



Challenge Report

Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images



Vaanathi Sundaresan^{a,b,c,*}, Giovanna Zamboni^{a,d,e}, Peter M. Rothwell^d, Mark Jenkinson^{a,f,1}, Ludovica Griffanti^{a,g,1}

^a Wellcome Centre for Integrative Neuroimaging, Oxford Centre for Functional MRI of the Brain, Nuffield Department of Clinical Neurosciences, University of Oxford, UK

^b Oxford-Nottingham Centre for Doctoral Training in Biomedical Imaging, University of Oxford, UK

^c Oxford India Centre for Sustainable Development, Somerville College, University of Oxford, UK

^d Centre for Prevention of Stroke and Dementia, Nuffield Department of Clinical Neurosciences, University of Oxford, UK

^e Dipartimento di Scienze Biomediche, Metaboliche e Neuroscienze, Università di Modena e Reggio Emilia, Italy

^f Australian Institute for Machine Learning (AIML), School of Computer Science, The University of Adelaide, Adelaide, Australia

^g Wellcome Centre for Integrative Neuroimaging, Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford, Oxford, UK

ARTICLE INFO

Article history:

Received 29 October 2020

Revised 10 March 2021

Accepted 16 July 2021

Available online 18 July 2021

Keywords:

Deep learning

White matter hyperintensities

U-Nets

Segmentation

MRI

ABSTRACT

White matter hyperintensities (WMHs) have been associated with various cerebrovascular and neurodegenerative diseases. Reliable quantification of WMHs is essential for understanding their clinical impact in normal and pathological populations. Automated segmentation of WMHs is highly challenging due to heterogeneity in WMH characteristics between deep and periventricular white matter, presence of artefacts and differences in the pathology and demographics of populations. In this work, we propose an ensemble triplanar network that combines the predictions from three different planes of brain MR images to provide an accurate WMH segmentation. In the loss functions the network uses anatomical information regarding WMH spatial distribution in loss functions, to improve the efficiency of segmentation and to overcome the contrast variations between deep and periventricular WMHs. We evaluated our method on 5 datasets, of which 3 are part of a publicly available dataset (training data for MICCAI WMH Segmentation Challenge 2017 - MWSC 2017) consisting of subjects from three different cohorts, and we also submitted our method to MWSC 2017 to be evaluated on the unseen test datasets. On evaluating our method separately in deep and periventricular regions, we observed robust and comparable performance in both regions. Our method performed better than most of the existing methods, including FSL BIANCA, and on par with the top ranking deep learning methods of MWSC 2017.

© 2021 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

1. Introduction

White matter hyperintensities (WMHs) of presumed vascular origin appear as bright localised areas in T2-weighted and fluid-attenuated inversion recovery (FLAIR) images (Wardlaw et al., 2013), and could appear hypointense on T1-weighted images. WMHs occur commonly in patients with cerebrovascular diseases (Li et al., 2013; Simoni et al., 2012; Zamboni et al., 2019) and have been associated with cognitive decline, atrophy and neurodegenerative diseases such as dementia (DeBette et al., 2010; Prins

and Scheltens, 2015; Pantoni et al., 2005). However, they are also found commonly in healthy elderly subjects (Wardlaw et al., 2013). Therefore, the relationship between the occurrence of WMHs and various clinical factors is not yet fully understood. While various visual rating scales are available (Fazekas et al., 1987; Scheltens et al., 1993; Wahlund et al., 2001) and can provide qualitative or categorical information regarding WMHs, these scales provide limited information regarding their spatial distribution. Voxel-wise WMH maps, on the other hand, enable more precise quantification of WMHs and open up the possibility of studying the relationship between the spatial location/distribution of WMHs and various clinical factors. In turn, this helps to identify patterns of normal and pathological ageing (Rostrup et al., 2012; Biesbroek et al., 2013). Hence, location and volume-based lesion characterisation are being increasingly considered in clinical setting (Wardlaw et al.,

* Corresponding author.

E-mail address: vaanathi.sundaresan@ndcn.ox.ac.uk (V. Sundaresan).

URL: <https://www.ndcn.ox.ac.uk/team/vaanathi-sundaresan> (V. Sundaresan)

¹ Contributed equally to this work

2013; Smith et al., 2019). Given the importance of analysing the clinical impact of WMHs, especially in large cohorts, manual segmentation of WMHs is time consuming and is prone to intra/inter-rater variability. Hence, an automated method to provide exact voxel-level localisation and accurate quantification of WMHs would be highly useful.

Several automated WMH segmentation methods have been proposed (Caligiuri et al., 2015) using features based on intensity (Ong et al., 2012; Damangir et al., 2012), combined with anatomy (Ong et al., 2012; Damangir et al., 2012) and appearance (such as shape, contrast etc.) (Samaille et al., 2012; Griffanti et al., 2016). Among the existing methods using hand-crafted features, unsupervised methods such as clustering (Admiraal-Behloul et al., 2005; Ong et al., 2012; Samaille et al., 2012), supervised classification algorithms (Anbeek et al., 2004; Damangir et al., 2012; Yoo et al., 2014; Griffanti et al., 2016) and probabilistic approaches (Yang et al., 2010) have been proposed. Despite the large amount of methods developed for WMH segmentation, only a few of them are publicly available (Damangir et al., 2012; Lao et al., 2008; Schmidt et al., 2012; Griffanti et al., 2016). Using hand-crafted features might not be sufficient to capture the lesion patterns and to overcome noise and artefacts. The lesion characteristics of WMHs show high variability depending on their location, making their segmentation challenging. For instance, between periventricular WMHs (PWMHs) and deep WMHs (DWMHs) (Griffanti et al., 2017), PWMHs usually appear brighter, larger and often form confluent lesions, with higher contrast when compared to DWMHs that usually occur as small punctate lesions. Also, the lesion load and distribution are influenced by demographic factors (e.g. age) and clinical conditions (e.g. hypertension). Additionally, artefacts (e.g. Gibbs ringing, motion artefacts) and noise that occur during image acquisition also affect the segmentation performance. Also, most of the methods have been evaluated on a limited number of subjects (Anbeek et al., 2004; De Boer et al., 2009; Yang et al., 2010; Steenwijk et al., 2013; Khademi et al., 2011; Yoo et al., 2014) or on a specific population with limited demographic and pathological characteristics (Gibson et al., 2010; Jeon et al., 2011).

Deep learning (DL) allows computational models with multiple layers to learn data representations at different layers of abstraction (LeCun et al., 2015), thus utilising more contextual information compared to the hand-crafted features. With increase in computational resources, such as graphical processing units (GPUs) and techniques such as data augmentation (Russakovsky et al., 2015), DL has been quickly emerging as a reliable segmentation tool in biomedical imaging, especially for lesion segmentation (Guerrero et al., 2018; Ghafoorian et al., 2017). For instance, in the MICCAI WMH Segmentation Challenge 2017 (MWSC 2017, <https://wmh.isi.uu.nl>, Kuijff et al. (2019)), out of 20 methods competing in the initial call, 14 of them (including the top ranking methods) were based on DL. Existing DL methods for WMH segmentation have used convolutional neural network (CNN) models, including various ensemble models (Li et al., 2018; Kuijff et al., 2019), encoder-decoder models (Li et al., 2013; Guerrero et al., 2018; Kuijff et al., 2019; Zhang et al., 2018) - especially U-Nets (proposed by Ronneberger et al. (2015)), 3D multi-dimensional gated recurrent networks (Andermatt et al., 2016) and ResNets (Guerrero et al., 2018). The common choices of inputs used for these networks have been 2D slices (Li et al., 2018) or small 2D/3D patches at multiple scales (Ghafoorian et al., 2017; Andermatt et al., 2016). In general, the choice of the model dimension is affected by various factors, such as size and distribution of the lesions and the amount of data available. Various architectures of 3D CNN models have been successfully used in the segmentation of various types of larger lesions (Kamnitsas et al., 2017; Havaei et al., 2017; Oktay et al., 2018). However, images with WMHs include small lesions in the deep regions, with low contrast and poor context, often with

poor resolution along the z-dimension. This constrains the accurate detection of WMH boundary voxels using 3D CNNs (Li et al., 2018). From the implementation point of view, 2D models use far fewer parameters when compared to 3D models. However, using 2D models can cause discontinuities in segmentation across the z-dimension since individual slices are considered separately. Therefore, to leverage the advantages of both 2D and 3D models, methods utilizing a 2.5D architecture, or with 2D models applied individually on each of the 3 planes of the image, have been proposed for various segmentation applications (Aslani et al., 2019; Piantadosi et al., 2020; Prasoon et al., 2013; Roth et al., 2014). These triplanar approaches have shown to avoid parameters explosion (Piantadosi et al., 2020), as in the case of 3D models, and provide better segmentation and continuity across slices than the 2D models. Another important aspect that changes widely between different datasets and can impact the model performance is voxel resolution. Most of the DL methods (including 3D and 2D methods) have been evaluated on images with isotropic voxels, mainly with axial acquisitions (Ghafoorian et al., 2017; Kuijff et al., 2019), while acquisition protocol variations (with anisotropic voxels and/or different axis of acquisition) are common in the real-world clinical datasets. Regarding the method accessibility, in the case of DL methods, not many are publicly available. In fact, even if most of the pre-trained models from MWSC 2017 challenge are publicly available for testing, there are no independent DL tools available that allow users to train/fine-tune the models on their own data for improving segmentation performance on various datasets from different scanners/centres.

We aimed to develop a DL tool that provides highly accurate segmentation in both periventricular and deep regions, that is publicly available, and that has the flexibility to change training hyperparameters options on various datasets. In this work, we propose TrUE-Net (Triplanar U-Net ensemble network), a DL method for segmentation of WMHs, consisting of an ensemble of U-Nets, each applied to one of the three planes (axial, sagittal and coronal) planes of structural brain MR images. We aim to improve WMH segmentation irrespective of lesion location and lesion load by training the TrUE-Net using loss functions that take into account the anatomical location and distribution of WMHs. We evaluate the proposed model on 5 different datasets with different acquisition (scanner and MRI protocol) and lesion characteristics: one from a study on neurodegeneration in prodromal and manifest Alzheimer's disease, one from a vascular cohort (coronal images with anisotropic voxels), and three from a publicly available training dataset from MWSC 2017. Additionally, our method was also evaluated on an unseen test data in MWSC 2017, which includes, in addition to data from the same centres as that of the publicly available MWSC training data, data acquired using two different scanners (a 1.5T and a PET-MR system). We compared the performance of TrUE-Net against methods using hand-crafted features and DL methods, with respect to manual segmentations. Initially, we performed a direct comparison between the results of TrUE-Net and BIANCA, the existing FSL (FMRIB software library) WMH segmentation tool (Griffanti et al., 2016). Later, we compared the TrUE-Net results with those of the top performing method (Li et al., 2018) from the MWSC 2017 on various datasets. Finally, we performed indirect comparisons of TrUE-Net results against various existing methods.

2. Materials and methods

2.1. Triplanar U-Net Ensemble Network (TrUE-Net)

2.1.1. Preprocessing

We used both T1-weighted and FLAIR images as inputs for the model. We reoriented the images to the standard MNI space, per-

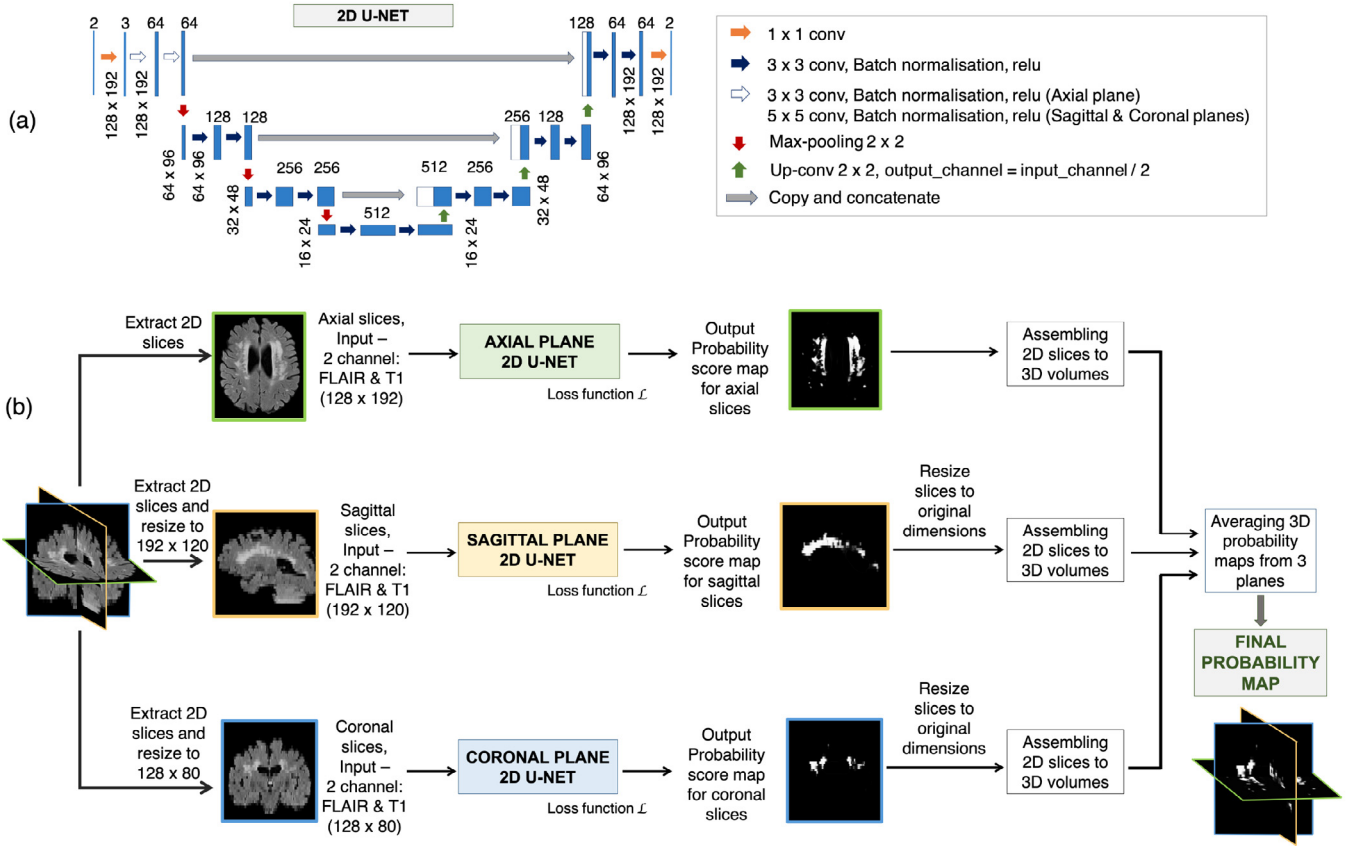


Fig. 1. Triplanar U-Net ensemble network (TrUE-Net). (a) U-Net model used in individual planes, (b) Overall TrUE-Net architecture.

formed skull-stripping with FSL BET (Smith, 2002) and bias field correction using FSL FAST (Zhang et al., 2001). We registered the T1-weighted image to the FLAIR using linear rigid-body registration (Jenkinson and Smith, 2001) and cropped the field of vision (FOV) close to the brain and applied Gaussian normalisation to normalise the intensity values. We then extracted 2D slices from the volumes from all three planes. For the axial plane, we cropped the slices to a dimension of 128×192 voxels. For sagittal and coronal slices, we cropped and resized the extracted slices to 192×120 and 128×80 voxels respectively, using bilinear interpolation.

2.1.2. TrUE-Net architecture

The proposed triplanar architecture consists of three 2D networks, each one detecting WMHs from a different plane. The triplanar network reduces discontinuities in WMH segmentation across slices and provides better and comprehensive lesion boundary delineation, using fewer parameters compared to a 3D CNN.

In TrUE-Net, we combined three 2D U-Nets in parallel within an ensemble model. In the ensemble architecture, variation in the individual probability maps (due to noise or spurious structure) is reduced when they are combined in the ensemble network.

Fig. 1 shows the architecture of the proposed TrUE-Net. For each plane, the 2D model takes FLAIR and T1-weighted slices as input channels and provides the probability map in the corresponding plane. In each plane, we trimmed the depth of the classic U-Net (Ronneberger et al., 2015) to obtain a 3-layer deep U-Net model. This reduces the computational load and improves the model sensitivity towards small lesions. Our model mostly uses 3×3 convolutional kernels, except for the initial 5×5 convolutional kernels in the first layer of the sagittal and coronal U-Nets

(Fig. 1(a)), since larger receptive fields could aid in learning more generic lesion patterns in these planes, thus reducing discontinuities across slices. Each convolution layer is followed by a batch normalisation layer and an activation layer (using *ReLU* - rectified linear unit). We added a 1×1 convolutional kernel at the end, before the softmax layer for predicting the probability maps. In the ensemble model, training of U-Nets in the individual planes occurs independently, using the slices extracted from the corresponding planes from the resized training images. During testing, for each network we assembled the slices into a 3D probability map. We then resized each 3D map back to the original dimension and finally averaged the three 3D maps to obtain the final probability map.

2.1.3. Loss function

We used a weighted sum of the voxel-wise cross-entropy (CE) loss function and the Dice loss (DCL) as the total cost function. The CE loss aims to make the segmentation better at the image-level and is biased towards the detection of larger periventricular WMHs. Hence, we weighted the CE loss function using a spatial weight map (an example shown in Fig. 2) to up-weight the deep areas that have high class imbalance (i.e. contain fewer WMHs compared to the background). The addition of the Dice loss also helps with deep WMHs, since missing small WMHs would make more difference to the Dice component than to the CE loss component. Hence, training the network with the inclusion of a Dice component in the loss function would favour finding small lesions and reducing false negatives, especially in the areas of high class imbalance (Milletari et al., 2016; Li et al., 2018).

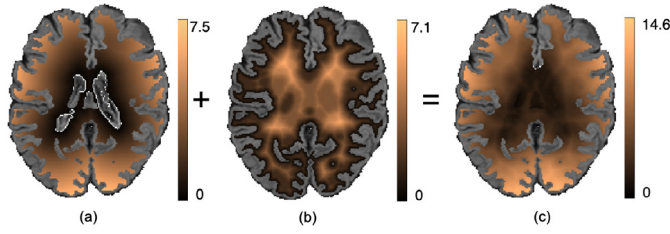


Fig. 2. Example weight maps for weighting the voxel-wise cross-entropy loss function. Maps showing the sum of (a) Distance from ventricles D_{vent} and (b) Distance from gray matter (GM) D_{GM} to get (c) Final weight map D_{wei} . The spatial weights (distances) are calculated for the whole brain. Here, the weight maps are shown after applying the WM mask from the post-processing step (see Section 2.1.4) for illustration purpose..

For each subject, we determined the distance from ventricles using the FSL `distancemap` command, $0 \leq D_{vent} \leq N_D$ (Fig. 2a), where N_D is the maximum distance from ventricles to the boundary of the brain mask, and the distance from the brain gray matter (GM, derived from the cortical CSF segmentation obtained with FSL `FAST`²), $0 \leq D_{GM} \leq M_D$ (Fig. 2b), where M_D is the maximum distance from the GM boundary to the centre of the brain. We then obtained the final weight map as the sum of both, $D_{wei} = D_{vent} + D_{GM}$, $0 \leq D_{wei} \leq (N_D + M_D)$ (Fig. 2c) (all distances in mm). The sum of these two distance values ensures that the deep region receives higher weights than the periventricular region, and simultaneously avoids relying on GM segmentation that could contain misclassified WMH voxels. Hence the total weighted CE loss for N voxels is given by,

$$CE = - \sum_{x=1}^N \sum_{i=1}^C y(x) D_{wei}(x) \log(p(x)) \quad (1)$$

where $p(x)$ denotes the output of the soft-max layer, C is the number of classes and $y(x) = 0$ or 1 , the value at each voxel x on the manual segmentation. Given the manual segmentation M and binary map P_{th} obtained by thresholding predicted probability maps, the Dice loss for N voxels is given by,

$$DcL = - \frac{2 \times \sum_{x=1}^N M(x) \cdot P_{th}(x)}{\sum_{x=1}^N M(x) + \sum_{x=1}^N P_{th}(x)} \quad (2)$$

Hence the total loss function L is given by,

$$L = CE + DcL \quad (3)$$

$$= - \sum_{x=1}^N \sum_{i=1}^C y(x) D_{wei}(x) \log(p(x)) - \frac{2 \times \sum_{x=1}^N M(x) \cdot P_{th}(x)}{\sum_{x=1}^N M(x) + \sum_{x=1}^N P_{th}(x)} \quad (4)$$

We chose equal weights on the basis of initial empirical experiments (not shown) for the Dice loss and the weighted CE loss while adding them to obtain the total loss. The range of CE loss function is ideally $0 \leq CE \text{ loss} \leq W$ (where W is determined from the spatial weight map) and the range of Dice loss is $-1 \leq \text{Dice loss} \leq 0$. Therefore, the range of the sum of these two loss values is $-1 \leq (CE \text{ loss} + \text{Dice loss}) \leq W$. Although the total loss function can assume negative values, in practice, the CE loss hardly converged at the value close to 0 and applying weight values (W) resulted in CE loss values > 0 at convergence. Therefore, even with the addition of Dice loss, the total loss value converged around 0 (or small

negative values, e.g. ≥ -0.1). However, it must be noted that during optimisation, it is not the value of the loss that is important but the change in loss across parameter space, since the weights of the model are updated using the gradient of the loss and not the loss value itself.

2.1.4. Post-processing

We padded the output probability maps with zeros to bring them back to their original dimensions. We later masked the probability map with the white matter mask obtained from a dilated and inverted cortical CSF tissue segmentation (using FSL `FAST` (Zhang et al., 2001)) combined with other deep grey exclusion masks (`make_bianca_mask` command in FSL `BIANCA` (Griffanti et al., 2016)). Finally, we thresholded the masked probability maps at 0.5.

2.1.5. Implementation details

We implemented the network in Python 3.6 using Pytorch 1.2.0. The network was trained on an NVIDIA Tesla V100, taking 45 seconds (for 3 planes) per epoch for 15,000 samples with the training/validation split of 90%/10%. For each leave-one-out (LOO) evaluation, we excluded the test subject and randomly sampled 10% of the remaining subjects as the validation dataset while using the 90% for training the model. We used a patience (number of epochs to wait for progress on validation set) value of 20 to determine the convergence (early stopping). The model converged at around 80 epochs for all the datasets. We used the Adam Optimiser with $\epsilon = 10^{-4}$. We used a batch size of 8, with an initial learning rate of 1×10^{-3} and reducing it by a factor 1×10^{-1} every 2 epochs, until it reaches 1×10^{-5} , after which we maintain the fixed learning rate value. We chose the above parameters empirically based on the model convergence (refer to Section 2.3.2 for more details). Data augmentation was applied using translation (x/y -offset $\in [-10, 10]$), rotation ($\theta \in [-10, 10]$), random noise injection (Gaussian, $\mu = 0$, $\sigma^2 \in [0.01, 0.09]$) and Gaussian filtering ($\sigma \in [0.1, 0.3]$), increasing the dataset by a factor of 10 and 6 for axial and sagittal/coronal planes respectively. The hyperparameter values for the data augmentation transformations were randomly sampled from the closed intervals specified above using a uniform distribution.

2.2. Datasets

2.2.1. Neurodegenerative cohort (NDGEN)

The dataset, used in Zamboni et al. (2013), includes MRI data from 9 subjects with probable Alzheimer's Disease, 5 with amnesic mild cognitive impairment and 7 cognitively healthy control subjects (age range 63 - 86 years; mean age 77.1 ± 5.8 years; median age 77 years; F:M = 10:11). Total brain volume range: 1,189,282 - 1,614,799 mm³, median: 1,424,669 mm³. Manual segmentation was available for all datasets (WMH load range: 1,878 - 89,259 mm³, median: 20,772 mm³). The images were acquired using a 3T Siemens Trio Scanner, with FLAIR (TR/TE = 9,000/89 ms, flip angle 150°, FOV 220 mm, voxel size $1.1 \times 0.9 \times 3$ mm, matrix size $256 \times 256 \times 35$ voxels) and T1-weighted sequence (3D MP-RAGE sequence, TR/TE = 2,040/4.7 ms, flip angle 8°, FOV 192 mm, voxel size 1 mm isotropic, matrix size $174 \times 192 \times 192$ voxels).

2.2.2. Vascular cohort - Oxford vascular study (OXVASC)

The dataset consists of 18 participants in the OXVASC study (Rothwell et al., 2004), who had recently experienced a minor non-disabling stroke or transient ischemic attack (age range 50 - 91 years; mean age 73.27 ± 12.32 years; median age 75.5 years; F:M = 7:11). Total brain volume range: 1,290,926 - 1,918,604 mm³, median: 1568233 mm³. Manual segmentation was available for all datasets (WMH load range: 3,530 - 83,391 mm³, median: 16,906 mm³). The images were acquired using a 3T

² A white matter mask was first obtained from a dilated and inverted cortical CSF tissue segmentation using FSL `FAST`. Then the WM was subtracted from the brain mask to obtain the GM mask.

Siemens Trio Scanner, with FLAIR (TR/TE = 9,000/88 ms, flip angle 150°, voxel size $1 \times 3 \times 1$ mm, matrix size $174 \times 52 \times 192$ voxels), T1-weighted sequence (3D MP-RAGE sequence, TR/TE = 2,000/1.94 ms, flip angle 8°, voxel size 1 mm isotropic, matrix size $208 \times 256 \times 256$ voxels) and diffusion-weighted imaging (TR/TE = 8,000/86 ms, GRAPPA factor 2, flip angle 16°, FOV 192 mm, voxel size $2 \times 2 \times 2$ mm, 32 directions, b value 1,500 s/mm²).

2.2.3. MICCAI WMH Segmentation challenge training dataset (MWSC)

The dataset consists of 60 subjects from three different sources (20 subjects each) provided as training sets for the challenge (<http://wmh.isi.uu.nl/>): UMC Utrecht, NUHS Singapore and VU Amsterdam. The brain volume ranges: 1,257,820 - 1,844,920 mm³ (median 1,473,389 mm³) for UMC Utrecht, 1,147,248 - 1,532,268 mm³ (median: 1,351,325 mm³) for NUHS Singapore and 1,219,614 - 1,787,321 mm³ (median: 1,441,201 mm³) for VU Amsterdam. Manual segmentations were available for all three datasets, with an additional exclusion label provided for other pathologies. In the challenge, the masks with exclusion labels were ignored during performance evaluation. However, we included these masks as parts of non-lesion tissue, during both training and testing, for the calculation of the performance metrics in order to get more stringent evaluation in the presence of other pathologies. The WMH volume ranges (excluding other pathologies) are 845 - 74,991 mm³ (median: 26,240 mm³) for UMC Utrecht, 786 - 61,332 mm³ (median: 17,795 mm³) for NUHS Singapore and 1,522 - 43,528 mm³ (median: 6,015 mm³) for VU Amsterdam. For more details regarding MRI acquisition parameters, refer to <http://wmh.isi.uu.nl/>.

2.2.4. MWSC Test data

This is an in-house dataset used only for evaluation by the organisers, consisting of 110 subjects from five different sources. Out of the total number of subjects, 90 subjects were from the three sources (UMC Utrecht, NUHS Singapore and VU Amsterdam), 30 subjects from each source, mentioned above for the training data. The remaining 20 images were from VU Amsterdam using different scanners - 3T Philips Ingenuity and 1.5T GE Signa HDxt, with 10 subjects from each scanner. For more details regarding MRI acquisition parameters, refer to <http://wmh.isi.uu.nl/>.

We used the NDGEN dataset for the initial model optimisation, to observe the effect of hyperparameters, model dimension and loss function components (sections 2.3.2–2.3.4). We then evaluated the optimised model on the NDGEN, OXVASC and MWSC datasets using leave-one-out evaluation (Section 2.3.5) and compared it with exiting methods (2.3.6–2.3.7). Finally, we performed external validation of our method on the MWSC 2017 unseen test dataset (Section 2.3.5).

2.3. Experiments

2.3.1. Performance evaluation metrics

We used the following performance metrics in our evaluation:

- **Dice Similarity Index (SI)**, calculated as $2 \times (\text{true positive WMH voxels}) / (\text{true WMH voxels} + \text{positive WMH voxels})$.
- **Voxel-wise true positive rate (TPR)** is the ratio of the number of true positive WMH voxels to the number of true WMH voxels.
- **Voxel-wise false positive rate (FPR)** is the number of false positive (FP) WMH voxels divided by the number of non-WMH voxels.
- **Cluster-wise TPR** is the number of true positive WMH clusters divided by the total number of true WMH clusters.
- **Absolute volume difference (AVD) (%)** is the absolute difference between the volume of manually segmented WMHs voxels and the volume of detected WMHs voxels, as a percentage of the manually segmented lesion voxels.

- **Cluster-wise F1-measure** = $\frac{2 \times (\text{Cluster-wise TPR} \times \text{Cluster-wise precision})}{(\text{Cluster-wise TPR} + \text{Cluster-wise precision})}$, where Cluster-wise precision is the number of true positive WMH clusters divided by the total number of detected WMH clusters.
- **95th percentile of Hausdorff distance measure (H95)**: We determined the set of the closest distances between the points on the detected WMH boundary and the manually segmented WMH boundary. We calculated the 95th percentile of this set of closest distance values, instead of the maximum value used for the standard Hausdorff value (since the former is less prone to noise).

It is important to note that there is some degree of uncertainty associated with the delineation of boundaries of diffuse lesions in both manual segmentations and the predicted outputs, which could affect voxel-wise evaluation metrics and H95. Hence, we also used various cluster-wise metrics for evaluating the segmentation performance. In cluster-wise metrics we adopted a 26-connected neighbourhood to define clusters.

2.3.2. Effect of training hyperparameters on model performance

We used the NDGEN dataset for the initial optimisation of network parameters, and for determining the number of epochs. We explored the effect of batch-size, learning rate and epsilon (ϵ) value of the Adam optimiser on model convergence. We experimented with three batch sizes: 8, 16 and 32, and three ϵ values: 1×10^{-2} , 1×10^{-4} and 1×10^{-6} . For learning rate, we tested the effect of following 3 settings: (i) **Higher**: Initially 1×10^{-2} , reducing by a factor 1×10^{-1} every 2 epochs until it reaches 1×10^{-4} and keeping it constant afterwards, (ii) **Medium**: Initially 1×10^{-3} , reducing by a factor 1×10^{-1} every 2 epochs until it reaches 1×10^{-5} and (iii) **Lower**: Initially 1×10^{-4} , reducing by a factor 1×10^{-1} every 2 epochs until it reaches 1×10^{-6} .

2.3.3. Ablation study: Effect of loss function components

We performed an ablation study to determine the effect of the components of the loss function on the segmentation results using the NDGEN dataset. For this experiment, we removed one component of the loss function at a time and compared the performance of TrUE-Net with three cases of loss functions: cross-entropy loss, weighted cross-entropy loss and weighted cross-entropy loss + Dice loss. We also compared the weighted cross-entropy loss + Dice loss with the case of using only the Dice loss component. Since we hypothesized that the weighting of the CE loss would specifically improve the performance in deep regions, we determined SI values in periventricular and deep regions separately, and compared the results between regions. Due to the small sample size and the non-Gaussian distribution of the data in most cases (Shapiro-Wilks test), we performed non-parametric statistics using Wilcoxon signed rank test. We adopted the 10 mm distance rule (DeCarli et al., 2005; Griffanti et al., 2018) for classification of PWMH and DWMH.

2.3.4. Effect of model dimension

We also determined the effect of the dimension of the model on the segmentation performance using the NDGEN dataset. We removed one aspect of model dimension at a time and compared three cases: 3D U-Net, 2D U-Net (on axial plane only) and TrUE-Net. Similar to the above study, we determined SI values in periventricular and deep regions separately and performed Wilcoxon signed rank test to compare between regions.

We maintained the same training and model parameters for all three options in both studies.

2.3.5. Evaluation of WMH segmentation

After optimising the proposed model, we performed leave-one-out (LOO) evaluation of our proposed model separately on the

following datasets: (i) MWSC (combined from Utrecht, Singapore and Amsterdam cohorts), (ii) NDGEN and (iii) OXVASC, based on the performance metrics specified in Section 2.3.1. We chose to perform LOO evaluation rather than fold validation since the former provides more unbiased performance estimates (as it uses a larger training data (Elisseff et al., 2003)) and provides more reliable estimates in smaller datasets (e.g. NDGEN and OXVASC). For comparison, the results of 3-fold cross validation on the MWSC dataset are reported in the supplementary material. Also, we submitted a docker container of our method (trained on the MWSC dataset) to the MWSC 2017 for evaluating our method on the unseen MWSC test data. The submitted container was validated on the test datasets by the organisers and the results were provided for the individual test datasets, along with the weighted average of performance metrics (weighted by the number of subjects in the dataset). On the MWSC, NDGEN and OXVASC datasets, we determined the performance of the model in the deep and the periventricular regions (using 10 mm distance rule as above) separately, and performed paired t-tests to compare the results between regions.

2.3.6. Comparison with BIANCA

We performed a direct comparison of our TrUE-Net with BIANCA using LOO evaluation on the MWSC, NDGEN and OXVASC datasets, based on the performance metrics specified in Section 2.3.1 and performed Wilcoxon signed rank test between TrUE-Net and BIANCA results.

BIANCA features and training options: For NDGEN, we used FLAIR and T1 as features and the options that provided the best results in the initial validation of BIANCA (Griffanti et al., 2016). Other than the default options, we used the following non-default options: location of training points = no border, number of training points = Fixed + unbalanced with 2,000 lesion points and 10,000 non-lesion points. The 'no border' option and fixed + unbalanced lesion and non-lesion points have been shown to provide the best results during initial validation of BIANCA (Griffanti et al., 2016) and also in our initial tests on the same data. For OXVASC we used FLAIR + T1 + mean diffusivity (MD) as features (refer to Zamboni et al. (2019) for the MD preprocessing details). Specific options used were: sw = 2, 3D patch with patch size of 3. Due to the anisotropic nature of the voxels, additional intensity features obtained by averaging over a smaller 3D patch provided better results during the initial tests. For the MWSC dataset, we trained BIANCA using the same features and BIANCA options used for NDGEN. For all the datasets, we applied a global threshold value of 0.9 on the BIANCA lesion probability maps and masked with the white matter mask (obtained in Section 2.1.4) to obtain the final binary lesion maps.

2.3.7. Comparison with the top-ranking method of MWSC 2017

On the MWSC dataset, we compared the LOO results of TrUE-Net with the subject-wise LOO performance metrics reported in the supplementary table S1 in Li et al. (2018), which is the top-ranking method in MWSC 2017 (Kuijff et al., 2019). Additionally, we performed LOO evaluation of the model proposed in Li et al. (2018)³, by training and evaluating their model separately on the OXVASC and NDGEN datasets for the comparison with TrUE-Net. For these experiments, we used the same training hyperparameters as specified in Li et al. (2018). We also performed Wilcoxon signed rank test between the performance measure of TrUE-Net and Li et al. (2018).

2.3.8. Comparison with other existing methods

Finally, we performed an indirect comparison of our results with those obtained by other WMH segmentation methods in the literature (until 2019). We included in our comparisons only methods that achieved minimum Dice overlap metric or voxel-wise TPR of 0.70.

3. Results

3.1. Effect of training parameters on model performance

Fig. 3 shows the loss function decay with different batch sizes, learning rates and epsilon (ϵ) values. In all the cases, the model started to converge at approximately 80 epochs.

At a batch size of 8, the loss function became noisier, but had lower values. Larger batch sizes resulted in over-fitting, as evident from higher validation loss values for batch sizes of 16 and 32 (Fig. 3a). Therefore, we chose a batch size of 8 for our experiments henceforth. As the ϵ value gets smaller in the denominator, the optimiser makes larger weight updates leading to unstable optimisation, as shown in Fig. 3b for an ϵ value of 1×10^{-6} . Since an ϵ value of 1×10^{-2} provided higher loss values (with slightly unstable validation loss), we set the ϵ value to the optimal value of 1×10^{-4} for all the subsequent experiments. From Fig. 3c, for a lower learning rate, the loss decay was slower and hence required more epochs for convergence. On the other hand, for a higher learning rate the loss values converged before 20 epochs, although with higher loss values. At the optimal learning, the loss decay was slow and converged to a much lower loss value for both training and validation datasets. Hence, we chose the optimal learning rate schedule to be from 1×10^{-3} to 1×10^{-5} .

3.2. Ablation study: Effect of loss function components on segmentation performance

The boxplots of the SI values for WMH segmentation in deep and periventricular areas are shown in Fig. 4. On performing a non-parametric Friedman test across three components of the loss function (CE loss, weighted CE loss and weighted CE + Dice loss), we found that the SI values significantly increase ($\chi^2 = 25.2$, $p < 0.0001$ for DWMHs, $\chi^2 = 22.9$, $p < 0.0001$ for PWMHs) with the addition of each component of the loss function. With the addition of each component of the loss function, we observed significant improvement in the SI value. Also, weighted CE loss + Dice loss provided significantly higher SI values when compared to CE loss component alone in both regions. As shown in Fig. 4, on applying the Dice loss component alone, we achieved median SI values of 0.73 (with IQR: 0.65 - 0.79) and 0.84 (with IQR: 0.77 - 0.89) for DWMHs and PWMHs respectively, which are significantly lower than those obtained with the combined loss function ($p = 0.04$ and $p = 0.03$ for DWMHs and PWMHs respectively).

Fig. 5 shows the effect of CE and Dice loss components on TrUE-Net segmentation for a subject with medium lesion load. From the boxplots and the visual results, the effect of the composition of the loss function was observable mainly along the edges of PWMHs and DWMHs. In general, we have a large imbalance between WMH and normal WM classes and aim to focus on detecting more true WMHs. Since the CE loss function aggregates the loss values at individual voxels into a global value, it is generally biased towards WMHs in periventricular regions (Fig. 5c) where the imbalance between WMH and non-WMH voxels is less. Therefore, in some cases, the segmentation results extended beyond the ventricle lining. Using weighted CE loss controls this behaviour, since it relies on the weight maps (described in Section 2.1.3) prepared with prior anatomical information. Overcoming this class imbalance, DWMHs are given weights that are more similar to

³ Training code available at: https://github.com/hongweilibran/wmh_ibbmTum

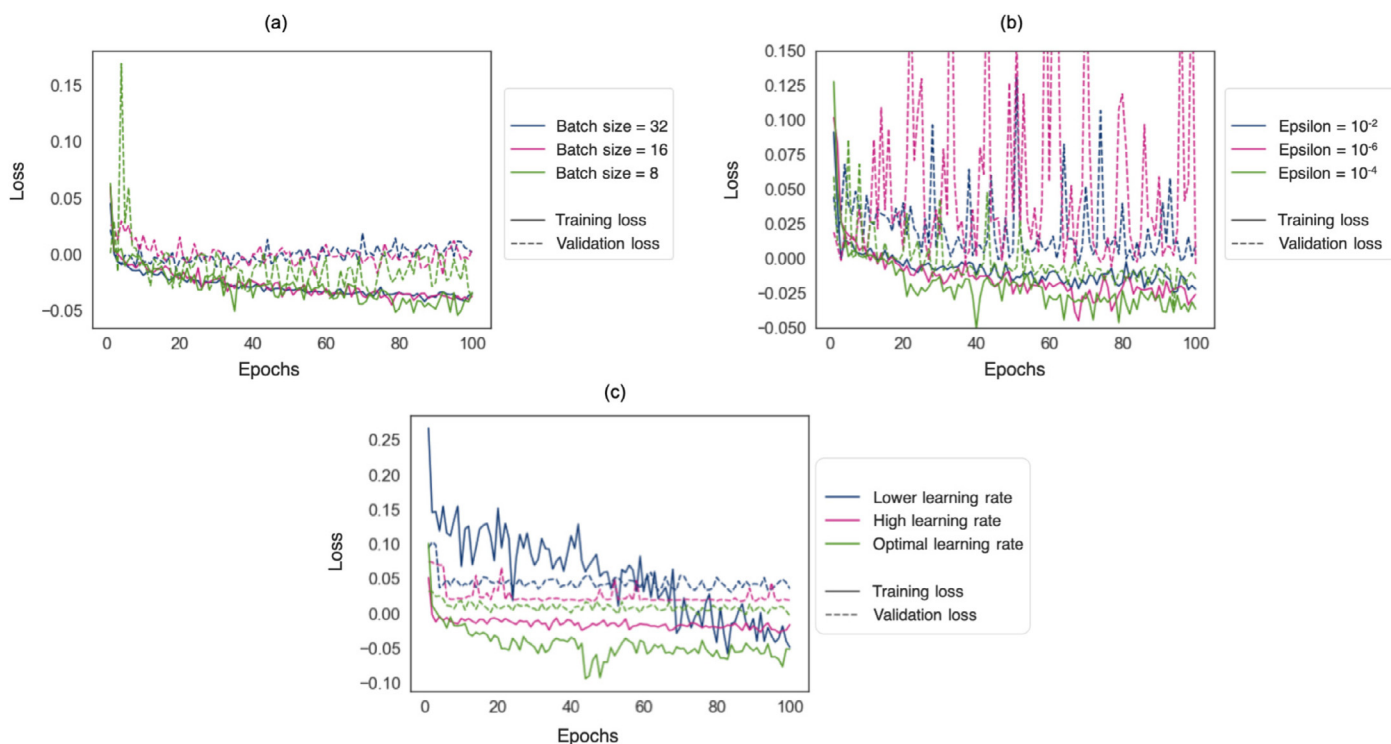


Fig. 3. Effect of training parameters on model convergence on the NDGEN dataset. Training and validation loss decays have been shown for different (a) batch sizes, (b) epsilon (ϵ) values and (c) learning rates. The plots shown in green correspond to the optimal values chosen for each parameter. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

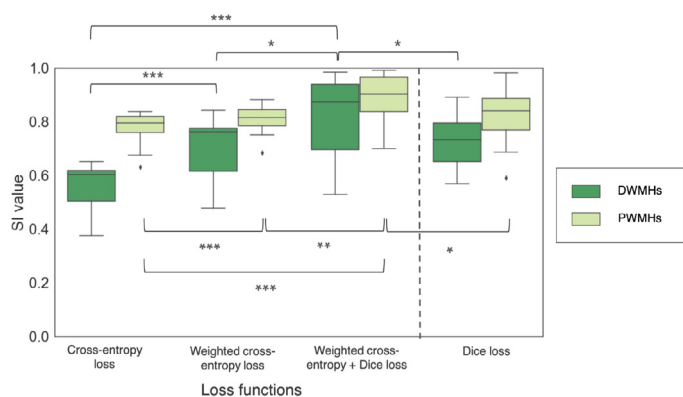


Fig. 4. Boxplots of SI values obtained for deep and periventricular regions on the NDGEN dataset for CE, weighted CE and weighted CE + Dice loss functions, compared against the Dice loss function only case. *** - $p < 0.0001$, ** - $p < 0.001$, * - $p < 0.01$.

those given to PWMHs. As a result, more DWMHs are detected and the over-segmentation in periventricular regions is avoided (Fig. 5d). Combining the advantages of both the Dice loss and the weighted CE loss components, we obtained better detection of DWMHs, along with accurate segmentation of PWMH borders (Fig. 5). The effect was particularly observable in low lesion load subjects, where the addition of the Dice loss component provided more precise segmentation, with fewer false positives. Fig. 6 shows the effect of lesion load on SI values of TrUE-Net segmentation for the three loss function compositions. In all three cases there was a significant correlation (indicated by Spearman correlation coefficient ρ_s) between SI and lesion load (CE loss: $\rho_s = 0.54$, $p = 0.02$; weighted CE loss: $\rho_s = 0.64$, $p = 0.001$; weighted CE + Dice loss: $\rho_s = 0.45$, $p = 0.09$). The combination of the weighted CE and

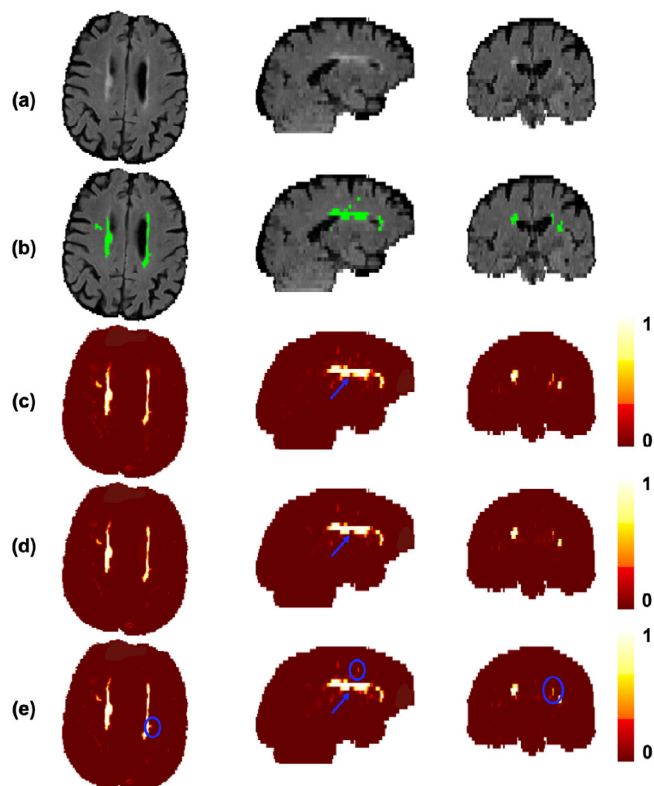


Fig. 5. Effect of loss functions on the segmentation performance in the NDGEN dataset. An example showing (a) FLAIR images, (b) manual segmentation against the results of (c) cross-entropy (CE) loss, (d) weighted cross-entropy loss and (e) weighted cross-entropy loss + Dice loss. Differences in lesion segmentations indicated by blue circles and arrows. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

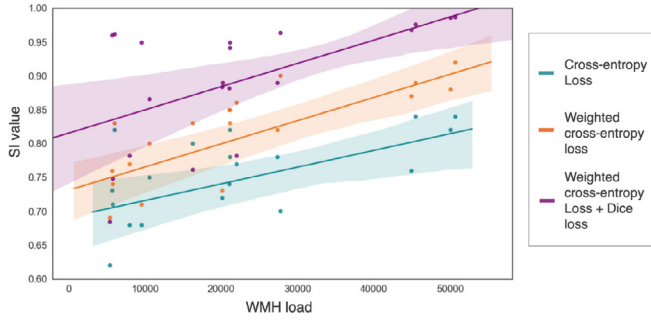


Fig. 6. Regression plot showing the impact of lesion load on SI values for weighted cross-entropy loss (magenta) and weighted cross-entropy loss + Dice loss (blue), on the NDGEN dataset. The shaded region represents the 95% confidence interval of the regression estimates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

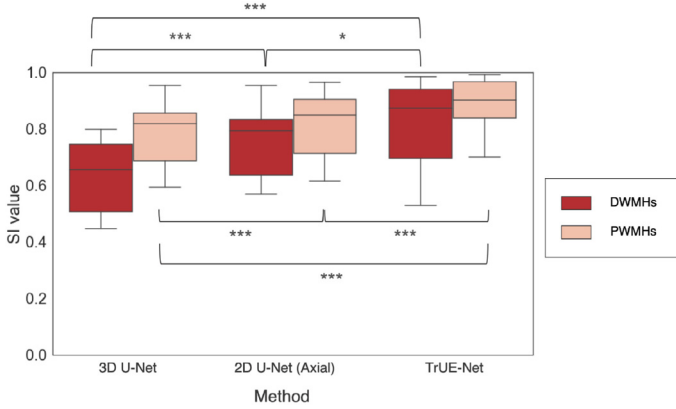


Fig. 7. Boxplots of SI values obtained for deep and periventricular regions on the NDGEN dataset for 3D U-Net, 2D U-Net (axial) and TrUE-Net. *** - $p < 0.0001$, ** - $p < 0.001$, * - $p < 0.01$.

Dice loss components achieved higher SI values and this case appears to be less affected by lesion load, as indicated by the slightly lower correlation value. This correlation between SI values and lesion load reflects a variation in performance with the amount of WMHs that are present and is undesirable for robust performance. However, the correlation coefficients of the cases with the CE loss component and weighted CE loss component were not significantly higher than the combination case (Fisher z-transformation scores, $\alpha = 0.45$ and $\alpha = 0.72$ respectively, where $\alpha < 0.05$ indicates that there is a significant difference).

3.3. Effect of model dimension on segmentation performance

The boxplots of the validation SI values for WMH segmentation in deep and periventricular areas for various cases of model dimensions are shown in Fig. 7. On performing non-parametric Friedman test across three different model dimensions (3D U-Net, 2D axial U-Net, TrUE-Net), we found that the SI values significantly increase for DWMHs ($\chi^2 = 29.1$, $p < 0.0001$) and PWMHs ($\chi^2 = 39.9$, $p < 0.0001$) across three different cases from 3D U-Net to TrUE-Net. Both 2D axial U-Net and TrUE-Net give significantly higher SI values than 3D U-Net in the deep and periventricular regions. Also, in both regions, TrUE-Net provided significantly higher SI values compared to 2D axial U-Net, due to the addition of 2D U-Nets in the sagittal and coronal planes in the TrUE-Net architecture.

Fig. 8 shows a sample prediction, from all three planes, of our proposed model compared with the 3D U-Net and the 2D axial U-Net. The 3D U-Net detected the most straightforward bright

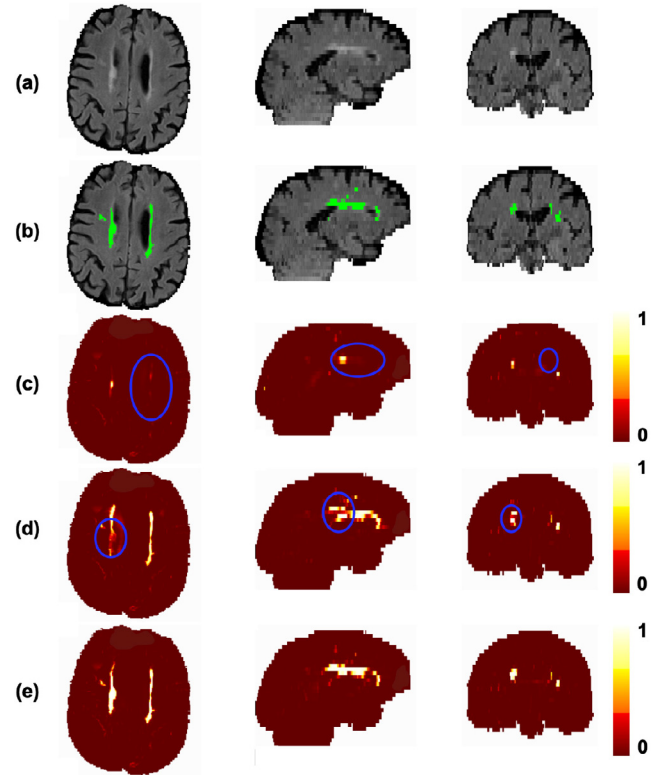


Fig. 8. Effect of model dimension on the segmentation in the NDGEN dataset. An example showing (b) manual segmentations against the results of (c) 3D U-Net, (d) 2D axial U-Net and (e) TrUE-Net. Differences in lesion segmentations indicated by blue circles. The colour bar indicates the WMH probability value at each voxel. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

PWMHs, which have high contrast. The model missed most of the DWMHs and provided under-segmentation or incorrect delineation of PWMHs (circled regions in Fig. 8c). We found that the 3D model missed more WMHs as the lesion load decreased, giving a poor segmentation in low lesion load subjects. The 2D axial U-Net gave a better segmentation when compared to the 3D model, but still lacked the contextual information across slices. Hence, it missed parts of WMHs in some of the contiguous slices, leading to discontinuities in WMH detection. An instance of this is shown in Fig. 8d, where the discontinuity is clearly visible in the sagittal and coronal planes. The proposed TrUE-Net model predicted WMHs using contextual information from all three planes and provided a continuous and more accurate segmentation (Fig. 8e and significant improvement in SI values as shown in Fig. 7) using fewer parameters compared to the 3D U-Net.

3.4. Evaluation of WMH segmentation performance

3.4.1. Leave-one-out validation

Fig. 9 shows the boxplots of the performance metrics for the leave-one-out (LOO) validation, performed separately, on the MWSC cohorts, NDGEN and OXVASC datasets. Table 1 shows the corresponding values. TrUE-Net achieved its best performance on the MWSC and OXVASC datasets. Within the MWSC cohorts, TrUE-Net achieved the best performance for the Utrecht cohort (SI: 0.92 ± 0.04 , voxel-wise TPR: 0.90 ± 0.07 , voxel-wise FPR: 0.7×10^{-4} , cluster-wise TPR: 0.88 ± 0.08 , cluster-wise F1-measure: 0.92 ± 0.07 , AVD: $10.95 \pm 6.6\%$, H95: 1.25 ± 0.72 mm). The model achieved the lowest performance on the NDGEN dataset, however with the highest cluster-wise TPR value of 0.87 ± 0.12 . This indicates that it detects more true positive lesions when compared to

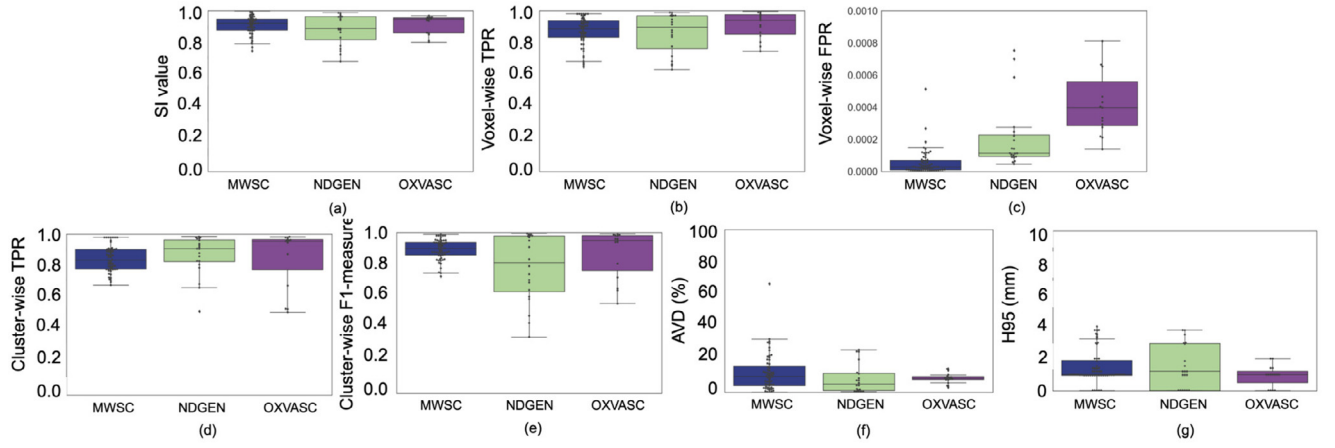


Fig. 9. Boxplots of performance metrics obtained from LOO evaluation on the MWSC, NDGEN and OXVASC datasets - (a) SI value, (b) voxel-wise TPR, (c) voxel-wise FPR, (d) cluster-wise TPR, (e) cluster-wise F1-measure, (f) Absolute volume difference (AVD), and (g) 95th percentile of Hausdorff distance.

Table 1

LOO evaluation of TrUE-Net on the MWSC, NDGEN and OXVASC datasets. The median values are provided with the interquartile range (IQR) between 25th and 75th percentiles reported in parentheses (the best median value for each measure across the datasets is highlighted in bold).

Perf. metrics	MWSC	NDGEN	OXVASC
SI	0.92 (0.88 - 0.95)	0.89 (0.82 - 0.96)	0.95 (0.86 - 0.96)
Voxel-wise TPR	0.89 (0.83 - 0.94)	0.89 (0.76 - 0.97)	0.94 (0.85 - 0.97)
Voxel-wise FPR	2.7 (0.9 - 6.8) × 10⁻⁵	1.1 (0.9 - 2.2) × 10 ⁻⁴	3.9 (2.8 - 5.6) × 10 ⁻⁴
Cluster-wise TPR	0.84 (0.78 - 0.90)	0.91 (0.83 - 0.97)	0.95 (0.78 - 0.97)
Cluster-wise F1 measure	0.90 (0.86 - 0.94)	0.81 (0.63 - 0.98)	0.95 (0.76 - 0.98)
AVD (%)	9.6 (3.9 - 15.9)	4.9 (0.9 - 11.5)	8.5 (7.7 - 9.4)
H95 (mm)	1 (0.96 - 1.89)	1.2 (0 - 2.94)	1 (0.5 - 1.2)

other datasets, while the lesion boundaries were better delineated in other datasets.

The visual results of LOO evaluations on the MWSC cohorts, NDGEN and OXVASC datasets are illustrated by a few examples shown in Fig. 10. In both high and low lesion load cases, TrUE-Net provided an accurate segmentation with respect to manual segmentation without detecting many false positives. In particular, TrUE-Net detected the deep lesions in the low lesion load cases, including subtle ones (Fig. 10f and j). It is also worth noting that, although manual segmentation is used as gold standard for evaluation, there might be inconsistencies and errors that would affect the performance evaluation. For instance, in a high lesion load example from the Utrecht cohort in Fig. 10a, some CSF voxels were included in the manual segmentation (indicated by a circle), while TrUE-Net successfully excluded them by providing lower probability values in those regions.

3.4.2. Results on MWSC 2017 unseen test dataset

On evaluating the docker container of our method on the unseen challenge test data, we obtained the following weighted average of performance metrics on the unseen test dataset: SI value - 0.77, H95 - 6.95, AVD - 20.49, cluster-wise TPR - 0.74 and cluster-wise F1 measure - 0.70. For the performance metrics on individual test datasets and the corresponding boxplots, please refer <https://wmh.isi.uu.nl/results/fmrib-truenet-2/>.

3.4.3. Performance in deep and periventricular regions

Fig. 11 shows boxplots of the performance metrics for DWMHs and PWMHs. Table 2 reports the corresponding descriptive statistics, along with the p-values of Wilcoxon signed rank test performed between DWMHs and PWMHs. Most of the performance metrics are not significantly different between PWMHs and DWMHs. Particularly, none of the differences in cluster-wise and voxel-wise TPRs are significant, indicating that TrUE-Net not only

successfully detects true lesions in both the periventricular and deep regions, but also segments the lesion boundaries accurately in both regions. The only significant differences can be found in the cluster-wise F1-measure in the MWSC dataset, as well as AVD and voxel-wise FPR values in the MWSC and OXVASC datasets. In the case of the MWSC dataset, the cluster-wise F1-measure in the deep regions were significantly higher than in the periventricular regions. This means that more true DWMHs were detected, with higher cluster-wise precision, compared to PWMHs. In the OXVASC and the MWSC datasets, while DWMHs showed significantly higher AVD percentage values compared to PWMHs, they still correspond to much lower lesion volumes when compared to PWMHs. For instance, AVD value of 25% in the deep region corresponds to a lesion volume of around 600 mm³ while the same value corresponds to a much higher value, of around 3800 mm³, in the periventricular region.

Overall, TrUE-Net provided performance metrics for DWMHs on par with PWMHs. This shows that the model provides good delineations of WMHs, with good sensitivity and specificity in both regions.

3.5. Comparison with BIANCA

Fig. 12 illustrates a few example segmentations obtained with TrUE-Net and BIANCA, with respect to manual segmentation, in the order of decreasing lesion load. TrUE-Net provided more accurate segmentations than BIANCA, especially in the low lesion load subjects. From the figure it can be observed that, as the lesion load decreases, BIANCA detected more false positives, particularly around the ventricles (shown in Fig. 12a, d and e) and oversegmented PWMHs (Fig. 12b and c).

Overall, TrUE-Net outperforms BIANCA in both voxel-wise and cluster-wise metrics. Fig. 13 shows the boxplots comparing the performance metrics between TrUE-Net and BIANCA. The correspond-

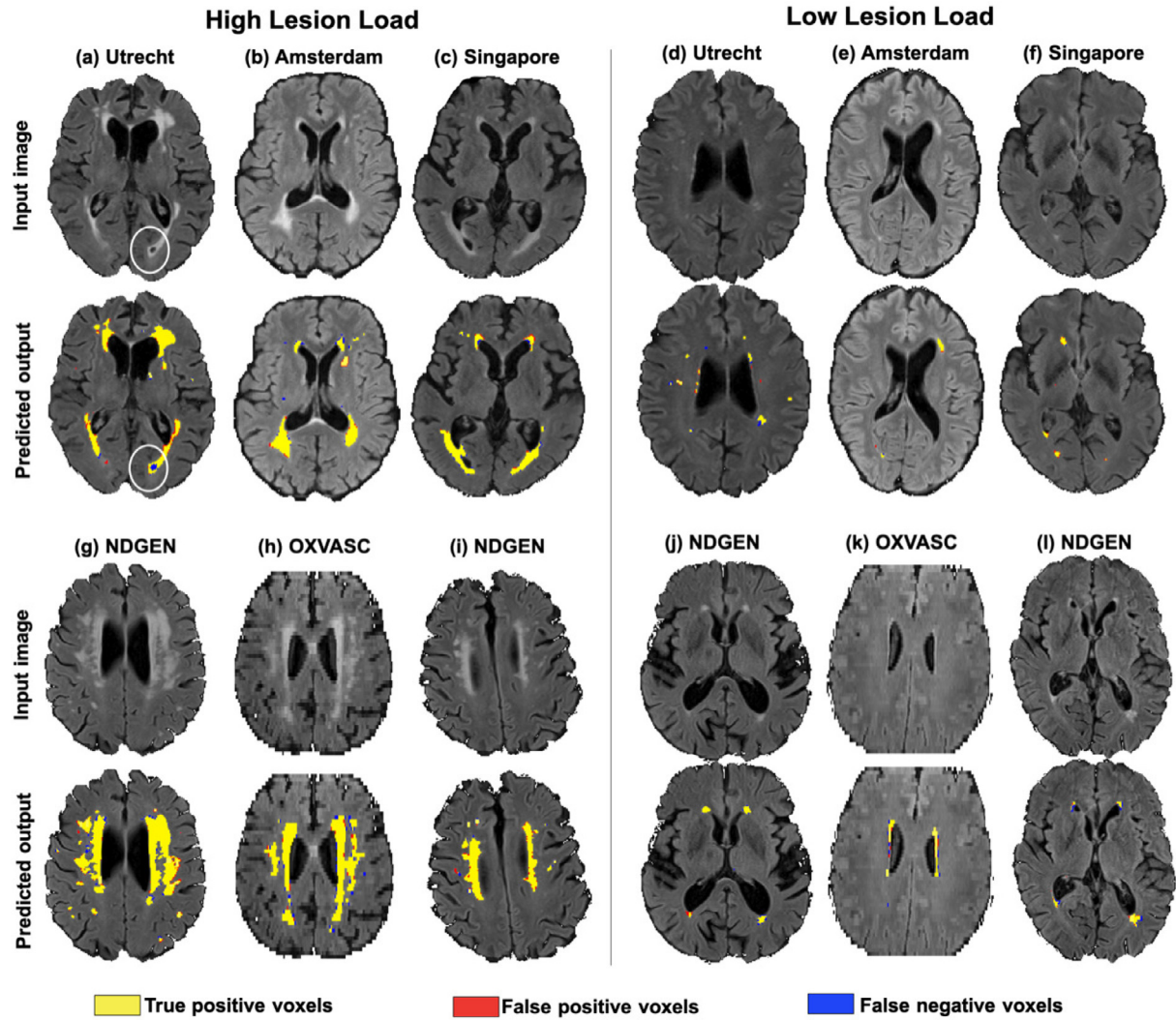


Fig. 10. Sample results of TrUE-Net segmentation on the high and low lesion cases - (a, d) Utrecht, (b, e) Amsterdam and (c, f) Singapore cohorts of MWSC dataset, (g, i, j, l) NDGEN and (h, k) OXVASC datasets. True positive, false positive and false negative voxels are indicated in yellow, red and blue respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

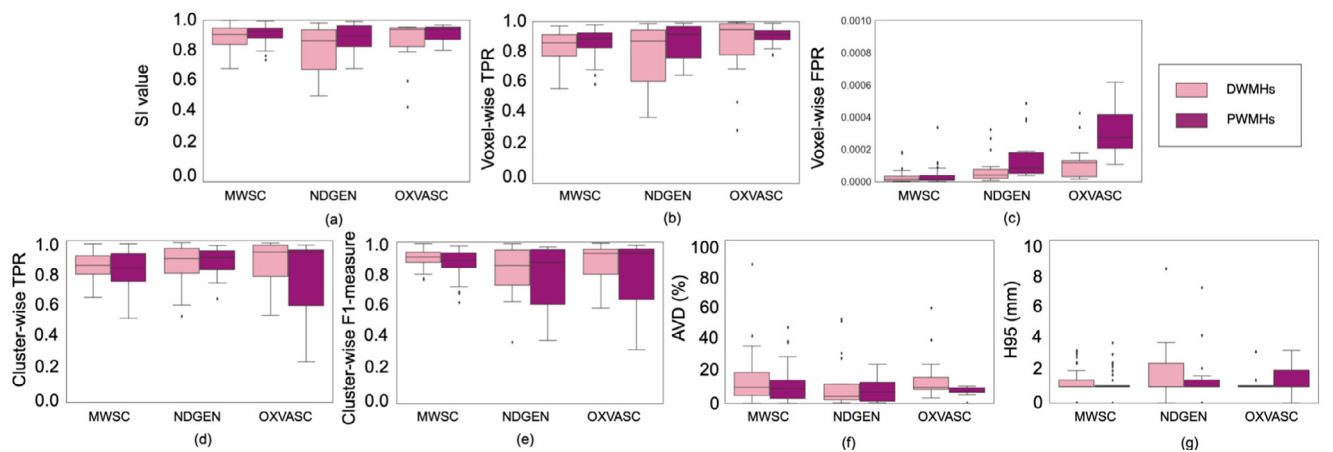


Fig. 11. Boxplots of performance metrics obtained in deep and periventricular regions for LOO evaluation on the MWSC, NDGEN and OXVASC datasets - (a) SI value, (b) voxel-wise TPR, (c) voxel-wise FPR, (d) cluster-wise TPR, (e) cluster-wise F1-measure, (f) Absolute volume difference (AVD), and (g) 95th percentile of Hausdorff distance.

Table 2

Comparison of performance metrics between PWMHs and DWMHs, along with p-values of Wilcoxon signed rank test results on the MWSC, NDGEN and OXVASC datasets (median and IQR values provided; significant p-values highlighted in bold).

		MWSC	NDGEN	OXVASC
SI	DWMHs	0.91 (0.85 - 0.95)	0.87 (0.69 - 0.94)	0.94 (0.84 - 0.95)
	PWMHs	0.93 (0.88 - 0.95)	0.90 (0.84 - 0.97)	0.95 (0.88 - 0.96)
	p-value	0.06	0.05	0.11
Voxel-wise TPR	DWMHs	0.87 (0.79 - 0.92)	0.88 (0.63 - 0.95)	0.95 (0.79 - 0.99)
	PWMHs	0.89 (0.84 - 0.93)	0.92 (0.77 - 0.97)	0.92 (0.89 - 0.95)
	p-value	0.08	0.10	0.45
Voxel-wise FPR	DWMHs	$1.3 (0.4 - 3.1) \times 10^{-5}$	$3.8 (2.0 - 7.5) \times 10^{-5}$	$1.1 (0.2 - 1.3) \times 10^{-4}$
	PWMHs	$1.5 (0.6 - 3.7) \times 10^{-5}$	$8.2 (5.0 - 17.9) \times 10^{-5}$	$2.7 (0.2 - 4.7) \times 10^{-4}$
	p-value	0.04	0.06	<0.001
Cluster-wise TPR	DWMHs	0.85 (0.79 - 0.91)	0.89 (0.80 - 0.95)	0.93 (0.78 - 0.97)
	PWMHs	0.83 (0.75 - 0.92)	0.90 (0.82 - 0.94)	0.93 (0.60 - 0.94)
	p-value	0.17	0.69	0.25
Cluster-wise F1-measure	DWMHs	0.91 (0.87 - 0.94)	0.85 (0.73 - 0.91)	0.93 (0.80 - 0.95)
	PWMHs	0.89 (0.84 - 0.93)	0.87 (0.61 - 0.95)	0.93 (0.64 - 0.96)
	p-value	0.02	0.08	0.17
AVD (%)	DWMHs	9.7 (4.8 - 18.8)	4.1 (1.9 - 11.8)	9.7 (8.6 - 15.7)
	PWMHs	9.0 (2.9 - 14.0)	6.7 (1.2 - 12.9)	7.3 (6.5 - 9.5)
	p-value	0.002	0.35	0.01
H95 (mm)	DWMHs	1 (1 - 1.41)	1 (1 - 2.45)	1 (1 - 1.7)
	PWMHs	1 (1 - 1.7)	1 (1 - 1.41)	1 (1 - 2.0)
	p-value	0.39	0.19	0.41

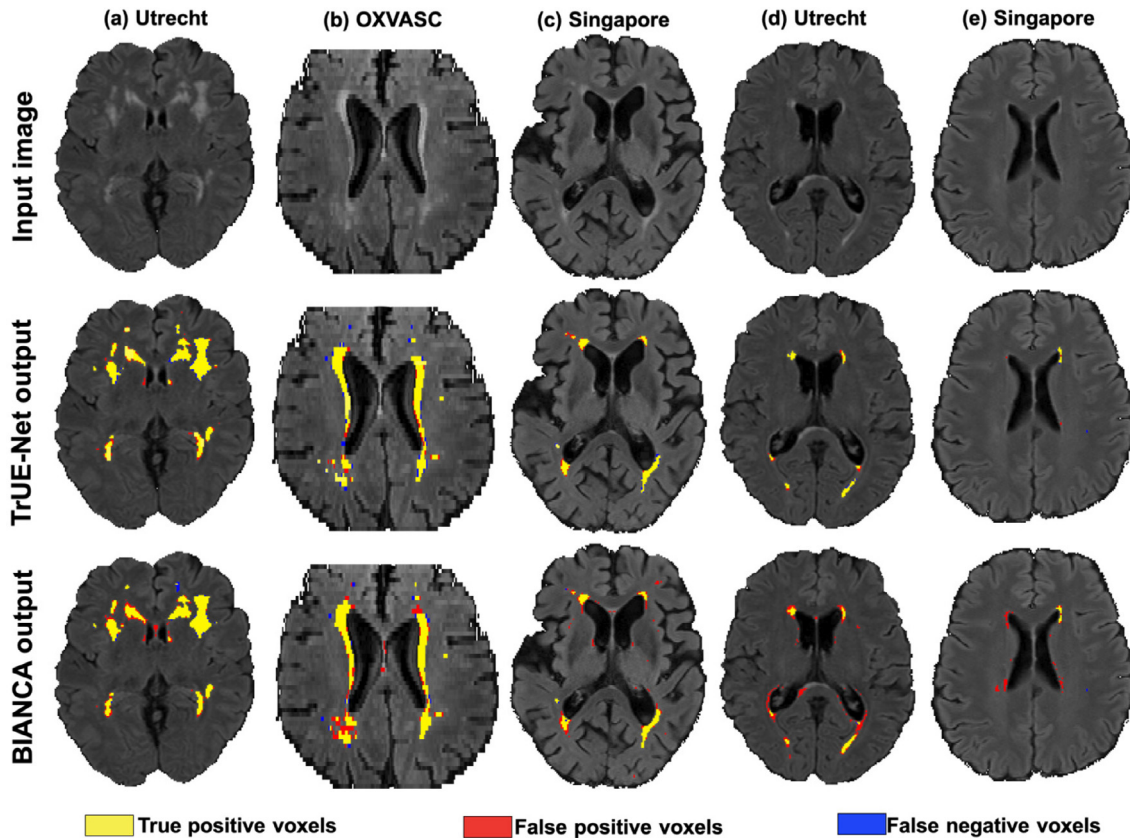


Fig. 12. Sample results for comparison of TrUE-Net segmentation with those of BIANCA. Left to right: decreasing order of lesion load from the Utrecht (a,d), Singapore (c, e) and OXVASC (b) datasets. True positive, false positive and false negative voxels are indicated in yellow, red and blue colour respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ing performance values and p-values are reported in Table 3. TrUE-Net provides significantly better results than BIANCA for almost all the performance metrics. BIANCA provided the worst performance on the MWSC dataset, with both more false positives and false negatives than TrUE-Net. In the NDGEN dataset, the voxel-wise TPR values are not significantly different for TrUE-Net and BIANCA, indicating that the two methods performs equally on this dataset.

Also voxel-wise FPR were not significantly different between the two methods in both NDGEN and OXVASC datasets. Fig. 14 shows the regression plots of SI values against lesion loads for the MWSC, NDGEN and OXVASC datasets. We determined the Spearman correlation coefficient of the SI values with lesion loads for the methods on the individual datasets and determined their statistical significance. Both TrUE-Net and BIANCA showed positive correlation with

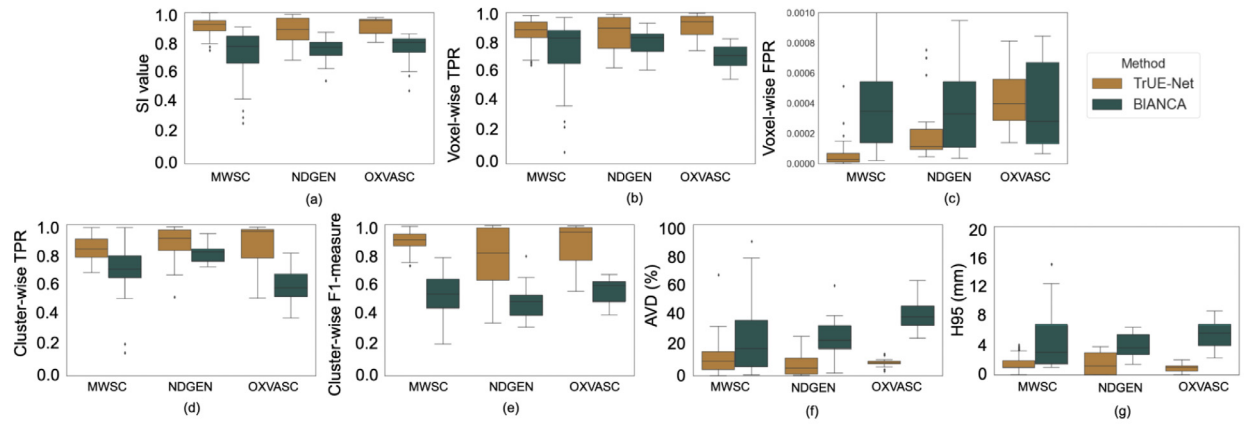


Fig. 13. Boxplots of performance metrics obtained for TrUE-Net and BIANCA on the Utrecht, Singapore, Amsterdam, NDGEN and OXVASC datasets - (a) SI value, (b) Absolute volume difference (AVD), (c) voxel-wise TPR, (d) voxel-wise FPR, (e) cluster-wise TPR, (f) cluster-wise F1-measure and (g) 95th percentile of Hausdorff distance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3

Comparison of TrUE-Net performance with BIANCA, along with p-values of Wilcoxon signed rank test results on the MWSC, NDGEN and OXVASC datasets (median and IQR values reported; significant p-values highlighted in bold).

		MWSC	NDGEN	OXVASC
SI	TrUE-Net	0.92 (0.88 - 0.95)	0.89 (0.82 - 0.96)	0.95 (0.86 - 0.96)
	BIANCA	0.77 (0.66 - 0.85)	0.77 (0.72 - 0.80)	0.80 (0.73 - 0.82)
	p-value	<0.001	0.001	<0.001
Voxel-wise TPR	TrUE-Net	0.89 (0.83 - 0.94)	0.89 (0.76 - 0.97)	0.94 (0.85 - 0.97)
	BIANCA	0.83 (0.66 - 0.88)	0.83 (0.74 - 0.86)	0.74 (0.65 - 0.78)
	p-value	<0.001	0.15	<0.001
Voxel-wise FPR	TrUE-Net	$2.7 (0.9 - 6.8) \times 10^{-5}$	$1.1 (0.9 - 2.2) \times 10^{-4}$	$3.9 (2.8 - 5.6) \times 10^{-5}$
	BIANCA	$3.4 (1.4 - 5.4) \times 10^{-4}$	$3.2 (1.0 - 5.4) \times 10^{-4}$	$2.6 (1.2 - 6.7) \times 10^{-4}$
	p-value	<0.001	0.07	0.98
Cluster-wise TPR	TrUE-Net	0.84 (0.78 - 0.90)	0.91 (0.83 - 0.97)	0.96 (0.78 - 0.97)
	BIANCA	0.70 (0.65 - 0.79)	0.82 (0.76 - 0.84)	0.58 (0.55 - 0.68)
	p-value	<0.001	0.09	<0.001
Cluster-wise F1-measure	TrUE-Net	0.90 (0.86 - 0.94)	0.81 (0.63 - 0.98)	0.95 (0.76 - 0.98)
	BIANCA	0.54 (0.45 - 0.64)	0.49 (0.40 - 0.53)	0.60 (0.52 - 0.63)
	p-value	<0.001	<0.001	<0.001
AVD (%)	TrUE-Net	9.6 (3.9 - 15.9)	4.9 (0.9 - 11.5)	8.5 (7.7 - 9.4)
	BIANCA	17.9 (5.9 - 36.7)	23.3 (18.1 - 32.9)	38.5 (33.1 - 48.4)
	p-value	<0.001	<0.001	<0.001
H95 (mm)	TrUE-Net	1 (0.9 - 1.9)	1.2 (0 - 2.94)	1 (0.5 - 1.2)
	BIANCA	3.0 (1.5 - 6.8)	3.6 (2.7 - 5.4)	5.5 (3.7 - 7.2)
	p-value	<0.001	<0.001	<0.001

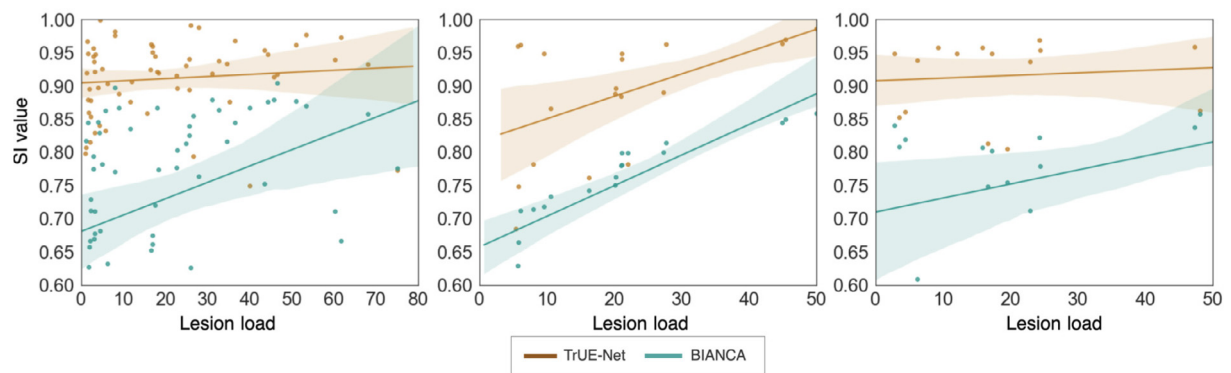


Fig. 14. Regression plot of SI values with respect to lesion load for TrUE-Net (brown) and BIANCA (green) on the MWSC, NDGEN and OXVASC datasets. The shaded region represents the 95% confidence interval of the regression estimates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

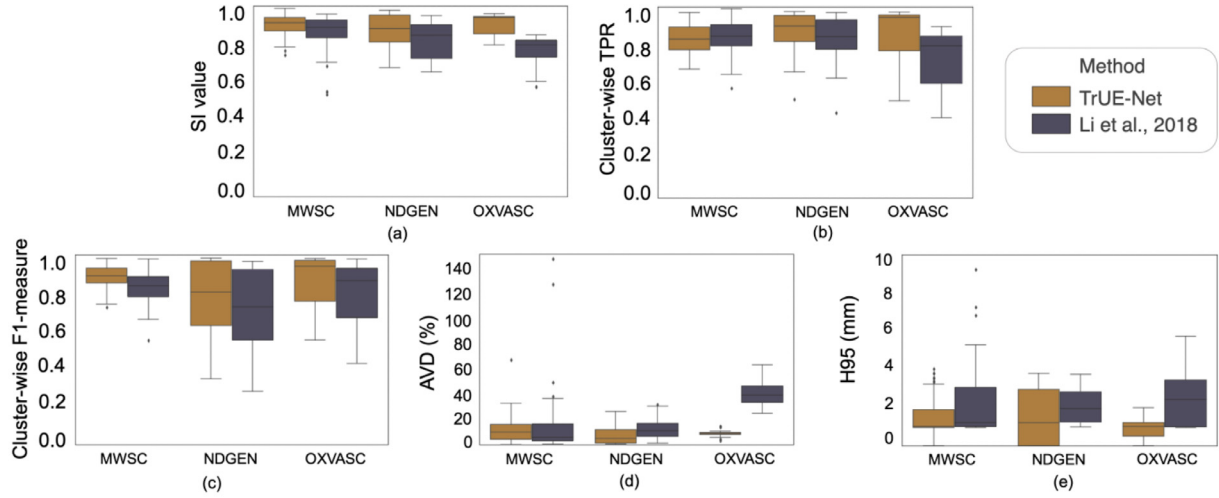


Fig. 15. Boxplots of performance metrics obtained for TrUE-Net and Li et al. (2018) on the MWSC dataset - (a) SI value, (b) Cluster-wise TPR, (c) Cluster-wise F1-measure, (d) Absolute volume difference (AVD) and (e) 95th percentile of Hausdorff distance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4

Comparison of TrUE-Net performance with Li et al. (2018), along with p-values of Wilcoxon signed rank test results on the MWSC, NDGEN and OXVASC datasets (median and IQR values reported; significant p-values highlighted in bold).

		MWSC	NDGEN	OXVASC
SI	TrUE-Net	0.92 (0.88 - 0.95)	0.89 (0.82 - 0.96)	0.95 (0.86 - 0.96)
	Li et al., 2018	0.90 (0.84 - 0.94)	0.85 (0.73 - 0.91)	0.80 (0.73 - 0.82)
	p-value	0.003	<0.001	<0.001
Cluster-wise TPR	TrUE-Net	0.84 (0.78 - 0.90)	0.91 (0.83 - 0.97)	0.96 (0.78 - 0.97)
	Li et al., 2018	0.85 (0.80 - 0.91)	0.85 (0.79 - 0.94)	0.82 (0.68 - 0.85)
	p-value	0.49	0.003	<0.001
Cluster-wise F1-measure	TrUE-Net	0.90 (0.86 - 0.94)	0.81 (0.63 - 0.98)	0.95 (0.76 - 0.98)
	Li et al., 2018	0.84 (0.79 - 0.89)	0.73 (0.56 - 0.93)	0.89 (0.71 - 0.94)
	p-value	<0.001	<0.001	0.003
AVD (%)	TrUE-Net	9.6 (3.9 - 15.9)	4.9 (0.9 - 11.5)	8.5 (7.7 - 9.4)
	Li et al., 2018	5.3 (2.8 - 16.2)	10.9 (6.5 - 16.4)	38.5 (33.1 - 48.4)
	p-value	0.48	<0.001	<0.001
H95 (mm)	TrUE-Net	1 (0.9 - 1.9)	1.2 (0 - 2.94)	1 (0.5 - 1.2)
	Li et al., 2018	1.2 (1 - 3.0)	1.9 (1.2 - 2.8)	2.4 (1.2 - 3.4)
	p-value	0.009	0.25	<0.001

lesion loads with significant correlations on the NDGEN dataset (TrUE-Net: $\rho_S = 0.64$, $p = 0.002$; BIANCA: $\rho_S = 0.99$, $p < 0.0001$), with significant difference between their correlations ($\alpha < 0.0001$). In general, the SI values of TrUE-Net showed less correlation with respect to lesion loads compared to those from BIANCA, indicating a consistent performance for all lesion loads.

3.6. Comparison with the top-ranking method of MWSC 2017

Fig. 15 shows the boxplots comparing the LOO performance metrics between TrUE-Net and the method proposed in Li et al. (2018) trained and evaluated separately on the MWSC, NDGEN and OXVASC datasets. The corresponding values and p-values of Wilcoxon signed rank test are reported in Table 4. On the MWSC dataset, TrUE-Net achieves significantly higher SI values and significantly lower H95 values compared to Li et al. (2018). However, Li et al. (2018) achieves better cluster-wise TPR values indicating that the method detects more true-lesions compared to TrUE-Net. On the other hand, the cluster-wise F1-measure value is significantly higher for TrUE-Net, which shows that Li et al. (2018) also detects more false positive clusters, while TrUE-Net provides more cluster-wise precision, resulting in a higher

cluster-wise F1-measure value. The performance metrics obtained for Li et al. (2018) on the NDGEN dataset were significantly lower (except H95) compared to those of TrUE-Net. Li et al. (2018) gave the worst performance on the OXVASC dataset, with significantly lower performance metrics when compared to TrUE-Net. This could be due to the fact that the OXVASC dataset consists of lower resolution in the axial plane, which is quite different from the other two datasets. We observed that Li et al. (2018) missed a few small lesions and undersegmented the lesions, missing voxels along the lesion boundaries.

3.7. Comparison with other existing methods

In order to contextualise the impact of the methods/improvements presented in this work, Table 5 illustrates an indirect comparison with other existing methods in the literature. The SI values obtained with TrUE-Net are higher than those reported for the non-DL methods in the existing literature, including BIANCA. In the case of DL methods, the performance of TrUE-Net is comparable to the top-performing DL methods in the challenge on the unseen test dataset.

Table 5
Comparison of existing methods (including BIANCA) with TrUE-Net^a.

Method	Type ^b	Population - study ^c (subjects)	Image modalities	SI value	Sens ^d	Spec ^e
Wang et al. (2012)	I,U	Singapore ageing cohort (272)	T1, T2, FLAIR	0.77	0.81	0.97
Gibson et al. (2010)	I,U	WM disease (18)	FLAIR	0.81	-	-
De Boer et al. (2009)	I,S	HC - Rotterdam scan study(6)	T1, PD, FLAIR	0.72	0.79	-
Steenwijk et al. (2013)	I,S	Hypertension (20)	T1, FLAIR	0.84	-	-
Damangir et al. (2012)	IA,S	HC, Alzheimer's, dementia - DemWest (102)	T1, FLAIR	-	0.90	0.99
Ghafoorian et al. (2016)	IA,S	Small vessel disease dataset (50)	PD, T2, FLAIR	-	0.73	-
Yoo et al. (2014)	I,S	Longitudinal/ dementia - Korean study (32)	FLAIR	0.76	-	-
Jeon et al. (2011)	IA,U	SVD - AMPETIS (45)	PD, T2, FLAIR	0.90	-	-
Yang et al. (2010)	IA,U	Dementia - LEILA (30)	FLAIR	0.81	-	-
Shi et al. (2013)	IAAp,U	Acute infarction (91)	DWI, T1, FLAIR	0.84	0.80	-
Samaile et al. (2012)	IAAp,U	MCI, CADASIL (67)	T1, FLAIR	0.72	-	-
Kruggel et al. (2008)	IAAp	HC, dementia - LEILA (116)	-	-	0.90	0.91
Khademi et al. (2011)	IAAp,U	Subjects with lesions (24)	FLAIR	0.83	0.82	0.99
Admiraal-Behloul et al. (2005)	IA,U	Vasc. disease - PROSPER (100)	PD, T2, FLAIR	0.75	-	-
Anbeek et al. (2004)	IA,S	Arterial vascular disease (20)	T1, IR, PD, T2, FLAIR	0.80	0.97	0.97
Li et al. (2018)	DL	MWSC TS	T1, FLAIR	0.80	-	-
Andermatt et al. (2016)	DL	MWSC TS	T1, FLAIR	0.78	-	-
Berseth, 2017	DL	MWSC TS	T1, FLAIR	0.77	-	-
Valverde et al. (2017)	DL	MWSC TS	T1, FLAIR	0.77	-	-
Kuijff et al. (2019)	DL	MWSC TS	T1, FLAIR	0.77	-	-
Kuijff et al. (2019)	DL	MWSC TS	T1, FLAIR	0.72	-	-
Xu et al. (2017)	DL	MWSC TS	T1, FLAIR	0.73	-	-
Ghafoorian et al. (2017)	DL	Elder SVD (50)	T1, FLAIR	0.78	-	-
BIANCA (Griffanti et al., 2016)	IAAp,S	NDGEN (21)	T1, FLAIR	0.77	0.80	-
		OXVASC (18)	T1, FLAIR, MD	0.74	0.71	-
		MWSC TrS (60)	T1, FLAIR	0.73	0.74	-
TrUE-Net	DL	MWSC TrS (60)	T1, FLAIR	0.91	0.87	-
		NDGEN (20)	T1, FLAIR	0.88	0.86	-
		OXVASC (18)	T1, FLAIR	0.91	0.91	-
TrUE-Net	DL	MWSC TS	T1, FLAIR	0.77	-	-

^a Mean values of evaluation metrics reported for all methods (including TrUE-Net).

^b Type: I - intensity, IA - intensity + anatomy, IAAp - intensity + anatomy + appearance, U - unsupervised, S - supervised, DL - deep learning

^c Population-study: MWSC TS - MICCAI WMH segmentation Challenge test dataset, MWSC TrS - MICCAI WMH segmentation Challenge training dataset (Kuijff et al., 2019)

^d Sensitivity or Voxel-wise TPR,

^e Specificity.

4. Discussion and conclusions

In this work, we proposed a DL model using an ensemble of U-Nets, named TrUE-Net, for accurate WMH segmentation. First, we investigated the effect of various training hyperparameters on model optimisation. We then studied the effect of various components of the loss function and of the model dimension on the segmentation performance. On data from 5 cohorts, we evaluated the overall segmentation performance as well as the performance in deep and periventricular regions separately. In addition, we directly compared our method with a non-DL method (BIANCA) and a DL method (the top ranking method of MWSC 2017). Finally, we provided an indirect comparison with various methods proposed in the literature.

When optimising our model, we observed that using a lower batch size resulted in noisy but lower loss values. The lower batch size could be advantageous due to two reasons: lower computation load per batch (hence higher speed) and faster convergence due to the regularisation effect of noisy gradient estimation (Wilson and Martinez, 2003; Keskar et al., 2016). Firstly, using smaller batches reduces the gradient estimation time per iteration. However, this would increase the overall number of iterations per epoch, which brings us to the second advantage. Due to the noisy estimate of mean gradient over batches, during subsequent iterations in our experiments we observed that the cost function fluctuates, getting out of some spurious local minima and converging to a better local minima quickly when using smaller batches. Additionally, in our experiments smaller batch sizes avoided over-fitting, as evident from the lower validation loss for a batch size of 8, compared to 16 and 32 (Fig. 3a). Regarding the parameter ϵ , we chose the

value 1×10^{-4} as an optimal value for further experiments, since it showed lower loss values. The parameter ϵ is used in the denominator (to avoid divide-by-zero error) in the determination of updates for weights. Having a very low ϵ value results in estimation of larger weight updates, leading to unstable optimisation as shown in the case of 1×10^{-6} (Fig. 3b). In the case of learning rate, choosing a higher value results in earlier convergence with higher loss values. On the other hand, lower values require more epochs to converge due to smaller steps of the updates in weights. Hence, we chose an optimal learning rate schedule between 1×10^{-3} and 1×10^{-6} , Fig. 3c, for our further experiments.

When comparing the results using different components of the loss function, we observed that the CE loss component is responsible for identification of the majority of lesions against the background voxels, while the Dice loss component is more sensitive to smaller lesions and precise boundaries. Using the Dice loss component individually provided significantly lower SI values when compared to the case of combined loss function, and provided results comparable to similar methods using Dice loss function for other segmentation applications (Piantadosi et al., 2020). While PWMHs and DWMHs differ in many aspects, like intensity, contrast and texture, these information are learnt by the model while training. Therefore, we used distance values (combination of distances from ventricles and GM) as an additional anatomical prior, derived in a data-driven manner, for weighting the CE loss function to make WMH segmentation better in the deep region. Also, these distance values are independent of lesion load and characteristics and depend only on anatomical structures such as ventricles and GM, which makes the segmentation better irrespective of lesion load. In addition, using the Dice loss component also results

in the detection of more subtle deep lesions, giving more accurate segmentation (Fig. 5e) and avoiding over-segmentation in low lesion load subjects. This results in higher SI values in these low lesion load subjects (shown in Fig. 6). Regarding the effect of model dimension, the 3D U-Net model detects most of the PWMHs lesions, since they are larger and have more contextual information, but it failed to detect smaller DWMHs (Figs. 8c and 7). The 2D model is more suitable for those cases, but the 2D model in a single plane alone is not sufficient to capture contextual information across slices and hence often misses lesions in contiguous slices, resulting in discontinuous segmentation (Fig. 8d). Hence, the triplanar ensemble of 2D U-Nets are better than a single plane 2D U-Net since the ensemble model has lower bias towards spurious detections compared to its constituent single-plane models. Moreover, the triplanar model sees data from different planes, thereby avoiding discontinuities in WMHs across consecutive slices and attenuating noise voxels. Hence using the triplanar model leads to both sensitive and precise detection (Figs. 8e and 7), with fewer parameters than the 3D U-Net model.

The TrUE-Net model gave good performance in all 5 datasets. In particular, in the MWSC dataset, TrUE-Net achieved the best cluster-wise F1-measure and the lowest voxel-wise FPR indicating precise and accurate segmentation. The MWSC cohort consists of subject with variations in lesion load and acquisition characteristics. Even though the SI value on the NDGEN dataset is lower than the other two datasets, TrUE-Net detects more small true lesions (best cluster-wise TPR) compared to the other datasets.

When comparing the performance of our model in the deep and the periventricular regions, the trend of performance metrics for DWMHs and PWMHs remained consistent across datasets. In general, we observed that SI values and voxel-wise TPR values were higher for PWMHs when compared to DWMHs. In the deep regions, cluster-wise TPR was higher than that in the periventricular regions. This means that more DWMHs are detected correctly (also evident from higher cluster-wise F1-measure values). However, the lesion boundaries are delineated better in PWMHs when compared to DWMHs, as indicated by slightly higher Hausdorff distance for DWMHs (Fig. 11 and Table 2). At this point, it is worth remembering that manual segmentation is our gold standard but not necessarily the absolute truth. Therefore these errors in the delineation of DWMHs could be due to inconsistencies in manual segmentation (due to low contrast and other confounders) rather than lower model sensitivity. Moreover, the application of the white matter mask in the post-processing step might affect the segmentation of DWMHs near the WM-GM interface. However, when specifically testing this on our data, we observed that differences in the performance metrics near the WM-GM interface with and without applying the WM mask were negligible and not significant (for more details refer to the supplementary material).

On comparing TrUE-Net with BIANCA, we found that TrUE-Net outperforms BIANCA with significant differences in the performance metrics in almost all datasets. Overall, we found that TrUE-Net achieves the highest SI values, cluster-wise F1-measures and the lowest Hausdorff distance values for all datasets, and provides better segmentation in various lesion load cases (Fig. 13). TrUE-Net performs well even in subjects with low lesion load, detecting fewer false positives than BIANCA. Moreover, the SI values achieved by TrUE-Net are higher and less affected by lesion loads when compared to BIANCA, especially on the MWSC and OXVASC datasets as shown in Fig. 14. This shows that, not only with high lesion loads, but also in the cases of small lesion loads where the lesions are quite small and looks like normal ventricle lining (Fig. 12e), TrUE-Net detects the lesions more accurately when compared to BIANCA. This is likely due to the fact that TrUE-Net learns the overall contextual information regarding the lesion distribution. On the other hand, BIANCA uses hand-crafted features

(e.g. intensity) that might be affected by contrast/texture variations, thus leading to a noisy segmentation. Among our validation datasets, only the MWSC dataset includes subjects with very low WMH load so further evaluations on bigger samples of low lesion data will be needed to generalise the performance of TrUE-Net. However, this comparison against BIANCA shows promising improvement in the low lesion load range. Also, TrUE-Net provides results comparable with the top ranking method of the MICCAI Challenge (Li et al. (2018)) on all datasets, with significantly higher mean SI values on the MWSC and OXVASC dataset. Particularly, the OXVASC dataset has lower resolution in the axial plane which is quite different from the characteristics of other datasets. TrUE-Net achieved better performance on this dataset due to its triplanar architecture, while Li et al. (2018) provided incorrect segmentation of lesion boundaries, even in PWMHs, due to lower axial resolution. When performing an indirect comparison of TrUE-Net with other existing methods, we observed that only a few methods have been evaluated on a variety of datasets (pathological population and/or healthy subjects). Most of the methods considered for the comparison analysis are tested on datasets from a specific population, and hence might require additional experimentation to validate/improve their generalisability (e.g. fine-tuning of parameters, change of cut-off/threshold values). Some of the multimodal methods in the literature allow the flexibility of choosing different input modalities, while others use fixed sets of modalities. There are both pros and cons associated with using either of the methods. Methods that are flexible allow users to use various available modalities and could handle datasets with missing modalities. On the other hand, the choice of input modalities becomes an additional parameter to tune when applied on an unknown dataset. Also, from our comparison analysis we observed that the use of more modalities might not always lead to better performance. For instance, for the OXVASC dataset, the best performance for BIANCA was achieved with T1 + FLAIR + MD, while TrUE-Net provided better SI values by using only T1 + FLAIR.

The results of evaluation of TrUE-Net on the unseen MWSC test datasets show that TrUE-Net performs well on data from different scanners, and in fact provides consistently high SI values, even for two additional datasets from unseen scanners (VU Amsterdam 3T Philips Ingenuity and VU Amsterdam 1.5T GE Signa HDxt). However, we observed that the cluster-wise performance metrics were lower for TrUE-Net when compared to the top ranking methods (table II in Kuijff et al. (2019)), despite the high SI values (<https://wmh.isi.uu.nl/results/fmrib-truenet-2/>). This shows that while TrUE-Net provides better segmentation of the detected true WMHs, it still misses small and subtle WMHs, especially in the unseen test datasets. Also, to further test generalisability, we trained TrUE-Net on the NDGEN dataset and tested it on the OXVASC dataset (different population, resolution and axis of acquisition). We observed that the NDGEN-trained model provides a segmentation performance comparable with the LOO evaluation on the OXVASC dataset, with significant increase in only voxel-wise FPR and H95 values (for more details and results, refer to supplementary material). Hence, an interesting future direction of research for this work could be to explore various domain adaptation techniques to improve the detection of WMHs across various unseen datasets, with the use of limited training data.

In conclusion, we proposed a model that provides accurate segmentation, with better performance than BIANCA and on par with the top ranking methods of MWSC 2017. We evaluated it on various datasets with different population and lesions characteristics. Regarding the tool availability, the python implementation of the training and evaluation codes for the TrUE-Net tool is currently available in <https://www.git.fmrib.ox.ac.uk/vaanathi/truenet> and the docker image of our method containing pretrained model (trained on MWSC) submitted to MWSC is available to download

from the docker hub (https://hub.docker.com/repository/docker/wmhchallenge/fmrib-truenet_2). Additionally, TrUE-Net will be integrated as an independent WMH segmentation tool in a future release of FSL, with options to facilitate easier training, testing and fine tuning of models on various datasets in a user-friendly manner.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Mark Jenkinson receives royalties from licensing of FSL to non-academic, commercial parties.

CRediT authorship contribution statement

Vaanathi Sundaresan: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft. **Giovanna Zamboni:** Supervision, Resources, Writing – review & editing. **Peter M. Rothwell:** Resources, Writing – review & editing. **Mark Jenkinson:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition, Project administration. **Ludovica Griffanti:** Conceptualization, Data curation, Supervision, Writing – review & editing, Project administration.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), Medical Research Council (MRC) [grant number EP/L016052/1] and Wellcome Centre for Integrative Neuroimaging, which has core funding from the Wellcome Trust (203139/Z/16/Z). The computational aspects of this research were funded from National Institute for Health Research (NIHR) Oxford BRC with additional support from the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. The Oxford Vascular Study is funded by the [National Institute for Health Research](#) (NIHR) Oxford Biomedical Research Centre (BRC), Wellcome Trust, Wolfson Foundation, the British Heart Foundation and the European Unions Horizon 2020 programme (grant 666881, SVDs@target). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. VS is supported by the Wellcome Centre for Integrative Neuroimaging. GZ is supported by the Italian Ministry of Education (MIUR) and by a grant "Dipartimenti di eccellenza 2018–2022", MIUR, Italy, to the Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia. PMR is in receipt of a NIHR Senior Investigator award. MJ is supported by the NIHR Oxford Biomedical Research Centre (BRC). LG is supported by the Oxford Parkinsons Disease Centre (Parkinsons UK Monument Discovery Award, J-1403), the MRC Dementias Platform UK (MR/L023784/2), and the National Institute for Health Research (NIHR) Oxford Health Biomedical Research Centre (BRC).

We acknowledge all the participants. For the NDGEN dataset, we are grateful to Prof. Gordon K. Wilcock and all the staff of Oxford Project to Investigate Memory and Ageing (OPTIMA) study. For the OXVASC dataset, we acknowledge the use of the facilities of the Acute Vascular Imaging Centre, Oxford. We also thank Dr. Chiara Vincenzi and Dr. Francesco Carletti for their help on generating the manual masks used in our experiments.

MJ receives royalties from licensing of FSL to non-academic, commercial parties. The authors report no potential conflicts of interest.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.media.2021.102184](https://doi.org/10.1016/j.media.2021.102184)

References

- Admiraal-Behloul, F., Van Den Heuvel, D., Olofsen, H., van Osch, M.J., van der Grond, J., Van Buchem, M., Reiber, J., 2005. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *Neuroimage* 28 (3), 607–617.
- Anbeek, P., Vincken, K.L., Van Osch, M.J., Bisschops, R.H., Van Der Grond, J., 2004. Probabilistic segmentation of white matter lesions in MR imaging. *Neuroimage* 21 (3), 1037–1044.
- Andermatt, S., Pezold, S., Cattin, P., 2016. Multi-dimensional Gated Recurrent Units for the Segmentation of Biomedical 3D-Data. In: *Deep Learning and Data Labeling for Medical Applications*. Springer, pp. 142–151.
- Aslani, S., Dayan, M., Storelli, L., Filippi, M., Murino, V., Rocca, M.A., Sona, D., 2019. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *Neuroimage* 196, 1–15.
- Biesbroek, J.M., Kuijff, H.J., van der Graaf, Y., Vincken, K.L., Postma, A., Mali, W.P., Biessels, G.J., Geerlings, M.I., Group, S.S., et al., 2013. Association between subcortical vascular lesion location and cognition: a voxel-based and tract-based lesion-symptom mapping study. the SMART-MR study. *PLoS ONE* 8 (4), e60541.
- Caligiuri, M.E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., Cherubini, A., 2015. Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review. *Neuroinformatics* 13 (3), 261–276.
- Damangir, S., Manzouri, A., Oppedal, K., Carlsson, S., Firbank, M.J., Sonnesyn, H., Tysnes, O.-B., O'Brien, J.T., Beyer, M.K., Westman, E., et al., 2012. Multispectral MRI segmentation of age related white matter changes using a cascade of support vector machines. *J. Neurol. Sci.* 322 (1–2), 211–216.
- De Boer, R., Vrooman, H.A., Van Der Lijn, F., Vernooij, M.W., Ikram, M.A., Van Der Lugt, A., Breteler, M.M., Niessen, W.J., 2009. White matter lesion extension to automatic brain tissue segmentation on MRI. *Neuroimage* 45 (4), 1151–1161.
- DeBette, S., Beiser, A., DeCarli, C., Au, R., Himali, J.J., Kelly-Hayes, M., Romero, J.R., Kase, C.S., Wolf, P.A., Seshadri, S., 2010. Association of MRI markers of vascular brain injury with incident stroke, mild cognitive impairment, dementia, and mortality: the framingham offspring study. *Stroke* 41 (4), 600–606.
- DeCarli, C., Fletcher, E., Ramey, V., Harvey, D., Jagust, W.J., 2005. Anatomical mapping of white matter hyperintensities (WMH) exploring the relationships between periventricular WMH, deep WMH, and total WMH burden. *Stroke* 36 (1), 50–55.
- Elisseeff, A., Pontil, M., et al., 2003. Leave-one-out error and stability of learning algorithms with applications. *NATO science series sub series iii computer and systems sciences* 190, 111–130.
- Fazekas, F., Chawluk, J.B., Alavi, A., Hurtig, H.I., Zimmerman, R.A., 1987. MR Signal abnormalities at 1.5T in Alzheimer's dementia and normal aging. *American Journal of Neuroradiology* 8 (3), 421–426.
- Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I.W., Sanchez, C.I., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., Platel, B., 2017. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Sci Rep* 7 (1), 5110.
- Ghafoorian, M., Karssemeijer, N., van Uden, I.W., de Leeuw, F.-E., Heskes, T., Marchiori, E., Platel, B., 2016. Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. *Med Phys* 43 (12), 6246–6258.
- Gibson, E., Gao, F., Black, S.E., Lobaugh, N.J., 2010. Automatic segmentation of white matter hyperintensities in the elderly using flair images at 3T. *J. Magn. Reson. Imaging* 31 (6), 1311–1322.
- Griffanti, L., Jenkinson, M., Suri, S., Zsoldos, E., Mahmood, A., Filippini, N., Sexton, C.E., Topiwala, A., Allan, C., Kivimäki, M., et al., 2017. Classification and characterization of periventricular and deep white matter hyperintensities on MRI: a study in older adults. *Neuroimage*.
- Griffanti, L., Jenkinson, M., Suri, S., Zsoldos, E., Mahmood, A., Filippini, N., Sexton, C.E., Topiwala, A., Allan, C., Kivimäki, M., et al., 2018. Classification and characterization of periventricular and deep white matter hyperintensities on MRI: a study in older adults. *Neuroimage* 170, 174–181.
- Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U.G., Kuker, W., Battaglini, M., Rothwell, P.M., et al., 2016. BIANCA (Brain intensity abnormality classification algorithm): a new tool for automated segmentation of white matter hyperintensities. *Neuroimage* 141, 191–205.
- Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joles, R., Wolz, R., Valdés-Hernández, M.d.C., Dickie, D., Wardlaw, J., et al., 2018. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical* 17, 918–934.
- Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. *Med Image Anal* 35, 18–31.
- Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. *Med Image Anal* 5 (2), 143–156.
- Jeon, S., Yoon, U., Park, J.-S., Seo, S.W., Kim, J.-H., Kim, S.T., Kim, S.I., Na, D.L., Lee, J.-M., 2011. Fully automated pipeline for quantification and localization of white matter hyperintensity in brain magnetic resonance image. *Int J Imaging Syst Technol* 21 (2), 193–200.

- Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal* 36, 61–78.
- Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P., 2016. On large-batch training for deep learning: generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.
- Khademi, A., Venetsanopoulos, A., Moody, A.R., 2011. Robust white matter lesion segmentation in FLAIR MRI. *IEEE Trans. Biomed. Eng.* 59 (3), 860–871.
- Kruggel, F., Paul, J.S., Gertz, H.-J., 2008. Texture-based segmentation of diffuse lesions of the brains white matter. *Neuroimage* 39 (3), 987–996.
- Kuijff, H.J., Biesbroek, J.M., de Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities: results of the WMH segmentation challenge. *IEEE Trans Med Imaging*.
- Lao, Z., Shen, D., Liu, D., Jawad, A.F., Melhem, E.R., Launer, L.J., Bryan, R.N., Davatzikos, C., 2008. Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Acad Radiol* 15 (3), 300–313.
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521 (7553), 436.
- Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.-S., Menze, B., 2018. Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *Neuroimage* 183, 650–665.
- Li, L., Simoni, M., Küker, W., Schulz, U.G., Christie, S., Wilcock, G.K., Rothwell, P.M., 2013. Population-based case-control study of white matter changes on brain imaging in transient ischemic attack and ischemic stroke. *Stroke* 44 (11), 3063–3070.
- Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571.
- Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention U-net: learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.
- Ong, K.H., Ramachandram, D., Mandava, R., Shuaib, I.L., 2012. Automatic white matter lesion segmentation using an adaptive outlier detection method. *Magn Reson Imaging* 30 (6), 807–823.
- Pantoni, L., Basile, A.M., Pracucci, G., Asplund, K., Bogousslavsky, J., Chabriat, H., Erkinjuntti, T., Fazekas, F., Ferro, J.M., Hennerici, M., et al., 2005. Impact of age-related cerebral white matter changes on the transition to disability—the LADIS study: rationale, design and methodology. *Neuroepidemiology* 24 (1–2), 51–62.
- Piantadosi, G., Sansone, M., Fusco, R., Sansone, C., 2020. Multi-planar 3d breast segmentation in mri via deep convolutional neural networks. *Artif Intell Med* 103, 101781.
- Prason, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M., 2013. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 246–253.
- Prins, N.D., Scheltens, P., 2015. White matter hyperintensities, cognitive impairment and dementia: an update. *Nature Reviews Neurology* 11 (3), 157.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.
- Rostrup, E., Gouw, A., Vrenken, H., van Straaten, E.C., Ropele, S., Pantoni, L., Inzitari, D., Barkhof, F., Waldemar, G., Group, L.S., et al., 2012. The spatial distribution of age-related white matter changes as a function of vascular risk factors—results from the LADIS study. *Neuroimage* 60 (3), 1597–1607.
- Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M., 2014. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 520–527.
- Rothwell, P., Coull, A., Giles, M., Howard, S., Silver, L., Bull, L., Gutnikov, S., Edwards, P., Mant, D., Sackley, C., et al., 2004. Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004 (Oxford vascular study). *The Lancet* 363 (9425), 1925–1933.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. *Int J Comput Vis* 115 (3), 211–252.
- Samaille, T., Fillon, L., Cuingnet, R., Jouvent, E., Chabriat, H., Dormont, D., Colliot, O., Chupin, M., 2012. Contrast-based fully automatic segmentation of white matter hyperintensities: method and validation. *PLoS ONE* 7 (11), e48953.
- Scheltens, P., Barkhof, F., Leys, D., Pruvo, J.P., Nauta, J., Vermersch, P., Steinling, M., Valk, J., 1993. A semiquantitative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging. *J. Neurol. Sci.* 114 (1), 7–12.
- Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., et al., 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage* 59 (4), 3774–3783.
- Shi, L., Wang, D., Liu, S., Pu, Y., Wang, Y., Chu, W.C., Ahuja, A.T., Wang, Y., 2013. Automated quantification of white matter lesion in magnetic resonance imaging of patients with acute infarction. *J. Neurosci. Methods* 213 (1), 138–146.
- Simoni, M., Li, L., Paul, N.L., Gruter, B.E., Schulz, U.G., Küker, W., Rothwell, P.M., 2012. Age- and sex-specific rates of leukoaraiosis in TIA and stroke patients: population-based study. *Neurology* 79 (12), 1215–1222.
- Smith, E.E., Biessels, G.J., De Guio, F., de Leeuw, F.E., Duchesne, S., Düring, M., Frayne, R., Ikram, M.A., Jouvent, E., MacIntosh, B.J., et al., 2019. Harmonizing brain magnetic resonance imaging methods for vascular contributions to neurodegeneration. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 11, 191–204.
- Smith, S.M., 2002. Fast robust automated brain extraction. *Hum Brain Mapp* 17 (3), 143–155.
- Steenwijk, M.D., Pouwels, P.J., Daams, M., van Dalen, J.W., Caan, M.W., Richard, E., Barkhof, F., Vrenken, H., 2013. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage: Clinical* 3, 462–469.
- Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramíó-Torrentà, L., Rovira, A., Oliver, A., Lladó, X., 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *Neuroimage* 155, 159–168.
- Wahlund, L., Barkhof, F., Fazekas, F., Bronge, L., Augustin, M., Sjögren, M., Wallin, A., Ader, H., Leys, D., Pantoni, L., et al., 2001. A new rating scale for age-related white matter changes applicable to MRI and CT. *Stroke* 32 (6), 1318–1322.
- Wang, Y., Catindig, J.A., Hilal, S., Soon, H.W., Ting, E., Wong, T.Y., Venkatasubramanian, N., Chen, C., Qiu, A., 2012. Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts. *Neuroimage* 60 (4), 2379–2388.
- Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., O'Brien, J., Barkhof, F., Benavente, O.R., et al., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology* 12 (8), 822–838.
- Wilson, D.R., Martinez, T.R., 2003. The general inefficiency of batch training for gradient descent learning. *Neural networks* 16 (10), 1429–1451.
- Xu, Y., Géraud, T., Puybureau, É., Bloch, I., Chazalon, J., 2017. White matter hyperintensities segmentation in a few seconds using fully convolutional network and transfer learning. In: International MICCAI Brainlesion Workshop. Springer, pp. 501–514.
- Yang, F., Shan, Z.Y., Kruggel, F., 2010. White matter lesion segmentation based on feature joint occurrence probability and χ^2 random field theory from magnetic resonance (MR) images. *Pattern Recognit Lett* 31 (9), 781–790.
- Yoo, B.I., Lee, J.J., Han, J.W., Lee, E.Y., MacFall, J.R., Payne, M.E., Kim, T.H., Kim, J.H., Kim, K.W., et al., 2014. Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images. *Neuroradiology* 56 (4), 265–281.
- Zamboni, G., Griffanti, L., Mazzucco, S., Pendlebury, S.T., Rothwell, P.M., 2019. Age-dependent association of white matter abnormality with cognition after TIA or minor stroke. *Neurology* 10–1212.
- Zamboni, G., Wilcock, G.K., Douaud, G., Drazich, E., McCulloch, E., Filippini, N., Tracey, I., Brooks, J.C., Smith, S.M., Jenkinson, M., et al., 2013. Resting functional connectivity reveals residual functional activity in alzheimers disease. *Biol. Psychiatry* 74 (5), 375–383.
- Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging* 20 (1), 45–57.
- Zhang, Y., Chen, W., Chen, Y., Tang, X., 2018. A post-processing method to improve the white matter hyperintensity segmentation accuracy for randomly-initialized U-net. In: 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP). IEEE, pp. 1–5.
- Berseth, M., 2017 WMH Segmentation Challenge, MICCAI.