Neuroimage: Reports 1 (2021) 100006

Contents lists available at ScienceDirect

Neuroimage: Reports

journal homepage: www.journals.elsevier.com/neuroimage-reports

Intracranial volume segmentation for neurodegenerative populations using multicentre FLAIR MRI☆

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

[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
ICV segmentation
Brain extraction
Skull-stripping
Deep learning

ABSTRACT

Intracranial volume (ICV) segmentation, also known as brain extraction or skull-stripping, is a critical preprocessing step in analytical pipelines for studying neurodegenerative diseases in magnetic resonance imaging (MRI). While the fluid-attenuated inversion recovery (FLAIR) MRI modality has emerged as an important sequence for analyzing cerebrovascular and neurodegenerative disease, most existing automated ICV segmentation methods have been developed for T1-weighted or multi-modal inputs. Additionally, many methods have been designed using single centre data of healthy subjects and encounter difficulties using images with varying acquisition parameters and neurodegenerative pathology. In this work, we develop and evaluate 2 traditional and 8 deep learning algorithms for ICV segmentation in FLAIR MRI. Training and testing were completed on 175 vol (8317 images) from 2 dementia and 1 vascular disease cohort. A human phantom FLAIR MRI dataset from a repeatedly scanned, healthy individual was also utilized for reliability analysis. Images were acquired from 47 imaging centres with varying scanners and parameters. To measure and compare performance, we present a novel framework for evaluating the effectiveness of computer generated segmentations on multicentre datasets. The evaluation framework includes assessments of algorithm accuracy, generalization capabilities, robustness to pathology and spatial location, and volumetric measurement reliability – all important dimensions for establishing proof of effectiveness (a prerequisite to clinical translation). The top performing method was a multiple resolution U-Net (MultiResUNet), which achieved a mean Dice similarity coefficient greater than 98% and was robust across pathology levels and spatial locations. Our results confirm a FLAIR-based ICV analytical pipeline can alone be utilized for large-scale neurodegenerative disease research. The presented evaluation framework can be deployed by other researchers to assess the viability of tools proposed for automated analysis of diverse, clinical MRI datasets.

1. Introduction

In 2010, there were 36.5 million people living with dementia and 7.7 million new cases occurring each year (Sosa-Ortiz et al., 2012). In Canada alone, it is estimated that over 1 million people will be living with dementia by 2031 with the annual cost of care exceeding $16 billion (Chambers et al., 2016). Alzheimer's disease (AD) is the leading cause of dementia (Mayeux and Stern, 2012). Cerebrovascular disease (CVD) has been identified as the second most common contributor to dementia risk (Smith et al., 2017) and early CVD may represent a pivotal stage for intervention before irreversible brain injury and disability occurs (Smith et al., 2017). Early identification allows interventions to alter disease

☆ * 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: justin.digregorio@ryerson.ca (J. DiGregorio).

https://doi.org/10.1016/j.ynirp.2021.100006
Received 7 January 2021; Received in revised form 24 February 2021; Accepted 26 February 2021
Available online 11 March 2021

trajectory with the aim of improving quality of life and cost to healthcare systems. Neuroimaging biomarkers are potentially powerful candidates for disease detection, quantification, and monitoring.

Magnetic resonance imaging (MRI) is ideal for neuroimaging biomarker measurement due to its ability to image the soft tissue in the brain with high detail. MRI visualizes structural abnormalities like cerebral atrophy and ischemic pathology such as white matter lesions (WML) (Brant-Zawadzki et al., 1996), which have been associated with dementia (Oppedal et al., 2015). With MRI, patients can be serially monitored to identify increasing WML volumes and accelerated brain aging patterns such as gray matter (GM) loss, white matter (WM) loss, and expansion of cerebrospinal fluid (CSF) spaces in the ventricular system, cortical sulci, and gyri which are higher in subjects with AD (Ntiri et al., 2020), (Scahill et al., 2003), (Silbert et al., 2003), (Gunter et al., 2003), (Sigurdsson et al., 2012), (Aribisala et al., 2013), (Leung et al., 2013), (Rocca et al., 2017). Neuroimaging biomarkers can therefore be leveraged to inform diagnostic criteria, to monitor risk, and to determine optimal intervention times (Struyfs et al., 2020). Unfortunately, performing manual biomarker analysis is expensive, laborious, and prone to inter-observer variability (Squitieri et al., 2009), (Nordenskjöld et al., 2013), (Hansen et al., 2015). Automated image analysis systems are a viable alternative that can compute quantitative biomarkers for subjects accurately and efficiently.

T1-weighted (T1) imaging is the most extensively used MRI modality for automated structural brain analysis. This is because T1 images provide strong contrast between the prominent healthy tissue classes of GM, WM, and CSF (Rehman et al., 2020a). Recently, fluid-attenuated inversion recovery (FLAIR) MRI has emerged as an important modality for neurodegenerative disease imaging and patient monitoring (Wardlaw et al., 2015), (García-Lorenzo et al., 2013), (Khademi et al., 2011), (Heinen et al., 2019), (Narayana et al., 2020). FLAIR MRI is preferred for WML analysis since the usually high T2 signal from CSF is suppressed (Soltanian-Zadeh and Peck, 2001), emphasizing the high white matter disease signal. This is due to the increased water content associated with ischemia and demyelination which can be more robustly seen with FLAIR than with T1 or T2-weighted (T2) imaging (Soltanian-Zadeh and Peck, 2001).

Intracranial volume (ICV) segmentation, also known as brain extraction or skull stripping, is the process of removing non-cerebral tissues like the skull, skin, and orbital cavities from brain MRI. ICV segmentation is a critical preprocessing step for biomarker extraction algorithms (i.e., tissue and lesion segmentation) and downstream analysis (i.e., image registration) where non-brain tissues are known to be a source of system interference (Khademi et al., 2020), (Manjón et al., 2014), (Kalavathi and Prasath, 2016), (Rehman et al., 2020b). Importantly, ICV is also used to normalize volumetric biomarkers of the brain (Rehman et al., 2020b), (Malone et al., 2015), (Schwarz et al., 2016) which can differentiate between normal and cognitively impaired subjects (Ntiri et al., 2020), (Leung et al., 2013), (Schwarz et al., 2016) by compensating for pre-morbid brain size, gender differences, and inter-subject head size variation (Nordenskjöld et al., 2013), (Hansen et al., 2015), (Rehman et al., 2020b), (Malone et al., 2015), (Schwarz et al., 2016).

There are a variety of ICV segmentation methods to analyze MRI (Kalavathi and Prasath, 2016), (Iglesias et al., 2011), (Ségonne et al., 2004), (Smith, 2002), (Eskildsen et al., 2012), (Datta and Narayana, 2011). Most of these methods have been designed for T1 MRI and cannot be directly translated to FLAIR sequences (DiGregorio, 2018). There are few works for FLAIR ICV segmentation in the literature. One such approach is semi-automated (Khademi et al., 2011), which remains laborious and subjective. Other approaches require multi-modal inputs (De Boer et al., 2007), (Hah et al., 2014), which increases acquisition costs and registration errors. In (Zhong et al., 2012), a FLAIR ICV segmentation method was designed based on edge detection, local moments of inertia, and morphology. Although promising, the authors acknowledge that this method was developed using single centre data (same scanner and parameters) and may not generalize to new data. Variabilities in scanner hardware and software across imaging centres creates inter-scan variability in tissue class intensities (Reiche et al., 2019), which causes challenges for automated approaches. In (Khademi et al., 2020), the authors tried to overcome this multicentre effect with a machine learning (ML) approach that utilized an intensity standardization preprocessing framework. This method improved generalization across different scanner vendors but had challenges in images with pathology. While such methods are promising, convolutional neural networks (CNN) can adapt to highly variable biomedical imaging data, are achieving state-of-the-art performance for a variety of MRI segmentation applications (Akkus et al., 2017), (Ali et al., 2019), (Bernal et al., 2019), do not require handcrafted feature engineering, and are easily translatable to FLAIR MRI. Therefore, the utility of CNN-based methods for ICV segmentation in multicentre FLAIR MRI is promising and requires further examination.

To this end, we implemented and evaluated 10 ICV segmentation algorithms using multicentre, multi-disease FLAIR MRI databases acquired with varying scanners and protocols. Both traditional approaches, being comprised of an unsupervised thresholding technique and a random forest classifier (Khademi et al., 2020), and deep learning approaches that utilize prominent CNN architectures for medical image segmentation were developed. To validate and compare algorithm performance, an evaluation framework that can be used to establish proof of effectiveness for automated analysis algorithms in large, multicentre datasets with diverse neurodegenerative pathology is proposed. Several dimensions related to clinical implementation including accuracy, generalizability, robustness to pathology and spatial location, and volume measurement reliability are introduced. The proposed tools and evaluation framework were realized using 175 vol (8317 image slices) from 35 international imaging centres and an additional 62 vol from a human phantom dataset (Duchesne et al., 2019a).

2. Materials and methods

2.1. Data

Experimental data for this work comes from 4 multicentre FLAIR MRI datasets. The first dataset is from the Alzheimer's Disease Neuroimaging Initiative (ADNI), one of the largest open repositories for the study of AD and dementia disease (Jack et al., 2008). ADNI includes 900 subjects with longitudinal follow up resulting in 4126 imaging volumes. The second data repository 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). CCNA has currently recruited 200 patients with additional time points for a total of 561 imaging volumes. The volumes from these three clinical datasets are accompanied by Montreal Cognitive Assessment (MoCA) scores (Nasreddine et al., 2005). MoCA is a cognitive screening assessment for neurodegenerative illnesses such as AD and vascular cognitive impairment. The last dataset, called the Single Individual volunteer for Multiple Observations across Networks (SIMON), is a longitudinal series acquired from one healthy male, scanned over many sessions between the ages of 29 and 46 at multiple centres and with various scanner models (Duchesne et al., 2019a). Each dataset contains FLAIR MRI that were acquired in the axial plane at 3T from General Electric (GE), Philips, and Siemens scanners.

For training and testing, 175 FLAIR MRI volumes (8317 image slices) were sampled from CAIN, ADNI, and CCNA. In total, 125 were sampled from CAIN, 25 from ADNI, and 25 from CCNA while stratifying across scanner vendors and across centres or disease if possible. Patient information and FLAIR imaging parameters for the sampled ADNI, CAIN, and CCNA volumes are listed in Table 1 to demonstrate the diversity of the

Table 1
Summary of the multicentre FLAIR MRI data used for algorithm evaluation. All volumes were acquired at 3T. Values for repetition time (TR), echo time (TE), inversion time (TI), and pixel spacing are represented by the range found in the data. CAIN, ADNI, and CCNA have ground truth delineations for ICV. SIMON is the same subject imaged multiple times.

Patient Information

Database	Disease	No. Volumes	No. Images	No. Patients	No. Centres	Age ± SD (yrs.)	F (%)
ADNI	Dementia	25	917	25	21	73.11 ± 5.30	44
CAIN	Vascular	125	6205	125	9	73.52 ± 8.40	36
CCNA	Dementia	25	1195	25	5	73.12 ± 8.37	56
SIMON	Normal	62	2976	1	12	–	0

Acquisition Parameters

Database	Scanner Vendors	Mag Field (T)	TR (ms)	TE (ms)	TI (ms)	Pixel Spacing (mm)	Slice Thickness (mm)
ADNI	GE, Philips, Siemens	3	9000–11000	90–154	2250–2500	0.8594	5
CAIN	GE, Philips, Siemens	3	9000–11000	117–150	2200–2800	0.4295–1	3
CCNA	GE, Philips, Siemens	3	9000–9840	125–144	2250–2500	0.9375	3
SIMON	GE, Philips, Siemens	3	9000	125	2500	0.9375	3

data. All volumes used for training and testing were accompanied by manual ground truth annotations of the ICV region which were generated by a biomedical student trained by a radiologist using ITK-SNAP[1] (Yushkevich et al., 2006) and Pathcore Sedeen[2]. ADNI volumes were sampled from 21 different centres and at least 3 vol from each ADNI disease classification were included. Specifically, 5 normal, 4 early mild cognitive impairment (EMCI), 7 late mild cognitive impairment (LMCI), 3 subjective memory concern (SMC), and 6 AD volumes were selected. CAIN and CCNA volumes were sampled from 9 to 5 imaging centres respectively, with approximately equal representation of GE, Siemens, and Philips scans.

The SIMON human phantom dataset was used to examine the reliability of ICV measurements. Our subset of 62 SIMON scans was acquired from 12 centres using the Canadian Dementia Imaging Protocol (CDIP) (Duchesne et al., 2019b) over all scanner vendors. See Table 1 for the CDIP imaging parameters.

2.2. Pre-processing

Intensity standardization was performed to remove variability caused by the multicentre effect (Zhong et al., 2012), (Reiche et al., 2019). All volumes were preprocessed using an inhouse pipeline (Reiche et al., 2019), which incorporated noise removal, bias field correction, and intensity normalization. The volumes were denoised using 3 × 3 median filtering followed by homomorphic filtering for bias field correction. Intensity standardization was then achieved through histogram based scaling and peak alignment with the use of an atlas. Image and atlas histograms were first normalized so bin frequencies were given as percentiles. This helped to ensure that smaller images (i.e., 256 × 256) with significantly less pixels had similar magnitudes to larger images (i.e., 560 × 560). In FLAIR histograms, the GM and WM tissues manifest as a single mode in the intensity histogram and the GM/WM peak was robustly detected by searching for the maximum bin count. The detected peaks then underwent alignment by applying a global, linear re-scaling of the entire histogram with a factor derived from the difference between the mode of the atlas and the target image. As shown in (Reiche et al., 2019), the intensity intervals of tissues in 350K FLAIR MRI are more consistent across multicentre data using this approach.

2.3. ICV segmentation algorithms

Two categories of algorithms were evaluated. The first category of algorithms was based on traditional approaches, including image processing and ML. The second category was based on deep learning methods that use CNNs. CNNs model multi-resolution relationships between pixels and use non-linear boundaries for excellent segmentation performance (Hwang et al., 2019), (Guerrero et al., 2018), (Thakur et al., 2020). CNN architectures were selected based on their proven feasibility for medical image segmentation. Apart from Kleesiek's method (section 2.3.3), all the selected CNN architectures utilize 2D convolutions rather than 3D. A recent study comparing 2D and 3D architectures for brain extraction found comparable results between the two configurations (Thakur et al., 2020).

2.3.1. Thresholding

Many traditional algorithms have used thresholding and mathematical morphology to isolate cerebral tissue (Rehman et al., 2020b). In line with these approaches, an intensity thresholding-based approach was investigated for multicentre FLAIR MRI. The intensity standardization applied during preprocessing removed intensity variability in the brain tissue regions across scanning devices and allowed predefined thresholds to isolate the intracranial cavity. Small structures connecting cerebral and non-cerebral regions, and small clusters of isolated pixels were removed through morphological erosion and connected component analysis. Dilation and hole filling were applied during post-processing to refine the initial segmentations.

2.3.2. Random forest

The method presented in (Khademi et al., 2020) is an ICV segmentation algorithm based on intensity standardization (Reiche et al., 2019), hand-crafted feature design, and a random forest classifier (RFC) for multicentre FLAIR MRI. An intuitive and interpretive feature set was designed, which included variants of local intensities, gradient magnitudes, and Gaussian derivatives. A novel sampling strategy was employed to include training samples from regions that are difficult to classify. Specifically, 75% of training pixels laid within 10 mm of the ICV border where there is feature overlap between brain and surrounding dura. Since the intensity ranges were standardized, local features were consistent across patients. An RFC was selected over other ML classifiers because of their ability to avoid overfitting (Khademi et al., 2020). In total, 200 individual learners were used, and the number of features analyzed at each node was set to 2 to reduce correlation between trees. RFC outputs were refined using the morphological operations described in section 2.3.1.

2.3.3. Kleesiek

The work by Kleesiek in (Kleesiek et al., 2016) is one of the first CNNs for brain extraction. This 3D architecture contains 8 fully convolutional layers consisting of 5 × 5 kernels. The outputs of certain layers are down-sampled via max-pooling to increase the receptive field and

[1] www.itksnap.org.
[2] https://pathcore.com/sedeen/.

introduce translation invariance. Learning is achieved by minimizing the Kullback-Leibler divergence loss function (Kleesiek et al., 2016). The authors explored 3D architectures because of their ability to automatically learn representative features of the brain while accounting for its volumetric nature (continuity between individual MRI slices). This algorithm was designed to operate on any combination of MRI modalities and was evaluated using T1 and multi-modal inputs. In this work, the containerized GitHub version[3]of this method was used with default parameters (i.e., optimizer, learning rate, number of training iterations).

2.3.4. *FCN8*

CNNs for classification can be reconfigured for semantic segmentation by replacing the final, fully connected layers with convolutions. This yields fully convolutional networks (FCN) and was first introduced in (Long et al., 2015). Since then, FCNs have been used for medical image segmentation tasks such as liver delineation in computed tomography (Ben-Cohen et al., 2016). Like the liver, the brain is a gross anatomical structure and may be suitable for segmentation by FCNs. In this work, the architecture in (Long et al., 2015) that uses a VGG 16-layer (VGG16) classifier base and upsamples stride 8 predictions back to pixels was used. This architecture, known as FCN8, uses skip connections to combine the final prediction layer with the 8x upsampled lower layer. By combining lower, fine layers with higher, course layers the network can generate locally detailed segmentations that consider global structure. We modified the work in (Long et al., 2015) by removing all VGG16 dropout layers since no overfitting occurred without them.

2.3.5. *U-Net*

Of the 8 CNN-based methods that were investigated, 4 represented the U-Net (Ronneberger et al., 2015) and some if its variations. 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 (Hwang et al., 2019), (Thakur et al., 2020), (Ronneberger et al., 2015). This architecture is composed of 2 paths: an encoder and a decoder. The encoding path contains units of convolutional and max pooling layers that perform feature extraction. The decoding path contains units of convolutional and transposed convolutional layers that are used in combination with skip connections to recapture the spatial context of images (Long et al., 2015), (Ronneberger et al., 2015). U-Net contains 5 levels where the filter depth is doubled during each encoding “descension” (via max pooling) and halved during each decoding “ascension” (via transposed convolution). In this work, the U-Net and U-Net variants (sections 2.3.6–2.3.8) were implemented with batch normalization layers succeeding convolutional layers. Batch normalization automatically standardizes layer inputs and stores relevant statistics so that they may be updated and applied to future inputs (Ioffe and Szegedy, 2015). This accelerates convergence and improves generalization via a modest regularization effect. In the original U-Net implementation, the initial filter depth was 64, but in this work, it was set to 32 for all U-Net variants. This greatly reduced the number of parameters without diminishing performance. The structure of our U-Net encoding and decoding units are shown in Fig. 1.

2.3.6. *SC U-Net*

The first U-Net variant we included was the skip connection U-Net (SC U-Net) proposed in (Wu et al., 2019). Skip connections are additional paths between the shallow and deep layers of a CNN architecture. SC U-Net includes all skip connections from the original U-Net while adding 4 additional connections between the up-convolution and down-convolution layers. Specifically, the authors added the outputs from each max-pooling layer in the encoder to the inputs for each transposed convolution layer in the decoder. These additional skip connections were included to 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). We deviated from the SC U-Net implementation shown in (Wu et al., 2019) by including batch normalization after all convolutional layers.

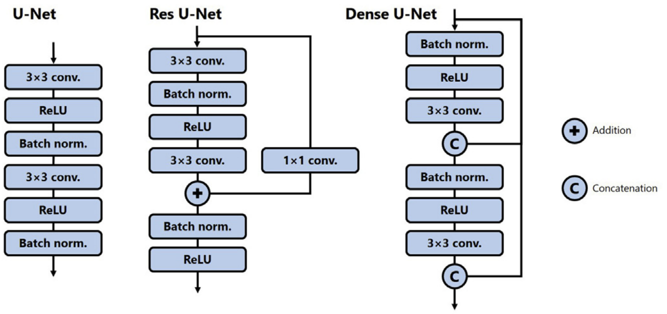

Fig. 1. Encoding/decoding units for U-Net, Res U-Net, and Dense U-Net. Encoding units reside between max pooling layers and decoding units reside between transposed convolutional layers.

2.3.7. *Res U-Net*

Res U-Net utilizes residual connections; a type of skip connection that has demonstrated strong performance on image recognition tasks (He et al., 2016a). Using identity mappings and after addition activations, residual connections enable direct information propagation between network layers (He et al., 2016a). This eases optimization in deep architectures without drastically increasing computational complexity. Networks with residual connections have been used to accurately segment lesions (Guerrero et al., 2018), healthy brain tissues (Chen et al., 2018), and ICV (Kolařík et al., 2019). Our Res U-Net was implemented by replacing the encoder and decoder units in our U-Net with units containing residual connections. See Fig. 1 for the Res U-Net units used in this work, which are based on the “batch normalization after addition” setup presented in (He et al., 2016b). The 1 × 1 convolutional layer is included to change the input filter depth and make addition possible.

2.3.8. *Dense U-Net*

Dense U-Net utilizes dense connections; a design feature that connects each layer to every other layer in a feed-forward manner. Granting all parts of the network direct access to input and loss function gradients improves information flow and eases training (Huang et al., 2017). This boosts performance by avoiding the learning of redundant features and providing regularization to reduce overfitting (Li et al., 2018). Dense U-Net models have been used to segment gross anatomical structures such as the liver in computed tomography (Li et al., 2018) and ICV in MRI (Kolařík et al., 2019). Following these approaches, we replaced the encoding and decoding units from our U-Net with units containing dense connections. The configuration of our Dense U-Net units can be seen in Fig. 1. Confining the dense connections to encoding and decoding units in this manner avoided feature map explosion and the loss of too much low-level information (Jégou et al., 2017).

2.3.9. *CompNet*

The complementary segmentation network (CompNet) was proposed in (Dey and Hong, 2018) as a robust solution to ICV segmentation in T1 MRI with prominent pathology. This architecture has a segmentation branch for learning features related to the ICV region and a complementary branch for learning features of the non-cerebral region. Cerebral pathology such as WML, stroke, and tumours are known to disrupt ICV segmentation. Since non-cerebral regions are less afflicted by pathology (and are therefore more consistent), learning features from the non-cerebral regions can be used to improve the quality of ICV segmentations. An additional branch reconstructs the image using the results from segmentation and complementary branches which are both

[3] https://github.com/GUR9000/Deep_MRI_brain_extraction.

U-Net architectures. This network learns by maximizing the negative Dice coefficient between the predicted brain mask (BM) and its ground truth (GT), minimizing the Dice coefficient between the predicted complementary mask (CM) of the non-cerebral tissues and GT, and minimizing the mean squared error (MSE) between the reconstructed image (R) and the input image (I) as in:

$$CompNet\ Loss = -Dice(BM, GT) + Dice(CM, GT) + MSE(R, I)$$

where the loss corresponding to the reconstruction branch guides learning as a feedback mechanism which expects reasonable outputs from the other 2 branches. In this work, we used an optimal variant of the CompNet which includes dense connections (Dey and Hong, 2018). Since no overfitting was observed without them, we altered the original implementation by removing all dropout and regularization layers.

2.3.10. 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 that the non-linear operations within these modified skip connections would reduce the semantic gap between shallow encoder and deep decoder features. We altered the “MultiRes” block from (Ibtehaz and Rahman, 2020) to follow the “batch normalization after addition” layer structure (He et al., 2016b). Fig. 2 shows the “MultiRes” units used in this work.

2.4. Post-processing

The traditional ICV segmentation methods required morphological post-processing to refine ICV masks after thresholding and classification. Such refinement was not considered to be a post-processing technique as it is a necessary component in each of the traditional algorithms. In contrast to these approaches, the ICV masks generated by the CNN-based methods did not require post-processing to generate high quality segmentations. Thus, only connected component analysis and automated, morphological hole filling were applied to obtain the best possible ICV masks.

2.5. Evaluation metrics

Evaluation metrics that are commonly used to quantify segmentation performance in medical imaging were used. Firstly, the Dice similarity coefficient (DSC) (Wu et al., 2019) was used which measures the spatial overlap between the automatically predicted brain mask (BM) and corresponding manual ground truth (GT):

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

where the DSC ranges between 0 and 1, and a value of 1 implies perfect overlap.

Hausdorff distance (HD) is a metric that compliments DSC by accounting for local differences between the predicted brain mask and ground truth. HD measures the distance between two subsets of points in a metric space and is mathematically defined as:

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

where smaller distances imply a greater degree of similarity. To improve robustness and reduce sensitivity to noisy segmentations, the 95th percentile HD was used.

To investigate the false positive rate, 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. If there are no false positives detected, EF will be equal to 0. Over-segmentation of cerebral tissues (i.e., inclusion of skull tissue, ocular orbits) would result in higher EF rates.

Average volume difference (AVD) quantifies the difference between the ground truth ICV (V_{GT}) and predicted ICV (V_{BM}). AVD is mathematically defined as:

$$AVD = \frac{|(V_{GT} - V_{BM})|}{V_{GT}}$$

where brain volumes are computed in millilitres and a smaller AVD implies a better segmentation. Additionally, Bland-Altman (B-A) plots were used to measure the agreement between the manual ground truth ICV volumes and the automatically predicted ICV volumes. By plotting difference against mean, B-A plots allowed assessment of the measurement bias associated with each method.

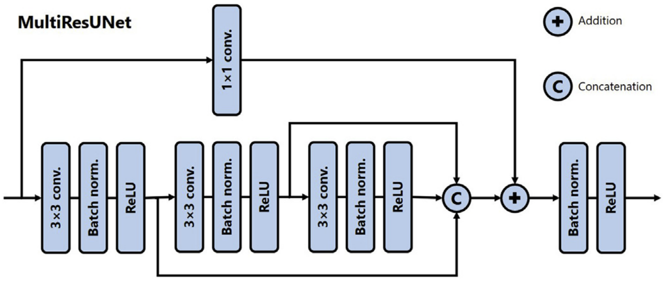

Fig. 2. Encoding/decoding unit for MultiResUNet. Encoding units reside between max pooling layers while decoding units reside between transposed convolutional layers.

2.6. Experimental design

In this work, we considered several performance dimensions that represent a novel framework for evaluating the effectiveness of automated tools for clinical application. For a tool to be widely adopted, it must be accurate, generalize to new datasets and scanners, robust to challenges in data (such as pathology or spatial location), and produce reliable volume measurements. As a result, to analyze the performance of each algorithm for ICV segmentation, four aspects of performance were assessed: (1) accuracy, (2) generalization capabilities, (3) robustness to pathology and spatial location, and (4) volumetric measurement reliability. The experiments corresponding to each of these tests are elaborated upon below, along with the data splits that were used to train models. A breakdown of these data splits and the overall data preparation process can be seen in Fig. 3.

2.6.1. Data splits

Data split #1 was a 50/20/30 training/validation/testing ratio of all 175 FLAIR MRI volumes resulting in 85 vol for training, 37 vol for hold-out validation, and 53 vol for testing. Stratified splitting was used so the proportions of CAIN, ADNI, and CCNA volumes in the training, validation, and test sets mirrored that of the overall population. The proportion of each scanner type (GE, Siemens, and Philips) was approximately stratified as well. Data split #1 was designed with two purposes in mind.

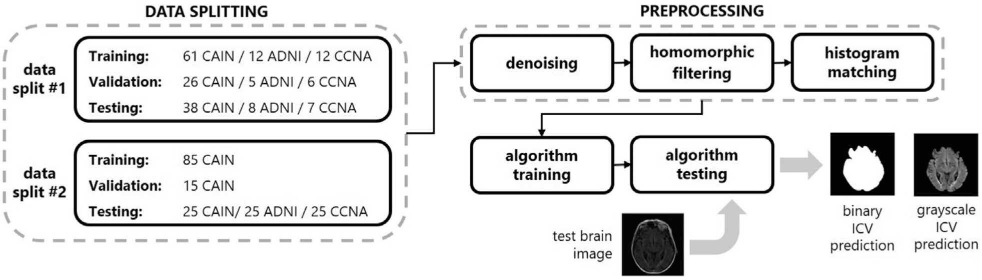

Fig. 3. Summary of the data split strategies and data preparation process prior to algorithm training and testing. All algorithms were trained, validated, and tested using the same subsets of data.

First, in an ideal scenario, training data would be available for each of the datasets being analyzed. Segmentation models exhibit improved performance when tested on data from the same distribution as the training data. Data split #1 was therefore used to establish ideal performance benchmarks. Second, data split #1 ensured that there was broad representation of the cerebral tissues and pathology in the training data. These datasets contain different diseases (dementia, vascular disease), diverse pathology, and acquisition parameters which were used to improve generalization capabilities of the approaches.

Data split #2 was used to mimic real-world models. In usual development scenarios, designers have access to a single dataset which is used to generate ground truths and train models. Upon clinical implementation, the algorithms would then be expected to generalize to new, unseen data. To test this scenario, CAIN data was exclusively used for training and validation since it is comprised of the most ground truths. The remaining 25 ADNI and 25 CCNA volumes, in addition to 25 held-out CAIN volumes, were used for testing. This resulted in 85 vol being used for training, 15 vol being used for hold-out validation, and 75 vol being used for testing. Since CAIN is vascular, ADNI is AD-type dementia, and CCNA is all types of dementia, this data split also evaluated algorithm generalization to differing diseases. This is important as morphological characteristics of the brain can be different depending on the type of risk factors and pathology present. By using CAIN for training, we were also able to match the training set size from data split #1, thus allowing fair analysis and comparison.

2.6.2. Accuracy

As a first step, the average accuracy of all methods was investigated. Since biomarkers can be used to guide clinical decisions and treatment options, accuracy is of utmost importance. Assessments of overall accuracy were conducted using models trained and tested from data splits #1 and #2. Results generated from data split #1 models were emphasized since it represents the ideal setup of increased data diversity. Our tests for accuracy included means and standard deviations of all evaluation metrics, as well as B-A plots between the true and predicted ICV measurements.

2.6.3. Generalization

To maximize translation opportunities, ICV segmentation algorithms should not experience drastic drop-offs in performance when the inputs come from different centres or scanning devices. To assess generalization across vendors, we observed evaluation metrics generated from data split #1 as a function of scanner vendor. To assess generalization to unseen databases, we observed evaluation metrics generated from data split #2 as a function of test database. See Fig. 4 for the variation exhibited by images from different databases and scanner vendors.

2.6.4. Robustness to pathology and spatial location

In this phase, the methods were evaluated in terms of their ability to segment ICV for varying levels of disease and spatial location. Data split #1 was used for all these experiments. Only the traditional (Thresholding, RFC) and the top CNN methods were compared.

Tools should have equal performance (robustness) irrespective of the amount of pathology found in the brain. To test this, evaluation metrics were measured as a function of pathological load. Neurodegenerative diseases are characterized by increased prevalence of WMLs and atrophy (expansion of CSF spaces) (Oppedal et al., 2015), (Ntiri et al., 2020), (Scahill et al., 2003), (Silbert et al., 2003), (Gunter et al., 2003), (Sigurdsson et al., 2012), (Aribisala et al., 2013), (Leung et al., 2013), (Rocca et al., 2017). For an estimate of WML and CSF loads, the standardized volumes were masked using the ICV ground truths and thresholded using the intensity ranges shown in Fig. 5. Standardization enabled thresholds to isolate WML and CSF regions and obtain an estimate of pathological load in millilitres. WML and CSF loads were classified as low, medium, or high using the guidelines described in (de Sitter et al., 2017) and (Squitieri et al., 2009) respectively. Table 2 summarizes the volume ranges that were used to categorize pathological load and the number of test subjects in each category. Fig. 6 shows sample images for low, medium, and high pathological loads classified according to this scheme. Algorithm performance as a function of MoCA score categorization was also observed. A MoCA score of 26 or greater defined cognitively normal patients, while a MoCA score less than 26 defined impaired patients (Nasreddine et al., 2005).

Algorithm robustness across spatial locations was also evaluated as certain anatomical regions are known to cause segmentation challenges. For example, poorer segmentations typically occur in extreme superior and inferior slices where it is difficult to distinguish between brain and surrounding dura. To quantify segmentation performance as a function of spatial location, the DSC was measured for five segments of each test volume, where each segment comprised 20% of the total (contiguous) slices from the ICV mask. This resulted in a single bottom region, three middle regions, and a single top region. To analyze these regions further,

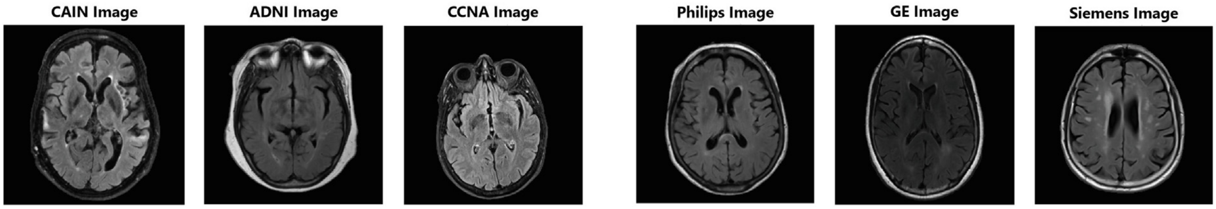

Fig. 4. Sample images showing variation of scans from different datasets and different scanner vendors.

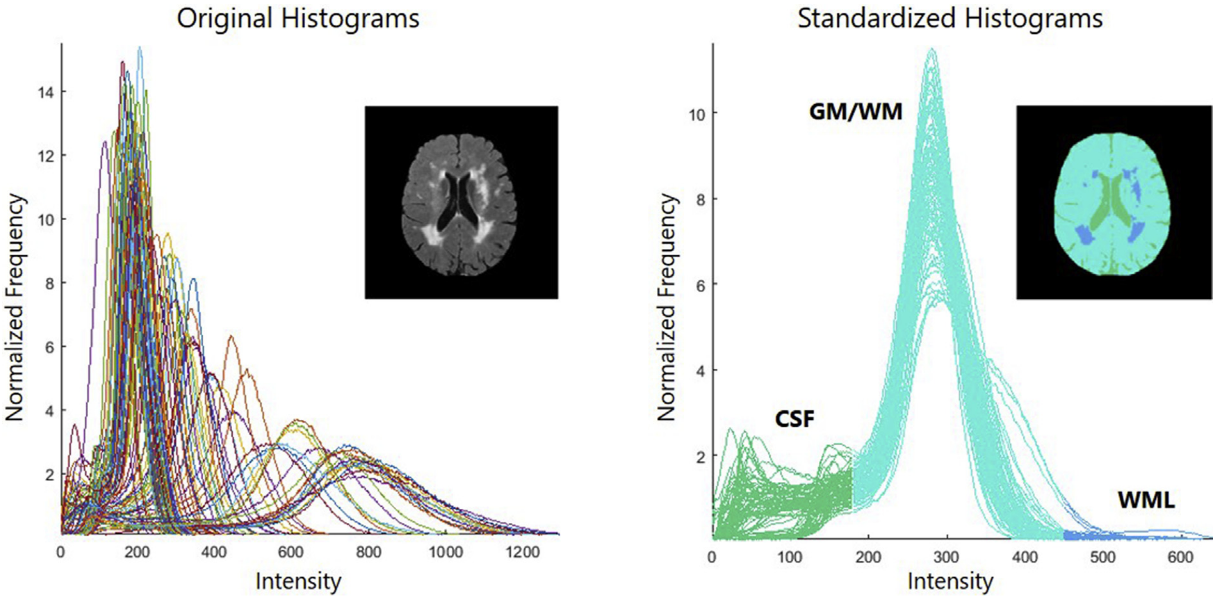

Fig. 5. Volume histograms before and after intensity standardization. The intensity ranges used to estimate different tissue volumes via thresholding are shown by the colours on the standardized histograms. Pixels representing the brain (GM/WM), CSF, and WML are shown as turquoise, green, and blue, respectively.

Table 2
Summary of the estimated volumes used to categorize WML and CSF loads as low, medium, or high. Pathological loads were estimated for the 53 test subjects in data split #1.

Pathology	Categorization	Volume	No. Test Subjects
WML	Low	<5 mL	18
	Medium	5–15 mL	19
	High	>15 mL	16
Atrophy (CSF)	Low	<205 mL	30
	Medium	205–275 mL	14
	High	>275 mL	9

the spatial distribution of FPs and FNs were visualized using error maps. Error maps were generated by registering the volumes, ground truths, and predicted ICV masks to a common spatial reference (Winkler, Kochunov, Glahn) using ANTs (Avants et al., 2011). Within the ANTs toolkit, the diffeomorphic image registration tool, called symmetric normalization (SyN), was used due to its proven efficacy for handling large deformation problems (Avants et al., 2008). Registration converted the volumes to a size of 256 × 256 × 55 with an isotropic voxel resolution of 1.0 × 1.0 × 1.0 mm^3. The error was then be computed as the absolute difference between the ground truths and ICV estimations for all slices. Error map generation was completed by averaging all the individual error maps (per slice) for each algorithm.

2.6.5. Volumetric reliability and reproducibility

The traditional (Thresholding, RFC) and top performing CNN methods from previous analyses were used to generate segmentations on the SIMON dataset. Data split #1 models were used. See Fig. 7 for example SIMON scans from each vendor, which demonstrates how differences in scanning hardware can cause the same brain to appear different despite a uniform protocol. A reliable tool would be able to output reproducible ICV measurements for the same subject despite this variation. Descriptive statistics, including the coefficient of variation (CoV), were used to compare each algorithm's ICV measurements between scanner vendors (see section 2.7 for more details).

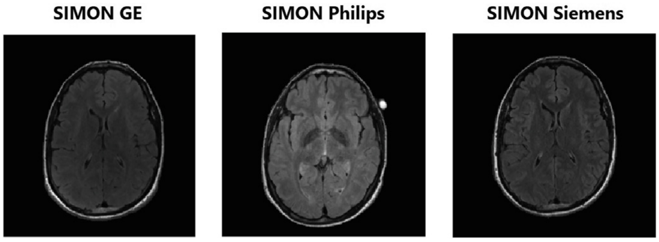

Fig. 7. Slices from SIMON volumes acquired on GE, Philips, and Siemens scanners. Despite all being acquired with the CDIP protocol (Yushkevich et al., 2006), differences in scanning hardware manifest themselves in the output volumes.

2.7. Statistical analysis

To analyze trends in the methods, statistical analysis was performed alongside the evaluation tests. It would be of interest to the clinical community to demonstrate which (if any) of the tools provide similar performance across the proposed dimensions of generalization (scanner types, datasets), pathology (CSF load, WML load), and spatial location (bottom, middle and top slices), and whether they provide reliable ICV measurements in the SIMON human phantom data.

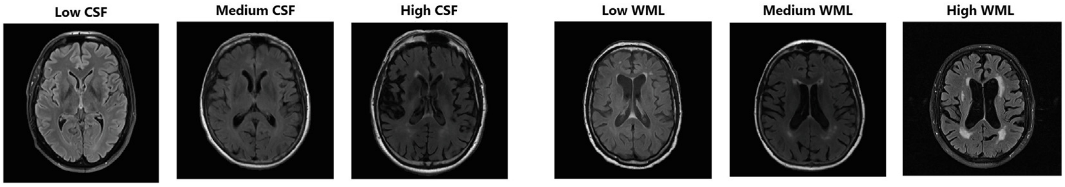

Fig. 6. Sample images showing variation of scans with different CSF loads and different WML loads.

To compare the same method over different groups (i.e., scanner vendor, dataset, etc.) a single evaluation metric was chosen to simplify analysis. DSC was the main evaluation metric selected for statistical analysis as it is the strongest indicator of overall performance in addition to being highly interpretative. For each of the test criteria, the mean DSC was statistically compared across groups using analysis of variance (ANOVA) tests. These tests demonstrate whether an algorithm has similar performance outcome variables (DSC) over the different predictor variables (dataset, scanner, pathology load, spatial location). 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 ICV segmentation methods. The raw DSC values were found to be non-normal and negatively skewed. As a result, the DSC values underwent reflection and a double logarithmic transformation to improve linearity and homogeneity of variance prior to analysis (Dobson and Barnett, 2018). For instances where the results of ANOVA tests were significant, Tukey-Kramer post-hoc adjustment for multiple comparisons were used to determine the sources of difference. Adjusted DSC was used as the primary outcome variable but results for adjusted HD and EF as outcome variables were also generated and can be found in the Appendix.

To assess the reliability of ICV analysis over repeated measurements, the estimated ICV from each method/scanner combination was computed from the human phantom dataset (SIMON) and analyzed. The mean and standard deviation of ICV measurements for each method/scanner combination were used to obtain the corresponding CoVs. A small CoV indicates a more consistent measurement over repeated measures where a CoV of <5% is commonly deemed acceptable (Campbell et al., 2010). Since the number of scans corresponding to each vendor was unbalanced, which can bias these descriptive measurements and limit comparison, the number of samples used to compute CoV from GE, Siemens, and Philips groups was determined from the smallest class size.

3. Results

In this section, the experimental setup and evaluation experiment outcomes will be presented. For 2D CNN methods, image slices were resized to 256 × 256 using bilinear interpolation and the corresponding ICV segmentations were transformed back to native dimensions prior to performance assessment. Data augmentation in the form of moderate shearing, scaling, rotation, and translation was applied to training images for artificial data expansion and to reduce overfitting (Chollet, 2018).

Unless specified otherwise, all CNN-based methods shared common design choices. All convolutional operations used 3 × 3 filters with rectified linear unit (ReLU) activation. The final output layers were configured with a single 1 × 1 filter and sigmoid activation to account for the binary nature of the segmentation task. An Adam optimizer with a learning rate of 0.0001 was used over 50 epochs with a batch size of 16. This number of epochs allowed the models to reach a steady-state validation loss. Final model weights were selected from the epoch that yielded the optimal validation loss. The Dice loss function was used for

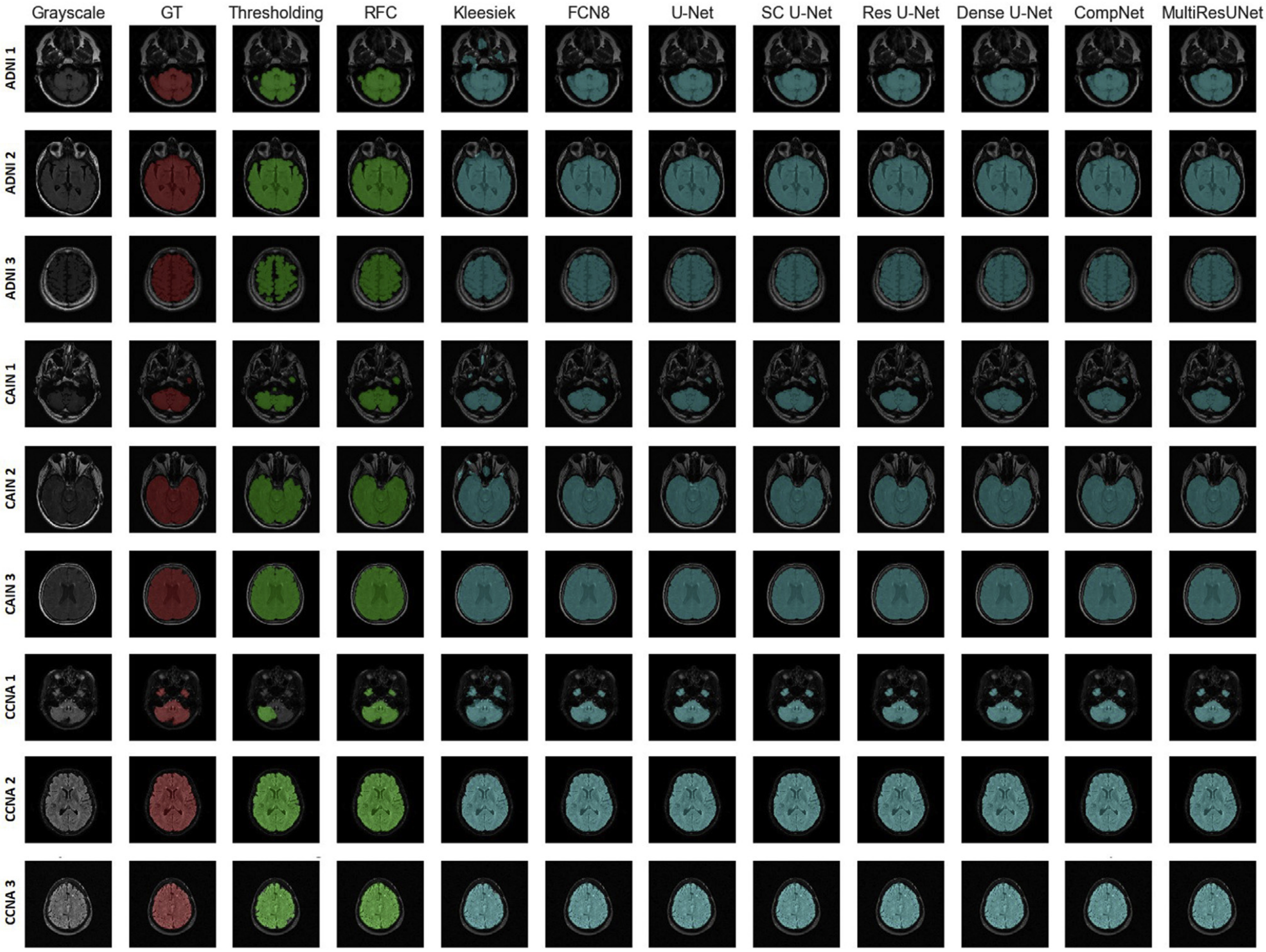

Fig. 8. Sample ICV segmentations from data split #1 models. The top rows are 3 different ADNI volumes, the middle rows are 3 different CAIN volumes, and the bottom rows are 3 different CCNA volumes. Red overlays show the ground truth (GT) annotations, green overlays show traditional method outputs, and turquoise overlays show deep learning method outputs.

most architectures. If $GT = \{gt_1, \cdots, gt_N\}$ are the ground truth masks over N slices and $BM = \{bm_1, \cdots, bm_N\}$ are the corresponding predicted brain masks, the Dice loss function can be mathematically defined as:

$$Dice(bm_n, gt_n) = \frac{\sum_{n=1}^{N} |bm_n \circ gt_n| + s}{\sum_{n=1}^{N} |bm_n| + |gt_n| + s}$$

where s is a smoothing factor used to avoid division by 0. Only Kleesiek (see section 2.3.3) and CompNet (see section 2.3.9) had differing loss functions and training setups. Due to its computational expense, CompNet was trained in (Dey and Hong, 2018) using an Adam optimizer with a learning rate of 0.001 over only 10 epochs with a batch size of 4. In this work, we use the same settings, but training ran for 15 epochs.

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).

3.1. Accuracy

Fig. 8 shows sample ICV segmentations generated by data split #1 models. The Thresholding method operated well in regions with large, constant intensities (i.e., middle slices) but missed brain tissue in upper and lower slices. The RFC method improved on the Thresholding method with some false positives and negatives being noted. In terms of the deep learning methods, FCN8, U-Net, SC U-Net, Res U-Net, Dense U-Net, CompNet, and MultiResUNet generated high quality segmentations over all datasets and spatial locations. For these methods, the ICV region was clearly delineated with smooth boundaries. There were some false positives noted in the Kleesiek method.

We first compared the ICV segmentation accuracy of the methods using the average evaluation metrics shown in Table 3. With respect to data split #1, MultiResUNet obtained the best overall DSC (98.12%), HD (1.44 mm), and EF (1.60%) while SC U-Net obtained the best AVD (1.20%). U-Net and SC U-Net were top 3 performers for at least 3 metrics each. Similar results were observed for data split #2 where MultiResUNet obtained the best overall DSC and HD. Evaluation metric distributions for data split #1 and #2 can be seen in Fig. 9 and Fig. 10, respectively. Each of the 2D CNNs had better medians, lower variance, and fewer outliers compared to the traditional methods. The worst performing algorithm in terms of DSC for data split #1 and #2 was Kleesiek, followed by Thresholding and RFC. These algorithms also exhibited high variability, indicating less predictable performance.

We also generated B-A plots between ground truth ICV and predicted ICV (in mL) as seen in Fig. 11. Considering data split #1, all 2D CNN methods showed measurement biases under 20 mL with U-Net, MultiResUNet, SC U-Net having the narrowest limits of agreement. The Thresholding, RFC, and Kleesiek methods had considerably larger biases (exceeding 80 mL) and broader limits of agreement, which suggests that they are less suitable proxies for manual segmentation. Similar trends were observed for data split #2 where U-Net, MultiResUNet, and SC U-Net were among top performers in respect to both bias and limits of agreement.

3.2. Generalization

To analyze method performance across scanner vendors, data split #1 was utilized. Data split #1 had testing data from the same distribution as the training data and enabled analysis of scanner effect on ICV segmentation while minimizing dataset effects. The data split #1 test set contained 17 GE scans, 19 Philips scans, and 17 S scans. Fig. 12 shows the DSC distributions as a function of scanner vendor. As can be seen, across the deep learning methods (except for Kleesiek), there was high accuracy, low variance, and good consistency in DSC across scanner vendors. From the traditional methods, Thresholding exhibited wider variability and lower DSCs compared to the RFC. Except for RFC, all methods had significant ANOVA models ($p < 0.001$) indicating differences in algorithm performance between vendors (see Table A in Appendix). Post-hoc testing revealed that the source of most differences ($p < 0.05$) was performance between GE versus Philips scans and Siemens versus Philips scans. Performance between GE and Siemens scans typically did not differ ($p > 0.05$). Although there was a statistical difference between scanners, it is noted by Fig. 12 that deep learning methods still had high accuracy across vendors.

To analyze generalization across datasets, data split #2 was utilized. Recall that this split had training data from a single dataset (CAIN) and test volumes from ADNI, CCNA, and CAIN to compare how methods would generalize to datasets within and outside their training distribution. This analysis also tested robustness across diseases since the source and target datasets were different (vascular and dementia disease). Fig. 13 shows the DSC for every method as a function of dataset. As can be seen, 2D CNNs performed slightly better on CAIN volumes but maintained high performance over all datasets. The traditional and Kleesiek methods showed much more variability across datasets and performed noticeably worse on ADNI brains. This indicates that these methods are less capable of generalizing to the differing AD related pathology. The null hypothesis was rejected ($p < 0.001$) for ANOVA models of all methods indicating that test dataset influenced algorithm performance (see Table B in the Appendix). Post-hoc analysis revealed that some methods (FCN8, SC U-Net, Dense U-Net, CompNet) had no performance difference ($p > 0.05$) between ADNI and CCNA volumes. All methods reported significant performance differences ($p < 0.05$) when CAIN was one of the groups. This indicates that the methods learned features related to vascular diseased brains that differ from those in the dementia datasets. Despite this, the 2D CNNs never report DSCs below 95% on the unseen data distributions indicating strong generalization potential. Only Thresholding did not report differences between ADNI and CAIN which can likely be attributed to the unsupervised nature of the algorithm.

Table 3
Average evaluation metrics across all ICV 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 values show the best result.

	DSC (%) ↑		HD (mm) ↓		EF (%) ↓		AVD (%) ↓	
	data split #1	data split #2	data split #1	data split #2	data split #1	data split #2	data split #1	data split #2
Thresh	91.72 ± 3.18	91.03 ± 3.79	9.35 ± 3.54	8.79 ± 3.76	3.92 ± 3.83	3.24 ± 2.63	11.27 ± 6.36	11.40 ± 7.51
RFC	94.79 ± 2.09	93.77 ± 3.09	4.38 ± 2.14	5.12 ± 2.76	8.38 ± 5.17	9.04 ± 7.53	6.35 ± 5.53	7.12 ± 6.93
Kleesiek	89.72 ± 3.35	88.81 ± 4.39	7.73 ± 3.87	8.33 ± 3.77	14.98 ± 10.94	10.74 ± 6.53	11.04 ± 12.38	9.16 ± 6.99
FCN8	97.90 ± 0.70	97.50 ± 0.89	1.68 ± 1.17	1.62 ± 0.81	2.83 ± 1.91	1.80 ± 0.89	1.86 ± 2.18	1.92 ± 1.60
U-Net	98.08 ± 0.47	97.70 ± 0.86	1.54 ± 0.82	1.57 ± 1.06	1.85 ± 1.02	1.52 ± 0.86	1.21 ± 1.08	1.87 ± 1.43
SC U-Net	98.07 ± 0.48	97.66 ± 1.04	1.59 ± 0.90	1.55 ± 0.56	1.83 ± 1.09	1.88 ± 1.04	**1.20 ± 1.15**	1.62 ± 1.30
Res U-Net	98.05 ± 0.43	97.71 ± 0.95	1.51 ± 0.79	1.50 ± 0.55	1.96 ± 1.11	1.72 ± 0.95	1.36 ± 1.11	1.65 ± 1.29
Dense U-Net	98.05 ± 0.42	97.52 ± 0.91	1.57 ± 0.92	1.61 ± 0.72	2.07 ± 1.10	**1.36 ± 0.91**	1.29 ± 1.08	2.42 ± 1.83
CompNet	97.72 ± 0.65	97.58 ± 1.18	1.82 ± 1.27	1.62 ± 0.78	3.05 ± 1.61	2.27 ± 1.18	1.90 ± 1.66	**1.56 ± 1.20**
MultiResUNet	**98.12 ± 0.41**	**97.76 ± 0.95**	**1.44 ± 0.68**	**1.44 ± 0.60**	**1.60 ± 0.90**	1.72 ± 0.95	1.28 ± 1.15	1.64 ± 1.24

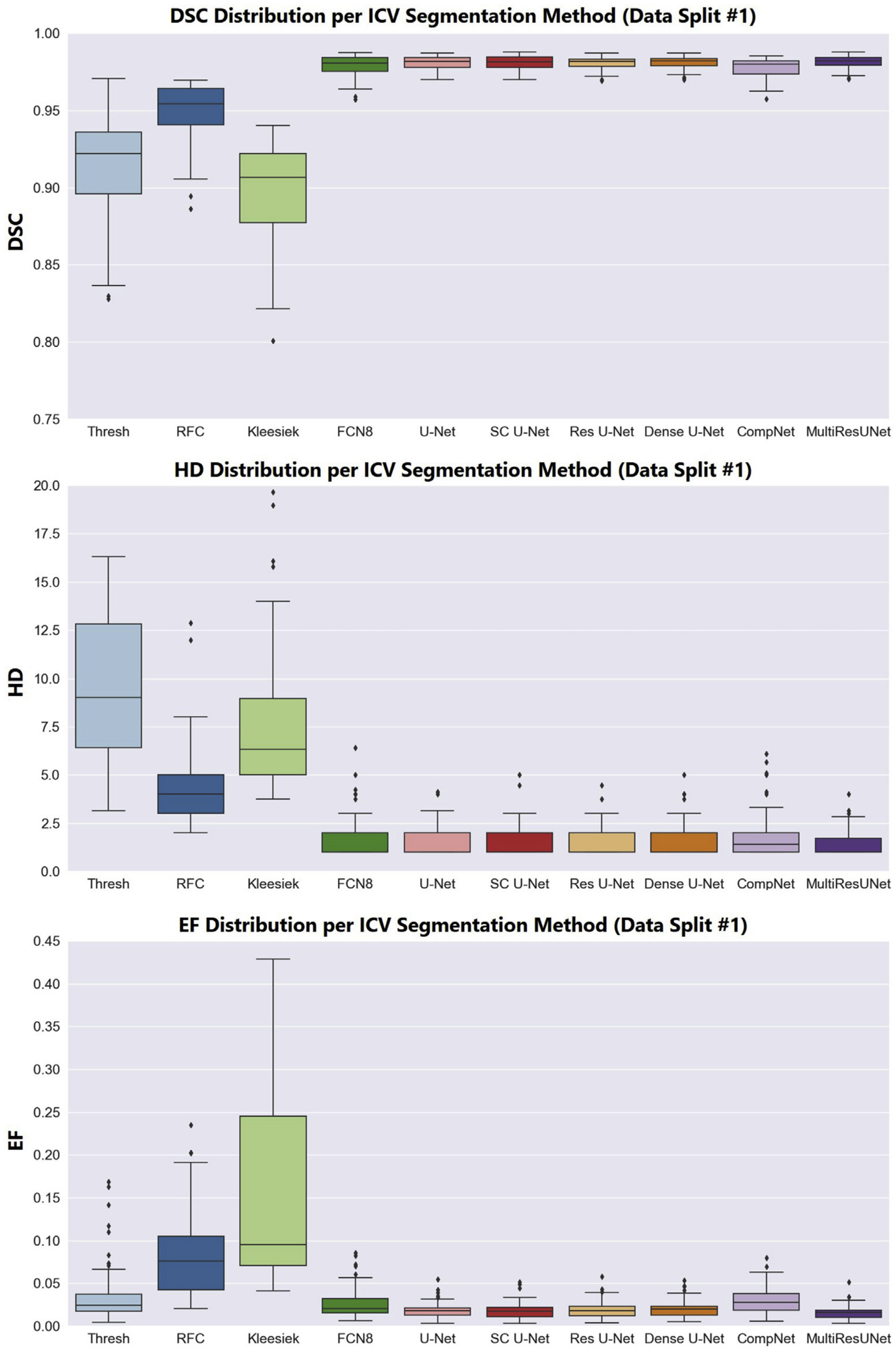

Fig. 9. DSC, HD, and EF distributions across all ICV segmentation methods for data split #1.

3.3. Robustness to pathology and spatial location

In this section, algorithm robustness to pathology and spatial location was analyzed. Comparison was limited to the top three performing deep learning methods according to accuracy (U-Net, SC U-Net, MultiResUNet) and the traditional methods (Thresholding, RFC). This enabled us to highlight robustness differences between traditional and deep learning methods and further assess which CNN is best suited for neurodegenerative population ICV segmentation. To ensure vascular and dementia causing pathologies were both accounted for, we used models trained from data split #1.

DSC distributions as a function of WML load, CSF load, and MoCA categorization are shown in Fig. 14. Of the 53 test volumes in data split #1, 24 were categorized as normal, 22 were categorized as impaired, and 7 did not have MoCA scores and were omitted. The 2D CNN methods had higher DSCs with lower variance over all WML and CSF loads compared to the traditional methods. Regarding performance on MoCA categorizations, the traditional algorithms performed worse on the impaired

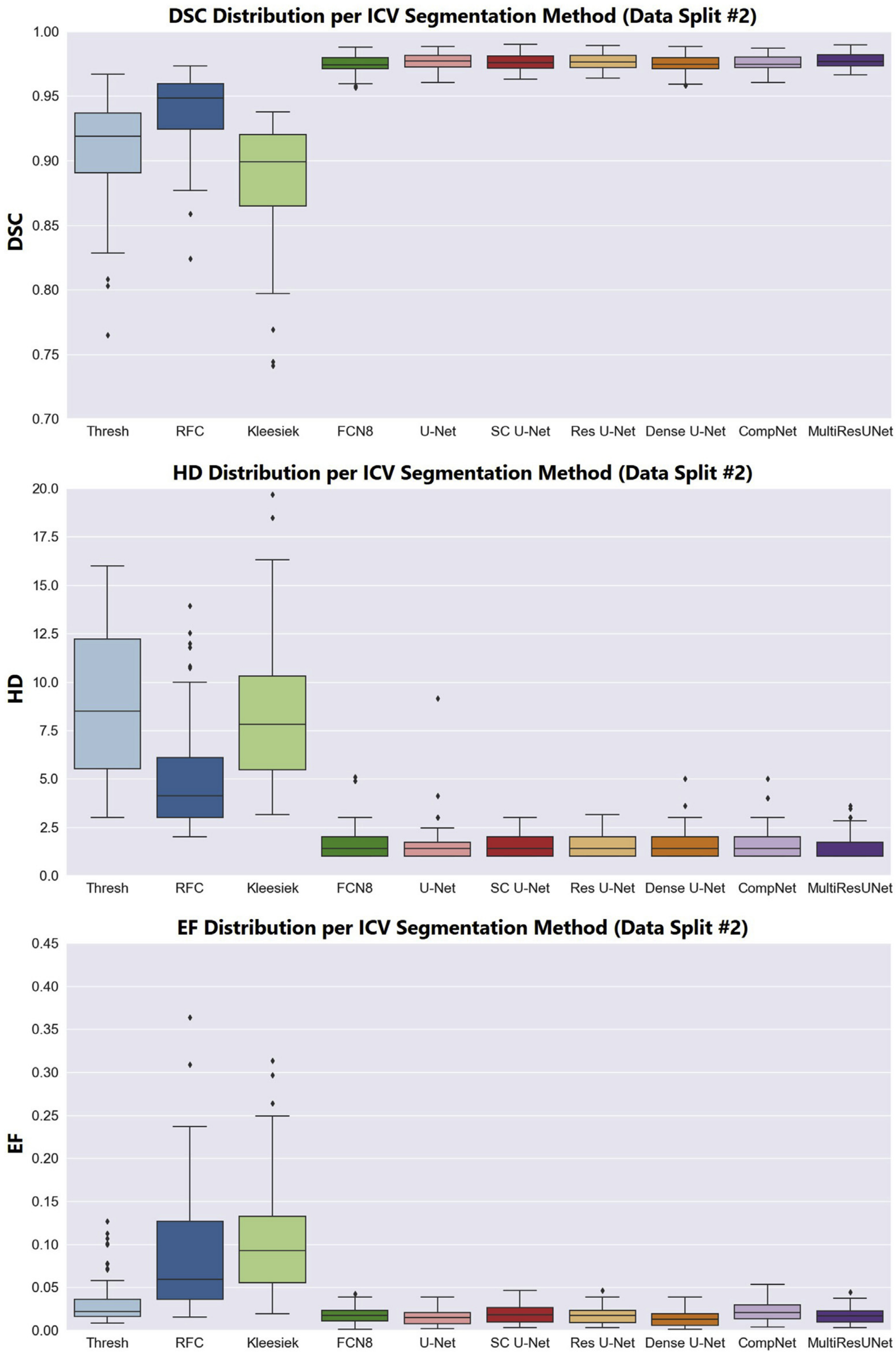

Fig. 10. DSC, HD, and EF distributions across all ICV segmentation methods for data split #2.

volumes, which likely contain more prevalent pathology, while the CNNs maintained performance with a small increase in variance on impaired volumes. The ANOVA models for 2D CNN methods reported no performance differences across all WML and CSF loads (see Tables C and D in the Appendix). Differences ($p < 0.001$) were only found for the traditional methods between WML loads. Post-hoc testing revealed differences ($p < 0.05$) between low versus medium WML loads for Thresholding and medium versus high WML loads for RFC. To further understand these differences, sample segmentations on 3 challenging cases (1 per database) are shown in Fig. 15. These cases were selected because of the large lesions and prominent atrophy. On the CAIN CVD case, traditional methods failed in proximity to the large lesion. On the CCNA MCI case, traditional methods generated many false negatives in proximity to the lesions and enlarged ventricles. On the ADNI AD patient, with atrophy characterized by a prominent central fissure, Thresholding omitted much of the CSF while RFC had difficulty discerning the ICV border. In contrast, segmentations from the 2D CNNs were smooth and accurately delineated the ICV region irrespective of atrophy or lesions.

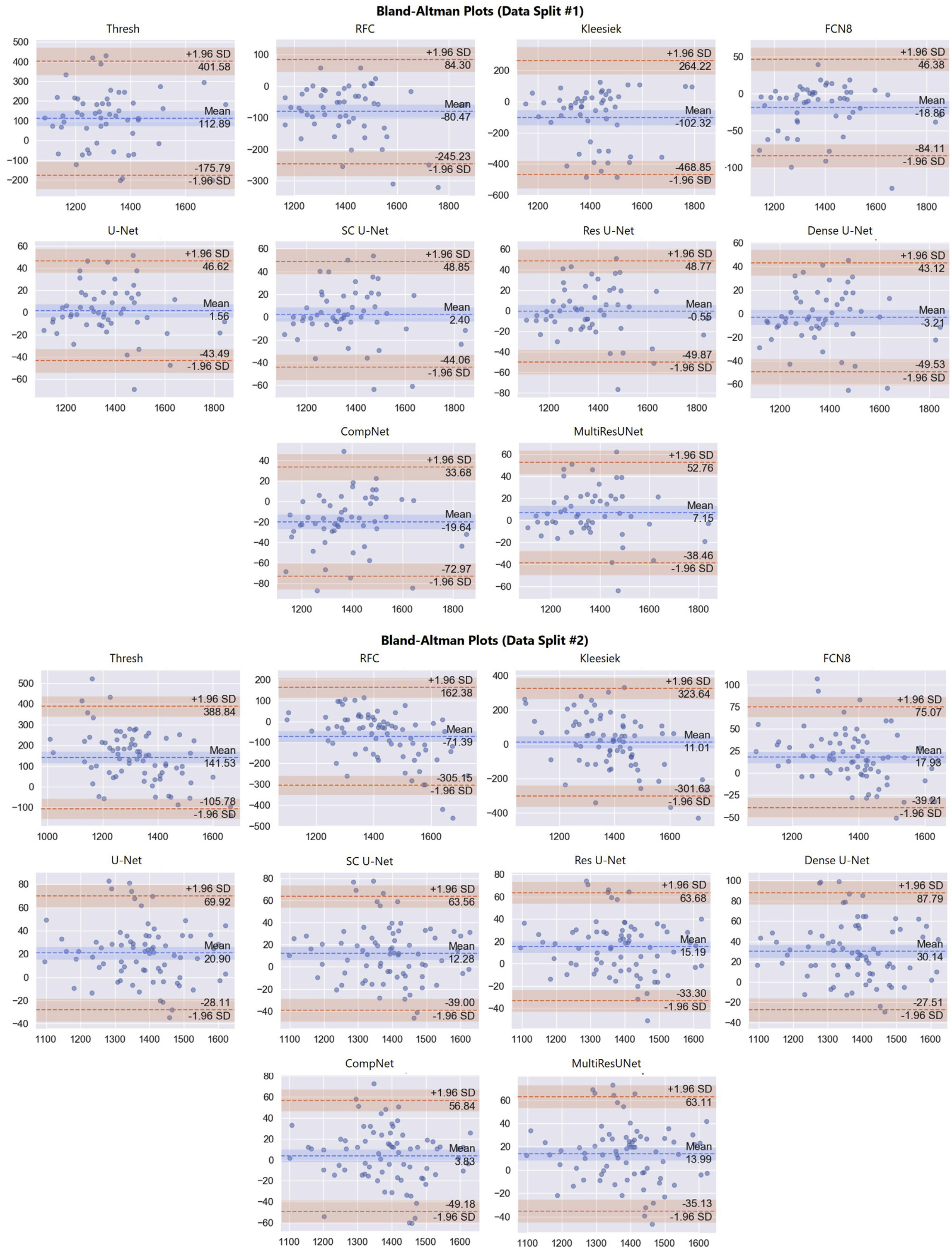

Fig. 11. B-A plots between manual ground truth and algorithm predicted ICV in millilitres for data split #1 and data split #2. Mean of methods is given on the x-axis and difference between methods is given on the y-axis.

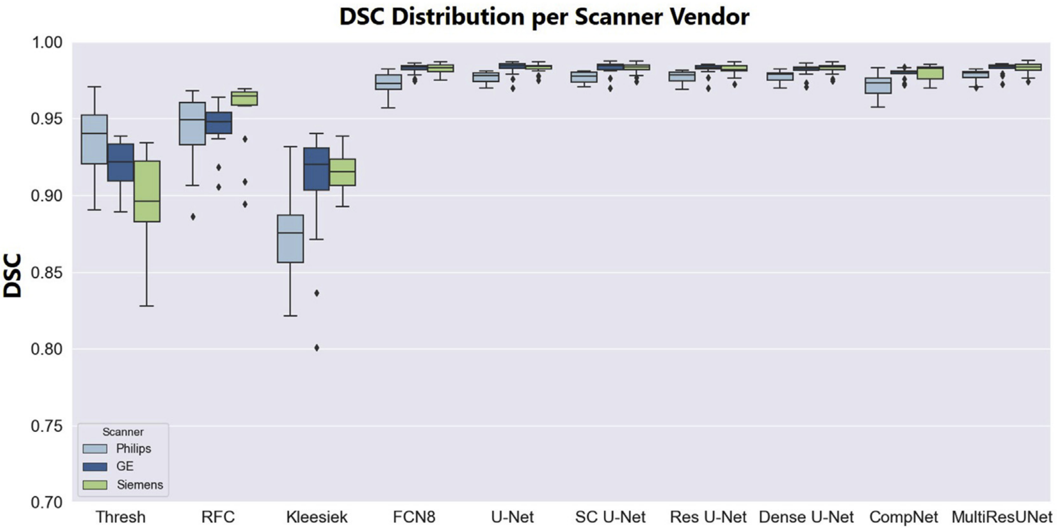

Fig. 12. DSC distributions across all ICV segmentation methods for data split #1 as a function of scanner vendor.

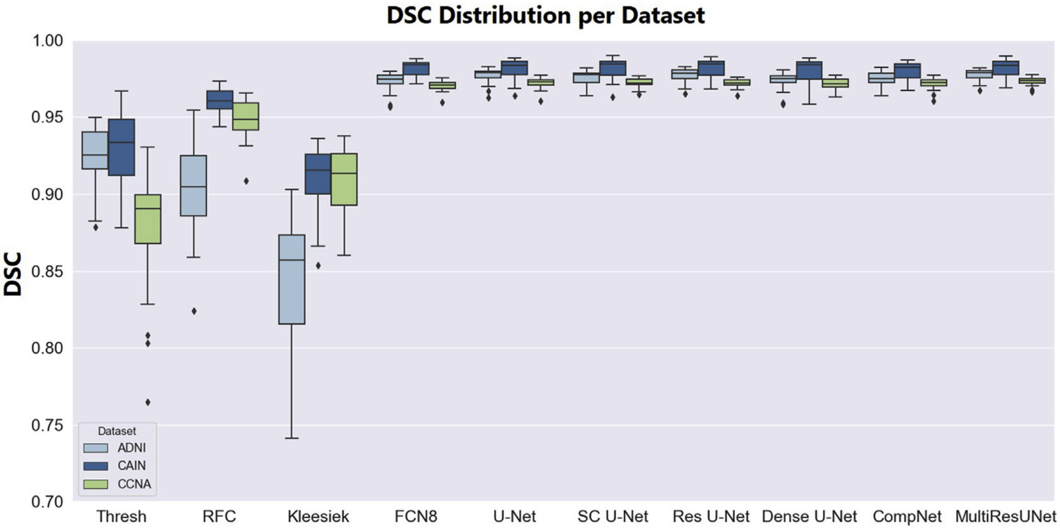

Fig. 13. DSC distributions across all ICV segmentation methods for data split #2 as a function of dataset.

Deviations were mainly exhibited by 2D CNNs when prominent pathology existed at the ICV border. For example, there were some false negatives along the periphery of the large CVD lesion and around the areas of extreme atrophy in the MCI patient.

The average regional DSCs for the traditional and top 2D CNN methods are shown in Fig. 16. Central regions of the brain are typically easier sites for automated segmentation as there is clear morphological separation between the brain and surrounding skull. All algorithms achieved comparable performance (average DSCs of at least 90% with low standard deviation) on middle regions. Superior and inferior slices present a challenge as there is increased tissue diversity and potential feature overlap with the brain. Performance on top and bottom slices varied across the methods. RFC for example, had similar performance on bottom and top slices, but there was more than a 10% drop in DSC compared to the middle slices. CNN methods by contrast had much more performance similarity across spatial locations. For the bottom 20% of slices, U-Net had a best DSC of 93.99% followed by MultiResUNet (93.93%) and SC U-Net (93.75%). For the top 20% of slices, U-Net also had a best DSC of 93.70% followed by SC U-Net (92.72%) and MultiResUNet (92.51%). Over all methods, ANOVA models rejected the null hypothesis ($p < 0.001$) indicating a difference in performance across space and anatomy (see Table E in the Appendix). Post-hoc analysis revealed there were performance differences ($p < 0.05$) between the middle versus top or bottom regions, and no performance differences ($p > 0.05$) between the top versus bottom regions as well as middle region 2 versus middle region 3 for all methods. To visualize segmentation error across spatial locations, error maps were generated with a few examples being shown in Fig. 17. For central slices, error maps were comparable aside from the CNNs providing better delineation of the central fissure and omitting ocular orbits. Thresholding exhibited a pattern of undersegmentation while RFC exhibited a pattern of oversegmentation (particularly in inferior slices where the ICV border is difficult to discern). Errors for the 2D CNNs were mainly concentrated in posterior regions, around the anterior sinuses, and at the ICV border.

3.4. *Volumetric reliability and reproducibility*

SIMON volumes came from the same, healthy subject and corresponding ICV measurements should not be significantly different across scans. In total, the same subject was scanned 62 times at 12 different centres (41 S, 14 Philips, and 7 GE). Data split #1 models of the traditional (Thresholding, RFC) and top performing CNN methods (U-Net, SC

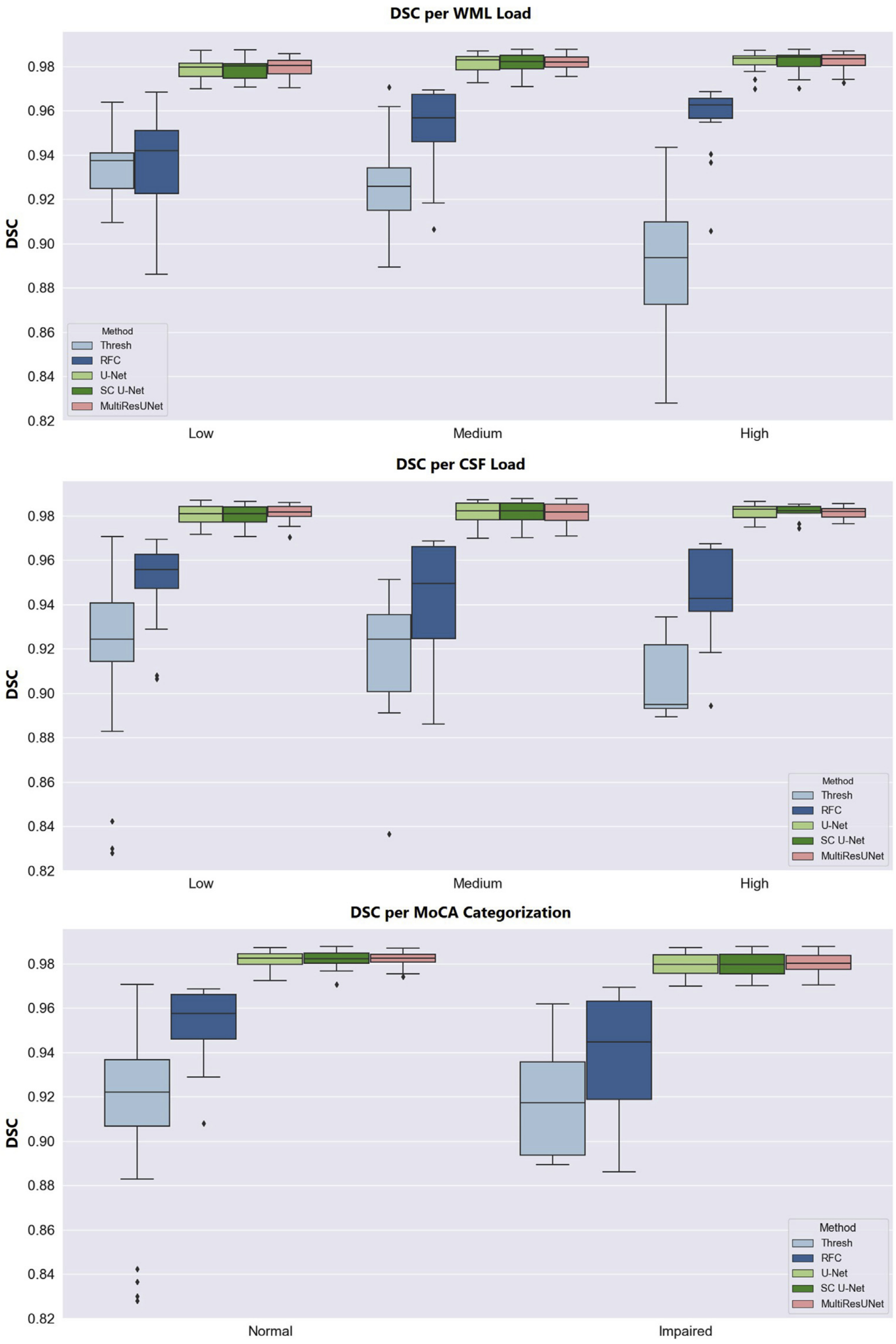

Fig. 14. DSC distributions across select ICV segmentation methods for data split #1 as a function of WML load, CSF load, and MoCA categorization.

U-Net, MultiResUNet) were used to generate ICV masks on the SIMON data. Table 4 shows the average SIMON ICV measurements organized by method and scanner vendor. Each method/scanner combination outputted ICV measurements hovering around 1400 mL which is in line with previous works measuring average ICV in healthy populations (Aribisala et al., 2013). Fig. 18 shows the SIMON ICV measurement distributions for each method/scanner combination. As can be seen, traditional methods not only had high variability, but generated notably different ICV measurements across scanner vendors. ICV measurements from the CNN methods had low variability and were more aligned across scanners. Siemens scans likely exhibited greater variability due to the larger sample size reflecting the true spread in the ICV measurements.

7 Philips scans and 7 S scans were randomly sampled prior to computing the CoV to balance the number of samples between scanner groups. The CoV values computed for each method/scanner combination on the balanced dataset are shown in Table 4. All method/scanner combinations, except for Thresholding/GE and Thresholding/Siemens, returned CoV values lower than 3% indicating a high degree of segmentation reliability overall. MultiResUNet yielded the lowest CoV for Siemens and Philips scans while SC U-Net yielded the lowest CoV for GE

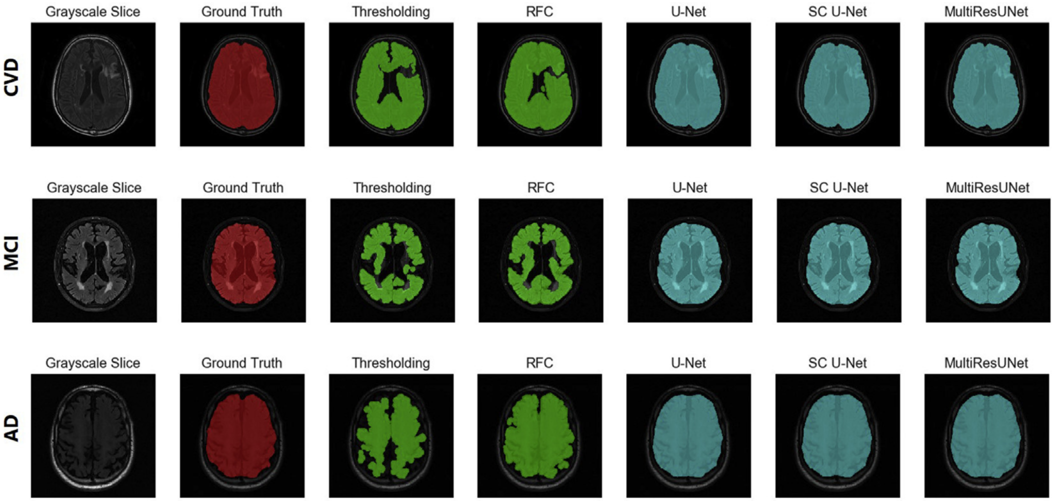

Fig. 15. Sample segmentations for challenging cases across select ICV segmentation methods. Red overlays show ground truth delineations, green overlays show traditional algorithm predictions, and turquoise overlays show 2D CNN predictions. The top row is a CVD case from the CAIN database, the middle row is an MCI case from the CCNA database, and the bottom row is an AD case from the ADNI database.

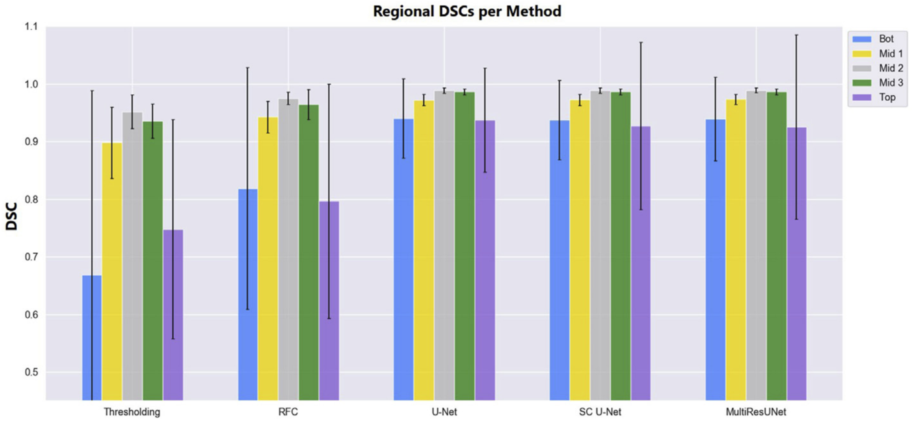

Fig. 16. Average regional DSC across select ICV segmentation methods. Each region represents progressive 20% increments of volume slices. Standard deviation is shown as black error bars.

scans. Over all scanners, the mean CoV of the CNNs was less than 1% (MultiResUNet was best at 0.62%) followed by RFC at 2.1% and Thresholding at 7.6%. To compare the cross-vendor CoVs per method, the absolute difference in CoV was taken for each scanner pair combination. For highly reproducible algorithms, the repeated measurement would yield very low differences in relative variability between scanners. As can be seen in Fig. 19, MultiResUNet has the lowest CoV differences (nearly 0%) for all scanner pairs.

4. Discussion

Although many structural biomarkers are extracted from T1 MRI, most WML segmentation algorithms require FLAIR MRI as the primary data input (García-Lorenzo et al., 2013), (Khademi et al., 2011), (Heinen et al., 2019). In (Narayana et al., 2020) it was demonstrated that the FLAIR modality is 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). Since FLAIR is routinely clinically acquired and highlights vascular disease with high specificity, there is benefit from developing methods that operate on this single sequence.

ICV segmentation is a crucial preprocessing step to FLAIR analysis, but most ICV segmentation algorithms have been designed for T1 MRI. The ROBEX algorithm (Iglesias et al., 2011) for example, uses a ML classifier trained on T1 data and cannot generalize to FLAIR due to differences in tissue class intensities. The brain extraction tool (BET) (Smith, 2002), another popular option, uses a deformable model that is initialized at the brain centre and expands until reaching a threshold represented by the skull in T1 inputs. BET also cannot generalize to FLAIR as the hyperintense appearance of WML cause the threshold to be prematurely crossed and the resultant brain masks to be under segmented (Khademi et al., 2020), (DiGregorio, 2018). Additionally, many methods are developed from normative data and may be sub-optimal for neurodegenerative populations with lesions and atrophy. To address this gap, this work proves the effectiveness of FLAIR specific ICV segmentation algorithms for multicentre, multi-disease data and presents a novel evaluation framework that can be used to establish proof of effectiveness for automated biomarker tools.

Our previous work (RFC) (Khademi et al., 2020) established that FLAIR is an effective sequence for ICV segmentation as DSCs exceeding 90% were obtained on multicentre datasets. However, due to the handcrafted nature of the features, pathology close to the brain periphery caused challenges. To overcome these challenges, it was postulated that

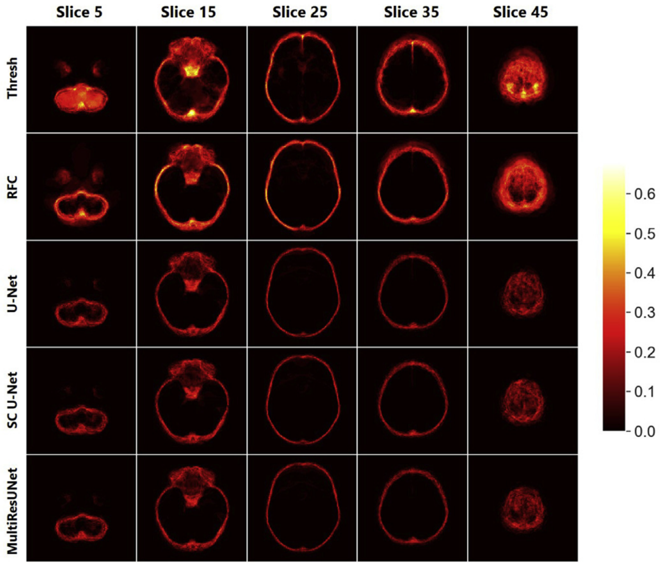

Fig. 17. Error maps for select ICV segmentation methods. Averaged error maps of segmentations are shown for slices 5, 15, 25, 35, and 45 in the 55 slice registration atlas space.

deep learning techniques that can adapt to wide variability in anatomy and pathology would further improve performance. As such, we utilized an established deep learning framework for brain extraction (Kleesiek et al., 2016) (Kleesiek) and adapted several CNN architectures for use on FLAIR MRI. Most of the adapted architectures were 2D. While 3D methods are gaining traction (Hwang et al., 2019), the additional computational expense and memory burden may not be worth it when used on FLAIR data. FLAIR sequences are often acquired with a high slice thickness and are of low resolution in the z-direction which may hinder the advantage of 3D networks (Guerrero et al., 2018).

Based on our experiments, architectures utilizing the "U-shape" (encoder and decoder arm) were the most effective for FLAIR specific ICV segmentation. During our assessment of accuracy, U-Net, SC U-Net, and MultiResUNet were the top performing methods with average DSCs greater than 98%, measurement biases below 8 mL, and narrow limits of agreement (see Fig. 11). The accuracy (average DSC) of traditional approaches was approximately 3–6% lower. In cases with more obvious pathology (atrophy, large lesions) and in the extreme superior and inferior slices the traditional approaches experienced a decrease in performance. This is likely due to the handcrafted feature design which has difficulties adapting to highly variable structures. In contrast, the deep learning methods had smooth ICV contours with no obvious false positive or negatives. The nonlinear nature of deep learning approaches can more effectively model the complex relationships between texture, shape, and structure found in the brain. Kleesiek was the only deep learning method that did not outperform traditional methods. In (Kleesiek et al., 2016), Kleesiek theorized that the proposed network may encounter difficulties on testing data with varying resolution (mm/voxel) from the training set. Given the resolution diversity of our data, this is a possible cause of reduced performance. Further, Kleesiek is a 3D architecture and the low z-plane resolution of FLAIR may have affected its performance. Overall, the accuracy exhibited by our best 2D CNNs was comparable to existing CNN-based ICV systems that use T1 inputs (Ntiri et al., 2020), (Hwang et al., 2019). The effectiveness of the 2D CNN architectures proves the viability of relatively computationally inexpensive systems with low prediction times.

Further experiments and statistical analyses were used to analyze

Table 4
Mean (μ) ± standard deviation (σ) of SIMON ICV measurements for each method across scanner vendors. Coefficients of variation (CoV) computed from balanced scanner groups are also shown where bold values indicate the best CoV.

	Thresholding		RFC		U-Net		SC U-Net		MultiResUNet	
	μ ± σ (mL)	CoV (%)	μ ± σ (mL)	CoV (%)	μ ± σ (mL)	CoV (%)	μ ± σ (mL)	CoV (%)	μ ± σ (mL)	CoV (%)
Siemens	1245.37± 106.27	8.90	1431.69± 32.49	2.11	1416.71± 24.05	1.41	1414.41± 21.95	1.33	1421.37± 16.88	**0.87**
Philips	1309.42± 52.37	2.70	1473.86± 26.38	1.56	1421.61± 8.08	0.69	1418.83± 21.95	0.73	1417.42± 7.98	**0.59**
GE	1306.43± 159.18	11.28	1489.71± 44.31	2.75	1417.81± 4.16	0.27	1421.21± 2.88	**0.18**	1426.14± 6.22	0.40

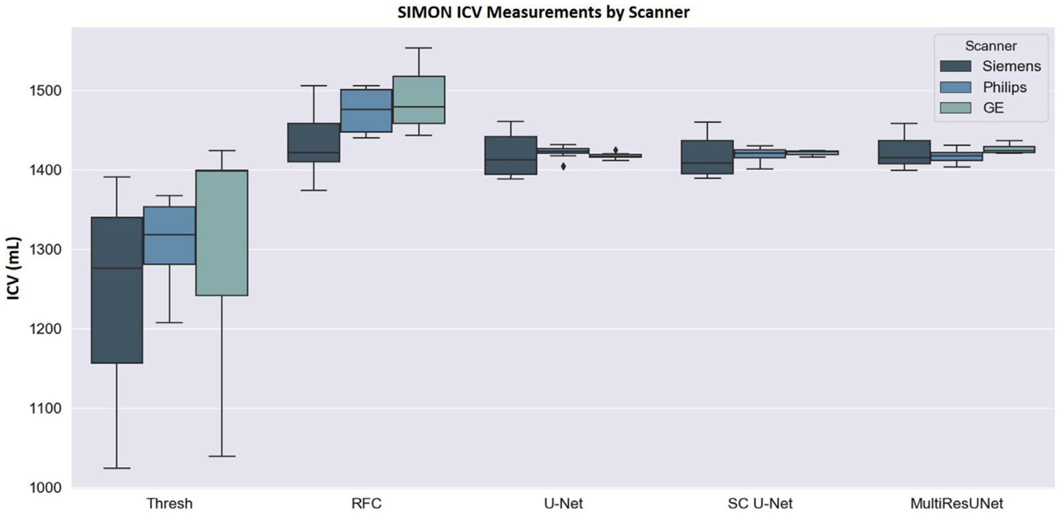

Fig. 18. ICV measurements computed on SIMON as a function of method and scanner vendor.

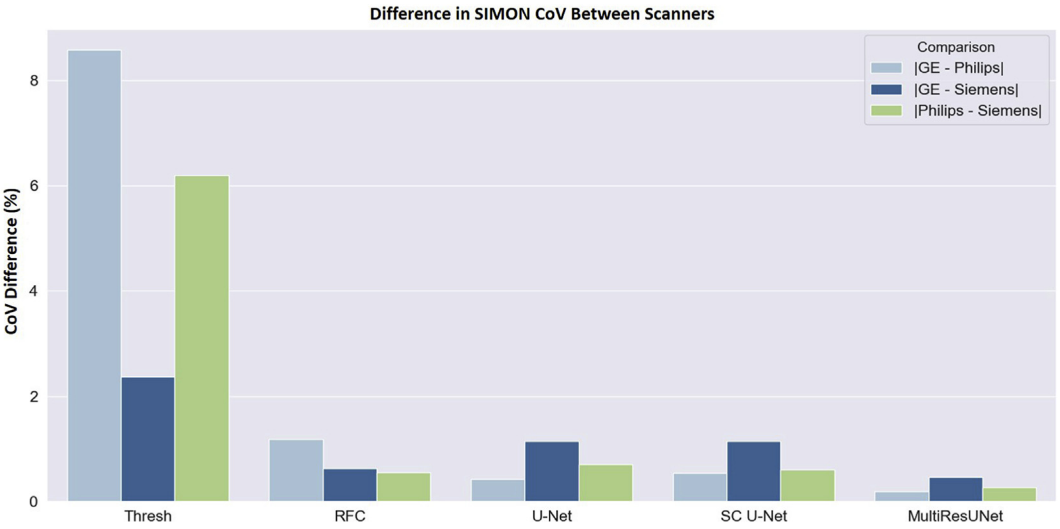

Fig. 19. Coefficients of variation (CoV) between scanners and methods.

performance metrics across different dimensions related to generalizability (scanners, datasets), robustness (pathology levels, spatial location) and reliability and reproducibility (SIMON). In general, if the error rate, or performance of the algorithms is statistically similar across each variable, there is a basis for believing the algorithm is robust across these dimensions. This is important to understand in terms of establishing proof of effectiveness for computer generated imaging biomarkers, which is a prerequisite to clinical implementation. To date, many proof of concept algorithms have been developed for neuroimaging applications (Akkus et al., 2017), (Ali et al., 2019), (Bernal et al., 2019) but there is a lack of performance testing across important parameters in multicentre datasets to determine clinical feasibility (Rehman et al., 2020a). Such an evaluation framework can be used to determine the optimal method for the task, to inform design decisions for method improvement, and to predict performance on new, prospective datasets.

During our assessments of generalization, all methods had significant differences in performance across scanners, except for the RFC method. This is likely due to the intensity standardization enabling fair comparison of features across multicentre datasets. Post-hoc analysis showed that Philips was a common source of difference across the methods, whereas similar DSC means were obtained in volumes from GE and Siemens scanners. This indicates that scanner vendor, particularly Philips, plays a role in algorithm performance. As the number of volumes from each group is relatively balanced in the training set (34% Philips, 25% GE, 41% Siemens), slightly inferior performance on Philips scans (see Fig. 12) is not the result of inadequate or imbalanced exposure during training. A possible reason for this is that the reconstruction algorithms or noise profiles of the scanners are most similar between GE and Siemens which allows networks to perform similarly across them. This rationale was supported by the average standardized histograms of all volumes per scanner vendor shown in Fig. 20. Intensity standardization was mostly able to align the intensity intervals of major tissue classes across vendors. However, Siemens and GE histograms had greater separation between the GM/WM peak and nulled tissues (i.e., CSF) compared to Philips. This may have made Philips scans more challenging from a classification perspective due to feature overlap between tissues. In terms of databases, all methods reported significant differences in performance. Recall that generalization to databases was tested using data split #2 which contained training data from a single dataset. Despite being tested on mostly unseen data in this scenario, our models did experience a drastic reduction in performance (i.e., MultiResUNet had a 0.36% decrease in mean DSC going from data split #1 to #2). Post-hoc analysis showed that across methods there was the most similarity in performance between ADNI and CCNA datasets which are the two dementia cohorts. This could indicate that the source of pathology in the training dataset plays a vital role in the performance of algorithms and should be considered when designing future algorithms for clinical use.

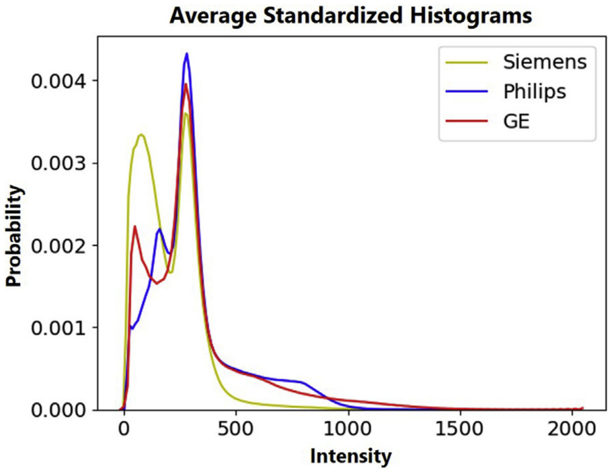

Fig. 20. Average intensity standardized histograms of all CAIN, ADNI, and CCNA volumes from each scanner vendor.

In terms of robustness to pathology, two disease burdens were investigated (WML and CSF load) that are related to vascular and dementia causing conditions. Considering WML loads, the deep learning methods were reliable across all levels (low, medium, high). This is very important from a clinical perspective, since WML are one of the most prominently studied types of pathology in FLAIR MRI. Based on these tests, there can be some confidence that performance will not diminish in the presence of ischemic and demyelinating pathology. In contrast, the performance of traditional methods fluctuated across lesion loads. Large lesions and lesions close to the ICV boundary created challenges for the traditional approaches, which is likely a major source of this variability. In terms of CSF loads, which are a proxy measurement of atrophy and ventricular enlargement, all methods exhibited statistically similar performance across levels, indicating that CSF presence does not strongly impact performance.

Mean DSCs across different spatial locations were compared to see algorithm robustness across anatomical regions. Central slices contain most of the cerebrum and portions of the orbital structures which are traditionally difficult to segment. In very superior slices, there are skull and head structures that create challenges since there is limited brain tissue to segment. In inferior slices, there is the cerebellum and brain stem, as well as other skeletal structures that can create challenges in ICV segmentation. Statistical analysis showed that all methods performed differently across spatial locations, with similar performance found in the top and bottom slices and in two of the middle regions. Similar performance in the top and bottom slices indicates that the relative error rate in these regions is similar. For the deep learning methods, this could be due to a data imbalance issue, as imaging volumes contain less slices from these areas (as compared to the middle regions). In future works, increasing the number of slices from these regions (extreme superior and inferior) that the networks see should improve consistency across the spatial locations. In the traditional approaches, lower and upper slices contain structures that are different from the middle cerebral tissues in terms of intensity, contrast, and edge-content, which can be difficult to differentiate with handcrafted features, and simple classifiers. In the middle regions, similar performance is expected since the structural tissue content is largely similar.

We also evaluated the volumetric measurement reliability and reproducibility of the algorithms on the human phantom dataset (SIMON). The 2D CNNs had low measurement variation, obtaining average CoVs of less than 1% across vendors. Regarding the traditional methods, RFC had a comparable average CoV of approximately 2% while Thresholding was beyond the accepted range of 5%. The CNN-based methods reliably produced similar ICV measurements across scanners with comparable variability as noted by the low CoV differences between vendors. The ICV measurement reliability exhibited by these methods is an important performance characteristic of a tool designed for longitudinal biomarker extraction. The human phantom was scanned between the ages of 29 and 46, making the brain tissue volumes subject to natural, age-related changes (Aribisala et al., 2013). This, compounded with hardware induced variability, creates expectation that there will some variation in the ICV measurements. However, it appears that the 2D CNNs were superior at minimizing this variability.

When comparing the deep learning-based methods, several observations can be made. Firstly, although there are some differences across spatial location, scanners, and databases, methods such as domain adaptation and increasing the training sample size can possibly mitigate these issues. Secondly, changes to the original “U-shaped” (U-Net) architecture only made minor performance differences for the coarse segmentation task of ICV segmentation. The encoder-decoder framework of U-Net seemed to maximize most of the performance potential and design features such additional skip connections, residual connections, and dense connections lead to minor or lateral performance differences. Other segmentation tasks in FLAIR, such as WML segmentation, deal with severe imbalance between positive and negative pixels (i.e., the % of pixels belonging to the WML class is small when considering an entire volume). Coupled with low training data availability, this can make learning difficult and the effect of architecture design differences more apparent. In ICV segmentation, the size of the brain provides better positive to negative pixel balance. This may be better for optimizing the potential of certain CNN architectures and in turn, makes measuring the effect of architecture differences more difficult. Despite comparable performance between our top proposed models, an objective of this work is to provide a system recommendation. As such, we recommend our implementation of the MultiResUNet architecture as an effective system for multicentre FLAIR ICV segmentation in neurodegenerative populations. MultiResUNet differentiated itself from other methods by obtaining the highest number of optimal validation metrics during our assessment of accuracy and having the lowest variability in SIMON ICV measurements between scanner vendors. Medical images contain both local and global structures and an architecture that captures multiple resolutions may be better at modeling fine and gross features in the brain. In our previous work (RFC) (Khademi et al., 2020), we found that the boundary between the CSF and the cranium was the most difficult to segment. By fine-tuning the pixel sampling strategy to better represent these regions, a substantial gain in performance was realized. Thus, the multi-resolution feature extraction of MultiResUNet may be effectively modeling the boundaries and brain tissue regions simultaneously for enhanced discrimination.

Regarding study limitations, 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. Similarly, due to its extreme computational expense, the CompNet architecture was only trained for 15 epochs which may explain the slightly lower performance compared to the other architectures. We also did not utilize a third data split to test the impact of not stratifying for scanner vendor. We expect that our models would also generalize under this scenario as data acquired from a single scanner vendor is analogous to data curated for a single dataset. Since our models did not exhibit a drastic performance drop-off when trained solely on CAIN, there is precedent for similar outcomes when trained solely on single vendor data. However, we would not expect strong generalization if data were acquired with the same scanner and acquisition settings as this would likely lead to extreme overfitting. Additionally, while our models demonstrated robustness to unseen neurodegenerative pathology, we did not test our models on other pathologies like gliomas. Though an aim of this study is to recommend an ICV segmentation solution for neurodegenerative populations, it may be insightful for other researchers to understand which architectural configurations can generalize to other categories of disease. We postulate that similar results will be found for other pathologies using these architectures. While we did not compare our models to those trained with multi-modal inputs, based on the high accuracy results of deep learning methods in FLAIR MRI, we were able to demonstrate that FLAIR is solely viable for high quality ICV segmentation. Comparing to T1 or multi-modal measurements could be a source of future investigation although in (Narayana et al., 2020) the authors found that FLAIR on its own is a viable sequence for tissue segmentation. In addition to technical development, future works will involve applying the proposed FLAIR ICV segmentation system to large, clinical datasets to enable subsequent biomarker extraction and normalization. Using the ICV to normalize biomarkers such as total brain volume, CSF volume, WML volume, and corresponding rates of change (i.e., atrophy, WML expansion) will enable the development of predictive models for neurodegenerative disease classification and for disease monitoring.

5. Conclusions

In this work, we adapted state-of-the-art techniques for segmentation of ICV in multicentre FLAIR MRI. ICV is an important structural biomarker for neurodegenerative disease diagnosis and management. Inaccurate ICV measurements induced by data variation or neurodegenerative pathology reduces the power of clinical studies due to downstream error propagation and insufficient correction for intersubject head size variation. To identify a strong ICV measurement system, we successfully designed an evaluation framework to compare the accuracy, generalization, robustness, and reliability of candidate methods with clinical application in mind. Using our framework, we were able to prove the viability of solely FLAIR ICV analysis, identify MultiResUNet as the best ICV segmentation method for FLAIR, and demonstrate algorithm robustness to multicentre, neurodegenerative disease data. This framework can be used to take proof of concept tools and demonstrate proof of effectiveness, which is a prerequisite to clinical translation. The evaluation framework can easily be expanded to other applications and methods and presents a mechanism for the assessment of computer generated biomarkers and tools.

Declaration of competing interest

This manuscript has not been published and is not under consideration for publication elsewhere. All authors have approved the manuscript and agree with its submission to Neuroimage: Reports. We have no conflicts of interest to disclose.

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 Alberta, 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

Table A
ANOVA analysis of effect of scanner vendor on algorithm performance for data split #1 (F-value and Pr > F) with the null hypothesis that the means of a metric is the same across vendors. Post-hoc analysis compared performance metrics across groups. For insignificant ANOVA tests, post-hoc testing is not performed. Bold p-values indicate no differences exist between scanners (assuming α = 0.05).

Method	Metric	F-Value	Pr > F	GE vs. Philips	GE vs. Siemens	Philips vs. Siemens
Thresh	DSC	13.35	<0.0001	0.0269	0.0493	<0.0001
	HD	13.85	<0.0001	0.0002	**0.9635**	<0.0001
	EF	28.98	<0.0001	0.0006	0.0027	<0.0001
RFC	DSC	3.44	0.0399			
	HD	0.76	0.4708			
	EF	9.21	0.0004	**0.9010**	0.0033	0.0007
Kleesiek	DSC	13.46	<0.0001	0.0004	**0.8394**	<0.0001
	HD	9.56	0.0003	0.0044	**0.7682**	0.0005
	EF	37.11	<0.0001	<0.0001	**0.0591**	<0.0001
FCN8	DSC	24.06	<0.0001	<0.0001	**0.9995**	<0.0001
	HD	29.15	<0.0001	<0.0001	**0.9871**	<0.0001
	EF	25.30	<0.0001	<0.0001	**0.5337**	<0.0001
U-Net	DSC	18.32	<0.0001	<0.0001	**0.9962**	<0.0001
	HD	25.37	<0.0001	<0.0001	**0.7602**	<0.0001
	EF	7.00	0.0021	**0.7661**	0.0197	0.0023
SC U-Net	DSC	16.36	<0.0001	<0.0001	**0.9999**	<0.0001
	HD	22.93	<0.0001	<0.0001	**0.9200**	<0.0001
	EF	8.13	0.0009	**0.5729**	0.0181	0.0008
Res U-Net	DSC	11.66	<0.0001	0.0003	**0.9968**	0.0004
	HD	21.58	<0.0001	<0.0001	**0.9243**	<0.0001
	EF	7.21	0.0018	**0.8522**	0.0028	0.0098
Dense U-Net	DSC	10.67	0.0001	0.0038	**0.6166**	0.0002
	HD	17.81	<0.0001	<0.0001	**0.9394**	<0.0001
	EF	9.91	0.0002	**0.9960**	0.0009	0.0009
CompNet	DSC	14.76	<0.0001	0.0003	**0.6847**	<0.0001
	HD	25.54	<0.0001	<0.0001	**0.9926**	<0.0001
	EF	17.41	<0.0001	0.005	0.0409	<0.0001
MultiResUNet	DSC	11.23	<0.0001	0.0003	**0.9646**	0.0008
	HD	13.00	<0.0001	<0.0001	**0.9048**	0.0004
	EF	7.10	0.0019	**0.6845**	0.0022	0.0173

Table B
ANOVA analysis of effect of dataset on algorithm performance for data split #2 (F-value and Pr > F) with the null hypothesis that the means of a metric is the same across datasets. Post-hoc analysis compared performance metrics across groups. For insignificant ANOVA tests, post-hoc testing is not performed. Bold p-values indicate no differences exist between databases (assuming α = 0.05).

Method	Metric	F-Value	Pr > F	ADNI vs. CAIN	ADNI vs. CCNA	CAIN vs. CCNA
Thresh	DSC	20.43	<0.0001	**0.5884**	<0.0001	<0.0001
	HD	35.30	<0.0001	<0.0001	<0.0001	0.0043
	EF	8.25	0.0008	**0.8739**	0.0011	0.0050
RFC	DSC	64.00	<0.0001	<0.0001	<0.0001	0.0142
	HD	3.78	0.0275	**0.3849**	**0.3350**	0.0204
	EF	59.28	<0.0001	<0.0001	<0.0001	0.0413
Kleesiek	DSC	37.97	<0.0001	<0.0001	<0.0001	**0.8656**
	HD	1.11	0.3359			
	EF	1.08	0.3442			
FCN8	DSC	41.28	<0.0001	<0.0001	**0.0945**	<0.0001
	HD	16.21	<0.0001	**0.4604**	<0.0001	0.0002
	EF	8.13	0.0007	**0.0932**	0.0004	**0.1425**
U-Net	DSC	23.00	<0.0001	0.0020	0.0050	<0.0001
	HD	19.32	<0.0001	0.0498	<0.0001	0.0010
	EF	17.50	<0.0001	**0.4079**	0.0001	<0.0001

(*continued on next column*)

Table B (*continued*)

Method	Metric	F-Value	Pr > F	ADNI vs. CAIN	ADNI vs. CCNA	CAIN vs. CCNA
SC U-Net	DSC	25.78	<0.0001	<0.0001	**0.0568**	<0.0001
	HD	23.34	<0.0001	**0.2317**	<0.0001	<0.0001
	EF	20.11	<0.0001	0.0347	0.0010	<0.0001
Res U-Net	DSC	29.22	<0.0001	0.0004	0.0015	<0.0001
	HD	24.56	<0.0001	**0.1305**	<0.0001	<0.0001
	EF	16.52	<0.0001	**0.0909**	0.0019	<0.0001
Dense U-Net	DSC	19.72	<0.0001	<0.0001	**0.4523**	<0.0001
	HD	21.19	<0.0001	**0.1378**	<0.0001	0.0001
	EF	22.77	<0.0001	**0.3864**	<0.0001	<0.0001
CompNet	DSC	25.57	<0.0001	<0.0001	**0.0624**	<0.0001
	HD	9.25	0.0003	**0.3832**	0.0002	0.0145
	EF	8.05	0.0007	**0.0661**	**0.2027**	0.0004
MultiResUNet	DSC	25.17	<0.0001	0.0002	0.0148	<0.0001
	HD	13.53	<0.0001	**0.0801**	<0.0001	0.0104
	EF	14.01	<0.0001	**0.0560**	0.0122	<0.0001

Table C

ANOVA analysis of effect of lesion load on algorithm performance for data split #1 (F-value and Pr > F) with the null hypothesis that the means of a metric is the same across lesion loads. Post-hoc analysis compared performance metrics across groups. For insignificant ANOVA tests, post-hoc testing is not performed. Bold p-values indicate no differences exist between lesion loads (assuming $\alpha = 0.05$).

Method	Metric	F-Value	Pr > F	High WML vs. Low WML	High WML vs. Medium WML	Low WML vs. Medium WML
Thresh	DSC	14.47	<0.0001	<0.0001	0.0005	**0.4495**
RFC	DSC	6.92	0.0022	0.0035	**0.8128**	0.0133
Kleesiek	DSC	6.45	0.0032	0.0029	**0.4830**	0.0459
FCN8	DSC	2.57	0.0870			
U-Net	DSC	2.47	0.0952			
SC U-Net	DSC	2.46	0.0955			
Res U-Net	DSC	2.03	0.1415			
Dense U-Net	DSC	2.04	0.1414			
CompNet	DSC	2.19	0.1227			
MultiResUNet	DSC	3.31	0.1101			

Table D

ANOVA analysis of effect of CSF load on algorithm performance for data split #1 (F-value and Pr > F) with the null hypothesis that the means of a metric is the same across CSF loads. Post-hoc analysis compared performance metrics across groups. For insignificant ANOVA tests, post-hoc testing is not performed. Bold p-values indicate no differences exist between CSF loads (assuming $\alpha = 0.05$).

Method	Metric	F-Value	Pr > F	High CSF vs. Low CSF	High CSF vs. Medium CSF	Low CSF vs. Medium CSF
Thresh	DSC	1.53	0.2256			
RFC	DSC	0.88	0.4203			
Kleesiek	DSC	0.89	0.4165			
FCN8	DSC	1.42	0.2511			
U-Net	DSC	0.23	0.7950			
SC U-Net	DSC	0.42	0.6613			
Res U-Net	DSC	0.14	0.8679			
Dense U-Net	DSC	0.02	0.9794			
CompNet	DSC	0.75	0.4784			
MultiResUNet	DSC	0.01	0.9939			

Table E

ANOVA analysis of effect of spatial location on algorithm performance for data split #1 (F-value and Pr > F) with the null hypothesis that the means of a metric is the same across spatial locations. Bottom = 1, Middle 1 = 2, Middle 2 = 3, Middle 3 = 4, Top = 5. Post-hoc analysis compared performance metrics across spatial locations. Bold p-values indicate no differences exist between spatial locations (assuming $\alpha = 0.05$).

Method	F-Value	Pr > F	1 vs. 2	1 vs. 3	1 vs. 4	1 vs. 5	2 vs. 3	2 vs. 4	2 vs. 5	3 vs. 4	3 vs. 5	4 vs. 5
Thresh	65.78	<.0001	<.0001	<.0001	<.0001	**0.9989**	<.0001	0.0072	<.0001	**0.0596**	<.0001	<.0001
RFC	120.58	<.0001	<.0001	<.0001	<.0001	**0.9697**	<.0001	<.0001	<.0001	**0.0744**	<.0001	<.0001
Kleesiek	228.08	<.0001	<.0001	<.0001	<.0001	0.0115	<.0001	<.0001	<.0001	0.0185	<.0001	<.0001
FCN8	101.35	<.0001	<.0001	<.0001	<.0001	**0.7647**	<.0001	<.0001	<.0001	**0.2106**	<.0001	<.0001
U-Net	125.18	<.0001	<.0001	<.0001	<.0001	**0.9864**	<.0001	<.0001	<.0001	**0.2574**	<.0001	<.0001
SC U-Net	121.39	<.0001	<.0001	<.0001	<.0001	**0.9536**	<.0001	<.0001	<.0001	**0.2897**	<.0001	<.0001
Res U-Net	145.27	<.0001	<.0001	<.0001	<.0001	**0.9536**	<.0001	<.0001	<.0001	**0.1107**	<.0001	<.0001
Dense U-Net	130.54	<.0001	<.0001	<.0001	<.0001	**1.0000**	<.0001	<.0001	<.0001	**0.2196**	<.0001	<.0001
CompNet	107.63	<.0001	<.0001	<.0001	<.0001	**0.5463**	<.0001	<.0001	<.0001	**0.1400**	<.0001	<.0001
MultiResUNet	131.06	<.0001	<.0001	<.0001	<.0001	**0.9433**	<.0001	<.0001	<.0001	**0.1834**	<.0001	<.0001

References

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.

Ali, H.M., Kaiser, M.S., Mahmud, M., 2019. Application of convolutional neural network in segmenting brain regions from MRI data. December. In: International Conference on Brain Informatics. Springer, Cham, pp. 136–146.

Aribisala, B.S., Hernández, M.C.V., Royle, N.A., Morris, Z., Maniega, S.M., Bastin, M.E., et al., 2013. Brain atrophy associations with white matter lesions in the ageing brain: the Lothian Birth Cohort 1936. Eur. Radiol. 23 (4), 1084–1092.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12 (1), 26–41.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54 (3), 2033–2044.

Ben-Cohen, A., Diamant, I., Klang, E., Amitai, M., Greenspan, H., 2016. Fully convolutional network for liver segmentation and lesions detection. In: Deep Learning and Data Labeling for Medical Applications. Springer, Cham, pp. 77–85.

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.

Brant-Zawadzki, M., Atkinson, D., Detrick, M., Bradley, W.G., Scidmore, G., 1996. Fluid-attenuated inversion recovery (FLAIR) for assessment of cerebral infarction: initial clinical experience in 50 patients. Stroke 27 (7), 1187–1191.

Campbell, M.J., Machin, D., Walters, S.J., 2010. Medical Statistics: a Textbook for the Health Sciences. John Wiley & Sons.

Chambers, L.W., Bancej, C., McDowell, I. (Eds.), 2016. Prevalence and Monetary Costs of Dementia in Canada: Population Health Expert Panel. Alzheimer Society of Canada in collaboration with the Public Health Agency of Canada.

Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A., 2018. VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. Neuroimage 170, 446–455.

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.

Chollet, F., 2018. Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek. MITP-Verlags GmbH & Co. KG.

Datta, S., Narayana, P.A., 2011. Automated brain extraction from T2-weighted magnetic resonance images. J. Magn. Reson. Imag. 33 (4), 822–829.

De Boer, R., Van Der Lijn, F., Vrooman, H.A., Vernooij, M.W., Ikram, M.A., Breteler, M.M., Niessen, W.J., 2007. Automatic segmentation OF brain tissue and white matter lesions IN MRI. April. In: 2007 4th IEEE International Symposium on Biomedical Imaging: from Nano to Macro. IEEE, pp. 652–655.

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.

Dey, R., Hong, Y., 2018. CompNet: complementary segmentation network for brain MRI extraction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp. 628–636. September.

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

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.

Duchesne, S., Chouinard, I., Potvin, O., Fonov, V.S., Khademi, A., Bartha, R., et al., 2019a. The Canadian dementia imaging protocol: harmonizing national cohorts. J. Magn. Reson. Imag. 49 (2), 456–465.

Duchesne, S., Chouinard, I., Potvin, O., Fonov, V.S., Khademi, A., Bartha, R., et al., 2019b. The Canadian dementia imaging protocol: harmonizing national cohorts. J. Magn. Reson. Imag. 49 (2), 456–465.

Eskildsen, S.F., Coupé, P., Fonov, V., Manjón, J.V., Leung, K.K., Guizard, N., Alzheimer's Disease Neuroimaging Initiative, 2012. BEaST: brain extraction based on nonlocal segmentation technique. Neuroimage 59 (3), 2362–2373.

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.

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.

Gunter, J.L., Shiung, M.M., Manduca, A., Jack Jr., C.R., 2003. Methodological considerations for measuring rates of brain atrophy. J. Magn. Reson. Imag.: An Official Journal of the International Society for Magnetic Resonance in Medicine 18 (1), 16–24.

Hah, T.T.T., Kim, J.Y., Choi, S.H., 2014. White matter hyperintensities extraction based T2-FLAIR MRI using non-local means filter and nearest neighbor algorithm. October. In: 2014 International Conference on IT Convergence and Security (ICITCS). IEEE, pp. 1–4.

Hansen, T.I., Brezova, V., Eikenes, L., Håberg, A., Vangberg, T.R., 2015. How does the accuracy of intracranial volume measurements affect normalized brain volumes? Sample size estimates based on 966 subjects from the HUNT MRI cohort. Am. J. Neuroradiol. 36 (8), 1450–1456.

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.

Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imag. 30 (9), 1617–1634.

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., 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 Official Journal of the International Society for Magnetic Resonance in Medicine 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.

Kalavathi, P., Prasath, V.S., 2016. Methods on skull stripping of MRI head scan images—a review. J. Digit. Imag. 29 (3), 365–379.

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., 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.

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., Biller, A., 2016. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. Neuroimage 129, 460–469.

Kolařík, M., Burget, R., Uher, V., Říha, K., Dutta, M.K., 2019. Optimized high resolution 3d dense-u-net network for brain and spine segmentation. Appl. Sci. 9 (3), 404.

Leung, K.K., Bartlett, J.W., Barnes, J., Manning, E.N., Ourselin, S., Fox, N.C., Alzheimer's Disease Neuroimaging Initiative, 2013. Cerebral atrophy in mild cognitive impairment and Alzheimer disease: rates and acceleration. Neurology 80 (7), 648–654.

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

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.

Malone, I.B., Leung, K.K., Clegg, S., Barnes, J., Whitwell, J.L., Ashburner, J., et al., 2015. Accurate automatic estimation of total intracranial volume: a nuisance variable with less nuisance. Neuroimage 104, 366–372.

Manjón, J.V., Eskildsen, S.F., Coupé, P., Romero, J.E., Collins, D.L., Robles, M., 2014. Nonlocal intracranial cavity extraction. Int. J. Biomed. Imag. 2014.

Mayeux, R., Stern, Y., 2012. Epidemiology of alzheimer disease. Cold Spring Harbor perspectives in medicine 2 (8), a006239.

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. Magn. Reson. Imag. 65, 8–14.

Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al., 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J. Am. Geriatr. Soc. 53 (4), 695–699.

Nordenskjöld, R., Malmberg, F., Larsson, E.M., Simmons, A., Brooks, S.J., Lind, L., et al., 2013. Intracranial volume estimated with commonly used methods could introduce bias in studies including brain volume measurements. Neuroimage 83, 355–360.

Ntiri, E.E., Holmes, M.F., Mojiri, P., Ramirez, J., Gao, F., Ozzoude, M., et al., 2020. Improved Segmentation of the Intracranial and Ventricular Volumes in Populations with Cerebrovascular Lesions and Atrophy Using 3D CNNs (bioRxiv).

Oppedal, K., Eftestøl, T., Engan, K., Beyer, M.K., Aarsland, D., 2015. Classifying dementia using local binary patterns from different regions in magnetic resonance images. Int. J. Biomed. Imag. 2015.

Rehman, H.Z.U., Hwang, H., Lee, S., 2020a. Conventional and deep learning methods for skull stripping in brain MRI. Appl. Sci. 10 (5), 1773.

Rehman, H.Z.U., Hwang, H., Lee, S., 2020b. Conventional and deep learning methods for skull stripping in brain MRI. Appl. Sci. 10 (5), 1773.

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

Rocca, M.A., Battaglini, M., Benedict, R.H., De Stefano, N., Geurts, J.J., Henry, R.G., et al., 2017. Brain MRI atrophy quantification in MS: from methods to clinical application. Neurology 88 (4), 403–413.

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.

Scahill, R.I., Frost, C., Jenkins, R., Whitwell, J.L., Rossor, M.N., Fox, N.C., 2003. A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. Arch. Neurol. 60 (7), 989–994.

Schwarz, C.G., Gunter, J.L., Wiste, H.J., Przybelski, S.A., Weigand, S.D., Ward, C.P., et al., 2016. A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity. Neuroimage: Clinical 11, 802–812.

Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. Neuroimage 22 (3), 1060–1075.

Sigurdsson, S., Aspelund, T., Forsberg, L., Fredriksson, J., Kjartansson, O., Oskarsdottir, B., et al., 2012. Brain tissue volumes in the general population of the elderly: the AGES-Reykjavik study. Neuroimage 59 (4), 3862–3870.

Silbert, L.C., Quinn, J.F., Moore, M.M., Corbridge, E., Ball, M.J., Murdoch, G., et al., 2003. Changes in premorbid brain volume predict Alzheimer's disease pathology. Neurology 61 (4), 487–492.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.

Smith, E.E., Cieslak, A., Barber, P., Chen, J., Chen, Y.W., Donnini, I., et al., 2017. Therapeutic strategies and drug development for vascular cognitive impairment. Journal of the American Heart Association 6 (5), e005568.

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

Sosa-Ortiz, A.L., Acosta-Castillo, I., Prince, M.J., 2012. Epidemiology of dementias and Alzheimer's disease. Arch. Med. Res. 43 (8), 600–608.

Squitieri, F., Cannella, M., Simonelli, M., Sassone, J., Martino, T., Venditti, E., et al., 2009. Distinct brain volume changes correlating with clinical stage, disease progression rate, mutation size, and age at onset prediction as early biomarkers of brain atrophy in Huntington's disease. CNS Neurosci. Ther. 15 (1), 1–11.

Struyfs, H., Sima, D.M., Wittens, M., Ribbens, A., de Barros, N.P., Vân Phan, T., et al., 2020. Automated MRI Volumetry as a Diagnostic Tool for Alzheimer's Disease: Validation of Icobrain Dm. NeuroImage: Clinical, 102243.

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, 117081.

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

Winkler, A.M., Kochunov, P., Glahn, D.C.. FLAIR templates. Available at. http://brainder.org.

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.

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.