Contents lists available at ScienceDirect

Neuroimage: Reports

journal homepage: www.sciencedirect.com/journal/neuroimage-reports

Segmentation of white matter lesions in multicentre FLAIR MRI☆

April Khademi [a,e,f], Adam Gibicar [a,*], Giordano Arezza [a], Justin DiGregorio [a], Pascal N. Tyrrell [b,c,d], Alan R. Moody [b]

[a] Image Analysis in Medicine Lab (IAMLAB), Department of Electrical, Computer, and Biomedical Engineering, Ryerson University, Toronto, Canada
[b] Department of Medical Imaging, University of Toronto, Toronto, Canada
[c] Department of Statistical Sciences, University of Toronto, Toronto, Canada
[d] Institute of Medical Science, University of Toronto, Toronto, Canada
[e] Keenan Research Center for Biomedical Science, St. Michael's Hospital, Unity Health Network, Toronto, Canada
[f] Institute for Biomedical Engineering, Science, and Technology (iBEST), A Partnership Between St. Michael's Hospital and Ryerson University, Toronto, Canada

ARTICLE INFO

Keywords:
FLAIR MRI
White matter lesions (WML)
Segmentation
Deep learning

ABSTRACT

White matter lesions (WML) in the brain are thought to be related to ischemic processes, demyelination, and axonal degeneration. The presence of WML predict cognitive decline, dementia, stroke, and death. Lesion progression increases these risks, making WML significant clinical biomarkers for investigation. To analyze WML objectively, consistently, and efficiently, automated WML segmentation methods for neurological MRI have been the focus of extensive research efforts. There have been many unsupervised and traditional machine learning methods proposed over the years. Recently, deep learning architectures have been utilized for WML segmentation with promising results. In this work, we evaluate seven WML segmentation tools for multicentre fluid attenuated inversion recovery (FLAIR) MRI. Two traditional methods were evaluated, one unsupervised method and the other a traditional machine learning approach. The traditional methods were compared to five deep learning-based approaches. FLAIR MRI have the advantage of highlighting WML lesions robustly and are used routinely in neurological workflows. Automated WML segmentation tools for FLAIR MRI could optimize clinical workflows and improve patient care. The WML segmentation algorithms were evaluated on a multicentre, multi-disease FLAIR MRI database acquired with varying scanners and protocols. In total 252 imaging volumes (~13 K image slices) with annotations, from 5 multicentre datasets (33 imaging centres) were used to train, validate and test the WML segmentation methods. Two clinical datasets, which include dementia and vascular disease pathologies, and three open-source datasets were used. To examine clinical utility of each algorithm and establish proof of effectiveness, algorithms were evaluated over several dimensions related to accuracy, generalizability, and robustness to pathology. This work presents a framework for evaluating the efficacy of WML segmentation algorithms for improved reliability, patient safety and clinical trials. Of all methods, SC U-Net was found to be the best algorithm for WML segmentation in terms of highest Dice similarity coefficient (DSC) over most dimensions (mean DSC = 0.71 over all volumes). Deep learning methods outperformed traditional methods, especially in lower lesion loads, but were not able to generalize across all disease categories or datasets.

1. Introduction

White matter lesions (WML), or leukoaraiosis, is routinely found in the aging brain and are established cerebral vascular disease (CVD) markers (Wardlaw et al., 2015, Pantoni, 2010, Azizyan et al., 2011). WML represent increased and altered water content in hydrophobic white matter fibers and tracts. Changes in white matter vasculature likely contributes to WML pathogenesis (Gorelick et al., 2011). WML may be the result of ischemic injury from decreases in regional cerebral blood flow (Pantoni and Garcia, 1997). Demyelination and axonal degeneration have also been suggested as probable mechanisms (Wardlaw et al., 2015). Typically, WML manifest as multifocal, diffuse

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

* Corresponding author.

E-mail address: agibicar@ryerson.ca (A. Gibicar).

https://doi.org/10.1016/j.ynirp.2021.100044
Received 31 March 2021; Received in revised form 22 June 2021; Accepted 31 July 2021
Available online 5 August 2021

periventricular or subcortical lesions of varying morphologies (Marek et al., 2018). The presence of WML is associated with cognitive decline, dementia, stroke, death, and lesion progression increases these risks (Debette and Markus, 2010, Alber et al., 2019). Therefore, WML are significant clinical biomarkers for investigation.

In T2-weighted and fluid-attenuated inversion recovery (FLAIR) magnetic resonance images (MRI), WML appear as hyperintense signals in the cerebral white matter (Marek et al., 2018). FLAIR MRI is preferred for WML analysis (Azizyan et al., 2011, Badji and Westman, 2020, Wardlaw et al., 2013), since the high signal from the cerebrospinal fluid (CSF) in T2 is suppressed, thus highlighting white matter disease (Lao et al., 2008). This is due to increased water content secondary to ischemia and demyelination and much more robustly seen in FLAIR than with T1/T2 (Gorelick et al., 2011). Characterization of WMLs is typically performed by a radiologist using visual rating systems such as the Fazekas scale (Fazekas et al., 1993) or by manual segmentation (Caligiuri et al., 2015). The Fazekas scale grades lesions by size, location and confluence but is subjective (Caligiuri et al., 2015). Manual segmentation is time-consuming, laborious, and has high inter and intra-variability (Caligiuri et al., 2015). For objective, consistent, and efficient WML analysis, automated WML segmentation methods have been the focus of extensive research efforts in recent decades.

Several unsupervised methods for WML segmentation have been proposed in the literature, including clustering and thresholding (Caligiuri et al., 2015). Jack et al. proposed a thresholding method for segmenting WML from FLAIR MRI based on step-wise regression (Jack et al., 2001). Statistical measures from the image histogram were used to find thresholds for separating cerebrospinal fluid (CSF), normal brain tissue (i.e. GM/WM) and WML. Admiraal-Behloul et al. proposed a two-level segmentation scheme (i.e. adaptive and reasoning) which combined information from proton density (PD), T2-weighted and FLAIR images (Admiraal-Behloul et al., 2005). In the adaptive stage, intensity values were mapped to linguistic variables such as bright and dark which was used with a fuzzy inference system to derive a label for each voxel (Admiraal-Behloul et al., 2005). Seghier et al. proposed a fuzzy classification algorithm for lesion segmentation based on outlier detection (Seghier et al., 2008) that identified outlier voxels in normalized GM and WM probability maps and had a high sensitivity for detecting lesions with varying characteristics (Seghier et al., 2008). Khademi et al. proposed an unsupervised method for segmenting WML with sub-voxel precision from FLAIR images by modeling the partial volume artifact (Khademi et al., 2011, Khademi et al., 2014, Khademi and Moody, 2015).

Supervised machine learning methods have also been prominent in the literature for WML segmentation. Anbeek et al. proposed a supervised method for segmenting WML from multi-modal MR images using a k-Nearest Neighbour (k-NN) classifier (Anbeek et al., 2004). The method incorporated both intensity and spatial information from registered T1-weighted (T1-w), inversion recovery (IR), proton density (PD), T2-weighted (T2-w) and FLAIR images. Lao et al. combined support vector machines (SVM) and AdaBoost (Lao et al., 2008) with a derived attribute vector (AV) comprising local intensity and spatial features obtained from FLAIR, PD, T2-w and T1-w images. AdaBoost was used to address the large class imbalance issue in WML segmentation by ensuring that classifier weights were more impacted by misclassified cases. In (De Boer et al., 2009), de Boer et al. proposed a method for segmenting CSF, GM and WM using an atlas-based k-NN classifier on multi-modal MRI data. The resultantGM segmentation was used to automatically find a threshold for segmenting WML in FLAIR MRI (De Boer et al., 2009). In the work by Simoes et al. (Simões et al., 2013), WML were segmented from 3-D FLAIR volumes using Gaussian mixture models (GMMs) which models each volume by three distinct classes: CSF, GM/WM and WML. Voxel-wise class probabilities were determined and subsequently thresholded to derive class-labels for each voxel (Simões et al., 2013). Knight et al. (2018) proposed a voxel-wise regression technique to segment WML in FLAIR MRI that built upon the open-source Lesion Segmentation Tool (LST) (Schmidt, 2017). A spatially parameterized logistic regression classifier was used to segment lesions on a per voxel basis. Grifftanti et al. proposed the Brain Intensity AbNormality Classification Algorithm (BIANCA), which manipulates different options for weighting spatial information, local spatial intensity averaging and different parameters for the number and location of training points in a k-NN classifier (Griffanti et al., 2016).

In light of these results, numerous deep learning (DL) architectures have been proposed for WML segmentation. Several of the top competitors in the MICCAI WML Segmentation Challenge used deep learning methods to robustly segment WML of presumed vascular origin, with the first-place team applying an ensemble of three U-Nets with different initializations (Kuijf et al., 2019, Ronneberger et al., 2015, Li et al., 2018a). Guerrero et al. proposed UResNet2 for simultaneous segmentation and differentiation of WML and stroke lesions (Guerrero et al., 2018). Conventional convolution blocks were replaced with residual blocks to improve stability and convergence (Guerrero et al., 2018). By training on 2-D patches, the authors were able to segment WML and stroke lesions at the same time. In (Moeskops et al., 2018), authors use multiple MRI modalities and CNNs with different patch sizes to capture multiresolution information for WML segmentation. A conventional U-Net CNN was applied on FLAIR MRI to segment WML and several other hyperintense pathologies in the brain (Duong et al., 2019). Recently, Wu et al. proposed an architecture called the Skip-Connection U-Net (SC U-Net) that added four additional skip connections to the original architecture (Wu et al., 2019). The authors compared SC U-Net to the U-Net proposed by Li et al. (2018a), without the use of ensembles, and found that SC U-Net outperformed U-Net on the WML Segmentation Challenge dataset (Kuijf et al., 2019). These works further demonstrate that deep learning can be successfully used for WML segmentation.

It is difficult to directly compare WML segmentation methods since they are typically evaluated on different datasets using different evaluation criteria (Caligiuri et al., 2015). The MICCAI WML Segmentation Challenge was developed to address this by allowing participants to directly compare techniques on a robust, open-source dataset using a standardized set of evaluation criteria (Kuijf et al., 2019). Since then there have been comparisons of WML algorithms, such as in (Heinen et al., 2019) where the performance of five automated WML segmentation methods were evaluated in a multicentre FLAIR and T1 dataset. The methods mainly consisted of traditional machine learning (ML) algorithms, including K-NN and LST. Performance is reported for sixty volumes from six different centres. Using similar (traditional) WML segmentation methods, in (de Sitter et al., 2017), the authors investigate five WML segmentation tools for multiple sclerosis (MS) lesion segmentation using FLAIR and T1 images. In total 70 patients from 6 centres were used to evaluate the methods. Many works have compared on small or moderate sized datasets that may not reflect the natural variability of clinical datasets. In (Vanderbecq et al., 2020), the authors considered seven open source traditional WML segmentation methods for T1 and FLAIR and studied performance on research and clinical datasets. In (Frey et al., 2019), the authors provide a meta-review of the current WML segmentation methods applied in large-scale MRI studies.

In this work, we evaluate seven WML segmentation tools for a large multicentre FLAIR MRI dataset. Two traditional methods (unsupervised, machine learning) and five deep learning-based approaches utilizing CNN architectures are selected and implemented due to their promise for WML segmentation in FLAIR MRI. Specifically, 2D patch-based approaches for U-Net and U-Net variants are evaluated in this work as the top performing architecture in the MICCAI challenge was based on 2D U-Nets (Li et al., 2018a). 2.5D and 3D architectures were not explored due to memory constraints and loss of lesion continuity due to thick slices in FLAIR MRI (DiGregorio et al., 2021). Although there are a variety of WML segmentation methods, most are dependent on a secondary T1 MRI or multiple sequences and cannot be directly translated to FLAIR (DiGregorio, 2018). FLAIR has the advantage of highlighting WML lesions better than other sequences (Azizyan et al., 2011, Gorelick et al., 2011, Badji and Westman, 2020, Wardlaw et al., 2013, Lao et al., 2008) and is used routinely in neurological workflows. In (Narayana et al., 2020), using FLAIR as a sole input to a CNN model was shown to provide similar WML segmentation performance as compared to models trained with other or multimodal sequences, and FLAIR on its own showed a lower false positive rate in the low lesion load cases. Therefore, automated WML segmentation tools using DL for FLAIR MRI could be valuable tools for clinical workflows.

Although many proof of concept biomarkers exist (tested in single centres with limited variability), there are only a few that are technically validated in large multicentre datasets to establish “proof of effectiveness” (Smith et al., 2019) which is a barrier to translation. Technical validation includes tests related to feasibility, accuracy, reproducibility and repeatability (Smith et al., 2019, Sullivan et al., 2015, Obuchowski et al., 2015), and should precede clinical validation, otherwise it is difficult to determine if biomarker changes are from the biological process or technical variability of the biomarkers (Smith et al., 2019). Biomarkers should be investigated on smaller, more controlled datasets; then scaled to large multi-centre sets to prove effectiveness. To examine the performance of each algorithm on clinical datasets and establish proof of effectiveness, the WML segmentation algorithms are evaluated over several dimensions related to accuracy, generalizability and robustness to pathology on a multicentre, multi-disease FLAIR MRI database acquired with varying scanners and protocols. In total 252 imaging volumes (~13 K image slices) with annotations, from 5 multicentre datasets (33 imaging centres) were used to train, validate and test the WML segmentation tools. There are three open-source datasets (MICCAI, ADNI, MRBrains) and two clinical datasets (CAIN, CCNA) which includes both dementia and vascular disease pathologies. The methodology presented to evaluate the effectiveness of a tool is novel and presents a unique framework for evaluating the efficacy of WML segmentation algorithms for robustness and reliability which translates into improved patient safety and more sensitive clinical trials.

2. Materials and methods

2.1. Data

Experimental data for this work comes from 5 multicentre FLAIR MRI datasets. The first dataset is from the Alzheimer’s disease Neuroimaging Initiative (ADNI) (Jack et al., 2008) and includes 900 subjects with longitudinal follow up (4126 imaging volumes). The second database is from the Canadian Atherosclerosis Imaging Network (CAIN) (Tardif et al., 2013), a pan-Canadian clinical study on vascular disease. There are 400 subjects in CAIN with follow up for a total of 871 volumes. The third dataset is from the Canadian Consortium on Neurodegeneration in Aging (CCNA), a pan-Canadian clinical study to analyze different types of dementia (Chertkow et al., 2019, Mohaddes et al., 2018). Currently, the FLAIR data from CCNA contains imaging volumes for 380 subjects, acquired at 20 imaging centers. Sixty volumes from the MICCAI WML Segmentation Challenge (Kuijf et al., 2019) are used, and 7 vol from the Challenge on MR Brain Segmentation at MICCAI 2018 (MRBrains). MICCAI and MRBrains have WML annotations for all the volumes. To generate annotations for CAIN, CCNA and ADNI, WML were manually segmented using ITK-SNAP[1] (Yushkevich et al., 2006) by three medical students trained by the same radiologist. There was a standard protocol employed and several review sessions before annotating began. One rater employed a semi-automated intensity-based region growing tool from ImageJ that fine-tuned boundaries due to partial volume averaging. It was a minor correction step that resulted in few pixel changes. In total there are 252 ground truth volumes, with 135 CAIN, 20 ADNI, 30 CCNA, 7 MRBrains and 60 from MICCAI. Each subject (imaging volume) is unique. Each dataset contains FLAIR MRI acquired in the axial plane at 3 T from General Electric (GE), Philips, and Siemens scanners. Ground truth datasets were stratified by center to ensure there was broad representation in the data from multiple centres and scanner models which could have varying acquisition protocols. Table 1 contains the demographic and imaging acquisition parameters for the sampled ADNI, CAIN, and CCNA volumes demonstrating the diversity of the data. The average and standard deviation of lesion load (LL) in mL is also reported, to highlight differences in the ground truth datasets.

2.2. Reliability of manual reference segmentation

To demonstrate the robustness of the WML manual segmentation protocol and training of the raters, an inter-rater agreement experiment was completed that included two raters and 20 unique FLAIR imaging volumes with varying WML loads. Both raters received the same training and conducted the WML annotations on the same 20 volumes. Both raters were blinded to each other’s results. A secondary dataset of 10 subjects was used to compare the manual segmentations from a single rater to that of the semi-automated correction in ImageJ. To measure inter-observer variability, the DSC, intraclass correlation coefficient (ICC), IAVD and EF are reported. ICC estimates and their 95 % confidence intervals were calculated based on a mean-rating ($k = 2$), absolute-agreement, 2-way random-effects model. To quantify agreement between the semi-automated and manual approach, a Bland-Altman plot was also used.

2.3. Pre-processing

Intensity standardization was performed to remove variability caused by the multicentre effect (Zhong et al., 2012, Reiche et al., 2019) for all deep learning methods and the unsupervised partial volume averaging (PVA) technique. The lesion prediction algorithm (LPA) has a built-in preprocessing pipeline so standardization was not applied for this method. For the unsupervised and deep learning methods, all volumes were preprocessed using (Reiche et al., 2019), which performs denoising, bias field correction, and intensity standardization. Imaging volumes are denoised with 3 × 3 median filtering followed by homomorphic filtering for bias field correction. The MICCAI dataset had bias field correction already applied so it was not used for this dataset. Intensity standardization is achieved through a novel scaling factor that aligns the histogram modes of volumes to that of an atlas. As shown in (Reiche et al., 2019), the intensity intervals of tissues in 350 K FLAIR MRI are more consistent across multicentre data using this approach. Skull-stripping is performed on the volumes using U-NET for intracranial volume (ICV) segmentation (DiGregorio et al., 2021).

2.4. WML segmentation algorithms

Two categories of algorithms were evaluated. The first category of algorithms was based on traditional approaches, including image processing and machine learning (ML) and included the partial volume averaging (PVA) modeling method and lesion prediction algorithm (LPA). The second category is based on DL and CNNs: U-Net, SC U-Net, MulitResUNet, UResNet2, and Tiramisu.

2.4.1. Partial volume average (PVA) modeling

An artifact found in magnetic resonance images (MRI) called partial volume averaging (PVA) has received much attention in the image processing community since accurate segmentation of cerebral anatomy and pathology is impeded by this artifact. For robust WML segmentation, a partial volume (PV) fraction estimation approach was developed for cerebral MRI that measures the proportion of each tissue in every voxel. The PV fraction is estimated directly from each image using a global edge metric that was shown to be proportional to the PV fraction. The estimated PVA fraction is used to compute intensity-based class memberships that are applied to segment anatomy and pathology with

[1] www.itksnap.org.

Table 1
FLAIR MRI ground truth datasets used for experimentation. Repetition time (TR), echo time (TE), inversion time (TI), and pixel spacing are represented by the range found in the data. Sex reported as percetage of women F (%) and age and LL reported as averages over the cohorts.

Patient Information

Database	Disease	Volumes	Images	Patients	Centres	Age ± SD (yrs.)	F (%)	LL±SD (mL)
ADNI	Dementia	20	700	20	14	76.0 ± 8.2	47	11.8 ± 10.1
CAIN	Vascular	135	6480	135	8	71.7 ± 6.0	27	12.2 ± 12.3
CCNA	Dementia	30	1440	30	7	77.5 ± 6.0	33.3	22.8 ± 18.8
MICCAI	Vascular	60	3580	60	3	–	–	17.6 ± 17.4
MRBRAINS	Normal/WML	7	336	7	1	–	–	22.0 ± 24.3
Total	All	252	12.5 K	252	33	–	–	15.0 ± 15.2

Acquisition Parameters

Database	**GE/Philips/Siemens**	**Mag Field (T)**	**TR (ms)**	**TE (ms)**	**TI (ms)**	**Pixel Spacing (mm)**	**Slice Thickness (mm)**
ADNI	7/6/7	3	9000–11,000	90–154	2250–2500	0.8594	5
CAIN	17/100/18	3	9000–11,000	117–150	2200–2800	0.4295–1	3
CCNA	2/3/25	3	9000–9840	125–144	2250–2500	0.9375	3
MICCAI	20/20/20	3	4800–11,000	82–279	1650–2500	0.9583–1.2	3
MRBRAINS	0/7/0	3	11,000	125	2800	0.958	3
Total	46/136/70	1.5–3	4800–11,000	82–279	1650–2800	0.4295–1.2	3–5

subvoxel accuracy. The PVA method was applied to FLAIR MRI lesions (Khademi et al., 2011, Khademi et al., 2014) and to segment normal anatomy (GM, WM) in T1 images (Khademi and Moody, 2015). The partial volume fraction assigns a value of 1 to voxels that are pure WML, 0 for voxels that do not contain any WML tissue, and an intermediate value for mixture (PVA) voxels (Caligiuri et al., 2015, Khademi et al., 2011). WML are segmented by thresholding this map.

2.4.2. *Lesion prediction algorithm (LPA)*

The lesion prediction algorithm (LPA) is an open-source WML segmentation tool that only requires a FLAIR image and does not require parameters from the user. The method was built by training a logistic regression model with the data of 53 MS patients with severe lesion patterns (Schmidt, 2017). The features that are considered are intensity and spatial location. A novel approach for fitting large-scale regression models was used to estimate the high dimensional model. The LPA algorithm utilizes a built in preprocessing pipeline, including intensity normalization, so additional preprocessing was not applied before applying LPA. The open source Matlab code is available from the author's personal website ("T – Lesion segmentatio, 2021).

2.4.3. *U-Net*

U-Net was proposed in 2015 and has been a mainstay in medical image segmentation research because of its ability to adapt to variable biomedical data (Ronneberger et al., 2015, Hwang et al., 2019, Thakur et al., 2020). The encoding path with units of convolutional and max pooling layers perform feature extraction. The decoding path contains units of convolutional and transposed convolutional layers and skip connections to recapture spatial context (Long et al., 2015). U-Net contains 5 levels where the filter depth is doubled during each downsampling block (via max pooling) and halved during each upsampling block (via transposed convolution). In this work, U-Net was implemented with batch normalization layers succeeding convolutional layers (Ioffe and Szegedy, 2015) to accelerate convergence and improve generalization via a modest regularization effect. The structure of the U-Net encoding and decoding units used in this work are shown in Fig. 1.

2.4.4. *SC U-Net*

The skip connection U-Net (SC U-Net) proposed in (Wu et al., 2019) adds additional paths (skip connections) between the shallow and deep layers of a CNN architecture. The outputs from each max-pooling layer in the encoder are inputs for each transposed convolution layer in the decoder. Skip connections ease training by improving information and back-propagation flow (Wu et al., 2019, Drozdzal et al., 2016). This has been shown to diminish the vanishing gradient problem that commonly occurs when training deep networks (Drozdzal et al., 2016). The SC U-Net architecture shown in (Wu et al., 2019) was implemented.

2.4.5. *UResNet2*

Semantic segmentation architectures can be equipped with residual connections; a type of skip connection that has demonstrated strong performance in image recognition tasks (He et al., 2016a). Residual connections enable direct information flow between network layers and ease optimization via identity mappings and after addition activations (He et al., 2016a). In this work, U-Net was modified to have residual connections as in the implementation of (Guerrero et al., 2018) which

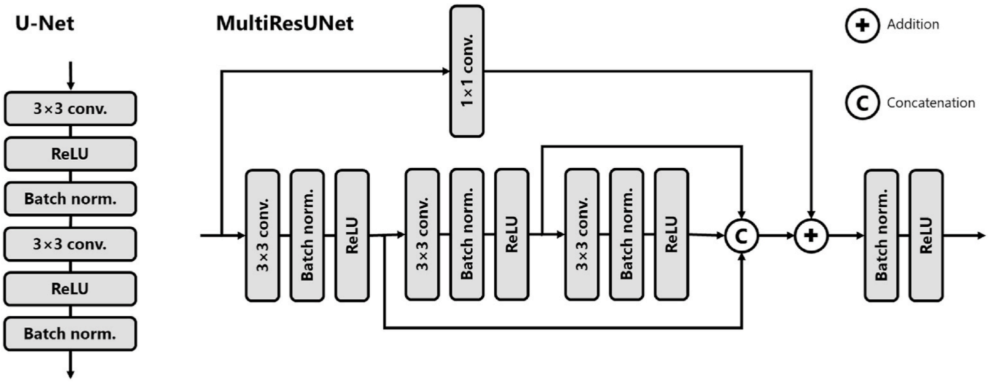


Fig. 1. Encoding/decoding units for U-Net, and MultiResUNet. Encoding units reside between max pooling layers and decoding units reside between transposed convolutional layers.

was shown to outperform other deep-learning methods for WML segmentation by a significant margin. UResNet2 replaces the encoding units of U-Net with residual elements containing two convolutional elements which was shown to improve convergence speeds (Guerrero et al., 2018).

2.4.6. *MultiResUNet*

The multiple resolution U-Net (MultiResUNet) was proposed in (Ibtehaz and Rahman, 2020) as an architecture better suited for analyzing images at multiple scales. In U-Net, the encoding/decoding units contain successive 3 × 3 convolutional layers, which are equivalent to a single 5 × 5 convolution (Szegedy et al., 2016). MultiResUNet expands on this concept by replacing all encoding/decoding units with "MultiRes" blocks; units that concatenate the outputs of 3 successive 3 × 3 convolutional layers and bind them with a residual connection. This efficiently captures features at the 3 × 3, 5 × 5, and 7 × 7 scales. MultiResUNet also replaces all skip connections with "Res" paths, sequences of residual convolutional layers. The authors theorized the non-linear operations within these modified skip connections would reduce the semantic gap between shallow encoder and deep decoder features. The "MultiRes" block from (Ibtehaz and Rahman, 2020) is modified to follow the "batch normalization after addition" layer structure (He et al., 2016b). Fig. 1 shows these "MultiRes" units.

2.4.7. *Tiramisu*

The Tiramisu architecture is a U-Net like network modified to contain dense connections (Jégou et al., 2017) where each layer is connected to every other layer in a feed-forward manner to grant the network direct access to input and loss function gradients and ease training (Huang et al., 2017). This often results in performance gains due to an avoidance of redundant feature learning and some provided regularization which reduces overfitting (Li et al., 2018b). The Tiramisu architecture strategically confines dense connections to the encoding and decoding units to avoid excessively large feature maps and low-level information loss (Jégou et al., 2017). In this work, we use the Tiramisu network proposed in (Jégou et al., 2017) which contains two transition blocks and two dense blocks in the encoding and decoding paths.

2.5. *Evaluation metrics*

To evaluate the performance of the WML segmentation algorithms several metrics are used. The Dice similarity coefficient (DSC) (Wu et al., 2019) is used to measure spatial overlap between the predicted WML mask (Seg) and corresponding manual ground truth (GT):

$$DSC = \frac{2|GT \cap Seg|}{|GT| + |Seg|}$$

where the DSC ranges between 0 and 1, and a value of 1 implies perfect overlap. To investigate the degree of over-segmentation, the extra fraction (EF) was computed:

$$EF = \frac{FP}{TP + FN}$$

where TP are the true positives, FP are the false positives, and FN is the false negatives of the automated ICV estimation as compared to the ground truth (Anbeek et al., 2004). Over-segmentation (i.e., inclusion of regions that are not WML) results in higher EF rates. Average volume difference in this work is measured using the absolute log-transformed volume difference (lAVD) (Kuijf et al., 2019). lAVD quantifies the difference between the ground truth (V_{GT}) and predicted volume (V_{seg}) by:

$$lAVD = \left| log \frac{V_{seg}}{V_{GT}} \right|$$

where smaller lAVD implies a better segmentation.

Hausdorff distance (HD) quantifies local differences between the predicted WML mask and ground truth by the distance between two subsets of points in a metric space:

$$HD_{95} = max_{xeGT} min_{yeBM} \|x - y\|$$

where smaller distances imply a greater degree of similarity. The 95th percentile HD was used to improve robustness and reduce sensitivity to noise. To consider variability in the metrics, the coefficient of variation is computed, as

$$CoV = \frac{\sigma_{metric}}{\mu_{metric}}$$

where σ_{metric} and μ_{metric} are the standard deviation and mean of a performance metric in an experiment. Bland-Altman plots were used to measure the volume difference between manual and automatically predicted volumes as a function of lesion load.

2.6. *Experimental design*

For a tool to be clinically adopted, it must be accurate, generalize to new datasets and scanners and robust to challenges in data (such as pathology). As a result, several dimensions related to the effectiveness of WML segmentation will be assessed: (1) accuracy, (2) generalization capabilities and (3) robustness to pathology.

2.6.1. *Data splits*

Datasplit1 is a 75/25 training/validation to testing ratio of all 252 FLAIR MRI volumes resulting in 189 vol for training/validation, and 62 vol for testing. Stratified splitting was used so the proportions of CAIN, ADNI, CCNA, MICCAI and MRBrains volumes in the training, validation, and test sets mirrored that of the overall population. Sampling was completed to stratify across each scanner type (GE, Siemens, and Philips) where possible. Using the training/validation volumes, 64 × 64 patches are extracted with 50 % overlap and 20 % of the patches were allocated for validation. This resulted in roughly 126 K training and 32 K validation patches. During test time, each test volume is patched, predictions are computed and the predicted WML volume is reconstructed. This process was repeated for each fold using four fold cross validation (with no overlap in testing data across the folds). This ensured every imaging volume of the 252 was tested on through one of the folds. Datasplit1 examines performance in an ideal scenario where training data is available for each dataset and can be used to establish ideal performance benchmarks.

Datasplit2 is used to mimic real-world models where a single dataset could be used to generate ground truths and train models. MICCAI data was used exclusively for training with a 80/20 split with 48 vol for training/validation. Patches were sampled for the 48 training and 20 % of the patches were used for validation, resulting in roughly 22 K training and 6 K validation patches. The remaining (held out) 20 ADNI, 135 CAIN, 30 CCNA, 7 MRBrains and 12 MICCAI volumes were used for testing. Since the datasets are of different pathology, centres and scanners this experiment can be used to examine generalizability. The datasplits and data processing pipeline for deep learning models can be seen in Fig. 2.

2.6.2. *Accuracy*

The average accuracy of all methods is investigated for models trained and tested from datasplit1 and datasplit2 and is conducted to collect descriptive statistics. Results generated from datasplit1 models were emphasized since it represents the ideal setup of increased data diversity. Accuracy included evaluation metric distributions, means, coefficient of variation (CoV), as well as correlations between the true and predicted volumes.

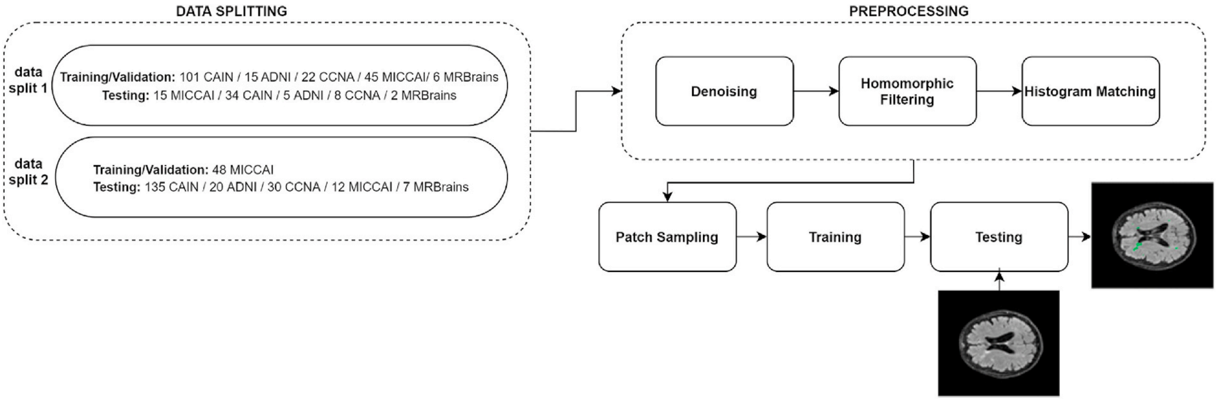

DATA SPLITTING
data split 1
Training/Validation: 101 CAIN / 15 ADNI / 22 CCNA / 45 MICCAI/ 6 MRBrains
Testing: 15 MICCAI / 34 CAIN / 5 ADNI / 8 CCNA / 2 MRBrains
data split 2
Training/Validation: 48 MICCAI
Testing: 135 CAIN / 20 ADNI / 30 CCNA / 12 MICCAI / 7 MRBrains
PREPROCESSING
Denoising
Homomorphic Filtering
Histogram Matching
Patch Sampling
Training
Testing

Fig. 2. Data splits and processing pipeline for training and testing the deep learning models. All methods use the same subsets of data.

2.6.3. *Generalization*

To maximize translation opportunities, WML segmentation algorithms should be robust across different centres, datasets and scanning devices. Evaluation metrics generated from datasplit1 and datasplit2 are used to assess generalization across scanner vendors, and unseen databases. See example images from different databases and scanner vendors in Fig. 3.

2.6.4. *Robustness to pathology*

WML segmentation algorithms may have variable performance depending on the lesion load or disease severity. For clinical use, tools ideally would have equal performance (robustness) across varying degrees of pathology. In these experiments WML segmentation performance is measured over varying levels of disease. datasplit1 was used for all experiments and only traditional (PVA, LPA) and top CNN methods are compared. To compare results across different lesion loads, the ground truth lesion loads were stratified by low (<5 mL), medium (5–15 mL), or high (>15 mL) (de Sitter et al., 2017) as shown in Table 2. Example images for each lesion category are shown in Fig. 4. To investigate WML segmentation performance as a function of overall neurodegeneration, the Montreal Cognitive Assessment score (MoCA) was used to categorize patients into cognitively normal and impaired, as shown in Table 3. MoCA is a clinical screening tool used to gauge cognitive impairment and overall neurodegeneration through various cognitive tests (Nasreddine et al., 1581).

2.7. *Statistical analysis*

To compare algorithm performance over different co-variates (i.e., scanner vendor, dataset, pathology level) statistical analysis is completed. A single evaluation metric (DSC) was chosen to simplify analysis as it is a strong indicator of overall performance and is interpretative. For each group, the mean DSC was statistically compared across groups using analysis of variance (ANOVA). ANOVA is used to determine whether an algorithm has similar performance (DSC outcome variables) over different predictor variables (dataset, scanner, pathology). ANOVA was selected based on descriptive statistics and goodness-of-fit-tests for normal distributions (i.e., Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling) on DSC values across all WML segmentation methods. DSC values underwent reflection with a single logarithmic transformation to improve linearity and homogeneity of variance prior to analysis (Dobson and Barnett, 2018). When ANOVA tests were significant, Tukey-Kramer post-hoc analysis for multiple comparisons were used to determine the sources of differences. Adjusted DSC was used as the primary outcome variable.

Table 2
Number of volumes with low, medium, or high lesion loads from datasplit1.

Pathology	Categorization	Volume	No. Test Subjects
WML	Low	<5 mL	68
	Medium	5–15 mL	99
	High	>15 mL	85

2.8. *Implementation details*

For all deep learning (DL) methods, the generalized dice loss was used (Sudre et al., 2017), Adam Optimizer with a learning rate of 1e-4 over 75 epochs, batch size of 64 (except for 32 for the Tiramisu network) and the images were patched into 64 × 64 regions with 50 % overlap, based on the findings of (Bernal et al., 2019). Moreover, 64 × 64 patches provided the best performance likely because they offer a more localized representation for lesions, while not being too small of an image size for the input. Slight data augmentations were applied for rotation, scaling, shearing, scaling and translation (Li et al., 2018a). For each architecture, the filter sizes were implemented as described in their respective papers. The number of trainable parameters for each model are: 8.23 M, 7.94 M, 2.11 M, 8.26 M, and 9.22 M for U-Net, SC U-Net, UResNet2, MultiResUnet and Tiarmisu, respectively. Models were trained on a computer with a NVIDIA Tesla P100 GPU with 16 GB of RAM. All presented results pertain to the test sets in the data splits (see Section 2.6.1). Traditional methods (PVA and LPA) were applied on the entire test volumes (without patching). The loss graphs for each of the methods and the datasplits are shown in Fig. A & B. in the Appendix showing convergence.

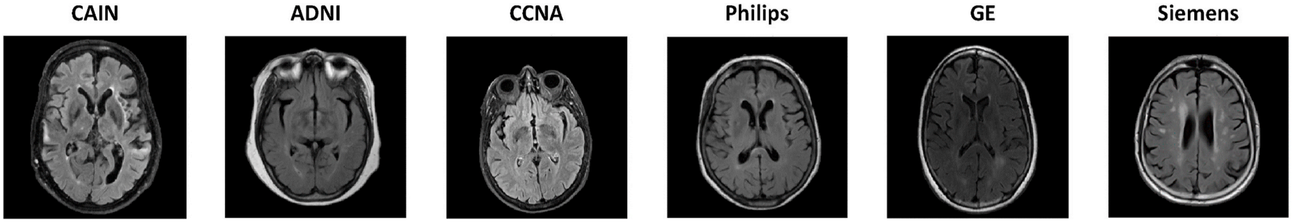


Fig. 3. Sample images from different datasets and scanner vendors.

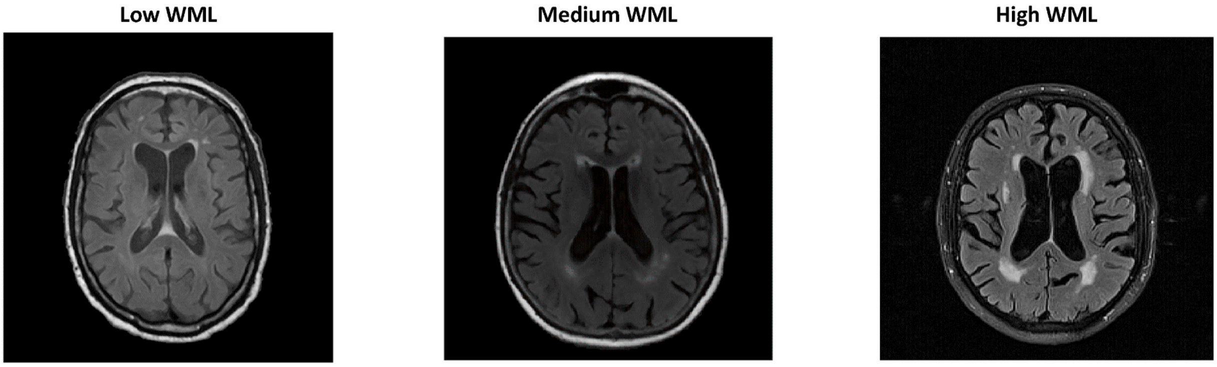


Fig. 4. Example slices from subjects with low, medium and high lesion loads.

Table 3
Number of normal and impaired volumes from datasplit1.

Pathology	Categorization	MoCA	No. Test Subjects
Cognitive Level	Normal	≥26	82
	Impaired	<26	91

3. Results

3.1. *Reliability of manual segmentations*

The ICC, IAVD, DSC, and EF of the manual segmentations for two separately trained raters across 20 of the same patients were computed to measure inter-rater variability. Across the 20 vol (approximately 1000 imaging slices) with manual annotations from both raters, the results showed very good volume agreement between the two raters with an ICC of 0.98 (95 % confidence interval with values ranging between 0.87 and 0.99) and mean lAVD = 0.25 ± 0.17, which indicates excellent reliability (Koo and Li, 2016). The regression plot for the volumes of the two raters is shown in Fig. C in the Appendix, which shows high correlation between the computed WML volumes for each value, with an R value of 0.99. Voxel level inter-rater agreement was also good with a mean DSC of 0.71 ± 0.11 and mean EF of 0.19 ± 0.17. Average mean lesion volume across both raters for the 20 patients was 14.45 ± 13.48 mL with a median of 7.18 mL and volumes per patient ranged from 1.07 mL to 41.70 mL which is a wide range of lesion loads. These experiments show the repeatability and reliability of the manual segmentation protocol and results. For the 10 subjects that had both manual and semi-automated corrected ground truths, the mean DSC = 0.6 ± 0.097, mean lAVD = 0.32 ± 0.21, mean EF = 0.66 ± 0.28. Average lesion volume for the manually rated 10 patients was 10.98 ± 7.59 mL with a median of 8.48 mL and volumes per patient ranged from 4.14 mL to 24.61 mL. As per (Dadar et al., 2017), a DSC > 0.5 is deemed to be very good agreement for WML segmentation in low lesion loads, since DSC values are smaller (and error is more noticeable) for objects with a high surface to volume ratio, as is the case for subjects with small lesion loads. As shown by Figure D in the Appendix, the average DSC of the subjects with≤5 mL (low lesion loads) is 0.51 indicating very good agreement. Additionally, in higher lesion loads > 5 mL, mean DSC = 0.65 which also indicates good agreement. Figure D in the Appendix also shows the Bland-Altman plot indicating a slight bias of −3.99 mL with 95 % confidence intervals of limits of agreement 2.39 to −10.36.

3.2. *Accuracy*

Fig. 5 shows sample WML segmentations for a slice from each dataset generated by datasplit1 models along with the respective DSC. DL methods detect WML with higher accuracy and precision compared to traditional methods. PVA segments confluent lesions, but has trouble in regions with diffuse (lower intensity) pathology. LPA detects a moderate number of lesions, but over-segments in some volumes – especially ADNI. In contrast, DL methods detect small, punctate lesions, diffuse and confluent pathology across all datasets, and seem to correlate closely with ground truth images.

The average evaluation metrics over all four folds and both data splits is shown in Table 4. The distributions of DSC and EF are shown in Fig. 6. All DL methods generated top performance over all metrics and data splits, with the top DSC performance for SC U-NET in both data splits (datasplit1: DSC = 0.71, datasplit2: DSC = 0.56), followed by U-Net (datasplit1: DSC = 0.70, datasplit2: DSC = 0.55) and MultiResUNet (datasplit1: DSC = 0.70, datasplit2: DSC = 0.55). The standard deviation of the metrics for the DL systems are also similar (and lower than traditional approaches) across the board. The traditional methods suffer from lower performance for both data splits with PVA DSC = 0.37 and 0.36 for datasplit1 and datasplit2, respectively, and LPA with higher performance with DSC = 0.46 and 0.41 for datasplit1 and datasplit2. All DL methods experienced deterioration in performance in datasplit2 while PVA and to some degree LPA maintained similar performance across data splits. In terms of EF, which measures the false positive rate, the DL methods consistently had the lowest EF (U-Net: EF = 0.32 for datasplit1 and SC U-Net: EF = 0.35 for datasplit2), lowest lAVD (U-Net, UResNet2, SC U-Net) and MultiResUNet had the lowest HD over both data splits. Traditional methods had larger EF, AVD and HD in general. Of all DL methods, Tiramisu had the worst performance over most metrics.

A common challenge of WML segmentation algorithms is segmentation in low lesion load scenarios. To examine performance of each method across lesion loads, DSC was stratified as a function lesion load ranges in Fig. 7, where increments of 5 mL were used to finely sample regions with low and moderate lesion loads. As can be seen all methods have lower DSC in low lesion loads that increase for larger lesion loads. The performance of DL methods are steadily higher than the traditional methods especially in the lower lesion load ranges (with similar performance in high lesion loads). The difference in performance between traditional and DL methods is less in datasplit2 due to a decrease in performance of the DL methods in this split. DSC values from the DL methods are also more variable in datasplit2.

Bland-Altman plots were generated to examine percent volume difference between ground truth and segmentation, as a function of the mean volume in Fig. 8. Traditional methods have much wider spread in the volume difference compared to DL methods with some large outliers especially in the larger lesion loads. In datasplit1, the DL methods have means close to 0 and spreads that are roughly the same (∼15), with the exception of Tiramisu which has the largest spread of 17.29. In datasplit2, the mean difference increased for the DL methods, and the spread has increased compared to datasplit1.

DL methods are performing similar to one another (but better than traditional methods even in datasplit2). Fig. 9 shows the averaged predictions for the DL methods and Fig. 10 has zoomed in regions. As can be seen, many of the DL methods predict the majority of the same lesions,

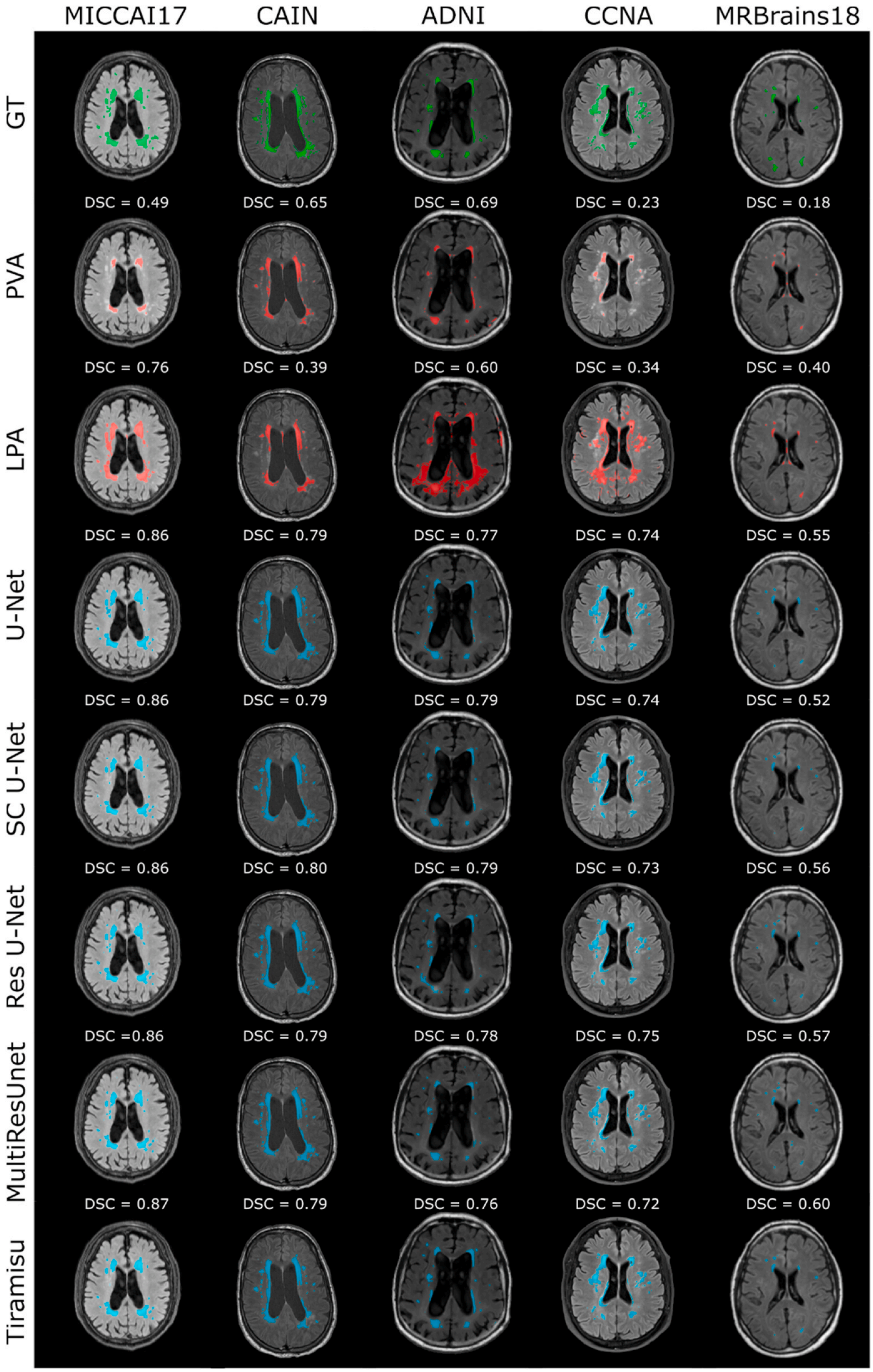


Fig. 5. Example WML segmentations and DSC volume scores across all methods from datasplit #1 models. Green overlays: ground truth (GT) annotations, red overlays: traditional methods, and turquoise overlays: deep learning systems. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 4
Average evaluation metrics across WML segmentation methods for both data splits. Metrics are shown as mean ± standard deviation. For each metric, ↑ means a higher value is better and ↓ means a lower value is better. Bold is best.

	DSC (%) ↑		EF (%) ↓		lAVD (%) ↓		HD (mm) ↓	
	datasplit1	datasplit2	datasplit1	datasplit2	datasplit1	datasplit2	datasplit1	datasplit2
PVA	0.37 ± 0.21	0.36 ± 0.20	5.89 ± 27.49	3.40 ± 17.08	0.71 ± 0.95	0.63 ± 0.77	29.74 ± 13.07	31.38 ± 13.68
LPA	0.46 ± 0.21	0.41 ± 0.20	1.95 ± 9.65	3.04 ± 14.07	0.61 ± 0.66	0.68 ± 0.74	22.23 ± 13.08	24.98 ± 12.69
U-Net	0.70 ± 0.13	0.55 ± 0.19	**0.32** ± 0.57	0.71 ± 1.0	**0.27** ± 0.28	**0.41** ± 0.37	8.44 ± 10.91	22.87 ± 18.58
SC U-Net	**0.71** ± 0.13	**0.56** ± 0.17	0.33 ± 0.81	**0.35** ± 0.49	**0.27** ± 0.28	0.42 ± 0.33	8.34 ± 11.34	21.20 ± 16.85
Res U-Net	0.70 ± 0.13	0.52 ± 0.18	0.39 ± 1.04	0.63 ± 0.84	**0.27** ± 0.29	0.53 ± 0.41	8.47 ± 11.27	22.14 ± 17.48
MultiResUNet	0.70 ± 0.13	0.55 ± 0.18	0.34 ± 0.64	0.47 ± 0.6	0.28 ± 0.27	0.44 ± 0.37	**8.33** ± 11.15	**21.18** ± 16.95
Tiramisu	0.69 ± 0.13	0.53 ± 0.21	0.38 ± 0.75	1.03 ± 1.34	0.29 ± 0.28	0.55 ± 0.46	12.54 ± 12.89	23.05 ± 18.38

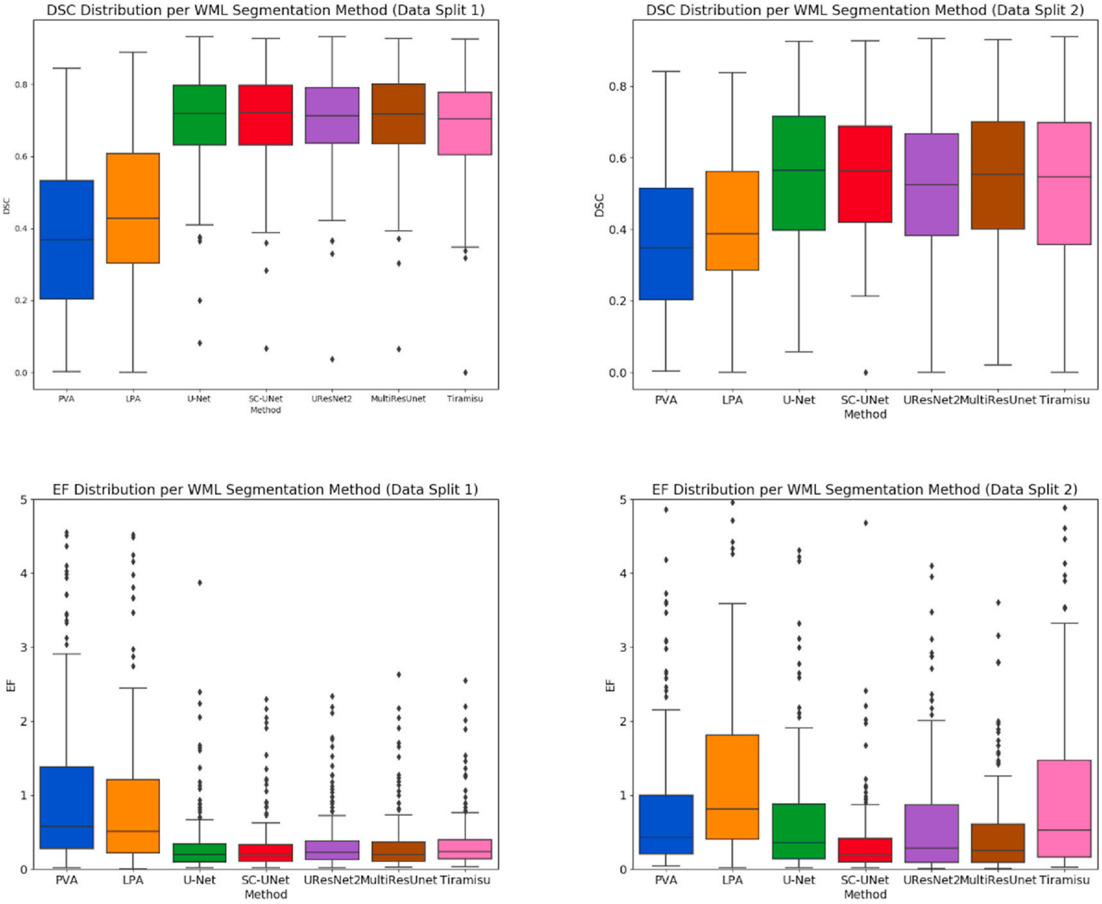


Fig. 6. DSC and EF distributions across all WML segmentation methods. Left: datasplit1. Right: datasplit2.

as shown by the brightest values in the heat map. There are small differences in the boundaries of these methods, and some small lesions. To take a closer look at the performance of the DL methods for WML segmentation, additional analysis is done here to choose the top three methods to analyze further. Table 5 summarizes the mean and CoV of the DSC. The mean DSC is highest for SC U-Net in both data splits (DSC = 0.706 in datasplit1 and DSC = 0.558 in datasplit2) followed by MultiResUNet (DSC = 0.705 in datasplit1 and DSC = 0.553 in datasplit2) and U-Net (DSC = 0.704 in datasplit1 and DSC = 0.550 in datasplit2). CoV is lowest for SC U-Net, UResNet2 and MultiResUNet (datasplit1) and lowest for SC U-NET, U-Net and MultiResUNet in datasplit2.

Fig. 11 contains the DSC distribution versus lesion load category for datasplit1 and datasplit2, which also is summarized in Table 6. Considering both datasplit1 and datasplit2, in the lower lesion loads, the top performer is SC U-Net (DSC = 0.61 in datasplit1 and DSC = 0.44 in datasplit2) followed by U-NET and MultiResUNet, and over all lesion loads, similar trends are seen (i.e top performance is obtained by SC U-Net, U-Net or MultiResUNet). Additional EF, lAVD and H95 metrics are included for all five deep learning methods for datasplit1 and datasplit2 in Figs. 11 and 12 and summarized in Table 6. In low lesion loads, there is higher EF for UResNet2 and Tiramisu with mean EF = 0.876 and 0.750, respectively, in datasplit1, indicating the best methods in terms of false positives in lower lesion loads are U-Net, MultiResUnet and SC U-Net. In datasplit2 the EF for UResNet2 and U-Net are similar, although the spread in UResNet2 is higher. These trends are supported by volume differences over both splits. In larger lesion loads the performance is comparable. The H95, which quantifies similarity in the boundaries between segmentation and ground truth lesions, is lowest for U-Net, MultiResUNet and SC U-Net especially in the lowest lesion loads, indicating that small lesions are being detected with good precision. In general, when comparing from datasplit1 and datasplit2, the performance is much more variable in lower lesion loads in datasplit2 (indicating the models struggle to detect smaller lesions) but in larger lesion

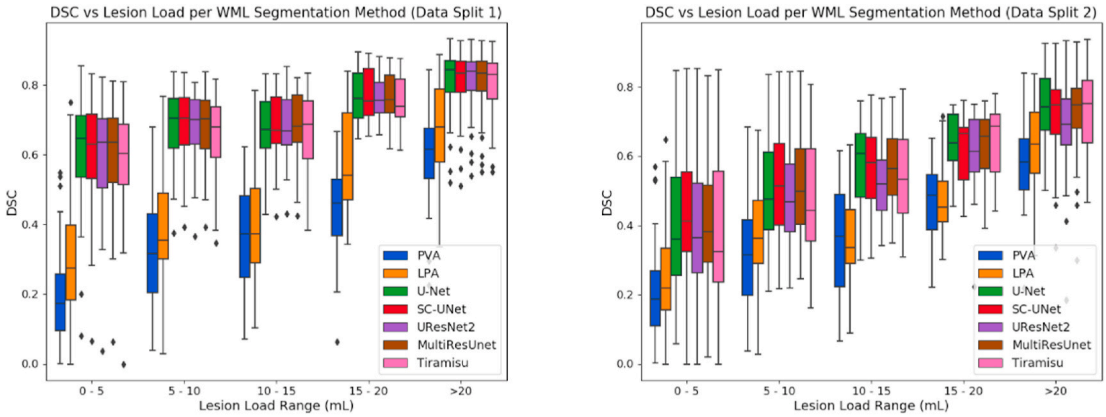


Fig. 7. DSC as a function of lesion load ranges for datasplit1 (left) and datasplit2 (right).

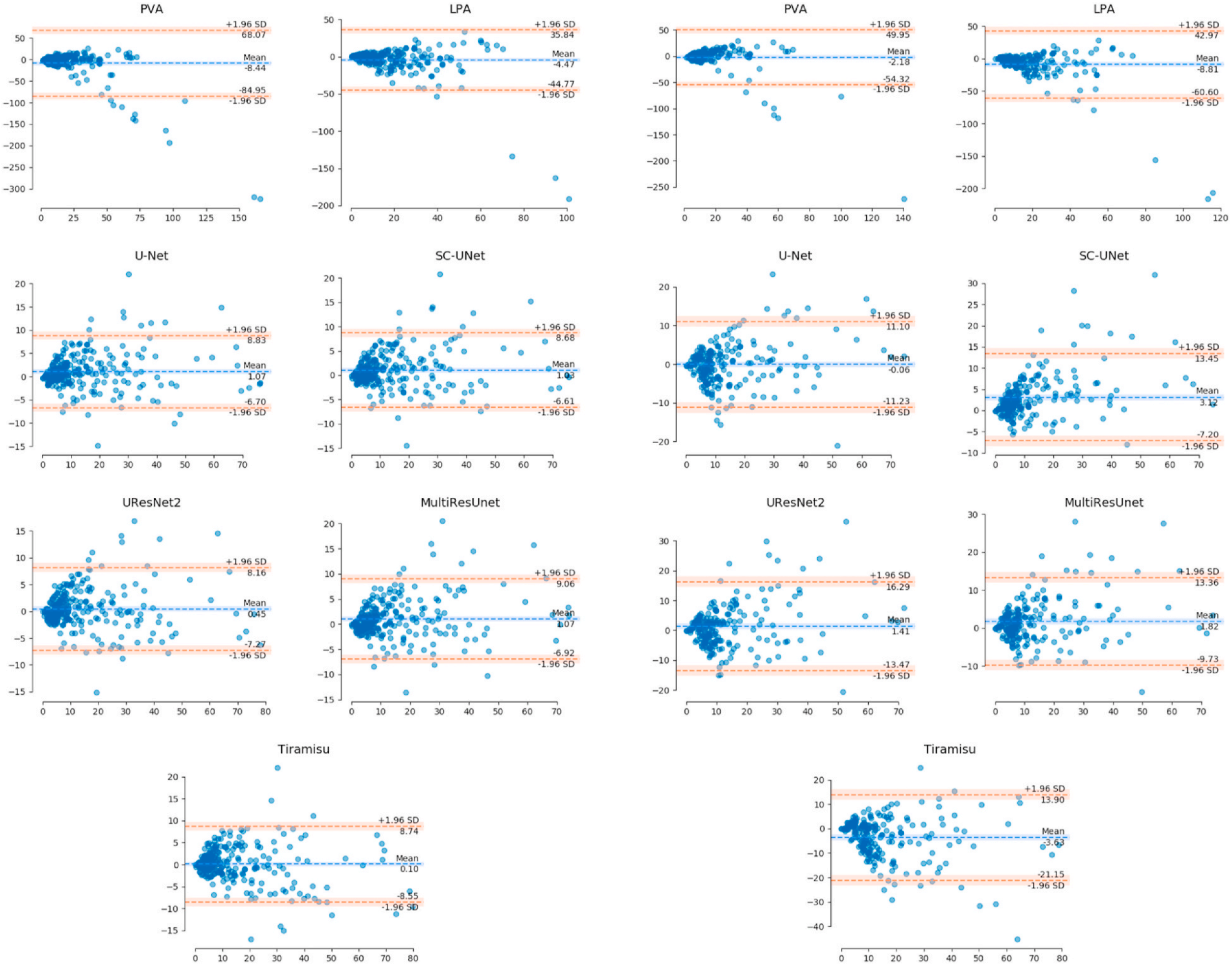


Fig. 8. Bland Altman plots for datasplit1 (left) and datasplit2 (right).

loads the performance is improved, and more consistent (lower standard deviation). Since U-Net, MultiResUNet and SC U-Net have the highest overlap (DSC), and perform well in terms of false positives, and volume differences, especially in lower lesion loads, these three DL methods are analyzed further.

3.3. Generalization

This section analyzes method performance across scanner vendors and datasets for both datasplit1 and datasplit2. Results are focused on the top three DL methods (SC U-Net, MultiResUNet, U-Net) and the traditional methods (PVA, LST). Fig. E (Appendix) contains the performance metrics for all DL methods. datasplit1 had testing data from the same distribution as the training set (i.e. all databases are represented in the training pool). Datasplit #1 contained 46 GE scans, 136 Philips scans, and 70 S scans and all data was tested through one of the folds. Models trained using datasplit2 had training data from a single dataset (MICCAI) and test volumes from ADNI, CCNA, CAIN, MICCAI and

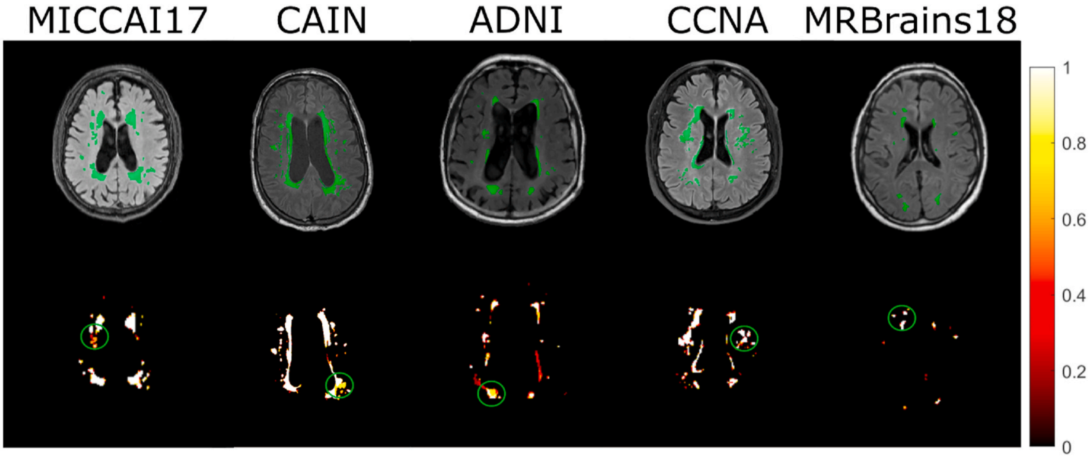


Fig. 9. Heatmaps representing average prediction across all DL methods for images in Fig. 5. Top: original with ground truths, bottom: averaged WML prediction over 5 DL methods.

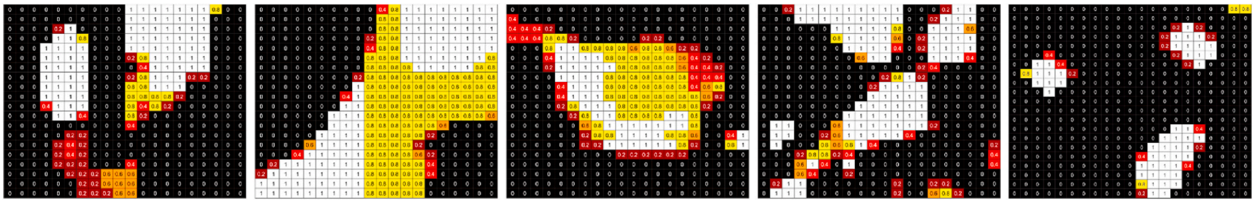


Fig. 10. Heatmaps representing average prediction across all DL methods. Zoomed in regions from circles in Fig. 9.

Table 5
Mean, standard deviation and CoV DSC for DL-based WML segmentation methods for both data splits. Bold is the best.

		U-Net	SC-U-Net	Res U-Net	MultiResUNet	Tiramisu
datasplit1	DSC	0.704 ± 0.134	**0.706** ± 0.132	0.703 ± **0.128**	0.705 ± 0.129	0.687 ± 0.133
datasplit2	DSC	0.550 ± 0.190	**0.558** ± **0.175**	0.522 ± 0.183	0.553 ± 0.179	0.532 ± 0.205
datasplit1	CoV DSC	0.190	0.186	**0.182**	0.184	0.194
datasplit2	CoV DSC	0.346	**0.314**	0.350	0.323	0.385

MRBrains to investigate how methods generalize to datasets within and outside their training distribution. The test set for datasplit2 contained 31 GE scans, 119 Philips scans, and 54 S scans. The training set had 16 GE, 16 Philips and 16 S scans from MICCAI in datasplit.

Table 7 and Fig. 13 summarizes DSC as a function of scanner vendor and data split. Over both datasplit1 and datasplit2 SC U-Net has the best mean DSC performance across GE, Philips and Siemens for datasplit1 with DSC = 0.69, 0.70, 0.74, respectively, and for Philips in datasplit2 with DSC = 0.51. U-Net has the best performance in datasplit2 for GE and Siemens (DSC = 0.64 and DSC = 0.68) and MulitResUnet had the lowest variance in DSC for GE and Philips (datasplit1). Deep learning performs better than traditional methods over all scanner vendors with an average improvement of 34 % and 21 % for datasplit1 and datasplit2 (respectively) compared to PVA, and an average improvement of 24 % (datasplit1) and 17 % (datasplit2) compared to LPA. There is less performance improvement using DL in datasplit2 since traditional methods have similar performance across scanners and data splits, but the DL methods drop off in performance for datasplit2. In datasplit1, DL methods have similar performance across the scanners, with Siemens having the best segmentation performance overall. Using datasplit2 the performance for the DL methods across scanner vendors is lower and more variable, with the most noticeable deterioration in Philips scans. Traditional methods exhibit variability across scanners in datasplit1 and datasplit2. The lowest performance for PVA was found in GE scanners, followed by Siemens and then Philips was the highest. The CoV of traditional methods across scanner types and data splits is similar but higher than the DL methods by more than double in datasplit1. DL methods have lower CoV compared to traditional methods and the lowest CoV in Siemens scanners for both datasplit1 and datasplit2. In datasplit1, across the DL methods, the most variability (highest CoV) in DSC comes from the GE scanner, while in datasplit2, the highest CoV is from the Philips scanner.

ANOVA was used to test similarity in DSC means across scanner groups for all the methods (PVA, LPA, U-Net, SC U-Net, Tiramisu, UResNet2, MultiResUNet). In datasplit1, except for the PVA method (traditional), all methods had significant ANOVA tests ($p < 0.05$) indicating differences in algorithm performance between vendors (see Table A in Appendix). Post-hoc analysis revealed the source of most differences ($p < 0.05$) was performance between GE versus Philips scans and GE versus Siemens scans. U-Net had significant differences across all three scanner vendors and Tiramisu and UResNet2 only had differences between GE and Philips. In datasplit2, similar trends are noted in that PVA (traditional method) has been found to have statistically similar DSC means across the scanner vendors, and this time so does LPA. All of the DL methods were found to have significantly different DSC means, which post hoc analysis revealed that differences were mainly between GE vs Philips. and Philips vs. Siemens. Differences in performance between GE and Siemens scans was not significant across all DL methods indicating consistent performance in these two scanner types.

To analyze generalization abilities of each method, the DSC performance was compared across datasets. Fig. 14 contains graphs of the DSC distributions, mean DSC and DSC CoV for WML segmentation methods in datasplit1 and datasplit2 (right) as a function of dataset. The metrics are summarized in Table 8 for the traditional methods and the top three

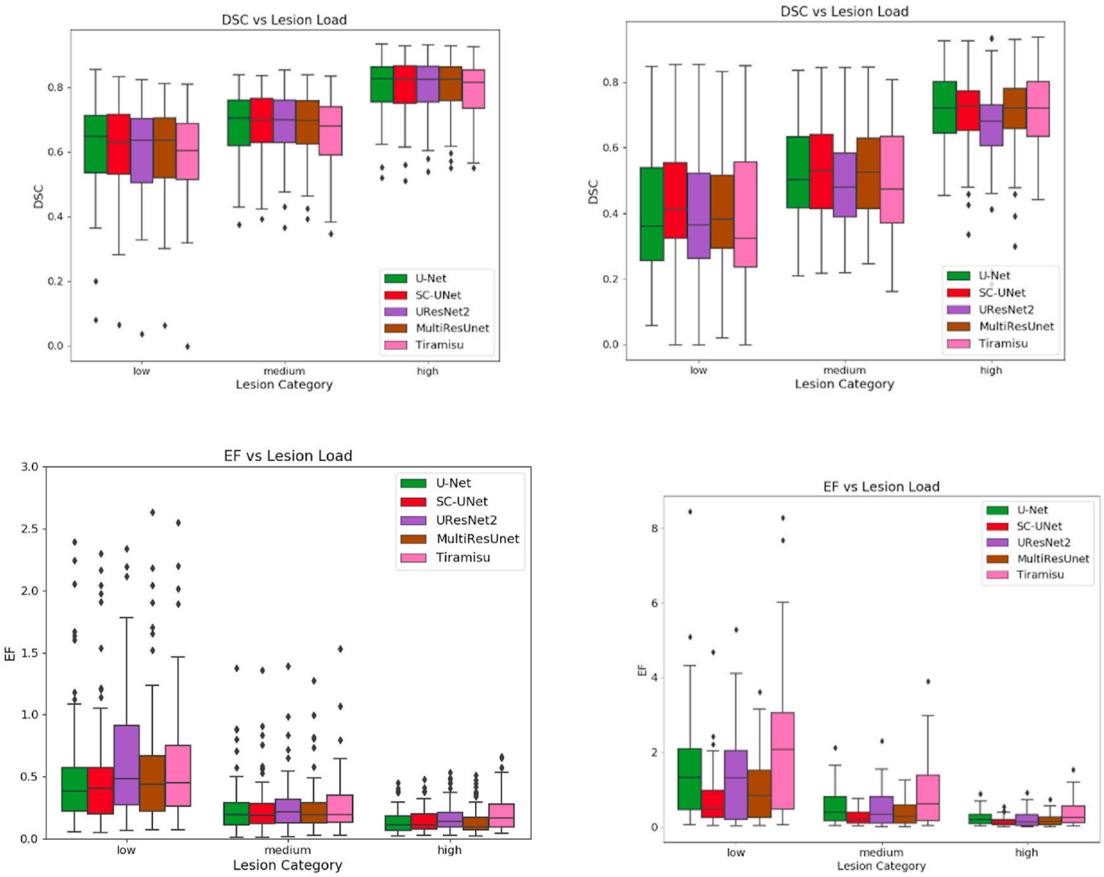


Fig. 11. DSC and EF versus lesion load category for the deep learning methods. Left: datasplit1. Right: datasplit2.

Table 6
Mean DSC, EF, IAVD and H95 for DL-based WML segmentation methods for both data splits. Bold is the best.

		Low LL		Medium LL		High LL	
		datasplit1	datasplit2	datasplit1	datasplit2	datasplit1	datasplit2
DSC	U-Net	0.611 ± 0.143	0.400 ± 0.178	0.680 ± 0.102	0.520 ± 0.153	0.803 ± 0.086	**0.716** ± **0.108**
	SC-U-Net	**0.613** ± 0.141	**0.437** ± 0.162	0.684 ± 0.098	**0.527** ± 0.147	**0.804** ± 0.085	0.699 ± 0.120
	Res U-Net	0.603 ± 0.131	0.392 ± 0.174	0.684 ± **0.092**	0.496 ± 0.146	0.803 ± **0.081**	0.665 ± 0.138
	MultiResUNet	0.610 ± 0.135	0.416 ± **0.161**	**0.685** ± 0.097	**0.527** ± **0.142**	0.801 ± 0.085	0.702 ± 0.123
	Tiramisu	0.589 ± **0.132**	0.376 ± 0.194	0.661 ± 0.104	0.499 ± 0.168	0.793 ± 0.085	0.707 ± 0.119
EF	U-Net	**0.669 ± 0.975**	1.613 ± 1.580	0.237 ± 0.206	0.530 ± 0.438	**0.142** ± 0.102	0.231 ± 0.190
	SC-U-Net	0.716 ± 1.476	**0.765 ± 0.802**	**0.237 ± 0.203**	**0.256 ± 0.182**	0.146 ± **0.101**	**0.146 ± 0.116**
	Res U-Net	0.876 ± 1.916	1.400 ± 1.250	0.260 ± 0.204	0.483 ± 0.458	0.169 ± 0.120	0.205 ± 0.209
	MultiResUNet	0.712 ± 1.112	1.021 ± 0.897	0.246 ± 0.204	0.357 ± 0.308	**0.142** ± 0.111	0.181 ± 0.157
	Tiramisu	0.750 ± 1.352	2.198 ± 1.992	0.265 ± 0.217	0.836 ± 0.786	0.206 ± 0.145	0.368 ± 0.354
IAVD	U-Net	**0.360** ± 0.372	0.727 ± 0.497	0.298 ± 0.257	**0.362 ± 0.244**	0.181 ± 0.170	**0.230 ± 0.197**
	SC-U-Net	0.366 ± 0.396	**0.451 ± 0.354**	0.295 ± 0.250	0.449 ± 0.326	0.180 ± 0.166	0.342 ± 0.296
	Res U-Net	0.400 ± 0.419	0.761 ± 0.439	**0.269 ± 0.232**	0.488 ± 0.351	**0.174** ± 0.163	0.408 ± 0.400
	MultiResUNet	0.370 ± **0.363**	0.586 ± 0.395	0.289 ± 0.247	0.431 ± 0.363	0.189 ± 0.171	0.322 ± 0.312
	Tiramisu	0.370 ± 0.375	0.946 ± 0.574	0.320 ± 0.272	0.507 ± 0.356	0.192 ± **0.154**	0.294 ± 0.223
H95	U-Net	**14.97 ± 15.28**	34.96 ± 21.49	8.58 ± 9.17	23.64 ± 16.33	3.18 ± 3.21	11.97 ± 11.51
	SC-U-Net	15.66 ± 16.60	**32.71 ± 18.47**	7.90 ± 8.46	22.11 ± 15.38	3.09 ± 3.21	**10.57 ± 9.50**
	Res U-Net	15.59 ± 16.97	33.12 ± 20.72	8.48 ± 7.79	22.67 ± 15.68	**2.89 ± 2.54**	12.51 ± 10.39
	MultiResUNet	15.54 ± 16.84	32.98 ± 20.40	**7.99 ± 7.59**	**21.14 ± 14.13**	3.07 ± 2.87	11.68 ± 10.61
	Tiramisu	22.01 ± 15.98	34.20 ± 21.49	13.38 ± 10.52	23.91 ± 16.71	4.21 ± 4.48	12.79 ± 12.43

DL methods. The performance for all DL methods is shown in Fig. F (Appendix). SC U-Net had the top performance over most datasets in datasplit1 (MICCAI: DSC = 0.78, CAIN: DSC = 0.67, ADNI: DSC = 0.71) and datasplit2 (CAIN: DSC = 0.50). The other top two performers were U-Net in datasplit2 (MICCAI: DSC = 0.79, CCNA: DSC = 0.69, ADNI: DSC = 0.67, MRBrains: DSC = 0.62) and MulitResUNet in datasplit1 (CCNA: DSC = 0.73, MRBrains: DSC = 0.65). In general, for all DL methods, there was a drop in mean DSC performance in datasplit2 as compared to datasplit1 over all datasets except for MICCAI which was included in both training sets. In the MICCAI dataset, the top performing DL methods had DSC>0.75 in datasplit1 and slightly higher with DSC>0.78 for datasplit2. MICCAI had the best test performance across all methods and datasplits except for PVA. The next best performing dataset across the methods is CCNA (30 vol), followed by ADNI (20 vol). Across both datasplits and most methods, CAIN (135 vol) has lowest performance, with noticeable DSC degradation in datasplit2 in the DL

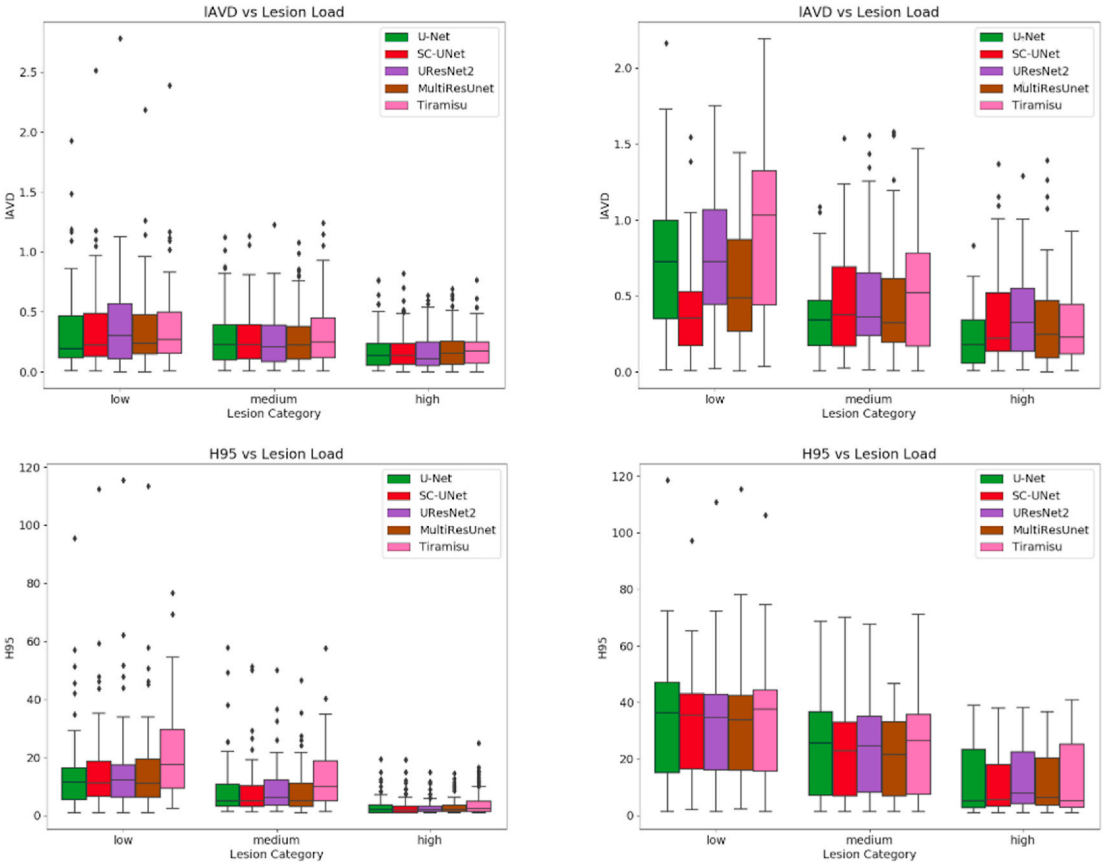


Fig. 12. IAVD and H95 versus lesion load category for the deep learning methods. Left: datasplit1. Right: datasplit2.

Table 7
Mean DSC for WML segmentation methods on both data splits as a function of scanner. Bold is best.

		PVA	LPA	U-Net	SC-U-Net	MultiResUNet
GE	datasplit1	0.361 ± 0.149	0.476 ± 0.205	0.686 ± 0.149	**0.687** ± 0.153	0.686 ± **0.147**
	datasplit2	0.404 ± 0.193	0.403 ± 0.186	**0.635 ± 0.167**	0.619 ± 0.181	0.615 ± 0.178
Philips	datasplit1	0.378 ± 0.211	0.423 ± 0.199	0.693 ± 0.135	**0.695** ± 0.133	0.693 ± **0.127**
	datasplit2	0.345 ± 0.195	0.397 ± 0.183	0.470 ± 0.180	**0.513 ± 0.174**	0.507 ± 0.175
Siemens	datasplit1	0.364 ± 0.212	0.503 ± 0.237	0.738 ± 0.115	**0.741 ± 0.107**	0.739 ± 0.107
	datasplit2	0.379 ± 0.204	0.446 ± 0.226	**0.679 ± 0.125**	0.622 ± 0.143	0.621 ± 0.155

methods. Mean DSC performance for MRBrains (7 vol) is low as well. In terms of traditional methods, PVA and LPA are both relatively consistent methods, despite their poor performance. LPA outperforms PVA on the MICCAI and MRBrains datasets, whereas PVA outperforms LPA on ADNI, and has similar performance in CAIN and CCNA. This is most likely due to LPA being trained on subjects with MS lesions, whereas ADNI, CAIN and CCNA are subjects with vascular and dementia disease (different pathologies). The CoV is variable across the datasets and over all methods, with lower CoV for the DL methods, especially in datasplit1. The highest CoV is for the MRBrains dataset in the DL methods. In datasplit2, the CoV has increased in the DL methods, with the CoV of CAIN being much higher than the other datasets, followed by MRBrains and CCNA.

To quantify the variability in performance across datasets, ANOVA testing was completed for all methods for datasplit1 and datasplit2. These results are summarized in Table C and Table D in the Appendix. The null hypothesis that the mean DSC is the same across datasets was rejected ($p < 0.05$) for all methods in datasplit1 and datsplit#2 except for the PVA method (which is a traditional method). In datasplit1, post-hoc analysis revealed the main differences in the DL methods are between CAIN and MICCAI. In datasplit2, post hoc analysis shows there are more differences across datasets and methods in this split. All methods had differences in mean DSC for ADNI vs. CAIN, CAIN vs. CCNA, CAIN vs. MICCAI. SC U-Net, UResNet2, and MultiResUNet had additional differences between CCNA vs. MICCAI, and UresNet2 and Tiramisu had additional DSC differences in CAIN vs. MRBrains. Overall, U-Net, MultiResUNet and SC U-Net had the least amount of differences across datasets.

3.4. Robustness to pathology

In this section, algorithm robustness to pathology is analyzed for mainly datasplit1 since this represents the ideal training scenario. Only the top three DL methods (U-Net, MultiResUNet, SC U-Net) along with the traditional methods (PVA, LST) are compared. DSC performance is analyzed as a function of WML lesion load category (low, medium, high) and impairment (MoCA≥26, MoCA<26). All 252 vol were used to categorize lesion loads through the four folds. Several example segmentations for each of the groups and methods are shown in Fig. 15. LPA oversegments especially in low lesion loads and PVA misses small, faint lesions in multiple categories. The DL systems seem to consistently detect small lesions, and majority of the lesions seen in the ground truth

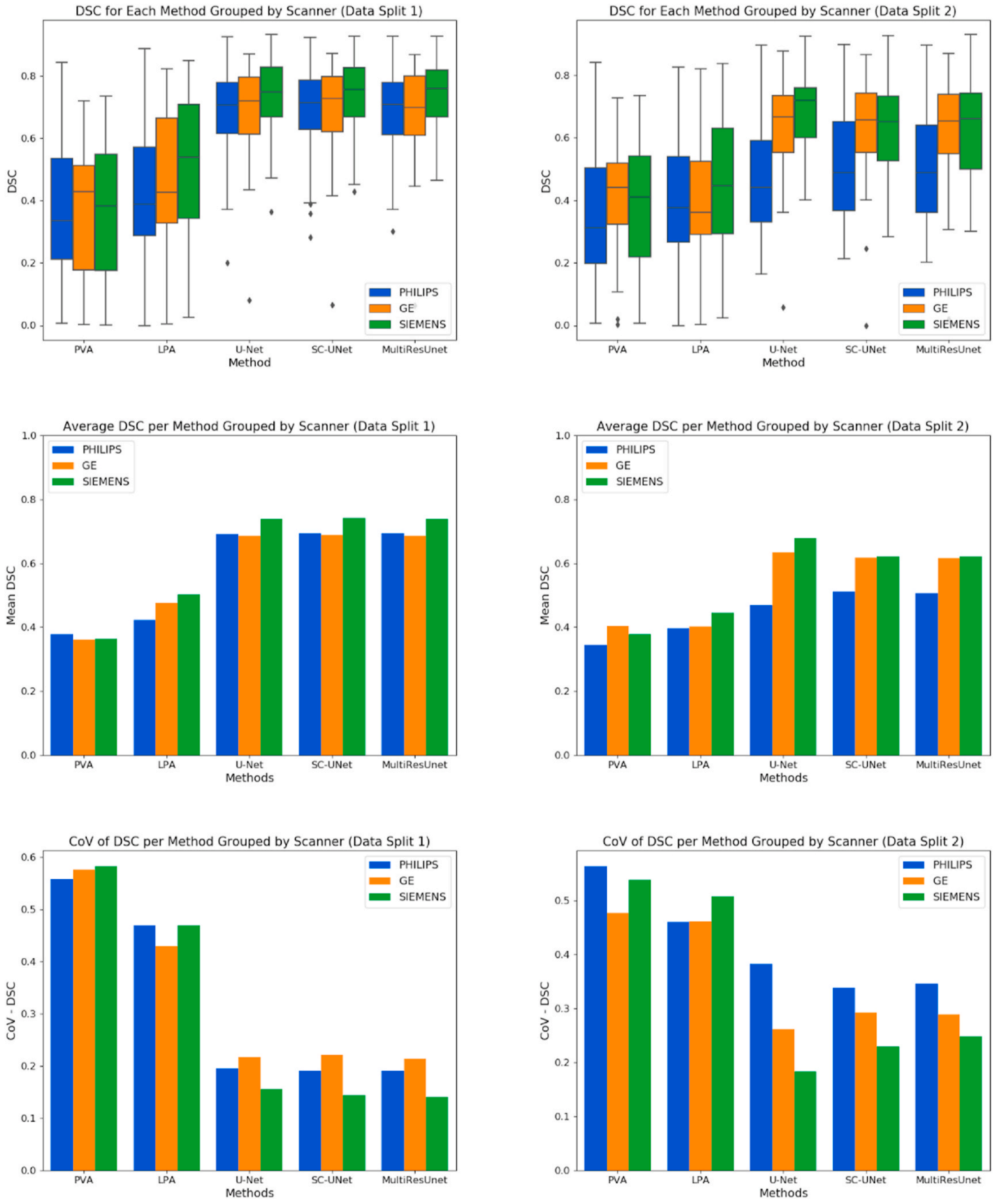


Fig. 13. DSC distributions, mean DSC and DSC CoV for WML segmentation methods for datasplit1 (left) and datasplit2 (right) as a function of scanner vendor.

delineations are detected for all lesion groups and cognitive status.

DSC distributions, mean DSC and DSC CoV as a function of WML lesion load category (low, medium, high) for datasplit1 are shown in Fig. 16 and summarized in Table 9. The average DSC for the DL methods are higher in all lesion categories with DSC>0.61 for low, DSC>0.68 for medium and DSC>0.80 for high. Of all methods, SC U-Net is the highest average DSC in two out of the three categories (low and high) with MultiResUNet being the highest for the medium lesion load (although only slightly higher than SC U-Net). The MultiResUNet had the lowest standard deviation for both low LL and medium LL cases. The traditional methods suffer especially in low lesion loads with DSC, with average DSC of 0.2 and 0.26 for PVA and LPA respectively. Performance of the traditional methods increases with lesion loads but is still lower than the DL methods. Although the DL methods are performing the best, there is still lower DSC performance in lower lesion loads. The CoV over all lesion loads is much lower in the DL methods indicating that there is more consistency in the predictions across lesion loads. The CoV is higher in the low lesion load category for all methods. The ANOVA testing results for DSC performance across lesion load levels is shown in Table E and Table F (Appendix) for datasplit1 and datasplit2, respectively. For all methods, both DL and traditional methods have significant differences in DSC means across the low, medium and high lesion loads categories across datasplit1 and datasplit2.

The same analysis is performed on test volumes that are normal and impaired. The DSC distributions, mean DSC and DSC CoV as function of cognitive status for datasplit1 is shown in Fig. 17 and summarized in Table 10. MultiResUNet has the top DSC performance in impaired and non-impaired subjects (DSC = 0.69 and DSC = 0.67, respectively), which is tied with SC U-Net for non-impaired subjects (may be due to low lesion loads). Over all methods, the performance went down for

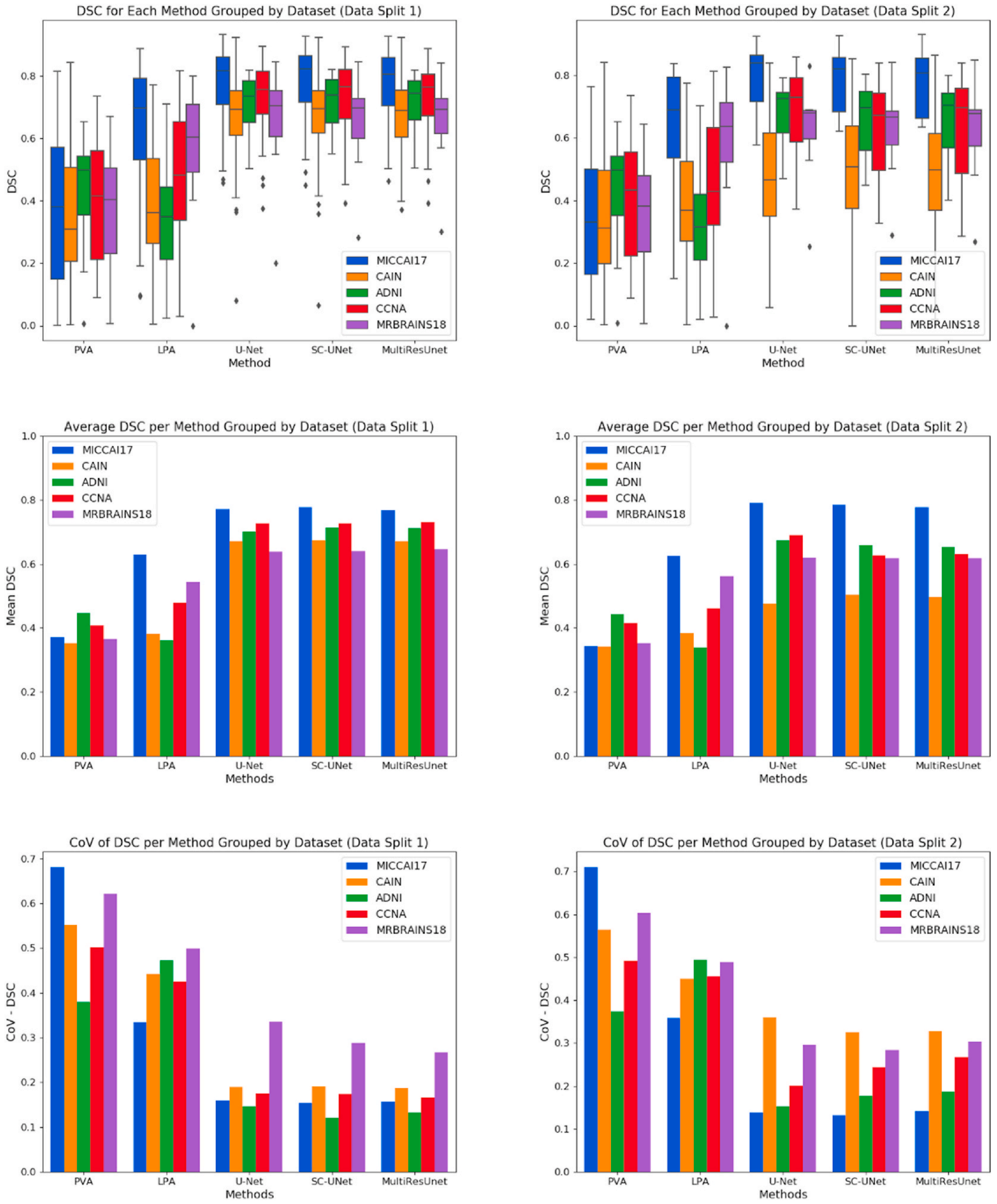


Fig. 14. DSC distributions, mean DSC and DSC CoV for WML segmentation methods across datasets. datasplit1 (left) and datasplit2 (right).

Table 8
Mean DSC (+/− std) for WML segmentation methods on both data splits as a function of dataset. Bold is the best.

		PVA	LPA	U-Net	SC-U-Net	MultiResUNet
MICCAI	datasplit1	0.370 ± 0.252	0.629 ± 0.211	0.773 ± 0.123	**0.777 ± 0.119**	0.769 ± 0.121
	datasplit2	0.343 ± 0.244	0.625 ± 0.224	**0.790** ± 0.109	0.785 ± **0.104**	0.777 ± 0.109
CAIN	datasplit1	0.352 ± 0.194	0.382 ± 0.169	0.672 ± 0.128	**0.673** ± 0.128	0.672 ± **0.126**
	datasplit2	0.342 ± 0.193	0.383 ± 0.172	0.477 ± 0.171	**0.504 ± 0.164**	0.498 ± 0.164
ADNI	datasplit1	0.446 ± 0.169	0.363 ± 0.172	0.701 ± 0.102	**0.714 ± 0.086**	0.713 ± 0.095
	datasplit2	0.444 ± 0.166	0.339 ± 0.167	**0.673 ± 0.102**	0.659 ± 0.117	0.654 ± 0.122
CCNA	datasplit1	0.408 ± 0.205	0.479 ± 0.204	0.727 ± 0.128	0.727 ± 0.127	**0.731 ± 0.121**
	datasplit2	0.415 ± 0.204	0.463 ± 0.211	**0.689 ± 0.139**	0.627 ± 0.153	0.631 ± 0.169
MRBrains	datasplit1	0.365 ± 0.227	0.543 ± 0.271	0.639 ± 0.214	0.640 ± 0.184	**0.647 ± 0.172**
	datasplit2	0.352 ± 0.213	0.562 ± 0.275	**0.619** ± 0.183	0.618 ± **0.176**	0.618 ± 0.187

impaired volumes, which likely contain more prevalent pathology, except for PVA which had a slight improvement in performance (and had the same performance in the impairment group over datasplit1 and datasplit2). The 2D CNN methods had higher DSCs with lower variance over MoCA groups compared to traditional methods, with almost 2x performance gains. The CoV is at least two-fold lower for DL methods

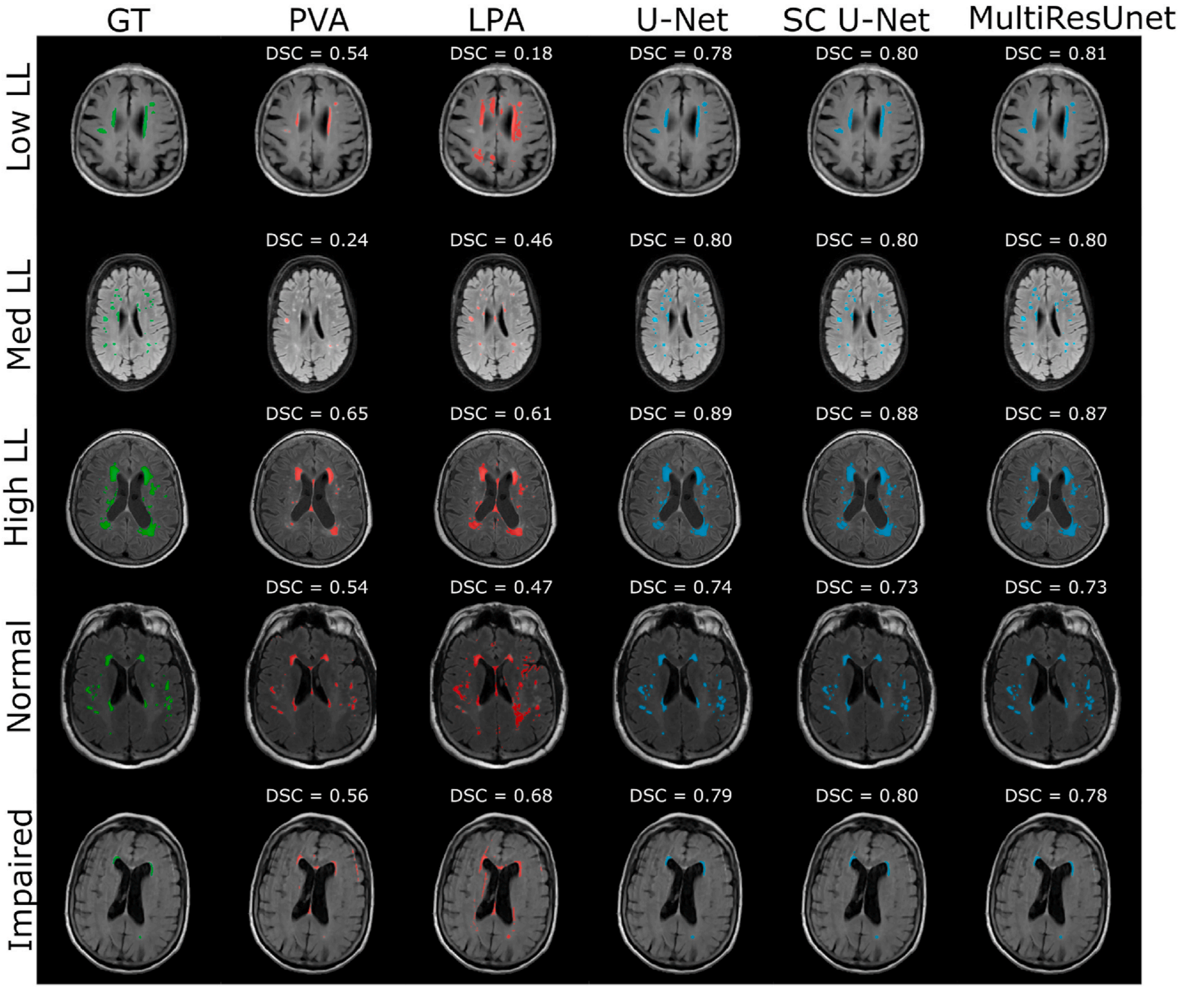


Fig. 15. Example segmentations and volume DSC scores for different levels of pathology and cognitive status. Green is the ground truth, red for traditional algorithm predictions, and turquoise are the top DL predictions. Top to bottom: ADNI (low LL), MICCAI (medium LL), CAIN (high LL), CAIN (normal), ADNI (impaired). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

(and similar for impaired and normal subjects) compared to traditional methods that had more variability in performance.

ANOVA testing was completed to test differences across normal and impaired groups for datasplit1 and datasplit2. Results are summarized in Table G and Table H (Appendix). In datasplit1, there were no significant differences in DSC means across MoCA groups for all traditional and DL methods ($p > 0.05$). In datasplit2, there were no significant differences across methods except for Tiramisu and U-Net.

There were no significant differences in DSC means across MoCA categories over all the methods for datasplit1. However, in datasplit2, there were significant differences across MoCA categories for U-Net and Tiramisu, indicating these models had difficulty generalizing in brains with differing levels of neurodegeneration. Note that MoCA was not available for MICCAI, MRBrains and 12 vol from CAIN, ADNI and CCNA. Therefore, the MoCA results only consider a subset of the total ground truth dataset. To investigate further, consider the frequency of lesion load category vs. MoCA in Fig. 18. Low, medium and high lesion load categories are found across different MoCA ranges (i.e., MoCA<26, MoCA≥26). Perhaps the average performance across lesion load groups is averaging out and making the mean performance similar for impaired and normal subjects. The MoCA score is a subjective rating and there could have been overlaps between normal and impaired patients which can also skew results.

3.5. *Post-hoc exploratory analysis*

A few outliers were examined by investigating the largest volume difference (maximum lAVD) on the fourth fold. A single volume was found to be an outlier across all methods. The ground truth and predictions for each method is shown in Fig. 19. As can be seen, PVA is too conservative on the lesion boundaries and faint lesions, which could indicate the PVA threshold is too high for this volume. LPA also grossly oversegmented the lesions and is likely due to different scanner types and intensity characteristics of the test data as compared to the training data, which is related to MS disease. DL methods show high correlation with the GT. As this is a low lesion load case (~5 mL), any pixel differences result in large errors.

We also investigated segmentation performance as a function of rater for datasplit1 and datasplit2 in Fig. 20. Rater1-3 is developed by the authors (CAIN, ADNI, CCNA) and as shown by Section 3.1, there was high agreement across raters for the WML ground truths. Rater4 is from MICCAI and Rater5 is from MRBrains competition and were developed by experts. In datasplit1, there is relatively high performance across all the methods, with the highest performance for Rater4, which is the MICCAI competition annotations developed through consensus. In datasplit2, there is much more variability in DSC performance over raters, and there is lower performance for Rater3, which corresponds to CAIN volumes from the Philips vendor. The lower performance on Rater3 in datasplit2 is likely attributed to differences in pixel resolutions

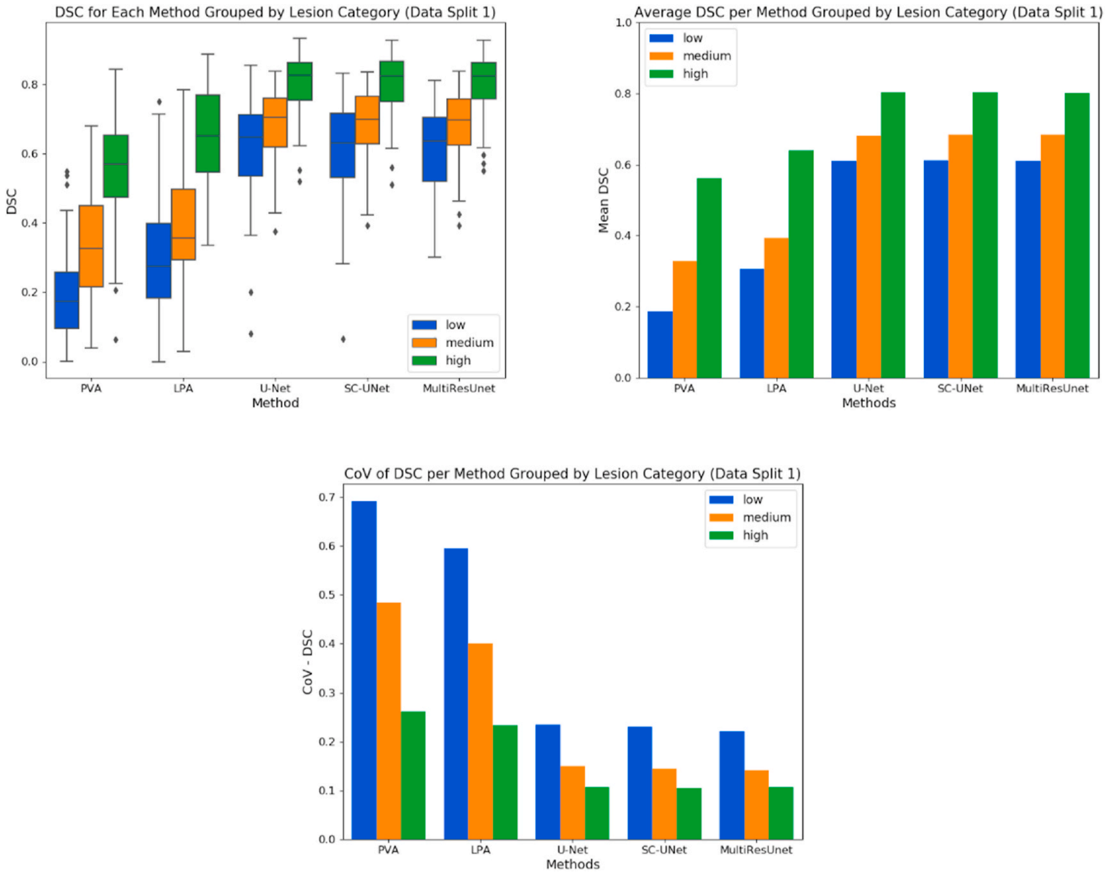


Fig. 16. DSC distributions, mean DSC and DSC CoV across WML segmentations methods for datasplit #1 vs. WML load.

Table 9
Mean DSC for WML segmentation methods on both data splits as a function of LL. Bold is best.

		PVA	LPA	U-Net	SC-U-Net	MultiResUNet
Low LL	datasplit1	0.204 ± 0.132	0.259 ± 0.154	0.611 ± 0.143	**0.613** ± 0.141	0.610 ± **0.135**
	datasplit2	0.204 ± 0.132	0.259 ± 0.154	0.400 ± 0.178	**0.437** ± 0.162	0.416 ± **0.161**
Medium LL	datasplit1	0.328 ± 0.159	0.393 ± 0.158	0.680 ± 0.102	0.684 ± 0.098	**0.685 ± 0.097**
	datasplit2	0.323 ± 0.161	0.369 ± 0.145	0.520 ± 0.153	**0.527** ± 0.147	0.527 ± **0.142**
High LL	datasplit1	0.562 ± 0.147	0.641 ± 0.150	0.803 ± 0.086	**0.804 ± 0.085**	0.801 ± 0.085
	datasplit2	0.550 ± 0.135	0.593 ± 0.148	**0.716 ± 0.108**	0.699 ± 0.120	0.702 ± 0.123

as this data is higher resolution (0.43 mm × 0.43 mm pixels) compared to the MICCAI training data which has 1 mm × 1 mm pixel dimensions. Physical dimension mismatch between training and testing datasets have been known to cause generalization issues in medical imaging and DL (Mahbod et al., 2021, Sabottke and Spieler, 2020, Liu et al., 2020). This is due to the fact that CNNs employ a fixed receptive field, i.e. 3 × 3, which determines the physical dimensions of the underlying objects that are analyzed. If the training data is not the same resolution as the test set, the scale of the objects analyzed would be different, and the model would learn features which will not generalize as well to other resolutions. Therefore, if the classifier is not exposed to examples from higher resolution data, it will not learn that representation. This was clearly demonstrated in datasplit1 results, which shows the classifier was able to learn the higher resolution representation when it is included in the training dataset. Similarly note that the held-out testing data for Rater1 and Rater2 in datasplit2, which has a median pixel size of 0.8594 mm × 0.8594 mm resolution, has better performance compared to Rater3, and this may be due to the fact that the resolutions in these raters are much more similar to the MICCAI training data resolution (1 mm × 1 mm).

To further investigate the impact of the annotations from Rater3, examine the WML segmentation performance that compares predictions to expert ground truths from MICCAI and MRBrains. In datasplit1 (trained on all data including Rater3) predictions on expert datasets had excellent performance with mean DSC = 0.78 in MICCAI and mean DSC = 0.64 in MRBrains. The performance in datasplit2 is roughly the same with DSC = 0.79 (MICCAI) and DSC = 0.62 (MRBrains) and since there is no drastic change in performance across datasplits we conclude that Rater3 is not negatively impacting results. Another issue is that the data is from the Philips scanner, and only 20 vol of Philips were in the training dataset. This is something we found in another work (DiGregorio et al., 2021), since after intensity standardization there is a slight misalignment of tissue intensities in the Philps scans, which could also be creating biases (see Discussion). Lastly, a semi-automated correction was used to touch up the manual segmentations for Rater3, which could add some bias, but we doubt this to be a major concern. As shown by the inter-rater agreement studies, there is very good agreement between the semi-automated and manual gold standard. This is supported by the high performance of the classifier over all other raters in datsplit1, which was trained using all the data including from Rater3. If Rater3's annotations were detrimental to the system we would see degradation in

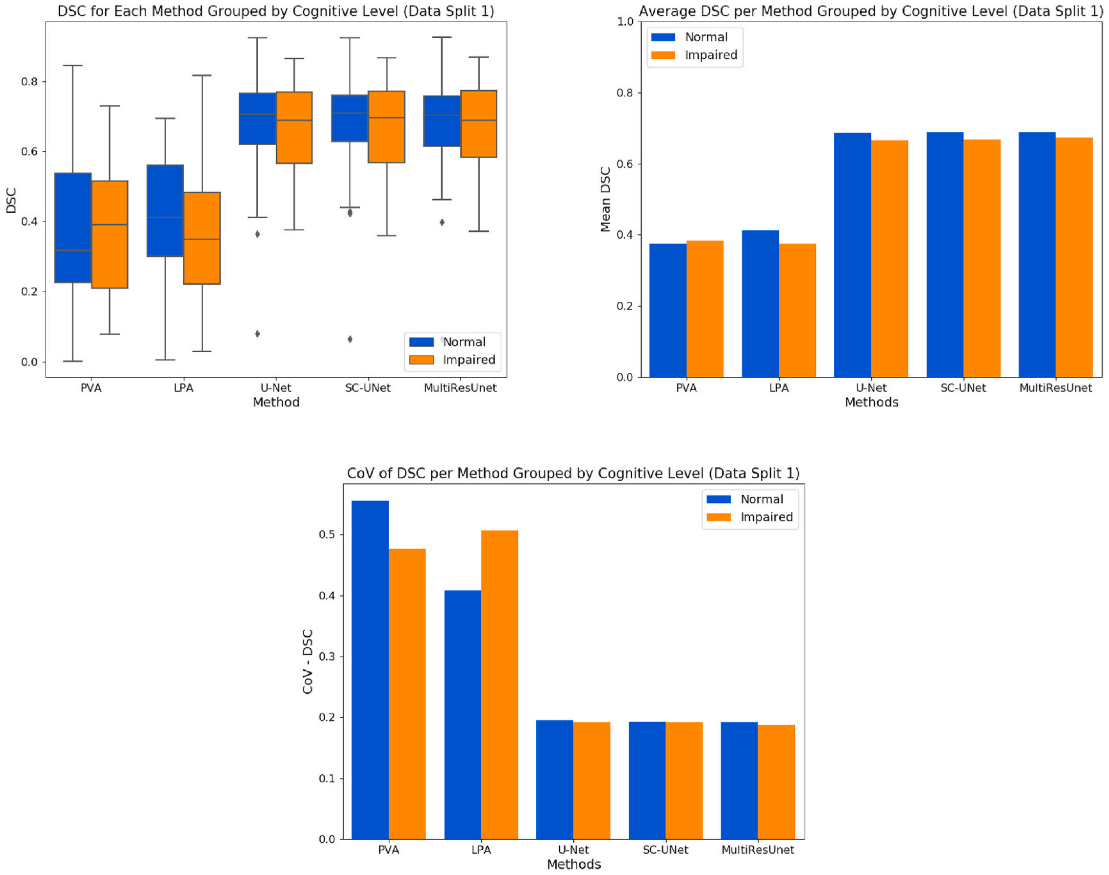


Fig. 17. DSC distributions, mean DSC and DSC CoV across WML segmentations methods for datasplit1 as a function of cognitive level (impaired: MoCA $<$ 26, normal: MoCA$\geq$26).

Table 10
Mean DSC for WML segmentation methods as a function of cognitive level. Bold is the best.

		PVA	LST	U-Net	SC-U-Net	MultiResUNet
Normal	datasplit1	0.376 ± 0.209	0.412 ± 0.168	0.687 ± 0.134	**0.688** ± 0.133	**0.688 ± 0.132**
	datasplit2	0.366 ± 0.207	0.412 ± 0.170	0.499 ± 0.191	**0.523 ± 0.179**	0.519 ± 0.182
Impaired	datasplit1	0.384 ± 0.183	0.375 ± 0.190	0.666 ± 0.127	0.667 ± 0.128	**0.671 ± 0.122**
	datasplit2	0.384 ± 0.183	0.359 ± 0.188	**0.591** ± 0.160	0.570 ± **0.153**	0.542 ± 0.166

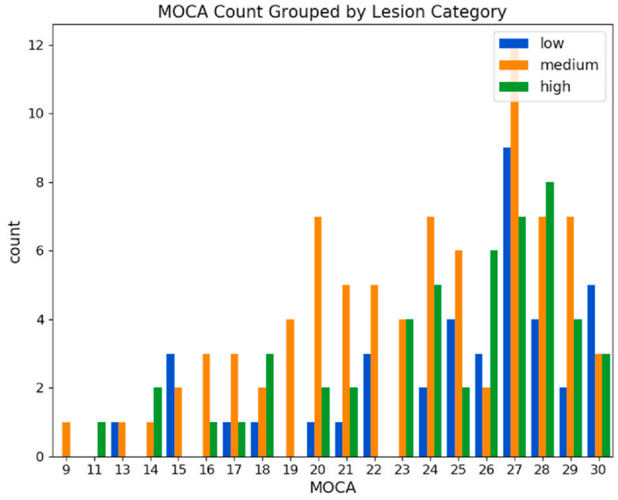


Fig. 18. Number of FLAIR imaging volumes with particular lesion load categories and MoCA.

performance by including them in the training pool, which was not observed. It is therefore postulated that scanner manufacturer, acquisition parameters and perhaps the annotations contributed to the reduction in DSC for datasplit2 for Rater3.

MICCAI has the highest overall performance again in datasplit2, probably because of the quality of annotations and also being trained on that dataset. Although there may be similarities between datasets that classifiers are learning and exploiting to achieve good performance in the respective datasets, classifiers may also be learning the annotating pattern of the dataset. This means the classifier learns based on the experience of a single set of delineations. Results from datasplit1 and datasplit2 clearly demonstrate that including annotations from multiple raters results in the best performance. This is because WML are hard to define and are done so based on a largely subjective criteria. Therefore, including more annotations from different raters can help to create more diversity in the training data which is better for learning wider representations. Expert annotations are costly and often only partially available and as algorithms get more data hungry it is becoming more commonplace to use annotations provided by two or more raters. Also, note MRBrains, despite being developed by experts, decreased in performance in datasplit2, which further supports the hypothesis that the

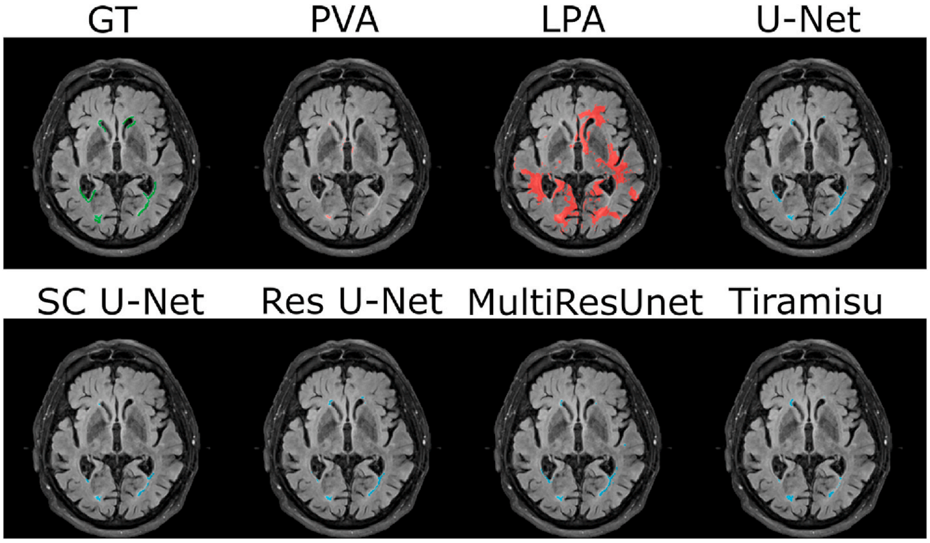


Fig. 19. Example outlier predictions determined by maximum lAVDAVD on the fourth fold in datasplit1. Green: ground truth, red: traditional methods, turquoise: deep learning methods. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

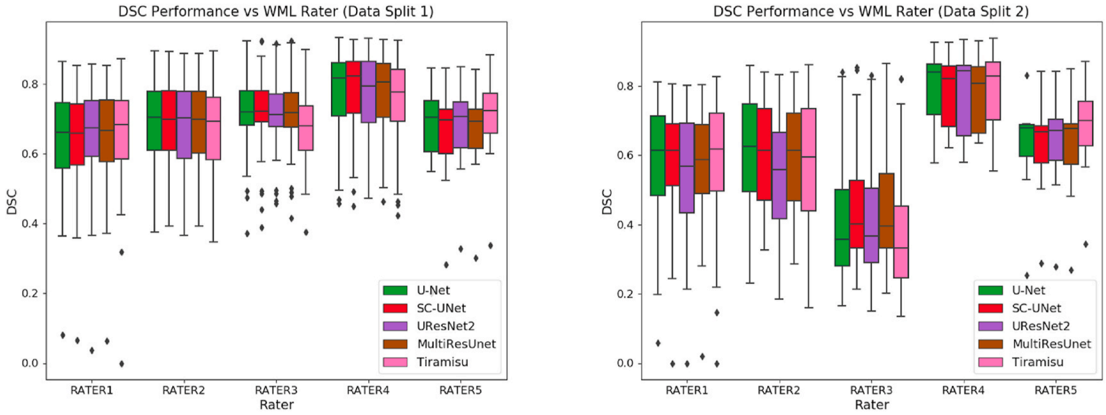


Fig. 20. DSC performance of the DL methods as a function of rater for datasplit1 and datasplit2.

systems could be partially learning annotation patterns and should be exposed to more raters.

4. Discussion

Although many structural biomarkers are extracted from T1 MRI, most WML segmentation algorithms require FLAIR MRI as a primary data input (Khademi et al., 2011, (Heinen et al., 2019, García-Lorenzo et al., 2013). In (Narayana et al., 2020), the FLAIR sequence was demonstrated to be the most crucial for lesion segmentation. Despite this, it is common to analyze FLAIR via multi-modal approaches that co-register FLAIR to T1 or T2 MRI (Soltanian-Zadeh and Peck, 2001, Khademi et al., 2020). This multiparametric approach prolongs scan times, increases acquisition costs, and can introduce registration errors across sequences (Narayana et al., 2020, Soltanian-Zadeh and Peck, 2001, Khademi et al., 2020) for WML segmentation approaches. Since FLAIR is routinely acquired clinically and highlights vascular disease with high sensitivity, there is benefit in developing methods that operate on this single sequence. Not only are there less integration hurdles using a single sequence, but also, WML based measurements can be readily used to augment neurological workflows. The results of this work demonstrate the value of FLAIR-based WML segmentation algorithms for multicentre, multi-disease data and presents a novel evaluation framework to establish proof of effectiveness. In contrast with some single center studies, a large multicenter dataset and comprehensive validation experiment was used in this work.

Descriptive and statistical analyses were used to analyze performance metrics across different dimensions related to accuracy, generalizability (scanners, datasets) and robustness (pathology levels). Investigating algorithms in larger groups of patients from multiple sites and over several dimensions is necessary to establish proof of effectiveness of computational tools and biomarkers. We believe in this manuscript, we have demonstrated proof of effectiveness, which is a prerequisite to clinical translation. To date, many proof of concept algorithms have been developed for neuroimaging (Akkus et al., 2017, Bernal et al., 2019) but there is a lack of performance testing across important dimensions in multicentre datasets to determine clinical feasibility. Therefore, the methodology and experiments presented in this work bridges that gap. The evaluation framework can be used on other tools to establish proof of effectiveness, to determine the optimal method for the task, to inform design decisions for method improvement, and to predict performance on new, prospective datasets. These evaluations translate into more reliable tools and ultimately better patient care.

In general, DL methods outperform traditional methods in terms of accuracy. In datasplit1, the performance was higher than traditional methods by 40 %, although the margin was narrowed between these two families of methods in datasplit2 (DL methods had a drop in performance while the performance of PVA and LPA remained relatively the same in datasplit2). The lower performance of DL methods in datasplit2 can be associated with the fact the training and testing distributions are different, which is a known challenge of deep learning (Zech et al., 2018). In terms of the lower performance of PVA and LPA; this indicates that intensity alone is not a strong enough feature for high segmentation performance, which may be due to diffuse regions of partial ischemia/demyelination, acquisition noise, and false positives such as CSF flow through. CNNs on the other hand are capable of modeling complex relationships between pixels and use non-linear boundaries for excellent segmentation performance (Guerrero et al., 2018). DL methods may be better suited for WML segmentation since WML are heterogeneous (variability in intensity, size, shape, texture) that could be better modeled using high dimensional approaches. Of all methods, SC U-Net exhibited the best mean performance over all test sets and dimensions including across lesion loads (low and high), scanners (GE, Siemens, Philips) and across datasets (MICCAI, CAIN, ADNI), with U-Net and MultiresUnet as runner ups. SC U-Net uses skip connections to maintain copies of higher resolution features in deeper layers which may help important features, such as those related to smaller lesions, persist in the network. MultiresUnet was also a leading method, likely due to the multiresolutional feature extractors (filters). In terms of comparison to literature, the proposed approach is comparable to the top performing models from the WMH segmentation challenge, with the SC U-Net architecture achieving a mean DSC of 0.78 and 0.79 on the MICCAI dataset for datasplit 1 and 2 respectively. The top 10 models from the challenge achieved a mean DSC of 0.78 or greater, with the top performing model achieving a DSC of 0.81 (Kuijf et al., 2019). The performance across U-Net variants was very similar to that of U-Net (see Fig. 11) and likely is due to the underlying U-Net model capturing majority of the information needed to robustly predict lesions on a per-voxel basis. In Figs. 9 and 10, it is seen that there is agreement in the predicted lesion masks across different U-Net variants. This highlights the potential for aggregation networks and ensemble systems for WML segmentation. For example, *Stack-Nets*, proposed by Li et al. achieved state-of-the-art performance on the MICCAI dataset (Li et al., 2019). *Stack-Nets* uses aggregated multi-scale U-Nets with multiple convolutional layers at different receptive fields (Li et al., 2019). In future work, a similar approach can be implemented by aggregating predictions from top performing U-Net variants to achieve better performance.

Statistical analysis of performance metrics across scanners and datasets found that the PVA method was able to provide the same mean performance across all scanner types and datasets, for dataplit1 and datasplit2, which indicates the PVA method is able to generalize across scanners and centres. That can largely be attributed to the intensity standardization framework and PVA-based model. LPA was able to generalize across many scanners and datasets, as well. In contrast, ANOVA testing showed the mean performance of DL methods varied for many scanner comparisons, with similar means found for Philips versus Siemens for many methods in datasplit1 and similarity in performance across scanners except if Philips was one of the comparisons in datasplit2. In datasplit1, GE was represented least in the training data, which may explain differences in performance when comparing Siemens or Philips to the performance of GE. In datasplit2 the Philips scanner has the lowest DSC performance and the most differences with other vendors. There could be a number of reasons for these performance trends.

Firstly, upon inspection of the standardized intensity histograms, there are slight differences between Philips and the other two scanners (which are similar to one another); see Fig. 21. Although intensity standardization creates a more consistent intensity interval over all, there is a slight misalignment in the CSF regions in Philips images as compared to GE and Siemens scanners. Since periventricular WML are neighbouring CSF regions, differences in contrast in these regions could cause performance degradation for Philips scans, especially in datasplit2 where there are less Philips training examples. Perhaps the reconstruction algorithms, contrast characteristics or noise profiles of GE and Siemens scanners are more similar to one another as compared to Philips allowing the systems to predict equally as well on both GE and Siemens scanner types. Additionally, as discussed in the rater analysis, the subset of data from Rater3 was composed of mainly Philips samples with higher pixel resolutions compared to all the other data from the training dataset. This likely limited generalization to these images due to the differing resolutions in the training and testing sets. CNNs employ a receptive field that is a fixed size for all images. If the underlying resolution of the images are changing, the physical dimensions of what is analyzed by the CNN changes also. If this representation is not present in the training set, the model will not learn it and performance suffers. In datasplit1 there is more Philips data in the training pool which lets the classifier effectively learn those patterns in the data (despite the intensity misalignment and differing pixel resolutions). However, in datasplit2 the classifier was not exposed to this data and therefore, the classifier could not effectively learn those patterns. These findings highlight the importance of evaluating tools using a two-datasplit strategy to fully appreciate the benefits and challenges of each approach. Many commercial and research systems are developed using a single dataset from a single institution. In this scenario, when translating tools to other centres there will be a performance reduction related to acquisition parameters and scanner type. Ideally, any model would be retrained on a dataset that includes some samples from the centre that it will be deployed at but this is not always practical. In datasplit1 there is similar mean DSC performance across the datasets, indicating that having a mix of the testing data distribution in the training set (and a more diverse representation) is critical for optimal performance. Scanner vendors have been shown to induce differences in many automated algorithms (Reiche et al., 2019, DiGregorio et al., 2021, Khademi et al., 2020) and therefore, future efforts should be pursued to balance

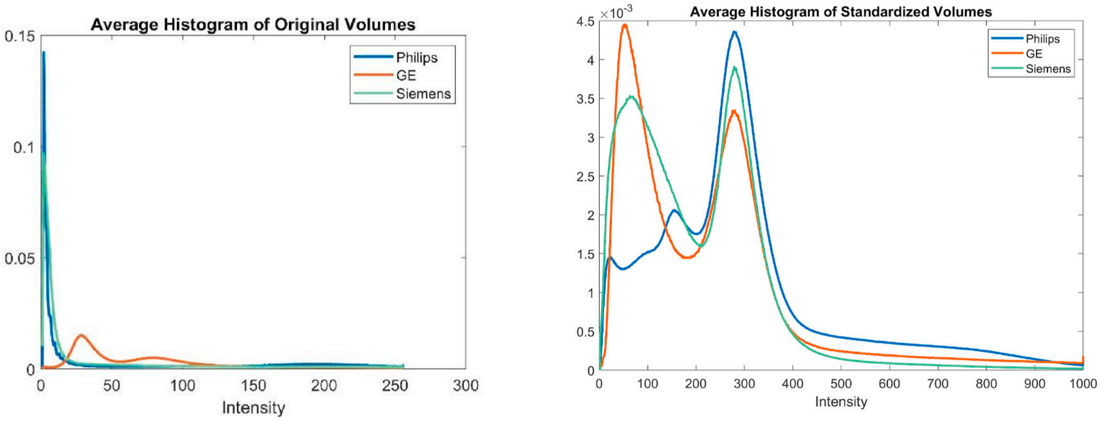


Fig. 21. Average intensity histograms over all 252 testing volumes for Philips, GE, Siemens scanners. Original (left) and standardized (right).

datasets, improve normalization techniques and investigate domain adaptation especially for DL methods.

In terms of generalization performance across datasets, the MICCAI dataset had the best performance in both datasplits for SC U-Net. In datasplit2, the high performance could be attributed to the training pool being strictly from MICCAI despite the number of training samples being lower. MICCAI annotations are also developed by experts, through consensus. Therefore high performance may be related to the annotating pattern, but as shown by the results for Rater5, which was developed by experts as well, there is still a drop in performance, so this is not the complete piece of the puzzle. This could indicate a preference for DL methods to predict better on datasets they are trained upon and that there are intra-dataset similarities (or bias) related to data characteristics and annotating patterns that DL systems are able to learn. It is challenging to delineate lesion boundaries precisely due to partial volume averaging, or diffuse (ischemic/demyelinating) pathology. Therefore, including multiple raters allows for multiple representations to be learned. Also, since it is common practice for less experienced raters to provide most of the annotated data, while having access to only a few expert annotations, this work demonstrates that combining our annotations (CAIN/ADNI/CCNA developed by medical students) with the ones from experts (MICCAI/MRBrains) is more beneficial for DL systems. As found in another work, the authors explored the impact of rater style on deep learning and found that training on a single rater leads to models that are strongly biased to only that single expert, which impedes generalization (Lucena et al., 2018). To combat this, the authors propose "silver" standard ground truths that are found using automated approaches that have been previously validated. Results showed that they outperformed (i.e., larger Dice coefficients) over state-of-the-art methods without using gold standard annotations for training CNNs indicating likely that using a larger sample is more beneficial, even if the annotations are not developed by experts. We found similar conclusions in this work and deduce it is better to include multiple raters in the training set to remove some of the bias that would be introduced by using only a single rater.

One of the important areas to be evaluated for WML segmentation tools is the ability to operate robustly over all lesion loads. There is growing evidence that certain lesion patterns have different etiological origins (Jung et al., 2020) and may differentiate between diseases such as dementia and CVD. In one of the lesion patterns identified in (Jung et al., 2020), there are periventricular lesions and in other patterns there are small punctate lesions in the deep WM. Tools should operate equally in these scenarios to quantify disease across neurodegenerative populations. As shown in the results, DL approaches stand out compared to traditional approaches in terms of performance, especially in the low lesion load categories. As visually seen, DL predictions are detecting small lesions that were hard to detect for years with traditional approaches. This is likely due to traditional approaches' large dependence on mainly intensity as a feature, while DL methods incorporate higher level features that may permit for detection of all types of lesions. Differences in neurodegeneration, as it manifests in the brain (atrophy, large ventricles, ischemia, lacunes, etc.), can arise with different disease levels and a robust tool should generalize across all levels. Despite the high performance, ANOVA tests revealed significant differences in DSC means across lesion load groups (low, medium and high) for DL as well as traditional approaches, for both datasplit1 and datasplit2, indicating that each of the tools' performance is highly dependent on lesion load. There were no significant differences across MoCA categories over all the methods for datasplit1. However, in datasplit2, there were significant differences across MoCA categories for U-Net and Tiramisu, indicating these models had difficulty generalizing in brains with differing levels of neurodegeneration.

Regarding study limitations, there are few that can be mentioned. Firstly, it is possible that architectures with more parameters (i.e., Dense U-Net) have even greater potential in a scenario where more training data is available. Also, the number of scanners by vendor type could be better balanced in future works as well. Currently, each centre from each dataset was roughly sampled equally, and because there was unequal distribution of scanner vendors across the centres, the result was unbalanced. However, this results in a diverse representation of acquisition parameters and scanner models which is still a benefit of the work. Another limitation is the amount of data available to perform inter-rater variability analysis. In the future, more samples will be generated. Additionally, in terms of comparison to other works, LPA was trained on MS data and on a moderately small sample size, which likely plays a large role in the reduced performance (and false positives shown in Fig. 19). While it was not our intention to fine-tune this method, it is possible to retrain this system which could possibly improve performance. Another limitation was the sample size of the manual segmentation comparisons, although the trends and analysis shows that there is good agreement between the raters. As was shown, using multiple raters increased performance significantly, likely through increased training data diversity and sample size - which is a strong benefit of this work. In the future we hope to include more annotations from additional raters. The last limitation is that we mainly investigated DSC and volume metrics to examine algorithm performance. In the future, we want to examine performance as a function of lesion size and location, which we believe are also important to investigate.

5. Conclusions

WML are important markers related to dementia and cerebrovascular disease (CVD) and automated tools can be used to measure them in an accurate, objective and efficient manner. In this work, seven WML segmentation techniques were evaluated for multicentre FLAIR MRI. Two methods consist of traditional approaches, which include a partial volume averaging technique and the lesion prediction algorithm (LPA) which uses regression on a per-voxel basis. The other five methods are based on CNNs and variants of the U-Net family, including U-Net, SC U-Net, UResNet2, MultiResNet and Tiramisu (dense nets) architectures. To train and test the algorithms, 252 FLAIR MRI volumes from 5 multicentre datasets (CAIN, ADNI, CCNA, MRBrains, MICCAI) from 33 centres with vascular and dementia pathology are utilized. Four folds were used to test over all 252 vol and performance is reported over all volumes. A second data splitting strategy was used to examine how algorithms generalized when trained on a single dataset and tested on the others. Algorithms are compared over a variety of dimensions related to clinical utility and safety, namely, accuracy, generalization across scanners and datasets, and robustness over different levels of disease. The evaluation framework may be used to determine the proof of effectiveness of computer-generated biomarkers related to WML segmentation and other tasks. Over all dimensions, deep learning methods consistently outperformed traditional methods, with SC U-Net obtaining top performance with a mean DSC = 0.71 over all 252 vol

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

● We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through the NSERC Discovery Grant program.

● The Canadian Atherosclerosis Imaging Network (CAIN) was established through funding from a Canadian Institutes of Health Research Team Grant for Clinical Research Initiatives (CIHR-CRI 88057). Funding for the infrastructure was received from the Canada Foundation for Innovation (CFI-CAIN 20099), with matching funds provided by the governments of [H]ta, Ontario, and Quebec.

● Data collection and sharing for this project was partially funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defence award number W81XWH-12-2-0012).

● We acknowledge the support from the Canadian Consortium on Neurodegeneration and Aging (CCNA) for data from the COMPASS-ND study, data storage, and management and standardized MRI protocols.

Appendix

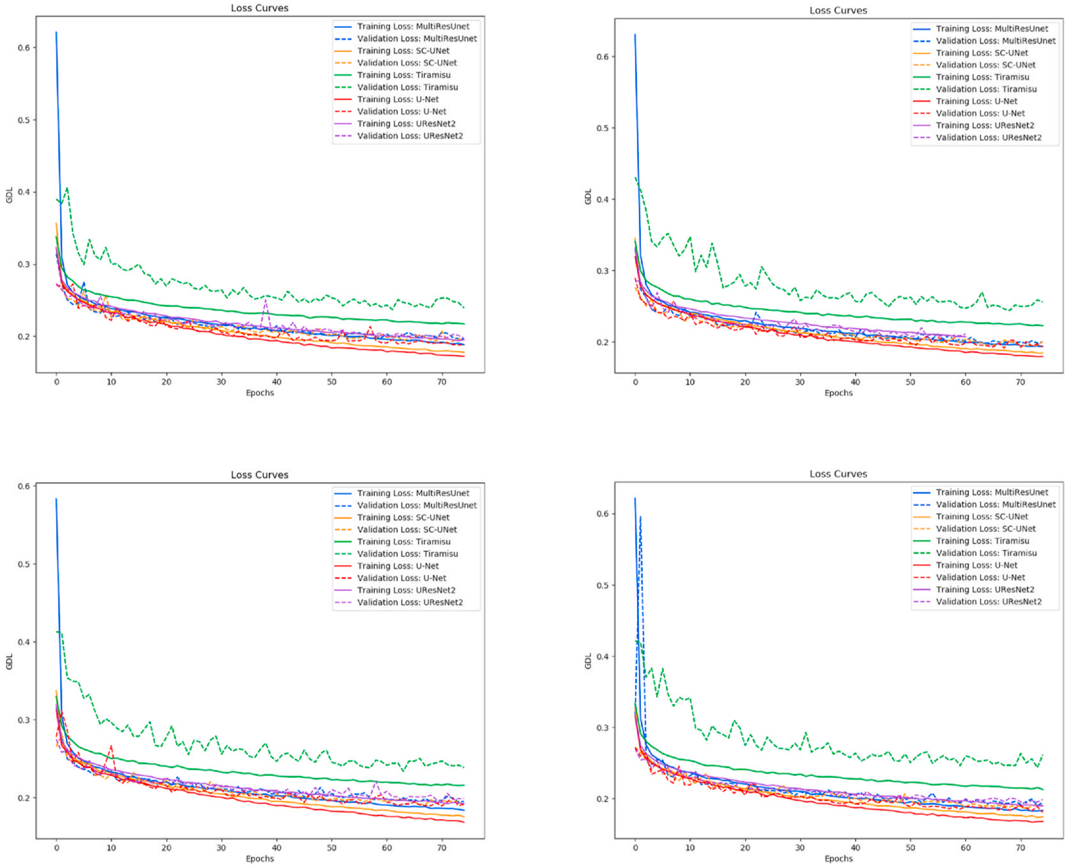


Fig. A. Training and validation loss curves for all methods for each fold of datasplit 1.

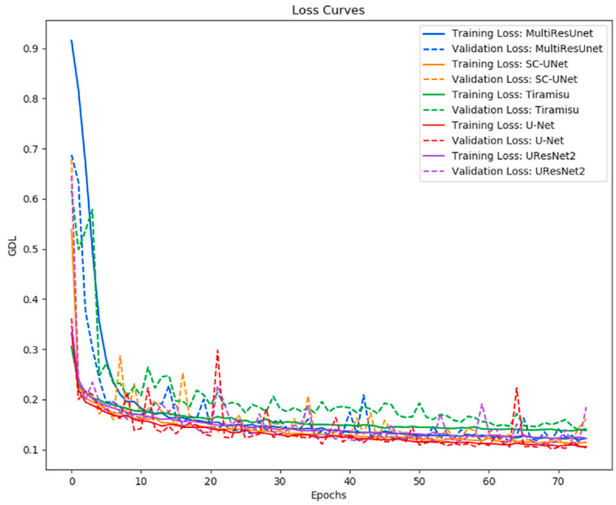


Fig. B. Training and validation loss curves for all methods for datasplit 2.

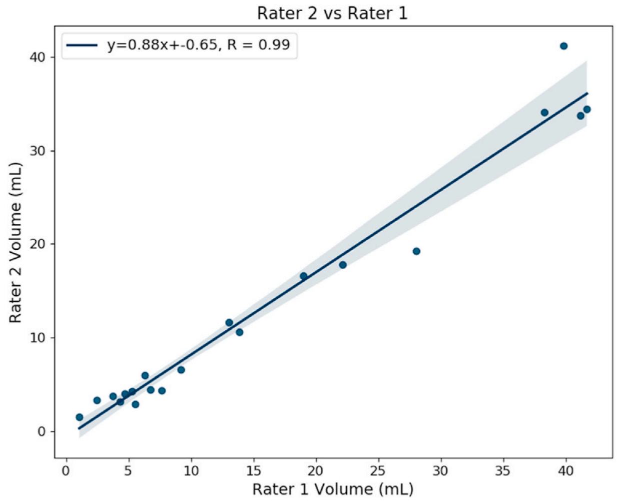


Fig. C. Regression plot of rater 2 vs rater 1 WML volume.

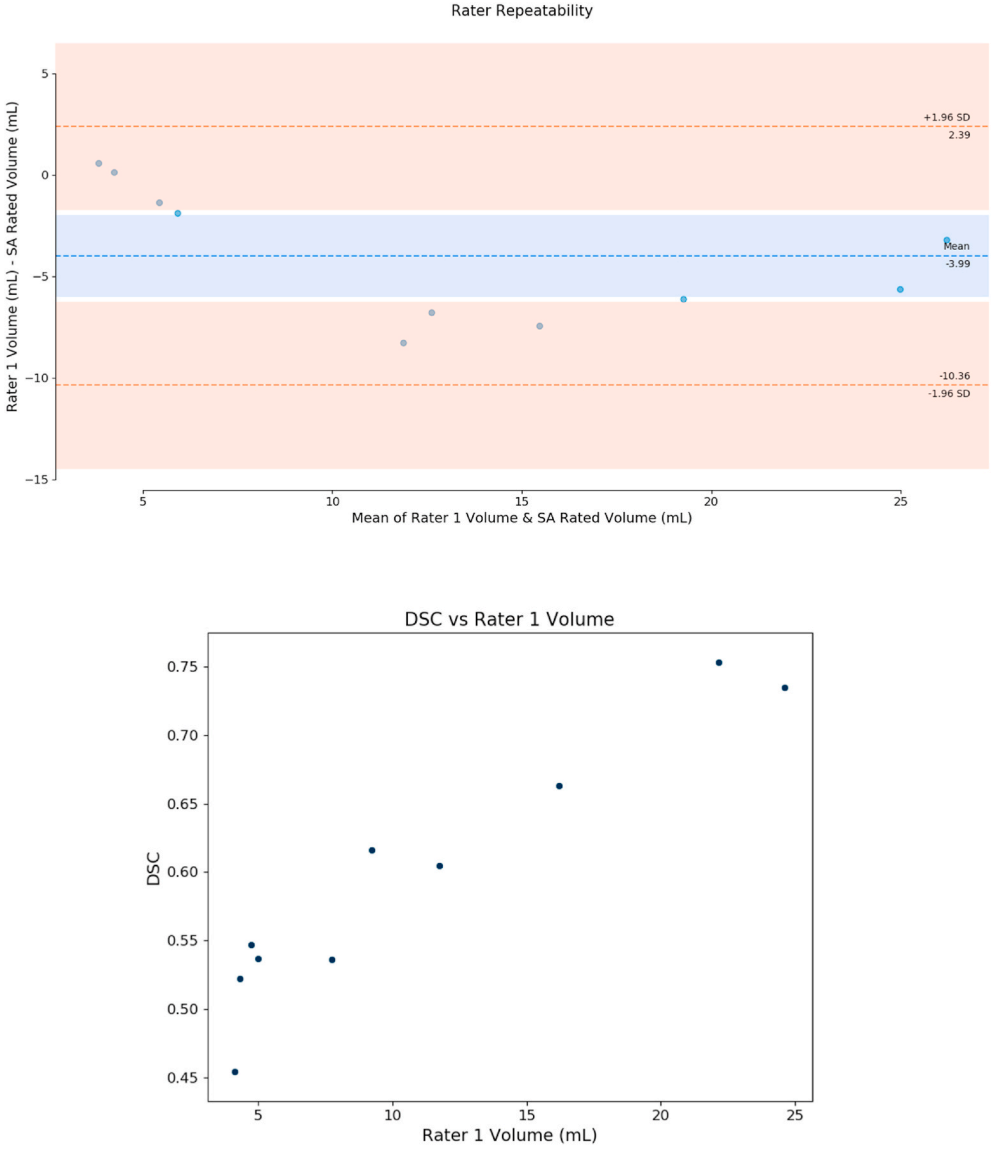


Fig. D. Bland-Altman plot and DSC vs Rater 1 vol for analyzing reliability of semi-automated ground truth protocol.

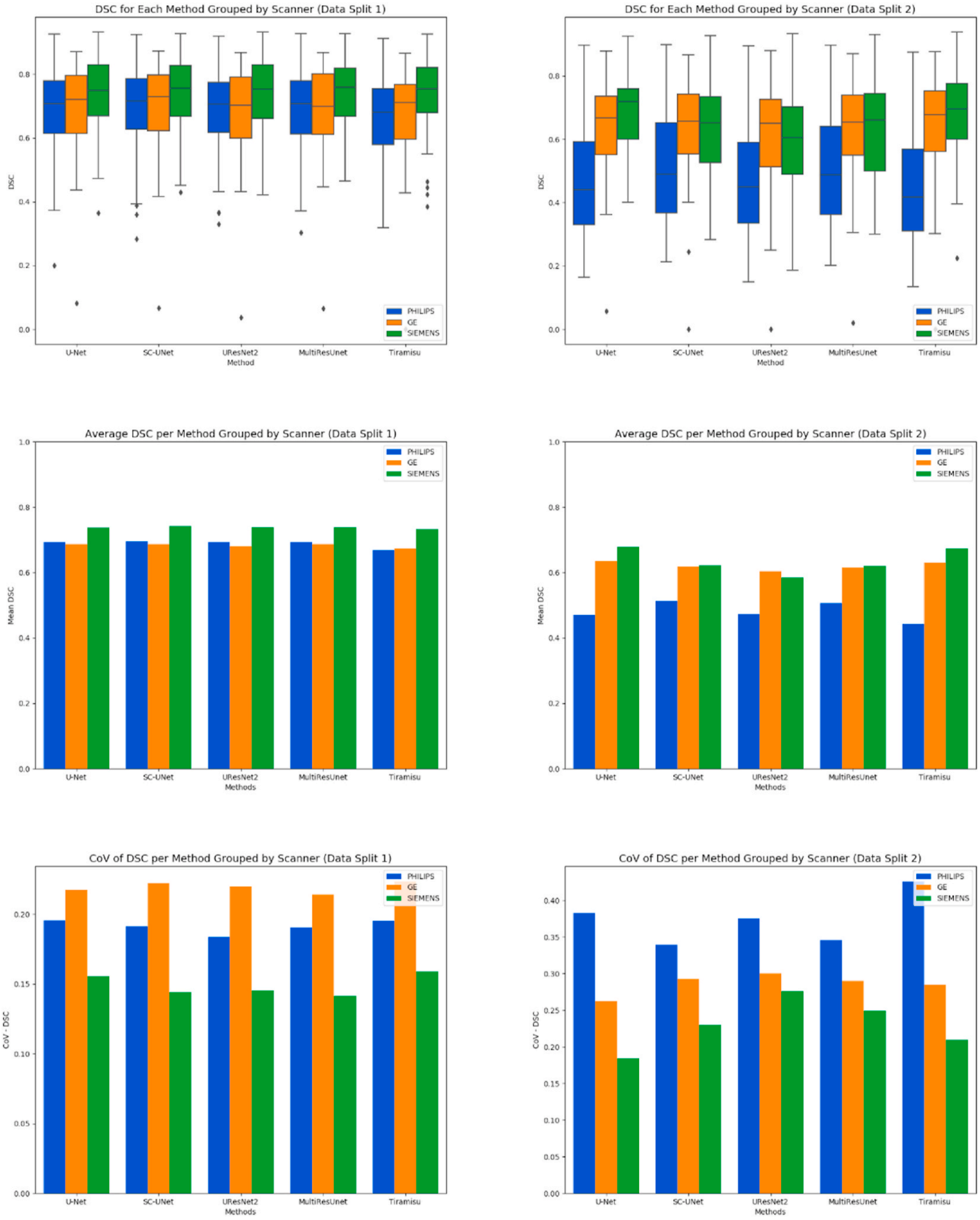
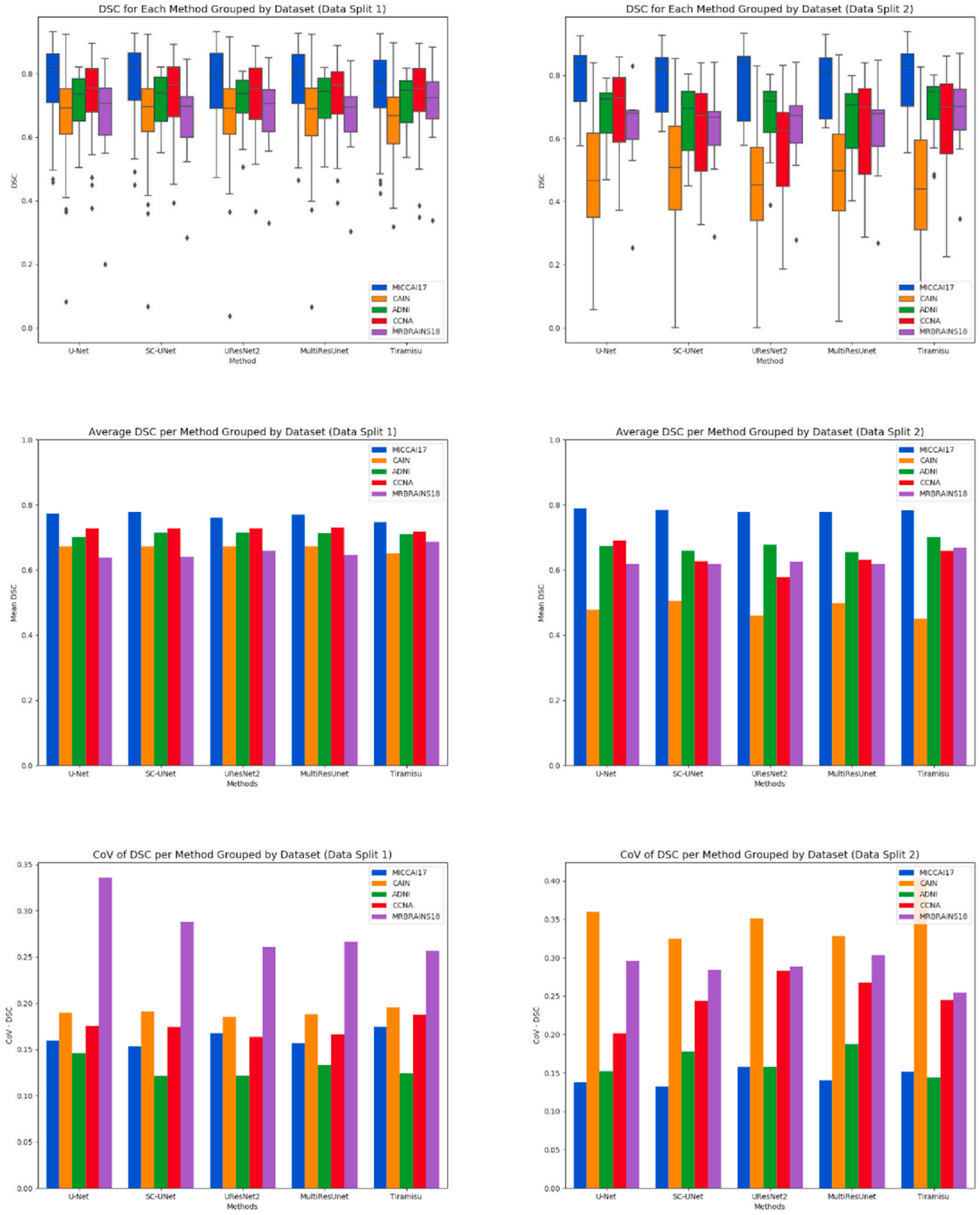


Fig. E. DSC distributions, mean DSC and DSC CoV for WML segmentation methods vs. scanner vendor. datasplit1 (left) and datasplit2 (right).


DSC for Each Method Grouped by Dataset (Data Split 1)

DSC for Each Method Grouped by Dataset (Data Split 2)

Average DSC per Method Grouped by Dataset (Data Split 1)

Average DSC per Method Grouped by Dataset (Data Split 2)

CoV of DSC per Method Grouped by Dataset (Data Split 1)

CoV of DSC per Method Grouped by Dataset (Data Split 2)

Fig. F. DSC distributions, mean DSC and DSC CoV for WML segmentation methods as a function of dataset. datasplit1 (left) and datasplit2 (right).

Table A
ANOVA analysis of effect of scanner vendor on algorithm performance for datasplit1 (F-value and Pr > F). Null hypothesis is that DSC means are the same across vendors. Post-hoc analysis compared DSC across groups for significant tests and bold p-values indicate significant differences between scanners (α = 0.05). Post-hoc results reported as differences between transformed DSC means of the two groups, and the p-value: diff (p-val).

Method	PVA	LPA	U-Net	SC-U-Net	UResNet2	Tiramisu	MultiResU
F-Value	0.19	3.97	3.13	3.48	4.08	5.99	3.35
Pr > F	0.828	0.0201	0.0454	0.0323	0.0181	0.0029	0.0367
GE vs. Philips		−0.04 (0.3016)	0.005 (0.9637)	0.006 (0.9594)	0.008 (0.8667)	−0.004 (0.9524)	0.005 (0.9535)
GE vs. Siemens		0.02 (0.7029)	0.04 (0.1075)	0.039 (0.0840)	0.044 **(0.0411)**	0.044 **(0.0481)**	0.038 (0.0891)
Philips vs. Siemens		0.06 **(0.0186)**	0.03 (0.0577)	0.034 **(0.0423)**	0.036 **(0.0320)**	0.048 **(0.0024)**	0.033 **(0.0487)**

Table B
ANOVA analysis of effect of scanner vendor on algorithm performance for datasplit2 (F-value and Pr > F). Null hypothesis is that DSC means are the same across vendors. Post-hoc analysis compared DSC across groups for significant tests and bold p-values indicate significant differences between scanners (α = 0.05). Post-hoc results reported as differences between transformed DSC means of the two groups, and the p-value: diff (p-val)

Method	PVA	LPA	U-Net	SC-U-Net	UResNet2	Tiramisu	MultiResUNet
F-Value	1.29	1.43	34.29	10.28	11.67	38	10.57
Pr > F	0.2788	0.2418	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
GE vs. Philips			−0.1 **(<0.0001)**	−0.07 **(0.0048)**	−0.09 **(0.0006)**	−0.13 **(<0.0001)**	−0.07 **(0.0050)**
GE vs. Siemens			0.03 (0.4640)	0.005 (0.9999)	−0.02 (0.8449)	0.03 (0.4895)	0.005 (0.9951)
Philips vs. Siemens			0.15 **(<0.0001)**	0.07 **(0.0003)**	0.08 **(0.0004)**	0.16 **(<0.0001)**	0.07 **(0.0002)**

Table C
ANOVA analysis of effect of dataset on algorithm performance for datasplit1 (F-value and Pr > F). Null hypothesis is that DSC means are the same across datasets. Post-hoc analysis compared DSC across groups for significant tests and bold p-values indicate significant differences between datasets (α = 0.05). Results reported as differences between transformed DSC means of the two groups, and the p-value: diff (p-val).

Method	PVA	LPA	U-Net	SC-U-Net	UResNet2	Tiramisu	MultiResU
F-Value	1.12	22.19	7.85	8.73	6.41	7.26	7.74
Pr > F	**0.3477**	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
ADNI vs. CAIN		0.01 (0.9937)	−0.02 (0.9005)	−0.03 (0.6908)	−0.03 (0.6479)	−0.05 (0.3249)	−0.03 (0.6797)
ADNI vs. CCNA		0.08 (0.1894)	0.02 (0.9241)	0.01 (0.9888)	0.01 (0.9934)	0.01 (0.9770)	0.02 (0.9769)
ADNI vs. MICCAI		0.19 **(<0.0001)**	0.06 (0.1199)	0.05 (0.1827)	0.04 (0.4721)	0.03 (0.6844)	0.05 (0.2905)
ADNI vs. MRBrains		0.13 (0.1407)	−0.04 (0.8711)	−0.05 (0.7312)	−0.04 (0.8836)	−0.015 (0.9983)	−0.05 (0.7845)
CAIN vs. CCNA		0.07 (0.0683)	0.04 (0.1722)	0.04 (0.1701)	0.04 (0.1733)	0.06 (0.0515)	0.05 (0.1172)
CAIN vs. MICCAI		0.18 **(<0.0001)**	0.08 **(<0.0001)**	0.08 **(<0.0001)**	0.07 **(<0.0001)**	0.08 **(<0.0001)**	0.08 **(<0.0001)**
CAIN vs. MRBrains		0.12 (0.1254)	−0.02 (0.9828)	0.03 (0.9758)	−0.01 (0.9996)	0.035 (0.9197)	−0.02 (0.9893)
CCNA vs. MICCAI		0.11 **(0.0013)**	0.04 (0.4317)	0.04 (0.3106)	0.03 (0.6581)	0.02 (0.8218)	0.03 (0.5626)
CCNA vs. MRBrains		0.047 (0.8850)	−0.06 (0.5164)	0.06 (0.4836)	−0.05 (0.7124)	−0.025 (0.9824)	−0.07 (0.5003)
MICCAI vs. MRBrains		−0.063 (0.7541)	0.1 (0.0719)	0.1 **(0.0458)**	−0.08 (0.2286)	−0.045 (0.7671)	−0.1 (0.0896)

Table D
ANOVA analysis of effect of dataset on algorithm performance for datasplit2 (F-value and Pr > F). Null hypothesis is that DSC means are the same across datasets. Post-hoc analysis compared DSC across groups for significant tests and bold p-values indicate significant differences between datasets (α = 0.05). Results reported as differences between transformed DSC means of the two groups, and the p-value: diff (p-val).

Method	PVA	LPA	U-Net	SC-U-Net	UResNet2	Tiramisu	MultiResUNet
F-Value	1.71	8.49	24.88	15.3	22.31	26.59	15.19
Pr > F	**0.1481**	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
ADNI vs. CAIN		0.03 (0.8618)	−0.13 **(<0.0001)**	−0.11 **(0.0005)**	−0.15 **(<0.0001)**	−0.17 **(<0.0001)**	−0.11 **(0.0006)**
ADNI vs. CCNA		0.08 (0.1202)	0.01 (0.9896)	−0.02 (0.9599)	−0.07 (0.1536)	−0.03 (0.9209)	−0.02 (0.9933)
ADNI vs. MICCAI		0.19 **(<0.0001)**	0.09 (0.1361)	0.1 (0.0915)	0.08 (0.2249)	0.07 (0.5072)	0.09 (0.1176)
ADNI vs. MRBrains		0.15 **(0.0287)**	−0.04 (0.9447)	−0.03 (0.9791)	−0.03 (0.9396)	−0.02 (0.9947)	−0.03 (0.9900)
CAIN vs. CCNA		0.05 (0.1626)	0.14 **(<0.0001)**	0.09 **(0.0010)**	0.08 **(0.0016)**	0.14 **(<0.0001)**	0.1 **(0.0003)**
CAIN vs. MICCAI		0.16 **(<0.0001)**	0.22 **(<0.0001)**	0.21 **(<0.0001)**	0.23 **(<0.0001)**	0.24 **(<0.0001)**	0.2 **(<0.0001)**
CAIN vs. MRBrains		0.12 (0.0522)	0.09 (0.1329)	0.08 (0.3115)	0.12 **(0.0410)**	0.15 **(0.0065)**	0.08 (0.2720)
CCNA vs. MICCAI		0.11 **(0.0414)**	0.08 (0.2230)	0.12 **(0.0112)**	0.15 **(0.0004)**	0.1 (0.1199)	0.1 **(0.0307)**
CCNA vs. MRBrains		0.07 (0.6022)	−0.05 (0.8015)	−0.01 (0.9999)	0.04 (0.9391)	0.01 (0.9999)	−0.01 (0.9996)
MICCAI vs. MRBrains		−0.04 (0.9428)	−0.13 (0.0972)	−0.13 (0.1046)	−0.11 (0.1430)	−0.09 (0.5078)	−0.12 (0.1551)

Table E
ANOVA analysis of effect of lesion load on algorithm performance for datasplit1 (F-value and Pr > F). Null hypothesis is that DSC means are the same across lesion loads. Post-hoc analysis compared DSC across groups for significant tests and bold p-values indicate significant differences between lesion loads (α = 0.05). Post-hoc results reported as differences between transformed DSC means of the two groups, and the p-value: diff (p-val).

Method	PVA	LPA	U-Net	SC-U-Net	UResNet2	Tiramisu	MultiResU
F-Value	131.44	94.42	68.1	68.77	85	84.6	72.34
Pr > F	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Low LL vs. High LL	0.24 **(<0.0001)**	0.22 **(<0.0001)**	0.14 **(<0.0001)**	0.14 **(<0.0001)**	0.15 **(<0.0001)**	0.15 **(<0.0001)**	0.15 **(<0.0001)**
Med LL vs. High LL	0.15 **(<0.0001)**	0.17 **(< 0.0001)**	0.09 **(<0.0001)**	0.09 **(<0.0001)**	0.09 **(<0.0001)**	0.1 **(<0.0001)**	0.09 **(<0.0001)**
Low LL vs. Med LL	0.09 (< **0.0001)**	0.05 **(0.0074)**	0.05 **(0.0004)**	0.05 **(0.0002)**	0.06 **(<0.0001)**	0.05 **(<0.0001)**	0.06 **(<0.0001)**

Table F
ANOVA analysis of effect of lesion load on algorithm performance for datasplit2 (F-value and Pr > F). Null hypothesis is that DSC means are the same across lesion loads. Post-hoc analysis compared DSC across groups for significant tests and bold p-values indicate significant differences between lesion loads (α = 0.05). Post-hoc results reported as differences between transformed DSC means of the two groups, and the p-value: diff (p-val).

Method	PVA	LPA	U-Net	SC-U-Net	UResNet2	Tiramisu	MultiResU
F-Value	86.04	79.36	68.73	52.47	48.29	62.29	61.45
Pr > F	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Low LL vs. High LL	0.21 (**< 0.0001**)	0.21 (**<0.0001**)	0.21 (**<0.0001**)	0.18 (**<0.0001**)	0.19 (**<0.0001**)	0.23 (**<0.0001**)	0.19 (**<0.0001**)
Med LL vs. High LL	0.13 (**<0.0001**)	0.15 (**<0.0001**)	0.14 (**<0.0001**)	0.12 (**<0.0001**)	0.12 (**<0.0001**)	0.15 (**<0.0001**)	0.12 (**<0.0001**)
Low LL vs. Med LL	0.07 (**<0.0001**)	0.06 **(0.0004)**	0.07 (**<0.0001**)	0.06 **(0.0023)**	0.07 **(0.0010)**	0.08 **(0.0002)**	0.07 **(0.0001)**

Table G
ANOVA analysis of effect of cognitive impairment on algorithm performance for datasplit1 (F-value and Pr > F). Null hypothesis is that the DSC means are the same across normal and impaired subjects. Bold values indicate significant differences exist across groups (α = 0.05).

Method	Metric	F-Value	Pr > F
PVA	DSC	0.09	0.7680
LPA	DSC	0.67	0.4154
U-Net	DSC	0.13	0.7208
SC-U-Net	DSC	0.09	0.7652
UResNet2	DSC	0.05	0.8279
Tiramisu	DSC	0.02	0.8989
MultiResUNet	DSC	0.04	0.8464

Table H
ANOVA analysis of effect of cognitive impairment on algorithm performance for datasplit2 (F-value and Pr > F). Null hypothesis is that the DSC means are the same across normal and impaired subjects. Bold values indicate significant differences exist across groups (α = 0.05).

Method	Metric	F-Value	Pr > F
PVA	DSC	0.01	0.9207
LPA	DSC	1.20	0.2738
U-Net	DSC	4.13	**0.0436**
SC-U-Net	DSC	1.21	0.2724
UResNet2	DSC	3.29	0.0713
Tiramisu	DSC	4.89	**0.0283**
MultiResUNet	DSC	1.10	0.2964

References

LST – Lesion Segmentation for SPM | Paul Schmidt – Freelance Statistician", 2021. Applied-statistics.de [Online]. Available. https://www.applied-statistics.de/lst.html, 08- Jun- 2021.

Admiraal-Behloul, F., Van Den Heuvel, D.M.J., Olofsen, H., van Osch, M.J., van der Grond, J., van Buchem, M.A., Reiber, J.H., 2005. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. Neuroimage 28 (3), 607–617.

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J., 2017. Deep learning for brain MRI segmentation: state of the art and future directions. J. Digit. Imag. 30 (4), 449–459.

Alber, J., Alladi, S., Bae, H.J., Barton, D.A., Beckett, L.A., Bell, J.M., et al., 2019. White matter hyperintensities in vascular contributions to cognitive impairment and dementia (VCID): knowledge gaps and opportunities. Alzheimer's Dementia: Transl. Res. Clin. Intervent. 5 (1), 107–117.

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.

Azizyan, A., Sanossian, N., Mogensen, M.A., Liebeskind, D.S., 2011. Fluid-attenuated inversion recovery vascular hyperintensities: an important imaging marker for cerebrovascular disease. Am. J. Neuroradiol. 32 (10), 1771–1775.

Badji, A., Westman, E., 2020. Cerebrovascular pathology in Alzheimer's disease: hopes and gaps. Psychiatr. Res. Neuroimaging 306, 111184.

Bernal, J., Kushibar, K., Asfaw, D.S., Valverde, S., Oliver, A., Martí, R., Lladó, X., 2019. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. Artif. Intell. Med. 95, 64–81.

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.

Chertkow, H., Borrie, M., Whitehead, V., Black, S.E., Feldman, H.H., Gauthier, S., et al., 2019. The comprehensive assessment of neurodegeneration and dementia: Canadian cohort study. Can. J. Neurol. Sci. 46 (5), 499–511.

Dadar, M., Maranzano, J., Misquitta, K., Anor, C.J., Fonov, V.S., Tartaglia, M.C., Carmichael, O.T., Decarli, C., Collins, D.L., 2017. Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging. Neuroimage 157, 233–249.

De Boer, R., Vrooman, H.A., Van Der Lijn, F., Vernooij, M.W., Ikram, M.A., Van Der Lugt, A., et al., 2009. White matter lesion extension to automatic brain tissue segmentation on MRI. Neuroimage 45 (4), 1151–1161.

de Sitter, A., Steenwijk, M.D., Ruet, A., Versteeg, A., Liu, Y., van Schijndel, R.A., et al., 2017. Performance of five research-domain automated WM lesion segmentation methods in a multi-center MS study. Neuroimage 163, 106–114.

Debette, S., Markus, H.S., 2010. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. BMJ 341.

DiGregorio, J., 2018. Brain extraction methods for neurological FLAIR MRI. In: 2018 Imaging Network Ontario Conference. IMNO.

DiGregorio, J., Arezza, G., Gibicar, A., Moody, A.R., Tyrrell, P.N., Khademi, A., 2021. Intracranial volume segmentation for neurodegenerative populations using multicentre FLAIR MRI. Neuroimage: Report 1 (1), 100006.

Dobson, A.J., Barnett, A.G., 2018. An Introduction to Generalized Linear Models. CRC press.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C., 2016. The importance of skip connections in biomedical image segmentation. In: Deep Learning and Data Labeling for Medical Applications. Springer, Cham, pp. 179–187.

Duong, M.T., Rudie, J.D., Wang, J., Xie, L., Mohan, S., Gee, J.C., Rauschecker, A.M., 2019. Convolutional neural network for automated FLAIR lesion segmentation on clinical brain MR imaging. Am. J. Neuroradiol. 40 (8), 1282–1290.

Fazekas, F., Kleinert, R., Offenbacher, H., Schmidt, R., Kleinert, G., Payer, F., et al., 1993. Pathologic correlates of incidental MRI white matter signal hyperintensities. Neurology 43 (9), 1683-1683.

Frey, B.M., Petersen, M., Mayer, C., Schulz, M., Cheng, B., Thomalla, G., 2019. Characterization of white matter hyperintensities in large-scale MRI-studies. Front. Neurol. 10, 238.

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

Gorelick, P.B., Scuteri, A., Black, S.E., DeCarli, C., Greenberg, S.M., Iadecola, C., et al., 2011. Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 42 (9), 2672–2713.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U.G., Kuker, W., Battaglini, M., Rothwell, P.M., Jenkinson, M., 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., Joules, R., et al., 2018. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. Neuroimage: Clinical 17, 918–934.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. October). Identity mappings in deep residual networks. In: European Conference on Computer Vision. Springer, Cham, pp. 630–645.

Heinen, R., Steenwijk, M.D., Barkhof, F., Biesbroek, J.M., van der Flier, W.M., Kuijf, H.J., et al., 2019. Performance of five automated white matter hyperintensity segmentation methods in a multicenter dataset. Sci. Rep. 9 (1), 1–12.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Hwang, H., Rehman, H.Z.U., Lee, S., 2019. 3D U-Net for skull stripping in brain MRI. Appl. Sci. 9 (3), 569.

Ibtehaz, N., Rahman, M.S., 2020. MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Network. 121, 74–87.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift arXiv preprint arXiv:1502.03167.

Jack Jr., C.R., O'Brien, P.C., Rettman, D.W., Shiung, M.M., Xu, Y., Muthupillai, R., et al., 2001. FLAIR histogram segmentation for measurement of leukoaraiosis volume. J. Magn. Reson. Imag.: An Off. J. Int. Soc. Magnet. Reson. Med. 14 (6), 668–676.

Jack Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imag.: An Off. J. Int. Soc. Magnet. Reson. Med. 27 (4), 685–691.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 11–19.

Jung, K.H., Stephens, K.A., Yochim, K.M., Riphagen, J.M., Kim, C.M., Buckner, R.L., Salat, D.H., 2020. Heterogeneity of Cerebral White Matter Lesions and Clinical Correlates in Older Adults. Stroke, STROKEAHA-120.

Khademi, A., Moody, A.R., Apr. 2015. Multiscale denoising and PVA estimation for WML segmentation in FLAIR MRI. In: IEEE International Symposium of Biomedical Imaging (ISBI).

Khademi, A., Venetsanopoulos, A., Moody, A.R., 2011. Robust white matter lesion segmentation in FLAIR MRI. IEEE Trans. Biomed. Eng. 59 (3), 860–871.

Khademi, A., Venetsanopoulos, A., Moody, A.R., 2014. Generalized method for partial volume estimation and tissue segmentation in cerebral magnetic resonance images. J. Med. Imag. 1 (1), 014002.

Khademi, A., Reiche, B., DiGregorio, J., Arezza, G., Moody, A.R., 2020. Whole volume brain extraction for multi-centre, multi-disease FLAIR MRI datasets. Magn. Reson. Imag. 66, 116–130.

Knight, J., Taylor, G.W., Khademi, A., 2018. Voxel-wise logistic regression and leave-one-source-out cross validation for white matter hyperintensity segmentation. Magnet. Reson. Imag. 54, 119–136.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropract. Med. 15 (2), 155–163.

Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., et al., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. IEEE Trans. Med. Imag. 38 (11), 2556–2568.

Lao, Z., Shen, D., Liu, D., Jawad, A.F., Melhem, E.R., Launer, L.J., et al., 2008. Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. Acad. Radiol. 15 (3), 300–313.

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.S., Menze, B., 2018a. Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. Neuroimage 183, 650–665.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A., 2018b. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imag. 37 (12), 2663–2674.

Li, H., Zhang, J., Muehlau, M., Kirschke, J., Menze, B.. Multi-scale convolutional-stack aggregation for robust white matter hyperintensities segmentation," arXiv.org, 27-feb-2019. [Online]. Available. https://arxiv.org/abs/1807.05153, 10-Jun-2021.

Liu, Z., Lian, T., Farrell, J., Wandell, B.A., 2020. Neural network generalization: the impact of camera parameters. IEEE Access 8, 10443–10454.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Lucena, O., Souza, R., Rittner, L., Frayne, R., Lotufo, R., 2018. Silver Standard Masks for Data Augmentation Applied to Deep-Learning-Based Skull-Stripping. In: IEEE 15th International Symposium on Biomedical Imaging. ISBI 2018), p. 2018.

Mahbod, A., Schaefer, G., Wang, C., Ecker, R., Dorffner, G., Ellinger, I., 2021. In: "Investigating and Exploiting Image Resolution for Transfer Learning-Based Skin Lesion Classification," 2020 25th International Conference on Pattern Recognition. ICPR).

Marek, M., Horyniecki, M., Frączek, M., Kluczewska, E., 2018. Leukoaraiosis–new concepts and modern imaging. Pol. J. Radiol. 83, e76.

Moeskops, P., de Bresser, J., Kuijf, H.J., Mendrik, A.M., Biessels, G.J., Pluim, J.P., Išgum, I., 2018. Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI. Neuroimage: Clinical 17, 251–262.

Mohaddes, Z., Das, S., Abou-Haidar, R., Safi-Harab, M., Blader, D., Callegaro, J., et al., 2018. National neuroinformatics framework for canadian consortium on neurodegeneration in aging (CCNA). Front. Neuroinf. 12, 85.

Narayana, P.A., Coronado, I., Sujit, S.J., Sun, X., Wolinsky, J.S., Gabr, R.E., 2020. Are multi-contrast magnetic resonance images necessary for segmenting multiple sclerosis brains? A large cohort study based on deep learning. Magnet. Reson. Imag. 65, 8–14.

Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H.. "The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment," journal of the American geriatrics society. [Online]. Available. https://pubmed.ncbi.nlm.nih.gov/15817019/, 10-Jun-2021.

Obuchowski, N.A., Reeves, A.P., Huang, E.P., et al., 2015. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons. Stat. Methods Med. Res. 24 (1), 68–106. https://doi.org/10.1177/0962280214537390.

Pantoni, L., 2010. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. Lancet Neurol. 9 (7), 689–701.

Pantoni, L., Garcia, J.H., 1997. Pathogenesis of leukoaraiosis: a review. Stroke 28 (3), 652–659.

Reiche, B., Moody, A.R., Khademi, A., 2019. Pathology-preserving intensity standardization framework for multi-institutional FLAIR MRI datasets. Magnet. Reson. Imag. 62, 59–69.

Ronneberger, O., Fischer, P., Brox, T., 2015. October). U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp. 234–241.

Sabottke, C.F., Spieler, B.M., 2020. The effect of image resolution on deep learning in radiography. Radiology: Artif. Intell. 2 (1).

Schmidt, P., 2017. Bayesian Inference for Structured Additive Regression Models for Large-Scale Problems with Applications to Medical Imaging. Doctoral dissertation,. lmu.

Seghier, M.L., Ramlackhansingh, A., Crinion, J., Leff, A.P., Price, C.J., 2008. Lesion identification using unified segmentation-normalisation models and fuzzy clustering. Neuroimage 41 (4), 1253–1266.

Simões, R., Mönninghoff, C., Dlugaj, M., Weimar, C., Wanke, I., van Walsum, A.M.V.C., Slump, C., 2013. Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images. Magnet. Reson. Imag. 31 (7), 1182–1189.

Smith, E.E., Biessels, G.J., De Guio, F., et al., 2019. Harmonizing brain magnetic resonance imaging methods for vascular contributions to neurodegeneration. In: Alzheimer's Dementia Diagnosis, Assessment and Disease Monitoring, 11, pp. 191–204. https://doi.org/10.1016/j.dadm.2019.01.002.

Soltanian-Zadeh, H., Peck, D.J., 2001. Feature space analysis: effects of MRI protocols. Med. Phys. 28 (11), 2344–2351.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, Cham, pp. 240–248.

Sullivan, D.C., Obuchowski, N.A., Kessler, L.G., et al., 2015. Metrology standards for quantitative imaging biomarkers. Radiology 277 (3), 813–825. https://doi.org/10.1148/radiol.2015142202 obuchowski.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Tardif, J.C., Spence, J.D., Heinonen, T.M., Moody, A., Pressacco, J., Frayne, R., et al., 2013. Atherosclerosis imaging and the Canadian atherosclerosis imaging network. Can. J. Cardiol. 29 (3), 297–303.

Thakur, S., Doshi, J., Pati, S., Rathore, S., Sako, C., Bilello, M., et al., 2020. Brain Extraction on MRI Scans in Presence of Diffuse Glioma: Multi-Institutional Performance Evaluation of Deep Learning Methods and Robust Modality-Agnostic Training. NeuroImage, p. 117081.

Vanderbecq, Q., Xu, E., Ströer, S., Couvy-Duchesne, B., Melo, M.D., Dormont, D., et al., 2020. Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. Neuroimage: Clinical 27, 102357.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., et al., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 12 (8), 822–838.

Wardlaw, J.M., Valdés Hernández, M.C., Muñoz-Maniega, S., 2015. What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. J. Am. Heart Assoc. 4 (6), e001140.

Wu, J., Zhang, Y., Wang, K., Tang, X., 2019. Skip connection U-Net for white matter hyperintensities segmentation from MRI. IEEE Access 7, 155194–155202.

Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31 (3), 1116–1128.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 15 (11).

Zhong, Y., Qi, S., Kang, Y., Feng, W., Haacke, E.M., 2012. June). Automatic skull stripping in brain MRI based on local moment of inertia structure tensor. In: *2012* IEEE International Conference on Information and Automation. IEEE, pp. 437–440.