highly severe WMH. Visual inspection of the case with the lowest DSC demonstrates some of the challenges of this segmentation task, such as overlooking hyperintensities in the parietal lobe and ventricle wall. The majority of these challenges were captured in the corresponding uncertainty maps as the high variance between different test-time models.

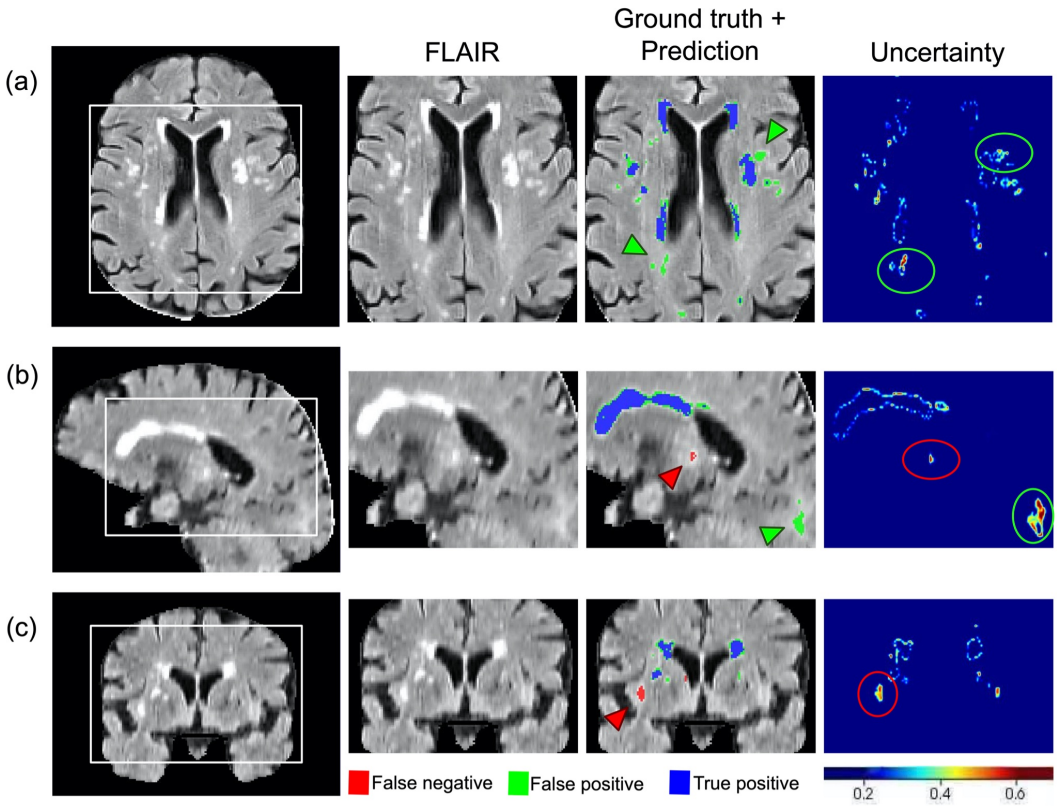

Figure 2. An example of WMH segmentation and uncertainty estimation, showing on a FLAIR scan, the Bayesian model's prediction and estimated epistemic uncertainty in a) axial b) sagittal, and c) coronal views. Blue represents the overlap between ground truth and prediction, red represents ground truth voxels missing in prediction (under-segmentation), green represents prediction voxels not in ground truth (over-segmentation).

Table 2. Evaluation of WMH segmentation on different methods with the following metrics: Dice similarity coefficient (DSC), Hausdorff distance in "mm" unit (modified as 95th percentile) (HD95), absolute volume difference (AVD%), sensitivity (Recall) and F-1 score for individual lesions. ↓ indicates that smaller values represent better performance.

	HyperMapper Baseline	**HyperMapper Bayesian**	**BIANCA**	**DeepMedic**
Dice similarity coefficient	0.892 (± 0.080)	**0.893 (± 0.080)**	0.604 (± 0.222)	0.858 (± 0.080)
Hausdorff distance (HD95) (mm) ↓	3.045 (± 4.417)	**2.979 (± 4.396)**	29.684 (± 17.105)	4.477 (± 6.322)
Absolute volume difference (%) ↓	**9.819 (± 12.173)**	9.843 (± 12.134)	108.970 (± 216.346)	13.846 (± 14.731)
Sensitivity (Recall)	0.762 (± 0.149)	0.762 (± 0.150)	**0.805 (± 0.168)**	0.764 (± 0.153)
F-1 score	0.752 (± 0.119)	**0.753 (± 0.119)**	0.199 (± 0.148)	0.652 (± 0.126)
Time (seconds)	**10s**	**16s**	24s	25s

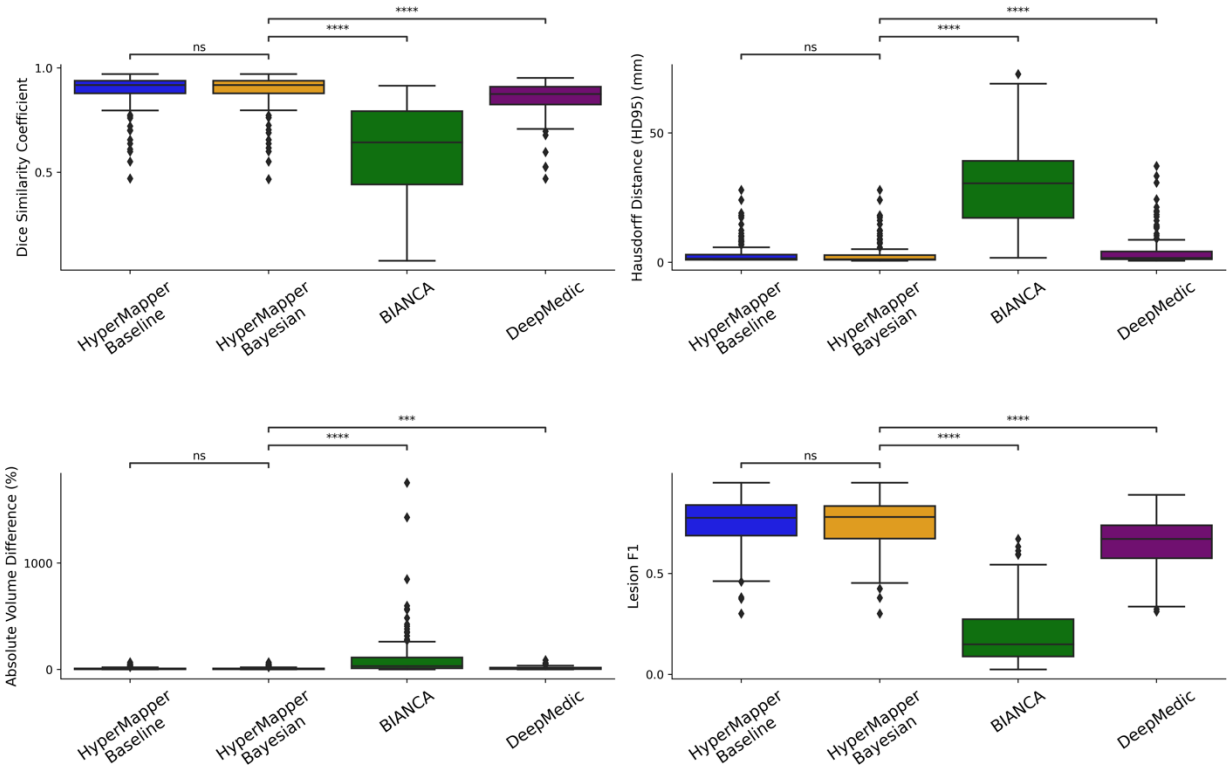

Figure 3. Evaluation of WMH segmentations across tested methods using the following metrics: Dice similarity coefficient, modified Hausdorff distance (HD95), absolute volume difference (%), and Lesion F1. not significant: ns, $p < 0.05$: *; $p < 0.01$: **; $p < 0.001$: ***; $p < 0.0001$: ****.

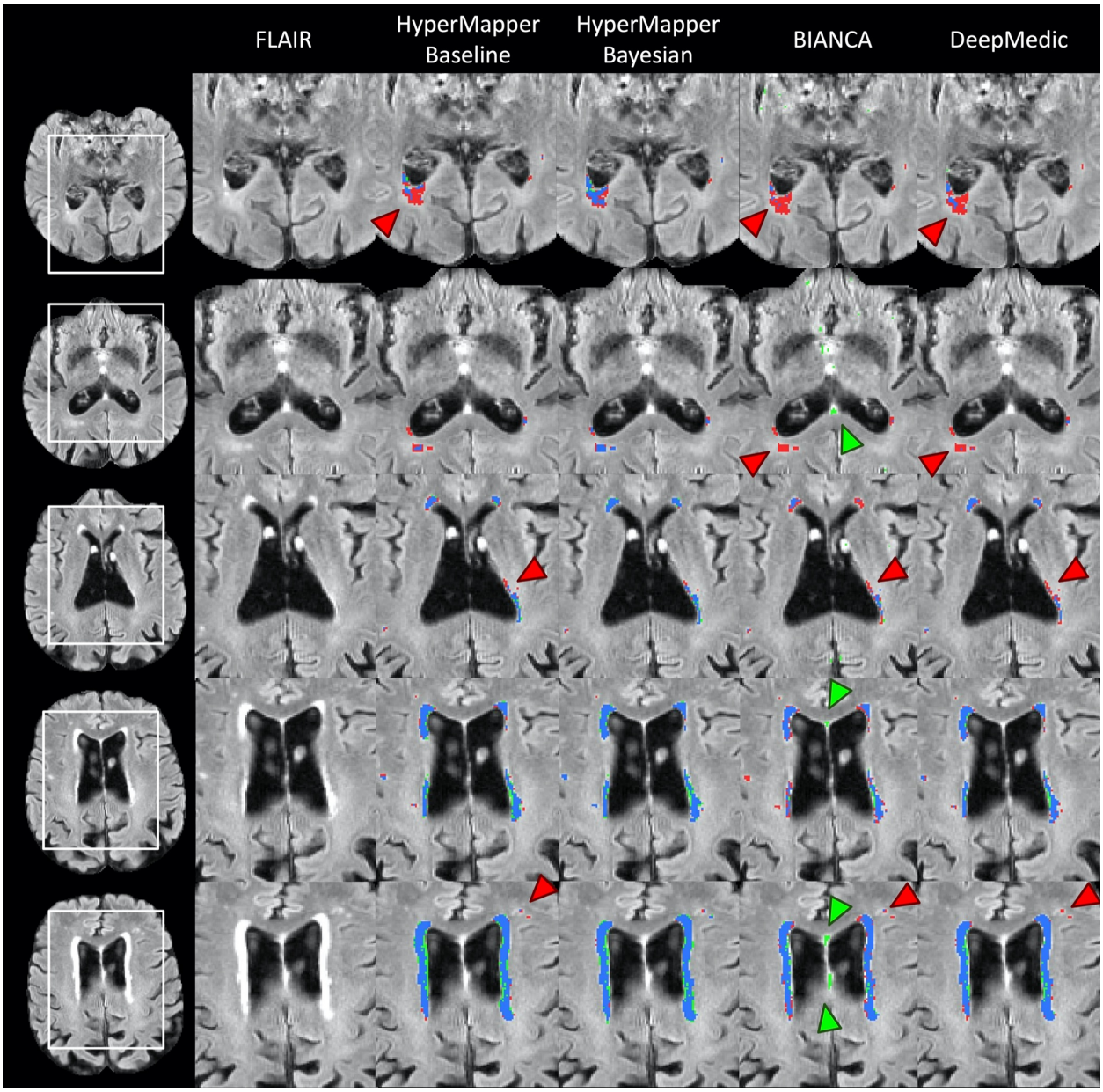

Figure 4. Visual comparison of the tested methods in an example subject. Blue represents the overlap between ground truth and prediction (true positive voxels), red (and red arrows) represents ground truth voxels missing in prediction (false negative voxels), green (and green arrows) represents prediction voxels not in ground truth (false positive voxels).

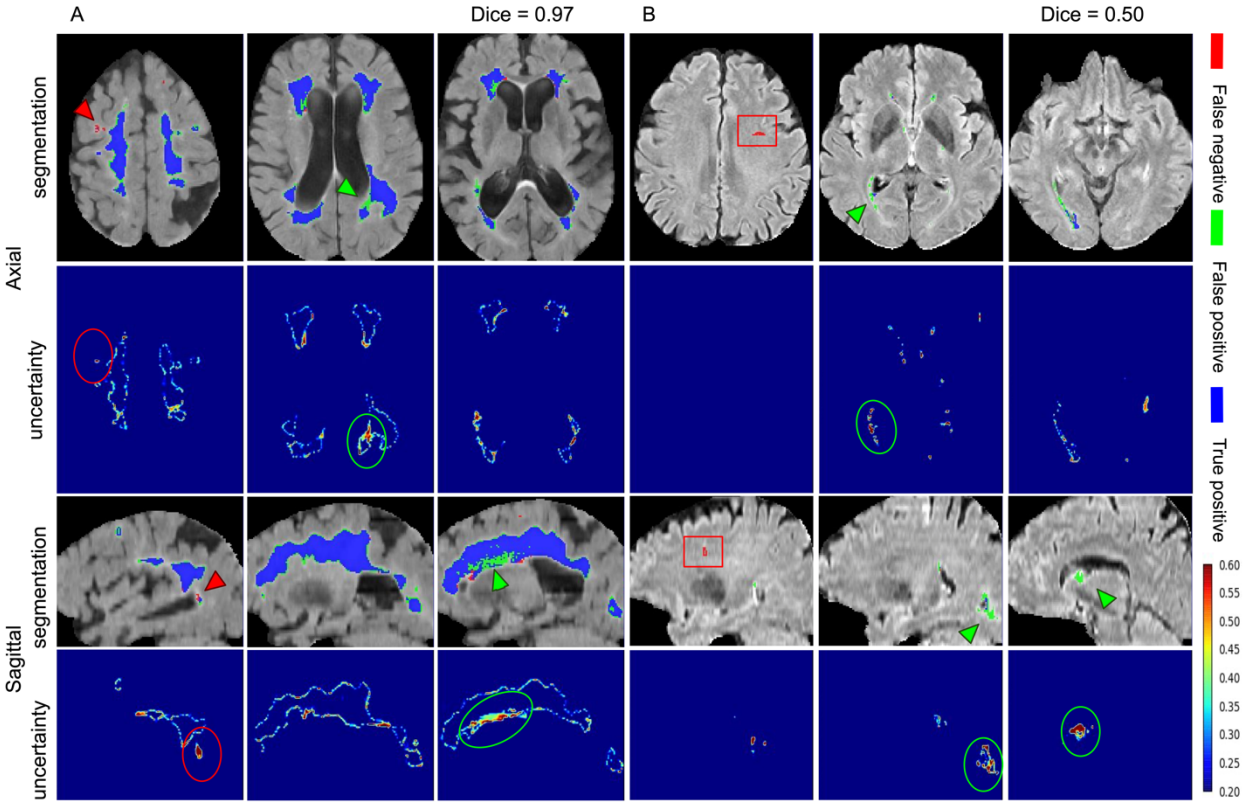

Figure 5. WMH segmentation and uncertainty maps of cases with the highest (A) and lowest (B) Dice similarity coefficients from the test set. Red arrowheads and circles highlight areas of under-segmented and green arrowheads and circles highlight areas that were over-segmented.

Evaluation on dWMH and pvWMH

We also evaluated our second model's capability of classifying both dWMH and pvWMH. This distinction is important in the clinical setting because accurate WMH labeling is a critical feature in classifying vascular dementia. Dice similarity coefficients on dWMH and pvWMH were 0.58 ± 0.22 and 0.87 ± 0.09, respectively. These results demonstrate that segmenting dWMH is more challenging due to their size and location. For instance, pvWMH usually appear larger, brighter, and often form confluent lesions with higher contrast, while dWMH usually appears as small punctate lesions. **Figure 6** highlights WMH multi-class segmentation results for an example subject with the first model's WMH segmentation, dWMH/pvWMH segmentation, and multi-class ground truth.

Evaluation on mild cases

To highlight the clinical and research utility of our models, we further evaluated their performance on cases with mild WMH burden, against other SOTA methods (**Table 3 and Suppl. Figure 2**). **Table 3** shows the evaluation of different methods on 50 cases with mild WMH burden (pvWMH (cc): 1.8 (1.1) and dWMH: 0.27 (0.22)). The results demonstrate that our Bayesian model had the highest performance as assessed by DSC, HD95, and F1-score (0.84 ± 0.10, 5.25 ± 6.51, and 0.721 ± 0.116, respectively). They also highlight the higher complexity of the task and the shortcomings of existing SOTA methods in extracting unique features from much smaller lesions with limited volumes and extent. Therefore, their early detection is clinically important for neurodegenerative populations.

Table 3. Evaluation of WMH segmentation on mild WMH cases. ↓ indicates that smaller values represent better performance.

	HyperMapper Baseline	**HyperMapper Bayesian**	**BIANCA**	**DeepMedic**
Dice similarity coefficient	0.841 (± 0.097)	**0.842 (± 0.097)**	0.340 (± 0.067)	0.779 (± 0.092)
Hausdorff distance (HD95) (mm) ↓	5.365 (± 6.450)	**5.252 (± 6.505)**	44.042 (± 9.564)	8.236 (±7.190)
Absolute volume difference (%) ↓	**11.298 (± 12.715)**	11.393 (± 12.812)	275.970 (±324.757)	22.087 (± 20.971)
Sensitivity (Recall)	0.712 (± 0.161)	0.712 (± 0.160)	**0.780 (± 0.187)**	0.749 (± 0.160)
F-1 score	0.720 (± 0.116)	**0.721 (± 0.116)**	0.096 (± 0.048)	0.594 (± 0.120)

Clinical adversarial cases

Our model's performance was evaluated on the effects of changes and perturbations in SNR, resolution, and contrast and compared against other SOTA methods (**Table 4)**. Cases with the lowest SNR led to a performance drop of our Bayesian model with an average drop of 9.7% in DSC and an average increase of 20% in HD95. While decreases in resolution resulted in an average drop of 9.4% in DSC and an increase of 96% in HD95. Changing image contrast produced a drop of 7.9% in DSC and an increase of 69% in HD95. WMH segmentation results and uncertainty maps for these simulated experiments are shown in **Figure 7** and **Suppl. Figure 8**. To further validate the performance of our model against changes in imaging protocols, we evaluated it against clinical adversarial attacks on a population with mild WMH burden (**Suppl. Table 1**). The results show that the mild WMH cases results in a higher performance drop in our Bayesian model on the DSC (lower SNR: 12.3%, lower resolution: 12.6%, and changing contrast: 10.4%), due to the complexity of small lesions and their higher sensitivity to different attacks such as higher noise and lower resolution.

All tested SOTA methods were compared to our Bayesian model on noisy input data with a sigma 0.2, downsampled data by a factor of 2 across all planes, as well as inputs with contrast changes using a gamma 0.5 (**Figures 8**, **Suppl. Figures 3** and **4**). DeepMedic failed to generate 62 (out of 158) WMH segmentation masks for downsampled adversarial cases, and 1 WMH segmentation for noise induced cases. Failed cases were not included in the analysis or generated figures. We observed that the clinical adversarial cases had significant effects on other SOTA methods as apparent by the substantial decreases in Dice coefficient and increases in Hausdorff distance values (**Table 4**). A qualitative comparison between all WMH segmentation methods on one subject for all three adversarial attacks is included in **Figure 9**. The performance of BIANCA and DeepMedic on the selected subject (**Figure 9**) were high before any attacks (DSC > 0.80) but dropped significantly on adversarial cases.

Our model had better performances across all metrics for the noise induced cases in comparison to other methods (**Figure 8)**. Specifically, the Bayesian model had a significantly higher DSC and lower HD95 relative to other SOTA methods (HyperMapper:- DSC: 0.81 ± 0.09, HD95: 3.59 ± 4.69mm; BIANCA:- DSC: 0.25 ± 0.13, HD95: 30.47 ± 18.45 mm; DeepMedic:- DSC: 0.64 ± 0.22, HD95: 14.09 ± 13.63 mm). Similar to noise induced cases, our Bayesian model had a better performance in comparison to other SOTA methods on adversarial attacks with lower resolution (DSC: 0.81 ± 0.10, HD95: 5.84 ± 6.48 mm). All other SOTA methods encountered a substantial decrease in performance as a result of downsampling the input

images. Although DeepMedic's performance was comparable on in-distribution data, its performance dropped significantly on adversarial cases with lower resolution (DSC: 0.31 ± 0.21, HD95: 50.05 ± 21.24 mm). A similar trend was observed in the contrast permutation adversarial experiments, where a slight drop in accuracy was observed using our Bayesian model (DSC: 0.82 ± 0.10, HD95 5.05 ± 6.17mm); in stark contrast to BIANCA (DSC: 0.39 ± 0.18, HD95: 29.50 ± 17.07 mm) and DeepMedic's performance (DSC: 0.67 ± 0.19, HD95: 10.74 ± 10.58 mm).

The performance of our models and SOTA methods were further evaluated on adversarial attacks in 50 mild WMH subjects (**Suppl. Table 1, Suppl. Figures 5, 6,** and **7**), demonstrating that our Bayesian model had the highest performance and robustness in comparison to tested SOTA methods across different adversarial attacks. DeepMedic failed to generate 44 (out of 50) WMH segmentation masks for downsampled adversarial cases, and 1 WMH segmentation for noise induced cases. DeepMedic experienced a substantial decrease in performance on adversarial cases with lower SNR (DSC: 0.42 ± 0.20, HD95: 26.64 ± 17.41 mm), lower resolution (DSC: 0.10 ± 0.09, HD95: 76.42 ± 6.63 mm), and contrast changes (DSC: 0.49 ± 0.18, HD95: 20.94 ± 12.35 mm).

Table 4. Evaluation of WMH segmentation on different adversarial attacks.

	Adversarial attacks	**HyperMapper Bayesian**	**BIANCA**	**DeepMedic**
Dice similarity coefficient	Noise (sigma=0.2)	**0.806 (± 0.086)**	0.252 (± 0.130)	0.639 (± 0.217) *
	downsampled (2x2x2)	**0.809 (± 0.101)**	0.348 (± 0.176)	0.306 (± 0.208) **
	Contrast (gamma=0.5)	**0.822 (± 0.102)**	0.389 (± 0.184)	0.677 (± 0.198)
Hausdorff distance (HD95) (mm) ↓	Noise (sigma=0.2)	**3.592 (± 4.686)**	31.521 (± 14.999)	14.092 (± 13.626) *
	downsampled (2x2x2)	**5.836 (± 6.475)**	30.469 (± 18.451)	50.049 (± 21.236) **
	Contrast (gamma=0.5)	**5.047 (± 6.172)**	29.488 (± 17.077)	10.742 (± 10.5854)
Absolute volume difference (%) ↓	Noise (sigma=0.2)	**15.037 (± 13.169**	159.602 (± 278.002)	42.627 (± 24.601) *
	downsampled (2x2x2)	**19.745 (± 18.260)**	156.712 (± 325.622)	77.265 (± 18.012) **
	Contrast (gamma=0.5)	**24.454 (± 23.178)**	176.860 (± 365.183)	40.007 (± 24.096)
Sensitivity (Recall)	Noise (sigma=0.2)	**0.720 (± 0.136)**	0.669 (± 0.168)	0.259 (± 0.124) *
	downsampled (2x2x2)	**0.677 (± 0.148)**	0.601 (± 0.217)	0.072 (± 0.040) **
	Contrast (gamma=0.5)	0.601 (± 0.200)	**0.759 (± 0.186)**	0.348 (± 0.191)
F-1 score	Noise (sigma=0.2)	**0.505 (± 0.160)**	0.139 (± 0.142)	0.376 (± 0.134) *
	downsampled (2x2x2)	**0.718 (± 0.010)**	0.316 (± 0.169)	0.131 (± 0.066) **
	Contrast (gamma=0.5)	**0.662 (± 0.135)**	0.213 (± 0.175)	0.469(± 0.166)

*DeepMedic failed on 1 subject with increased noise, **DeepMedic failed on 62 subjects with lower resolution

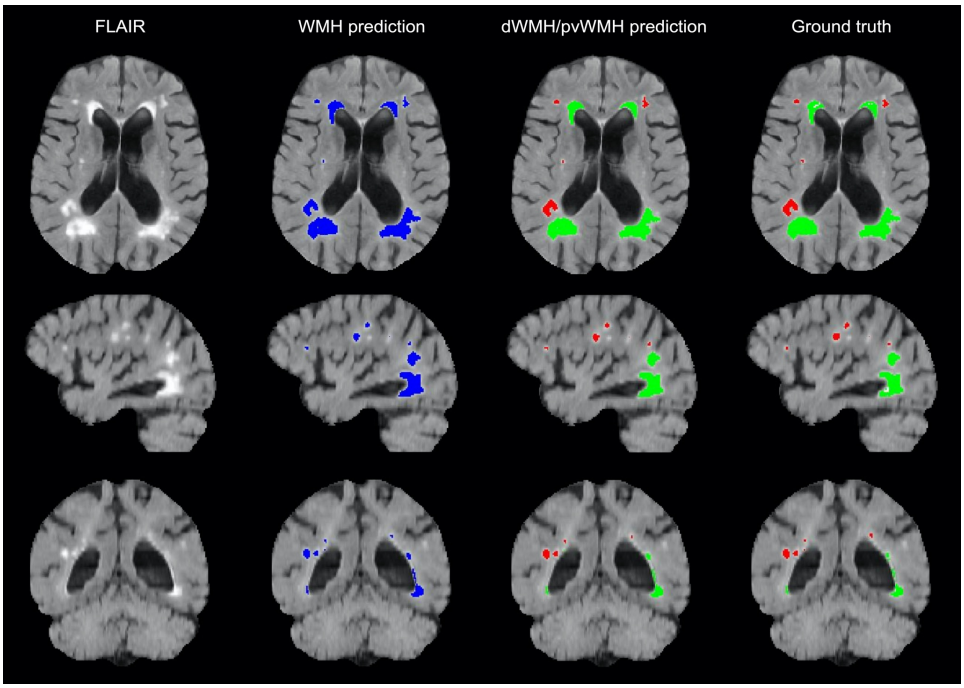

Figure 6. An example of WMH segmentation on a FLAIR scan (axial, sagittal, and coronal view), showing the Bayesian model's total WMH prediction, dWMH and pvWMH prediction, as well as ground truth labels. Blue labels represent Bayesian model WMH prediction, red labels represent dWMH, and green labels represent pvWMH.

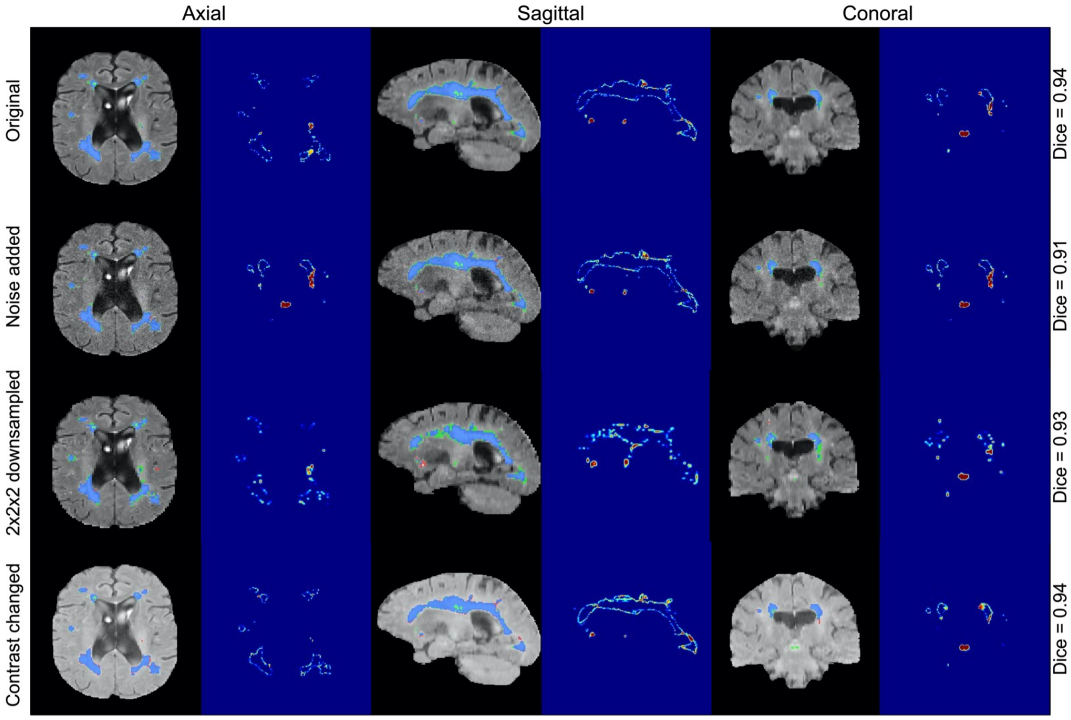

Figure 7. WMH segmentation and uncertainty estimates using our Bayesian model under three types of adversarial attacks applied to the same subject (the addition of noise with a sigma of 0.2, downsampling of resolution by a factor of 2x2x2, and changing contrast with 0.5 gamma).

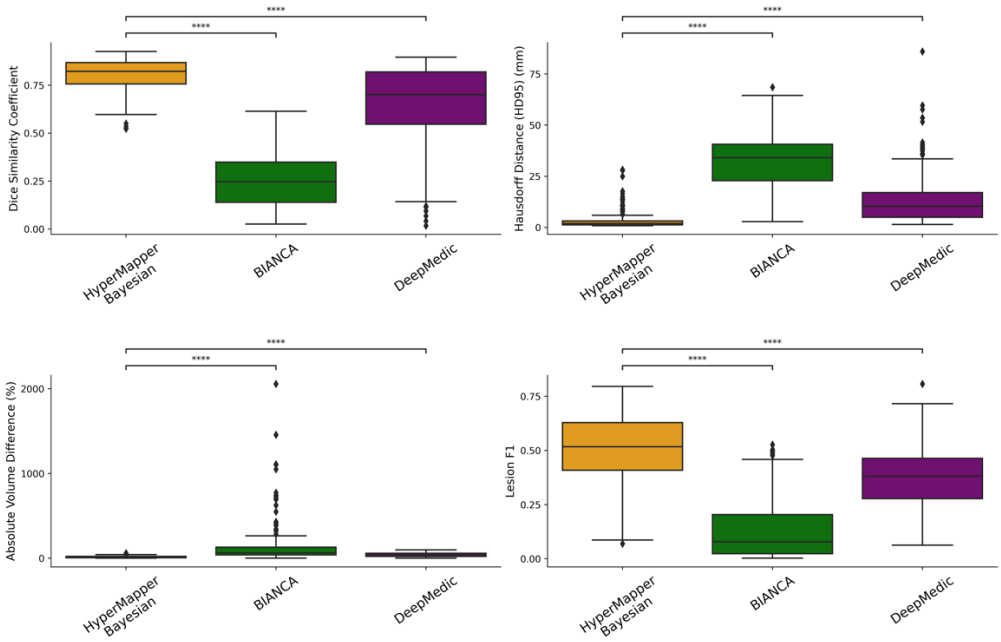

Figure 8. Evaluation of WMH segmentation on cases with increased noise. not significant: ns, $p < 0.05$: *; $p < 0.01$: **; $p < 0.001$: ***; $p < 0.0001$: ***.

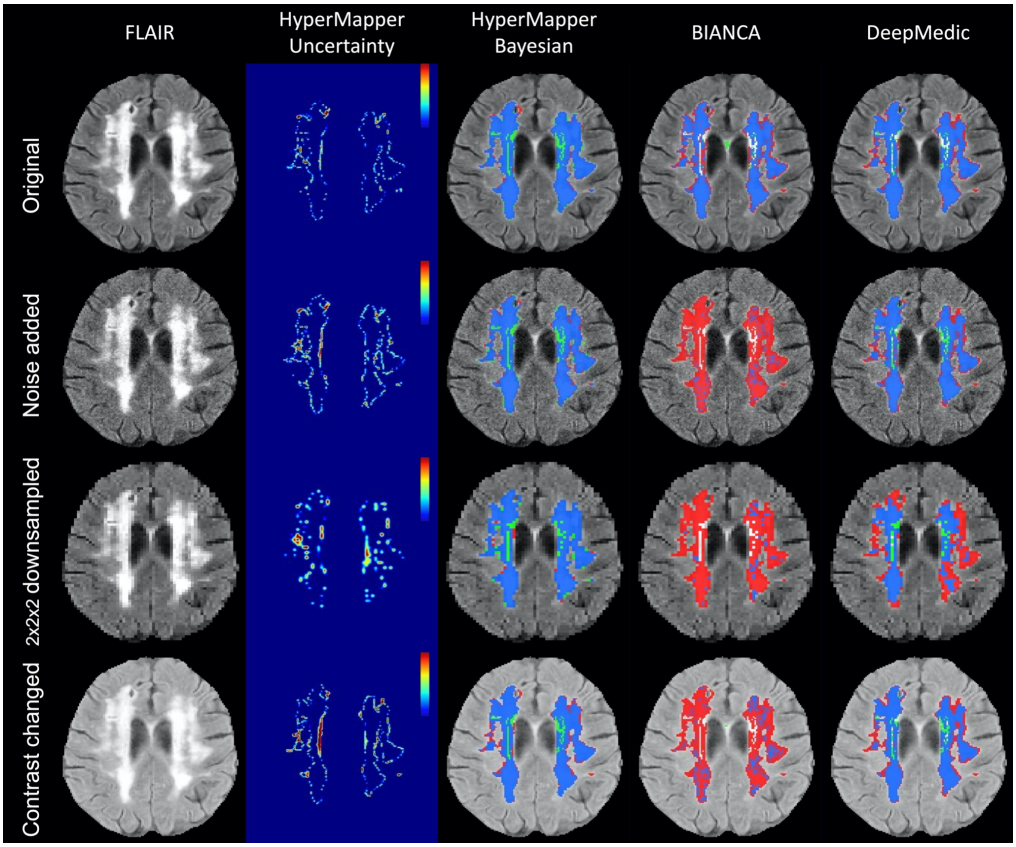

Figure 9. Visual comparison of the segmentation methods under three types of adversarial attacks (the addition of gamma noise with a sigma of 0.2, downsampling of resolution by a factor of 2x2x2, and changing contrast with 0.5 gamma).

Discussion

This work presented Bayesian 3D CNNs to segment total, deep and periventricular WMH, and estimate model uncertainty to provide a quantitative assessment of confidence in segmentation accuracy. Our models were trained using a heterogeneous multi-site, multi-scanner dataset with different augmentation schemes including the addition of noise, various resolutions, and changes in contrast, to make the model more robust to the challenges that commonly result from different MRI scanners and acquisition protocols. We first highlighted the utility of the resulting uncertainty maps for quality control. Then, we compared our models to available SOTA methods demonstrating our model's performance on a test dataset including an unseen multi-site study. We further validated the performance of all methods on populations with mild WMH burden. Finally, the performance of our model was validated against SOTA methods on clinical adversarial attacks including lower SNR, lower resolution, and changes in contrast.

Uncertainty maps for quality control

For estimating model uncertainty, we implemented a Bayesian approach, where MC samples from the posterior distribution were generated by keeping the dropout layers active at test time. Variance over the MC samples was used to provide a voxel-wise model uncertainty map, whereas maximum likelihood estimation over the MC predictions provided the final segmentation. Based on our experiments and visual assessments, the uncertainty maps reliably represented correctly segmented regions with low uncertainty and mis-segmented regions with high uncertainty. Generally, the areas of high uncertainties were localized in the periphery of the segmentations. Furthermore, the uncertainty increased when the model was tested on the out-of-distribution datasets (simulating data with low signal-to-noise ratio, low resolution, and different contrast). Uncertainty is a key concept to highlight scans with lower segmentation accuracy. Uncertainty could therefore be used to either guide the expert or combined with an uncertainty-aware postprocessing method to improve segmentation.

Evaluation of clinical datasets

While many WMH segmentation algorithms exist due to the importance placed on quantifying WMH in neuroimaging studies and neurodegenerative disorders, these algorithms commonly require manual parameter tuning and are computationally expensive. In addition, many current methods do not produce optimal results in populations with mild vascular lesions. Our models provide an open-source, accurate and robust solution to segment WMH and classifying dWMH and pvWMH that is fast and require no parameter optimization; highlighting their applicability to segment mild WMH burden. Our Bayesian model achieved an average DSC of 0.89 ± 0.08 and HD95 of 2.98 ± 4.40 mm for WMH segmentation in this difficult test dataset.

The Dice similarity coefficient is commonly used to evaluate the overlap between segmentation output and ground truth labels; however, it is less indicative of mismatch when segmenting larger ROIs or outlier voxels. Thus, it should be reviewed in conjunction with surface distance metrics such as the Hausdorff distance as a complementary evaluation metric for measuring boundary mismatches. Although the absolute volume difference (AVD) serves as a measurement of the similarity in voxel counts, it was included as an evaluation metric to provide a fuller context to the segmentation results. To assess over-segmentation and under-segmentation, the values on recall (sensitivity) and F1-score were calculated. Evaluating the Dice similarity coefficient and Hausdorff score, along with AVD, Recall and F1-score, we demonstrated that our proposed model improves on current SOTA methods in terms of accuracy while providing uncertainty maps, performing the task in seconds.

We chose a multi-site and multi-scanner dataset to train our model on ground truth data that were semi-automatically edited by experts rather than automated outputs produced by other algorithms as they are

expected to be more accurate. The consistently higher performance of our model on several multi-site studies highlights its robustness against different scan protocols, disease groups, and unseen datasets. It should be noted that while manual segmentation is often considered the gold standard, it may not necessarily represent the absolute truth. Therefore, some errors in segmentations could be due to inconsistencies in manual segmentation rather than lower model accuracy. **Figure 5(B)** highlights a case where another vascular lesion (PVS) was mislabelled as WMH in the frontal lobe in the ground truth data, highlighted by red squares, since our model accurately did not capture it as WMH.

Evaluation on d/pvWMH and mild cases

Distinguishing the dWMH and pvWMH is important due to different clinical implications such as classifying vascular dementia. While our model achieved a high performance on pvWMH, the task of segmenting dWMH was more challenging due to its shape, size and location. Our model's improvement in performance over SOTA methods was even more apparent in populations with mild WMH burdens. The ability to accurately capture smaller WMH is a more challenging task of lesion detection and quantification; however, it is important for studying and tracking preclinical stages of neurodegenerative diseases and small vessel disease, as it has been shown that WMH at baseline predicts future WMH and is associated with dementia.

Clinical adversarial cases

Our results on the "clinical adversarial cases" suggest that our Bayesian model is largely robust to cases with varying degrees of noise, downsampling and contrast changes. There are several reasons why our model may be more robust than SOTA methods that we compared against, such as augmentation strategies, residual blocks, dilated convolutions, and skip-connection architecture. Tested SOTA methods had widely varying results when segmenting lower SNR data, lower resolution, and data with different contrast than the training set, with substantial volume differences and drops in quantitative metrics or segmentation fidelity. Notably, they failed on the majority of cases with lower resolution, highlighting the sensitivity of some deep learning networks to out-of-distribution data and the need for further validation of the generalizability of these networks. Further improvements of our models may also be necessary in light of some of the weaknesses found, most notably on downsampled and noisy data. In addition to the simulated attacks we employed, there are other challenges that could be tested in future work, such as the simulation of significant subject motion that could have an effect on smaller dWMH.

Though our models performed well compared to other segmentation algorithms, there are still some future improvements that can be made. While our models are trained on a dataset without stroke lesions, the stroke's penumbra presents with similar contrast to WMH on FLAIR sequences, hence one of the main foci of future work is to train models to distinguish between the two lesions. Future work will also investigate the effect of a combined loss function (for example a weighted cross entropy loss and a weighted dice loss) to handle datasets with varying degrees of contrast and volume sizes.

Conclusion

We present a robust and efficient WMH segmentation model, which also generates an uncertainty map for quality control. In addition, we present a second model to classify dWMH and pvWMH using the initial total WMH segmentation. We trained our CNN models with expert manually edited segmentations from four large multi-site studies including participants with vascular lesions and atrophy, which represent challenging populations for segmentation techniques, and then tested them on an unseen multi-site study in addition to the four large multi-site datasets. Our segmentation models achieved high accuracy compared to SOTA algorithms on a wide spectrum of WMH burdens, especially mild WMH. Additionally, we used an augmentation scheme to make our model robust to simulated images with SNR, low resolution, and different contrasts. We are making our pipelines and models available to the research community and

developed an easy-to-use pipeline with a graphical user interface (GUI) and thorough documentation for making it accessible to users without programming knowledge at: https://hypermapp3r.readthedocs.io.

Acknowledgements

We are grateful for the support of the Medical Imaging Trial Network of Canada (MITNEC) Grant #NCT02330510, and the following site Principal investigators: Christian Bocti, Michael Borrie, Howard Chertkow, Richard Frayne, Robin Hsiung, Robert Laforce, Jr., Michael D. Noseworthy, Frank S. Prato, Demetrios J. Sahlas, Eric E. Smith, Vesna Sossi, Alex Thiel, Jean-Paul Soucy, and Jean-Claude Tardif. We are also grateful for the support of the Canadian Atherosclerosis Imaging Network (CAIN) (http://www.canadianimagingnetwork.org/), and the following investigators: Therese Heinonen, Rob Beanlands, David Spence, Philippe L'Allier, Brian Rutt, Aaron Fenster, Matthias Friedrich, Ben Chow, and Richard Frayne. This research was conducted with the support of the Ontario Brain Institute, an independent non-profit corporation, funded partially by the Ontario government. The opinions, results and conclusions are those of the authors and no endorsement by the Ontario Brain Institute is intended or should be inferred.

Declarations

Funding
This study was funded by the Canadian Institute for Health Research (CIHR) MOP Grant #13129, CIHR Foundation grant #159910, Ontario Brain Institute and the L.C Campbell Foundation. RHS is supported by a Heart and Stroke Clinician-Scientist Phase II Award. The work was also supported by the Medical Imaging Trial Network of Canada (MITNEC) Grant #NCT02330510. Matching funds were provided by participant hospital and research foundations, including the Baycrest Foundation, Bruyere Research Institute, Centre for Addiction and Mental Health Foundation, London Health Sciences Foundation, McMaster University Faculty of Health Sciences, Ottawa Brain and Mind Research Institute, Queen's University Faculty of Health Sciences, St. Michael's Hospital, Sunnybrook Health Sciences Centre Foundation, the Thunder Bay Regional Health Sciences Centre, University Health Network, the University of Ottawa Faculty of Medicine, and the Windsor/Essex County ALS Association. The Temerty Family Foundation provided the major infrastructure matching funds.

Conflicts of Interest
The authors declare that they have no conflicts of interest.

Data Accessibility
The developed algorithm and trained models (network weights) are publicly available at: https://hypermapp3r.readthedocs.io under the GNU General Public License v3.0. An example dataset is included for testing purposes. We have developed an easy-to-use pipeline with a GUI and thorough documentation for making it accessible to users without programming knowledge.

References

Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. "Weight Uncertainty in Neural Network." In *Proceedings of the 32nd International Conference on Machine Learning*, edited by Francis Bach and David Blei, 37:1613–22. Proceedings of Machine Learning Research. Lille, France: PMLR.

Carmichael, Owen, Christopher Schwarz, David Drucker, Evan Fletcher, Danielle Harvey, Laurel Beckett, Clifford R. Jack Jr, Michael Weiner, Charles DeCarli, and Alzheimer's Disease Neuroimaging Initiative. 2010. "Longitudinal Changes in White Matter Disease and Cognition in the First Year of the Alzheimer Disease Neuroimaging Initiative." *Archives of Neurology* 67 (11): 1370–78.

Çiçek, Özgün, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation." In *Lecture Notes in Computer Science*, 424–32.

Dadar, Mahsa, Josefina Maranzano, Simon Ducharme, Owen T. Carmichael, Charles Decarli, D. Louis Collins, and Alzheimer's Disease Neuroimaging Initiative. 2018. "Validation of T1w-Based Segmentations of White Matter Hyperintensity Volumes in Large-Scale Datasets of Aging." *Human Brain Mapping* 39 (3): 1093–1107.

Debette, Stéphanie, and H. S. Markus. 2010. "The Clinical Importance of White Matter Hyperintensities on Brain Magnetic Resonance Imaging: Systematic Review and Meta-Analysis." *BMJ* 341 (July): c3666.

Farhan, Sali M. K., Robert Bartha, Sandra E. Black, Dale Corbett, Elizabeth Finger, Morris Freedman, Barry Greenberg, et al. 2017. "The Ontario Neurodegenerative Disease Research Initiative (ONDRI)." *The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques* 44 (2): 196–202.

Gal, Yarin, and Zoubin Ghahramani. 2016. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In *International Conference on Machine Learning*, 1050–59.

Goubran, Maged, Emmanuel Edward Ntiri, Hassan Akhavein, Melissa Holmes, Sean Nestor, Joel Ramirez, Sabrina Adamo, et al. 2019. "Hippocampal Segmentation for Brains with Extensive Atrophy Using Three-dimensional Convolutional Neural Networks." *Human Brain Mapping*. https://doi.org/10.1002/hbm.24811.

Graves, Alex. 2011. "Practical Variational Inference for Neural Networks." In *Advances in Neural Information Processing Systems*, 2348–56. Citeseer.

Griffanti, Ludovica, Giovanna Zamboni, Aamira Khan, Linxin Li, Guendalina Bonifacio, Vaanathi Sundaresan, Ursula G. Schulz, et al. 2016. "BIANCA (Brain Intensity AbNormality Classification Algorithm): A New Tool for Automated Segmentation of White Matter Hyperintensities." *NeuroImage* 141 (November): 191–205.

Gudbjartsson, H., and S. Patz. 1995. "The Rician Distribution of Noisy MRI Data." *Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine* 34 (6): 910–14.

Guerrero, R., C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, et al. 2018. "White Matter Hyperintensity and Stroke Lesion Segmentation and Differentiation Using Convolutional Neural Networks." *NeuroImage. Clinical* 17: 918–34.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–78.

Hernández, M. C. V., R. J. Piper, and M. E. Bastin. 2014. "Morphologic, Distributional, Volumetric, and Intensity Characterization of Periventricular Hyperintensities." *American Journal*. http://www.ajnr.org/content/35/1/55.short.

Herrmann, L. L., M. Le Masurier, and K. P. Ebmeier. 2008. "White Matter Hyperintensities in Late Life Depression: A Systematic Review." *Journal of Neurology, Neurosurgery, and Psychiatry* 79 (6): 619–24.

Heuvel, D. M. J. van den. 2006. "Increase in Periventricular White Matter Hyperintensities Parallels Decline in Mental Processing Speed in a Non-Demented Elderly Population." *Journal of Neurology, Neurosurgery & Psychiatry*. https://doi.org/10.1136/jnnp.2005.070193.

Hong, Jisu, Bo-Yong Park, Mi Ji Lee, Chin-Sang Chung, Jihoon Cha, and Hyunjin Park. 2020. "Two-Step Deep Neural Network for Segmentation of Deep White Matter Hyperintensities in Migraineurs." *Computer Methods and Programs in Biomedicine* 183 (January): 105065.

Ioffe, Sergey, and Christian Szegedy. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1502.03167.

Jeon, Seun, Uicheul Yoon, Jun-Sung Park, Sang Won Seo, Jung-Hyun Kim, Sung Tae Kim, Sun I. Kim, Duk L. Na, and Jong-Min Lee. 2011. "Fully Automated Pipeline for Quantification and Localization of White Matter Hyperintensity in Brain Magnetic Resonance Image." *International Journal of Imaging Systems and*

Technology 21 (2): 193–200.
Jokinen, Hanna, Juha Koikkalainen, Hanna M. Laakso, Susanna Melkas, Tuomas Nieminen, Antti Brander, Antti Korvenoja, et al. 2020. "Global Burden of Small Vessel Disease-Related Brain Changes on MRI Predicts Cognitive and Functional Decline." *Stroke; a Journal of Cerebral Circulation* 51 (1): 170–78.
Kamnitsas, Konstantinos, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. 2017. "Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation." *Medical Image Analysis* 36 (February): 61–78.
Keith, Julia, Fu-Qiang Gao, Raza Noor, Alex Kiss, Gayathiri Balasubramaniam, Kelvin Au, Ekaterina Rogaeva, Mario Masellis, and Sandra E. Black. 2017. "Collagenosis of the Deep Medullary Veins: An Underrecognized Pathologic Correlate of White Matter Hyperintensities and Periventricular Infarction?" *Journal of Neuropathology and Experimental Neurology* 76 (4): 299–312.
Kendall, Alex, Vijay Badrinarayanan, and Roberto Cipolla. 2017. "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding." *Procedings of the British Machine Vision Conference 2017*. https://doi.org/10.5244/c.31.57.
Kendall, Alex, and Yarin Gal. 2017. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5574–84. Curran Associates, Inc.
Kilgore, Collin B., Jeremy F. Strain, Tammie L. S. Benzinger, John C. Morris, and Beau M. Ances. 2020. "Increases in Periventricular White Matter Hyperintensities Associate with Alzheimer Disease Biomarkers: Neuroimaging/multi-Modal Comparisons." *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 16: e041286.
Kingma, Diederik P., and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980.
Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2016. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1612.01474.
LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44.
Lee, Seonjoo, Fawad Viqar, Molly E. Zimmerman, Atul Narkhede, Giuseppe Tosto, Tammie L. S. Benzinger, Daniel S. Marcus, et al. 2016. "White Matter Hyperintensities Are a Core Feature of Alzheimer's Disease: Evidence from the Dominantly Inherited Alzheimer Network." *Annals of Neurology*. https://doi.org/10.1002/ana.24647.
Lee, Seung-Jae, Joong-Seok Kim, Ji-Yeon Yoo, In-Uk Song, Bum-Soo Kim, So-Lyung Jung, Dong-Won Yang, Yeong-In Kim, Du-Shin Jeong, and Kwang-Soo Lee. 2010. "Influence of White Matter Hyperintensities on the Cognition of Patients with Parkinson Disease." *Alzheimer Disease and Associated Disorders* 24 (3): 227–33.
Li, Hongwei, Gongfa Jiang, Jianguo Zhang, Ruixuan Wang, Zhaolei Wang, Wei-Shi Zheng, and Bjoern Menze. 2018. "Fully Convolutional Network Ensembles for White Matter Hyperintensities Segmentation in MR Images." *NeuroImage* 183 (December): 650–65.
Marcotte, Karine, Naida L. Graham, Kathleen C. Fraser, Jed A. Meltzer, David F. Tang-Wai, Tiffany W. Chow, Morris Freedman, Carol Leonard, Sandra E. Black, and Elizabeth Rochon. 2017. "White Matter Disruption and Connected Speech in Non-Fluent and Semantic Variants of Primary Progressive Aphasia." *Dementia and Geriatric Cognitive Disorders Extra* 7 (1): 52–73.
Milletari, F., N. Navab, and S. Ahmadi. 2016. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation." In *2016 Fourth International Conference on 3D Vision (3DV)*, 565–71.
Nalepa, Jakub, Michal Marcinkiewicz, and Michal Kawulok. 2019. "Data Augmentation for Brain-Tumor Segmentation: A Review." *Frontiers in Computational Neuroscience* 13 (December): 83.
Neal, Radford M. 2012. *Bayesian Learning for Neural Networks*. Springer Science & Business Media.
Ntiri, Emmanuel E., Melissa F. Holmes, Parisa M. Forooshani, Joel Ramirez, Fuqiang Gao, Miracle Ozzoude, Sabrina Adamo, et al. 2021. "Improved Segmentation of the Intracranial and Ventricular Volumes in Populations with Cerebrovascular Lesions and Atrophy Using 3D CNNs." *Neuroinformatics*, February. https://doi.org/10.1007/s12021-021-09510-1.
Paschali, Magdalini, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. 2018. "Generalizability vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples." In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 493–501. Springer International Publishing.
Pearson, Karl, and Francis Galton. 1895. "VII. Note on Regression and Inheritance in the Case of Two Parents." *Proceedings of the Royal Society of London* 58 (347-352): 240–42.
Prins, Niels D., and Philip Scheltens. 2015. "White Matter Hyperintensities, Cognitive Impairment and Dementia: An Update." *Nature Reviews. Neurology* 11 (3): 157–65.

Ramirez, J., E. Gibson, A. Quddus, N. J. Lobaugh, A. Feinstein, B. Levine, C. J. M. Scott, N. Levy-Cooperman, F. Q. Gao, and S. E. Black. 2011. "Lesion Explorer: A Comprehensive Segmentation and Parcellation Package to Obtain Regional Volumetrics for Subcortical Hyperintensities and Intracranial Tissue." *NeuroImage* 54 (2): 963–73.

Ramirez, Joel, Christopher J. M. Scott, Alicia A. McNeely, Courtney Berezuk, Fuqiang Gao, Gregory M. Szilagyi, and Sandra E. Black. 2014. "Lesion Explorer: A Video-Guided, Standardized Protocol for Accurate and Reliable MRI-Derived Volumetrics in Alzheimer's Disease and Normal Elderly." *Journal of Visualized Experiments: JoVE*, no. 86 (April). https://doi.org/10.3791/50887.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In *Lecture Notes in Computer Science*, 234–41.

Rozsa, A., M. Gunther, and T. E. Boult. 2018. "Towards Robust Deep Neural Networks with BANG." In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 803–11.

Smith, Eric E., Svetlana Egorova, Deborah Blacker, Ronald J. Killiany, Alona Muzikansky, Bradford C. Dickerson, Rudolph E. Tanzi, Marilyn S. Albert, Steven M. Greenberg, and Charles R. G. Guttmann. 2008. "Magnetic Resonance Imaging White Matter Hyperintensities and Brain Volume in the Prediction of Mild Cognitive Impairment and Dementia." *Archives of Neurology* 65 (1): 94–100.

Sundaresan, Vaanathi, Giovanna Zamboni, Campbell Le Heron, Peter M. Rothwell, Masud Husain, Marco Battaglini, Nicola De Stefano, Mark Jenkinson, and Ludovica Griffanti. 2019. "Automated Lesion Segmentation with BIANCA: Impact of Population-Level Features, Classification Algorithm and Locally Adaptive Thresholding." *NeuroImage* 202 (November): 116056.

Swardfager, Walter, Di Yu, Gustavo Scola, Hugo Cogo-Moreira, Parco Chan, Yi Zou, Nathan Herrmann, et al. 2017. "Peripheral Lipid Oxidative Stress Markers Are Related to Vascular Risk Factors and Subcortical Small Vessel Disease." *Neurobiology of Aging* 59 (November): 91–97.

Tustison, Nicholas J., Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. 2010. "N4ITK: Improved N3 Bias Correction." *IEEE Transactions on Medical Imaging* 29 (6): 1310–20.

Wang, Guotai, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. 2019. "Aleatoric Uncertainty Estimation with Test-Time Augmentation for Medical Image Segmentation with Convolutional Neural Networks." *Neurocomputing* 335 (September): 34–45.

Wang, Yanbo, Joseree Ann Catindig, Saima Hilal, Hock Wei Soon, Eric Ting, Tien Yin Wong, Narayanaswamy Venketasubramanian, Christopher Chen, and Anqi Qiu. 2012. "Multi-Stage Segmentation of White Matter Hyperintensity, Cortical and Lacunar Infarcts." *NeuroImage* 60 (4): 2379–88.

Wardlaw, Joanna M., Maria C. Valdés Hernández, and Susana Muñoz-Maniega. 2015. "What Are White Matter Hyperintensities Made Of?" *Journal of the American Heart Association*. https://doi.org/10.1161/jaha.114.001140.

Wu, Jiong, Yue Zhang, Kai Wang, and Xiaoying Tang. 2019. "Skip Connection U-Net for White Matter Hyperintensities Segmentation From MRI." *IEEE Access*. https://doi.org/10.1109/access.2019.2948476.

Yoo, Byung Il, Jung Jae Lee, Ji Won Han, San Yeo Wool Oh, Eun Young Lee, James R. MacFall, Martha E. Payne, Tae Hui Kim, Jae Hyoung Kim, and Ki Woong Kim. 2014. "Application of Variable Threshold Intensity to Segmentation for White Matter Hyperintensities in Fluid Attenuated Inversion Recovery Magnetic Resonance Images." *Neuroradiology* 56 (4): 265–81.

Yu, Fisher, and Vladlen Koltun. 2015. "Multi-Scale Context Aggregation by Dilated Convolutions." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1511.07122.

Zhang, Zicong, Kimerly Powell, Changchang Yin, Shilei Cao, Dani Gonzalez, Yousef Hannawi, and Ping Zhang. 2020. "Brain Atlas Guided Attention U-Net for White Matter Hyperintensity Segmentation." *arXiv [eess.IV]*. arXiv. http://arxiv.org/abs/2010.09586.

Zhu, Yinhao, and Nicholas Zabaras. 2018. "Bayesian Deep Convolutional Encoder–decoder Networks for Surrogate Modeling and Uncertainty Quantification." *Journal of Computational Physics* 366 (August): 415–47.