Unsupervised Domain Adaptation using Silver Standards and Transfer Learning for Ventricle Segmentation in FLAIR MRI

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Lateral ventricular volume (LVV) an important biomarker for clinical investigation. In this work, we present the first deep learning-based LVV segmentation method for fluid attenuated inversion recovery (FLAIR) MRI. To mitigate covariate shifts between source and target domains, this work proposes and unsupervised domain adaptation (UDA) method that optimizes target domain performance. Silver standard (SS) masks were generated from target datasets using a novel unsupervised ventricular segmentation algorithm and used to supplement the gold standard (GS) source domain data. Three models were tested on held-out test sets from four datasets: 1) trained on target SS masks and fine-tuned on source GS masks (All Layers), 2) trained on source gold standard (GS) masks (GS CAIN Only), 3) trained on target SS masks (SS Only). The All Layers model (mean DSC = 0.89) showed significantly ($p < 0.05$) higher DSC and more consistent results compared to the GS only model on three target domains. Results suggest pre-training with noisy labels from the target domain allows the model adapt to the dataset-specific characteristics of the target domain and provides for robust parameter initialization. This method has wide application to other medical imaging problems where labeled data is scarce, and can be used as a per-dataset calibration method to accelerate wide-scale adoption.

Keywords: Domain adaptation, transfer learning, ventricle segmentation, silver standards

1. Introduction

As of 2016, over 550,000 Canadians live with dementia and this number is expected to double within the next 15 years (Chambers et al., 2016). Alzheimer's disease (AD) is the most common form of dementia and is associated with rapid decline in cognition (Mayeux and Stern, 2012). Quantitative imaging biomarkers from neurological exams, such as total brain volume (TBV), cerebrospinal fluid (CSF) volume, and white matter lesion (WML) volume are important for monitoring disease progression (Thung et al., 2018) (Mofrad et al., 2021) (Popuri et al., 2020).

One of the most commonly studied neurological biomarkers for dementia is lateral ventricular volume (LVV). Enlarged ventricles correlate with decreased TBV and increased CSF volume and have been associated with deficits in cognition and memory (Ertekin et al., 2016). Since manual analysis of LVV is labourious and subjective, automated tools are the optimal alternative. Previous studies have primarily segmented the LVV using T1-weighted MRI images and achieved a mean Dice Similarity Coefficient (DSC) ranging from 0.84-0.88 using deep learning techniques. (Ertekin et al., 2016)(Ertekin et al., 2016).

While T1 MRI are used for the structural analysis of anatomy, fluid attenuated inversion

© 2023 Anonymous.

recovery (FLAIR) MRI is more readily analyzed for white matter lesions (WML), which are related to stroke, ischemia, and dementia (Wardlaw et al., 2015). Additionally, FLAIR MRI is routinely used in clinical practice by neuroradiologists, radiologists and neurologists, and therefore, biomarker tools for this sequence have high translation potential (Khademi et al., 2021). FLAIR may have the potential to better segment the LVV due the nulled CSF signal increasing the CSF-to-tissue contrast (Soltanian-Zadeh and Peck, 2001) (Narayanan et al., 2020). However, to our knowledge, there is no available ventricular segmentation technique for FLAIR MRI to-date. This work aims to develop and validate the first ever deep learning-based segmentation method for the LVV in mutlicentre FLAIR MRI.

Convolutional neural networks (CNN) excel at image segmentation due to their ability to extract high-dimensional and nonlinear features. The challenge, however, is when deep learning algorithms are applied to images from a distribution outside the training dataset, there is a drop in performance. This covariate shift is common in medical imaging due to different scanning devices (hardware/software), acquisition parameters and protocols (Ioffe and Szegedy, 2015b) (Huang et al., 2006) (Um et al., 2019). Generalization gaps create translational hurdles since algorithms created by vendors cannot be widely deployed and adopted. Annotations from the target site can supplement training sets, but generating gold standard (GS) labeled ground truths for medical imaging is laborious and expensive.

To overcome these challenges, this work proposes a novel unsupervised LVV method that automatically generates silver standard (SS) masks in the unlabeled target domain to supplement GS data to improve target domain performance. As will be shown, the method can be used to calibrate AI tools on a per-site or per-dataset basis, in an easy and unsupervised manner that improves generalization and mitigates multicentre variability (Khademi et al., 2020). The concept of using SS masks to increase the sample size of the training set has been successfully implemented in the application of intracranial volume (ICV) segmentation (Lucena et al., 2019). This work offers the following contributions: (1) We propose the first FLAIR-only deep learning-based and unsupervised LVV segmentation methods. (2) We demonstrate how to leverage SS masks from an unlabeled target domain to improve performance and consistency in the target domain(s) for unsupervised domain adaptation. (3) Propose a generic framework for per-site calibration for any medical imaging AI problem.

2. Methodology

2.1. Data

Four multicentre, multi-scanner datasets were used in this work. The first is from Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jr et al., 2008), which is a public repository for studying AD. The second dataset is from the Canadian Atherosclerosis Imaging Network (CAIN), which is a vascular disease study (Tardif et al., 2013). The third dataset is a dementia cohort from the Canadian Consortium on Neurodegeneration in Aging (CCNA)(Chertkow et al., 2019). The last dataset is from the Ontario Neurodegenerative Disease Research Initiative (ONDRI), which is also a dementia dataset (Ramirez et al., 2020). CAIN was used as the source domain and ADNI, CCNA, and ONDRI were used as target domains. Datasets have varying acquisition parameters and scanners (Table 1), resulting in a diverse clinical sample. There are 30 GS volumes for each of the three target domains (4500 images) and 80 GS volumes from CAIN, the source domain (3840 images).

Table 1: FLAIR MRI ground truth data. All data is 3T and 3-5mm slice thickness.

	Patient Information					
Database	**Disease**	**Volumes**	**Images**	**Patients**	**Centres**	**VV ± SD (mL)**
ADNI	Dementia	30	1260	30	12	64.67 ± 37.65
CAIN	Vascular	80	3840	80	8	30.22 ± 16.06
CCNA	Dementia	30	1440	30	9	42.26 ± 31.78
ONDRI	Dementia	30	1440	30	6	37.52 ± 13.31
Total	All	170	7980	170	33	39.71 ± 15.24
	Acquisition Parameters					
Database	**GE/Philips/Siemens**	**TR (ms)**	**TE (ms)**	**TI (ms)**	**X Spacing (mm)**	**Y Spacing (mm)**
ADNI	10/10/10	9000-11000	90-154	2250-2500	0.8594	0.8594
CAIN	20/20/20	9000-11000	117-150	2200-2800	0.4285-1	0.4285-1
CCNA	10/10/10	9000-9840	117-148	2250-2500	0.9375	0.9375
ONDRI	10/10/10	2250-100000	90-12810	2250-100000	0.92-0.94	0.89-0.94
Total	50/50/50	2250-100000	90-12810	2250-100000	0.4295-1.2	0.4295-1.2

2.2. Image Pre-Processing

All volumes underwent a series of standard image pre-processing techniques. Bias field correction was applied to remove any low frequencies corrupting the MRI images. Intensity standardization was then performed by aligning the intensity distributions (Reiche et al., 2019). ICV segmentation using a MultiResUNET CNN was used to remove all non-brain tissues (i.e., skull, orbital cavities, and skin) (DiGregorio et al., 2021).

2.3. Source Domain Gold Standard (GS) Annotations

The GS annotations were done by a medical student trained by a radiologist for ventricle annotations in ITKSnap. If the choroid plexus obscured the edge of the ventricle, the delineation included the choroid plexus within the ventricles. The septum pellucidum was excluded from the annotations. Areas below the anterior-posterior commissure line were eliminated to exclude the 3rd and 4th ventricles as well as the inferior horns (Dwyer et al., 2017).

2.4. Target Domain Silver Standard (SS) Masks

An unsupervised method was designed to generate SS LVV masks from the target domains. Initially, a threshold was applied to intensity standardized images to remove gray and white matter, and isolate a rough estimate of the total CSF (ventricles + subarachnoid space). An erosion kernel was applied to the brain tissue mask, which was then multiplied by the total CSF mask. Next, a largest object search was then implemented to remove the subarachnoid CSF, thus isolating the ventricles. Hole filling was applied to the largest object. To ensure the detected ventricles occurred along the midline, the distance to the mid-sagittal plane was calculated. The distance between centre of mass of the brain and the detected ventricles was determined to confirm ventricles were located near the centre of the brain with respect to the frontal plane. Pixels below the anterior-posterior commissure line were removed. This method is completely unsupervised and can be used to easily and efficiently to generate SS training data from any unlabeled target imaging centre.

2.5. Deep Learning-based Ventricular Segmentation

A 2D U-Net was used as the CNN architecture for this work as it is a fast and accurate deep learning technique that has achieved high performance on biomedical data (Ronneberger et al., 2015), (Hwang et al., 2019), (Thakur et al., 2020). The encoding path consists of convolutional and max pooling layers to perform feature extraction. The decoding path

contains units of convolutional and transposed convolutional layers and skip connections to recapture spatial context (Long et al., 2015). UNET contains five levels where the filter depth is doubled during each downsampling block (via max pooling) and halved during each upsampling block (via transposed convolution). In this work, UNET was implemented with batch normalization layers succeeding convolutional layers (Ioffe and Szegedy, 2015a) to accelerate convergence and improve generalization via a modest regularization effect. Adam optimizer was used with a learning rate of 1e-4, batch size of 16 with 50 epochs, and a generalized dice loss for all models and experiments. Data augmentations were applied for rotation, scaling, shearing, scaling and translation. Models were trained on an NVIDIA GeForce RTX 3090 Ti GPU with 32GB of RAM.

2.6. Transfer Learning

Transfer learning and fine-tuning neural networks is an effective method to enhance performance by transferring knowledge of a network trained on one dataset, onto another dataset or problem domain. Pre-trained networks from natural images may not be beneficial for medical image applications, due to differences in image properties and covariate shift that is pervasive in medical imaging. In this work, we use the unsupervised LVV segmentation to generate SS masks in the unlabeled target domain, and use the SS to pre-train a base model to learn low-level ventricular features and provide an optimal parameter initialization. The GS masks, which are carefully annotated, will be used to fine-tune the model to capture more precise details. In Amiri et al. the authors determined that fine-tuning all of the layers of the U-Net architecture yielded the optimal segmentation performance and therefore will be adopted in this work (Amiri et al., 2020).

2.7. Performance Metrics

To measure segmentation performance and consistency of segmentations, the Dice Similarity Coefficient (DSC) coefficient of variation (CoV) of the DSC were used. To investigate volume-based metrics, linear regression and correlation were used to compare predicted ventricular volume (LVV) and ground truth LVV, as well as Bland Altman plots.

2.8. Experimental Setup

Three types of models are examined for ventricular segmentation on target data. The first model is GS CAIN Only, which is trained with 50 CAIN (source) GS masks. The second model SS Only, is trained using SS masks from the target dataset. The last model is called All Layers and is pre-trained on the target SS dataset and fine-tuned using the GS CAIN data. For each target (ADNI, CCNA and ONDRI) and source (CAIN) datasets, 50/100/150/200 SS masks were extracted using the method of Section 2.4. A total of 33 deep learning models were trained (4 types of SS masks $\times$4 datasets (=16), for SS Only and All Layers (=32) and the GS Only model (=33)). Volumes were stratified according to scanner vendor (Siemens, GE, Phillips). For each target and source datasets, 30 imaging volumes were held-out (non-overlapping, unique subjects) for testing performance. Table 2 summarizes the data for each model. ANOVA and Tukey's post-hoc on DSC was used to determine whether All Layers model is performing better in target data compared to GS CAIN Only models and ANOVA for number of SS masks.

Table 2: Summary of datasplits of SS and GS masks for model training, tuning, and testing. GS masks are from source domain (CAIN).

Model	Datasets	SS Train	GS Tune	GS Train	Target Test
GS CAIN Only	ADNI/CAIN/CCNA/ONDRI	0	0	50	30/30/30/30
SS Only	ADNI/CAIN/CCNA/ONDRI	50/50/50/50	0	0	30/30/30/30
SS Only	ADNI/CAIN/CCNA/ONDRI	100/100/100/100	0	0	30/30/30/30
SS Only	ADNI/CAIN/CCNA/ONDRI	150/150/150/150	0	0	30/30/30/30
SS Only	ADNI/CAIN/CCNA/ONDRI	200/200/200/200	0	0	30/30/30/30
All Layers	ADNI/CAIN/CCNA/ONDRI	50/50/50/50	50	0	30/30/30/30
All Layers	ADNI/CAIN/CCNA/ONDRI	100/100/100/100	50	0	30/30/30/30
All Layers	ADNI/CAIN/CCNA/ONDRI	150/150/150/150	50	0	30/30/30/30
All Layers	ADNI/CAIN/CCNA/ONDRI	200/200/200/200	50	0	30/30/30/30

3. Results

On all held-out testing data from CAIN, CCNA, ADNI and ONDRI (120 imaging volumes), the unsupervised ventricular segmentation algorithm (SS masks) achieved a mean DSC=0.82 and CoV=0.10 (see Fig.2). Qualitative results for the unsupervised algorithm (SS masks), GS CAIN Only, SS Only and All Layers models are shown in Fig. 5. The most notable discrepancy between models is the false negatives from the GS CAIN Only model when testing on inferior slices (where LVs are more morphologically complex), and when the LVs were relatively narrow, while the All Layers robustly detected the LVV.

3.1. Source Domain Performance

SS masks from the source dataset (CAIN) along with GS (CAIN) masks were used to test performance on the 30 held-out CAIN volumes. Distributions and mean DSC and CoV for GS CAIN Only, SS Only and All Layers models is shown in Fig. 3. All Layers model yielded the most accurate (DSC = 0.89) and consistent (CoV = 0.04) results, compared to the GS CAIN Only model (DSC = 0.88 and CoV = 0.05). While segmentation performance and consistency was slightly better in the All Layers fine-tuned model, there was no significant difference ($p = 0.96$) when comparing to GS CAIN Only. This indicates performance of the model within distribution is maintained or slightly improved and this framework will be used to test on target datasets. There was do difference in performance across the models with different number of SS masks in the All Layers Model.

3.2. Target Domain Performance

To determine the effect of SS masks from the unlabeled target domain, All Layers models were pre-trained using SS masks for each of targets separately (CCNA, ADNI, ONDRI) and fine-tuned using the GS CAIN masks to generate dataset specific models. The same dataset-specific SS masks were used to train SS Only models for each dataset. The performance of these models, compared to the GS CAIN only models is shown in Table 3. All four All Layers models yielded the highest DSC and lowest CoV in comparison to the GS CAIN Only models when testing on target datasets. ANOVA and Tukey's post-hoc results deemed performance is significantly improved when comparing DSC between the All Layers model

to both GS CAIN Only and SS Only models ($p < 0.05$). The significant reduction in CoV (GS CAIN Only mean CoV = 0.11, vs. mean CoV = 0.05 for All Layers), and similar DSC values across datasets in the All Layers models indicates performance of the proposed algorithm on the target data is more accurate and consistent. ANOVA also indicated there is no difference ($p = 0.99$) in DSC performance for different amounts of SS masks in All Layer models so including any amount of SS seems beneficial.

Models with 100 SS masks (the best DSC) were retained and compared further to

Table 3: Performance on held-out testing volumes. Bold is best for that dataset and * indicates model is significantly ($p < 0.05$) different from GS CAIN Only.

Model (SS Masks)	**CAIN**		**ADNI**		**CCNA**		**ONDRI**	
	DSC	**CoV**	**DSC**	**CoV**	**DSC**	**CoV**	**DSC**	**CoV**
GS CAIN Only (0)	0.88	0.05	0.85	0.08	0.85	0.17	0.85	0.15
SS Only (50)	0.80	0.15	0.81	0.11	0.81	0.13	0.84	0.07
SS Only (100)	0.78	0.15	0.80	0.15	0.83	0.10	0.84	0.08
SS Only (150)	0.85	0.09	0.84	0.10	0.83	0.10	0.84	0.08
SS Only (200)	0.84	0.11	0.83	0.10	0.82	0.11	0.84	0.08
All Layers (50)	**0.89**	**0.04**	0.88*	**0.05**	0.89*	0.05	0.88	**0.05**
All Layers (100)	**0.89**	**0.04**	0.88	0.06	**0.90***	**0.04**	**0.89***	**0.05**
All Layers (150)	**0.89**	0.05	0.88	0.06	**0.90***	0.06	**0.89***	**0.05**
All Layers (200)	**0.89**	**0.04**	**0.89***	**0.05**	0.88*	0.05	0.88*	**0.05**

the GS Only models. All three models were tested on the held-out target datasets and linear regression and Bland-Altman analysis between the predicted and ground truth LVV is shown in Fig.4. Table 4 summarizes linear regression results on all datasets individually and together. The All Layers model yielded the highest correlation and best slope (close to a value of 1) for all three target datasets, and was close in the source (CAIN) dataset. Fig. 4 displays Bland Altman plots. Predictions from the All Layers model showed the least amount of difference in comparison to the GTs as based on the offset between the predicted LVV and true LVV on average it is lower than the GS Only model) as well as tighter bounds indicating less variability and better consistency.

4. Discussion

Although previous works primarily used T1-weighted MRI for LVV segmentation (Ertekin et al., 2016) (Wu and Tanga, 2021), results of the All Layers and GS CAIN Only models suggest that FLAIR is an effective imaging sequence for LVV segmentation using deep learning, as they achieved state-of-the-art mean DSC values. This can be attributed to the increased CSF-to-tissue contrast that FLAIR offers by nulling the CSF signal and combining the grey and white matter and strong performance of CNN-based segmentation schemes.

Previous technologies have demonstrated that transfer learning can be successfully used when the source and target data are from different domains (e.g., natural images to medical images) (Amiri et al., 2020) (Cheplygina, 2019) (Tajbakhsh et al., 2016). The results in this work show that pre-training with SS masks from the target domain and fine-tuning with source domain GS masks significantly improved overall performance and consistency in the target domain compared to source models. We hypothesize the model learns dataset- and problem-specific global trends from target SS masks that are injected into the model. As

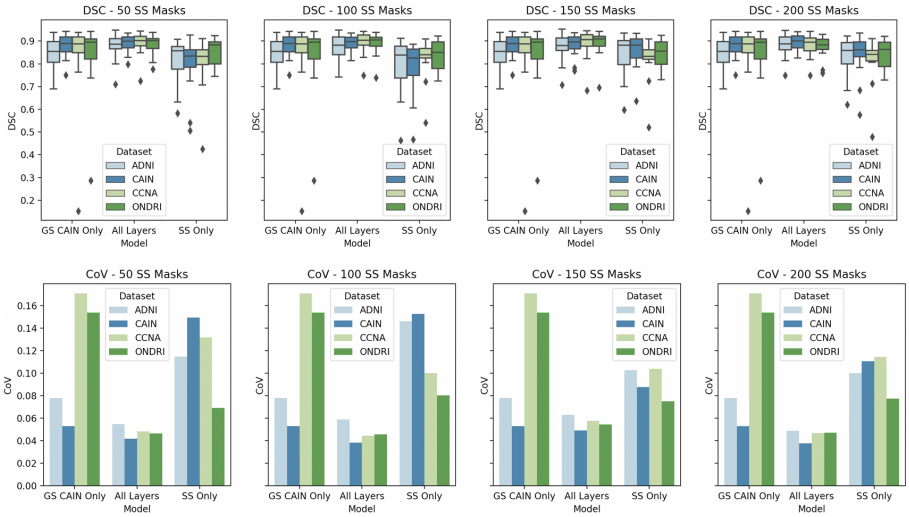

Figure 1: DSC (top) and CoV (bottom) for each model by number of SS masks used in training compared between target datasets.

Table 4: Linear regression and correlation coefficients for each model.

Dataset	Model	R^2	Intercept	Slope
CAIN	GS CAIN Only	0.95	**0.94**	**1.02**
	SS Only	0.94	-3.91	0.95
	All Layers	**0.97**	1.61	1.06
ADNI	GS CAIN Only	0.89	15.19	0.62
	SS Only	0.68	11.92	0.48
	All Layers	**0.95**	**5.13**	**0.81**
CCNA	GS CAIN Only	0.84	12.14	0.66
	SS Only	0.61	15.43	0.45
	All Layers	**0.86**	**9.95**	**0.74**
ONDRI	GS CAIN Only	0.68	**10.93**	**0.55**
	SS Only	0.57	16.22	0.42
	All Layers	**0.71**	14.31	**0.55**
All Target	GS CAIN Only	0.80	12.41	0.62
	SS Only	0.54	13.49	0.49
	All Layers	**0.85**	**9.91**	**0.71**
All Data	GS CAIN Only	0.80	11.72	0.64
	SS Only	0.68	11.92	0.48
	All Layers	**0.84**	**9.93**	**0.72**

shown by SS Only models, the target data is not enough, and that is because the SS masks can have noise, which is cleaned up by the GS on fine-tuning. Exposing the model to SS masks from the target domain helps to adapt to other domains. This also highlights that instead of using out-of-problem-domain pre-trained models like imageNet, this UDA poses a new way to generate pre-trained models that can be trained on a large amount of SS data. While we show good performance for LVV applications, this could have wide implications

for other pre-training applications.

This work was validated on three large, multicnetre clinical datasets (ADNI, CCNA, and ONDRI) which is a huge benefit of the current work. The method has been successfully applied to multiple target domains with similar results demonstrates the robustness of this domain-adaptation pipeline. Fig. 6 shows All Layers model has the best performance on the target domain (with roughly the same performance) and a 4% (significant) increase in DSC compared to the GS model. This improvement can be seen visually in the examples in Fig. 5. The GS CAIN Only models underestimate and miss small ventricular areas, which is exacerbated in the out-of-domain targets. The All Layers model improved segmentation performance specifically on narrow LVs and LVs located in both inferior and superior slices, where the morphological shape becomes quite complex and discontinuous. This was likely because the All Layers model was exposed to more site-specific variability, via the SS masks, in comparison to the GS CAIN Only model. Some examples also showed the SS Only model outperformed the GS CAIN Only model, more specifically, the SS Only model was able to better fill in the middle, dark region of the ventricles. This is due to the unsupervised algorithm's reliance on intensity-based thresholding. Therefore, using SS masks for pre-training provided a better parameter initialization, thus leading to a more generalized and accurate model. Results indicate the model benefits from being exposed to dataset-specific features and variables in training, albeit via noisy labels. These features/variables include many that could be related to covariate shifts (intensity distributions, imaging parameters, and/or image resolution).

5. Conclusion

This work proposes the first FLAIR MRI only lateral ventricuar volume (LVV) segmentation method with state-of-the-art performance. An unsupervised domain adaptation (UDA) technique for LVV segmentation in FLAIR MRI was designed to improve segmentation accuracy and mean volume difference across three multicentre target datasets by using SS masks and transfer learning. The proposed UDA method improved both the accuracy and consistency of the segmentation compared to predictions from the source domain model. The proposed framework is an easy and effective method to design dataset-specific pre-trained models using the vast amounts of unlabeled target data available in imagine centres.

Acknowledgments

Acknowledgments withheld.

References

Mina Amiri, Rupert Brooks, and Hassan Rivaz. Fine-Tuning U-Net for Ultrasound Image Segmentation: Different Layers, Different Outcomes. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 67(12):2510–2518, 2020. doi: 10.1109/TUFFC.2020.3015081.

Larry Chambers, Christina Bancej, and Ian McDowell. Prevalence and Monetary Costs of Dementia in Canada: Population Health Expert Panel. Toronto, ON, Canada: The

Alzheimer Society of Canada. *Alzheimer Society of Canada*, 36(10):231–232, 2016. doi: https://doi.org/10.24095/hpcdp.36.10.04.

Veronika Cheplygina. Cats or CAT scans: Transfer learning from natural or medical image source data sets? *Current Opinion in Biomedical Engineering 2*, 9:21–27, 2019. doi: https://doi.org/10.1016/j.cobme.2018.12.005.

Howard Chertkow, Michael Borrie, Victor Whitehead, Sandra E Black, Howard H Feldman, Serge Gauthier, David B Hogan, Mario Masellis, Katherine McGilton, Kenneth Rockwood, Mary C Tierney, Melissa Andrew, Ging-Yuek R Hsiung, Richard Camicioli, Eric E Smith, Jennifer Fogarty, Joseph Lindsay, Sarah Best, Alan Evans, Samir Das, Zia Mohaddes, Randi Pilon, Judes Poirier, Natalie A Phillips, Elizabeth MacNamara, Roger A Dixon, Simon Duchesne, Ian MacKenzie, and R Jane Rylett. The comprehensive assessment of neurodegeneration and dementia: Canadian cohort study. *Canadian Journal of Neurological Sciences*, 46(5):499–511, 2019. doi: 10.1017/cjn.2019.27.

Justin DiGregorio, Giordano Arezza, Adam Gibicar, Alan R. Moody, Pascal N. Tyrrell, and April Khademi. Intracranial volume segmentation for neurodegenerative populations using multicentre flair mri. *Neuroimage*, 1(1):100006, 2021. doi: https://doi.org/10.1016/j.ynirp.2021.100006.

Michael G Dwyer, Diego Silva, Niels Bergsland, Dana Horakova, Deepa Ramasamy, Jaqueline Durfee, Manuela Vaneckova, Eva Havrdova, and Robert Zivadinov. Neurological software tool for reliable atrophy measurement (NeuroSTREAM) of the lateral ventricles on clinical-quality T2-FLAIR MRI scans in multiple sclerosis. *NeuroImage: Clinical*, 15: 769–779, 2017. doi: 10.1016/j.nicl.2017.06.022.

Tolga Ertekin, Niyazi Acer, Emel Köseoğlu, Gökmen Zararsız, Ali Sönmez, Kazım Gümüş, and Erdal Kurtoğlu. Total intracranial and lateral ventricle volumes measurement in Alzheimer's disease: A methodological study. *Journal of Clinical Neuroscience*, 34:133–138, 2016. doi: 10.1016/j.jocn.2016.05.038.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schlkopf, and Alex Smola. Correcting Sample Selection Bias by Unlabeled Data. *Advances in Neural Information Processing Systems*, 19:601–608, 2006. doi: https://doi.org/10.48550/arXiv.1502.03167.

Hyunho Hwang, Hafiz Zia Ur Rehman, and Sungon Lee. 3D U-Net for Skull Stripping in Brain MRI. *Applied Sciences*, 9(3):569, 2019. doi: https://doi.org/10.3390/app9030569.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning*, 37:448–456, 2015a.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning*, 37:448–456, 2015b. doi: https://doi.org/10.48550/arXiv.1502.03167.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Gene Alexander Paul Thompson, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L Whitwell, Chadwick Ward, Anders M Dale, Joel P Felmlee amd Jeffrey L Gunter, Derek L G Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles S DeCarli, Gunnar Krueger, Heidi A Ward, Gregory J Metzger, Katherine T Scott, Richard Mallozzi, Daniel Blezek, Joshua Levy, Josef P Debbins, Adam S Fleisher, Marilyn Albert, Robert Green, George Bartzokis, Gary Glover, John Mugler, and Michael W Weiner. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Canadian Journal of Cardiology*, 27(4): 685–691, 2008. doi: 10.1002/jmri.21049.

April Khademi, Brittany Reiche, Justin DiGregorio, Giordano Arezza, and Alan R. Moody. Whole volume brain extraction for multi-centre, multi-disease FLAIR MRI datasets. *Magnetic Resonance Imaging*, 66:116–130, 2020. doi: https://doi.org/10.1016/j.cobme.2018.12.005.

April Khademi, Adam Gibicar, Giordano Arezzaa, Justin DiGregorio, Pascal N. Tyrrell, and Alan R.Moody. Segmentation of white matter lesions in multicentre FLAIR MRI. *Neuroimage: Reports*, 1(4):1–29, 2021. doi: https://doi.org/10.1016/j.ynirp.2021.100044.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3441–3440, 2015. doi: 10.1109/CVPR.2015.7298965.

Oeslle Lucena, Roberto Souza, Letícia Rittner, Richard Frayne, and Roberto Lotufo. Convolutional neural networks for skull-stripping in brain MR imaging using silver standard masks. *Artificial Intelligence in Medicine*, 98:48–58, 2019. doi: https://doi.org/10.1016/j.artmed.2019.06.008.

Richard Mayeux and Yaakov Stern. Epidemiology of Alzheimer Disease. *Cold Spring Harbor Perspectives in Medicine*, 2(8):1–18, 2012. doi: 10.1101/cshperspect.a006239.

Samaneh A. Mofrad, Astri J. Lundervold, Alexandra Vik, and Alexander S. Lundervold. Cognitive and MRI trajectories for prediction of Alzheimer's disease. *Scientific Reports*, 11(2122):1–10, 2021. doi: 10.1016/j.media.2018.01.002.

Sridar Narayanan, Kunio Nakamura, Vladimir S Fonov, Josefina Maranzano, Zografos Caramanos, Paul S Giacomini, D Louis Collins, and Douglas L Arnold. Brain volume loss in individuals over time: Source of variance and limits of detectability. *Neuroimage*, 214: 1–5, 2020. doi: 10.1016/j.neuroimage.2020.116737.

Karteek Popuri, Da Ma, Lei Wang, and Mirza Faisal Beg. Using machine learning to quantify structural MRI neurodegeneration patterns of Alzheimer's disease into dementia score: Independent validation on 8,834 images from ADNI, AIBL, OASIS, and MIRIAD databases. *Human Brain Mapping*, 41(14):4127–4147, 2020. doi: 10.1002/hbm.25115.

Joel Ramirez, Melissa F. Holmes, Christopher J. M. Scott, Miracle Ozzoude, Sabrina Adam, Gregory M. Szilagyi, Maged Goubra, Fuqiang Gao, Stephen R. Arnott, Jane M. Lawrence-Dewar, Derek Beaton, Stephen C. Strother, Douglas P. Munoz, Mario Masellis,

Richard H. Swartz, Robert Bartha, Sean Symons, Sandra E. Black, and the ONDRI Investigators. Ontario Neurodegenerative Disease Research Initiative (ONDRI): Structural MRI Methods and Outcome Measures. *Frontiers in Neurology*, 11(847):1–17, 2020. doi: 10.3389/fneur.2020.00847.

Brittany Reiche, Alan R. Moody, and April Khademi. Pathology-preserving intensity standardization framework for multi-institutional FLAIR MRI datasets. *Magnetic Resonance Imaging*, 62:59–69, 2019. doi: 10.1016/j.mri.2019.05.001.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

Hamid Soltanian-Zadeh and Donald J. Peck. Feature space analysis: effects of MRI protocols. *Medical Physics*, 28(11):2344–2351, 2001. doi: 10.1118/1.1414306.

Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016. doi: https://doi.org/10.1016/j.cobme.2018.12.005.

Jean-Claude Tardif, J David Spence, Therese M Heinonen, Alan Moody, Josephine Pressacco, Richard Frayne, Philippe L'allier, Benjamin J W Chow, Matthias Friedrich, Sandra E Black, Aaron Fenster ans Brian Rutt, and Rob Beanlands. Atherosclerosis imaging and the Canadian Atherosclerosis Imaging Network. *Canadian Journal of Cardiology*, 29 (3):297–303, 2013. doi: 10.1016/j.cjca.2012.09.017.

Siddhesh Thakur, Jimit Doshi, Sarthak Pati, Saima Rathore, Chiharu Sako, Michel Bilello, Sung Min Ha, Gaurav Shukla, Adam Flanders, Aikaterini Kotrotsou, Mikhail Milchenko, Spencer Liem, Gregory S. Alexander, Joseph Lombardo, Joshua D. Palmer, Pamela LaMontagne, Arash Nazeri, Sanjay Talbar, and Spyridon Bakas. Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *Applied Sciences*, 220:117081, 2020. doi: https://doi.org/10.1016/j.neuroimage.2020.117081.

Kim-Han Thung, Pew-Thian Yap, Ehsan Adeli, Seong-Whan Lee, and Dinggang Shen. Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion. *Medical Image Analysis*, 45:68–82, 2018. doi: https://doi.org/10.1016/j.media.2018.01.002.

Hyemin Um, Florent Tixier, Dalton Bermudez, Joseph O Deasy, Robert J Young, , and Harini Veeraraghavan. Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets. *Physics in Medicine and Biology*, 64(16):165011, 2019. doi: 10.1088/1361-6560/ab2f44.

Joanna M. Wardlaw, Maria C. Valdés Hernández, and Susana Muñoz-Maniega. What are White Matter Hyperintensities Made of? Relevance to Vascular Cognitive Impairment.

Journal of the American Heart Association, 4(6):1–19, 2015. doi: https://doi.org/10.1161/JAHA.114.001140.

Jiong Wu and Xiaoying Tanga. Brain segmentation based on multi-atlas and diffeomorphism guided 3D fully convolutional network ensembles. *Pattern Recognition*, 115(107904):1–12, 2021. doi: https://doi.org/10.1016/j.patcog.2021.107904.

Appendix A. Supplemental Data

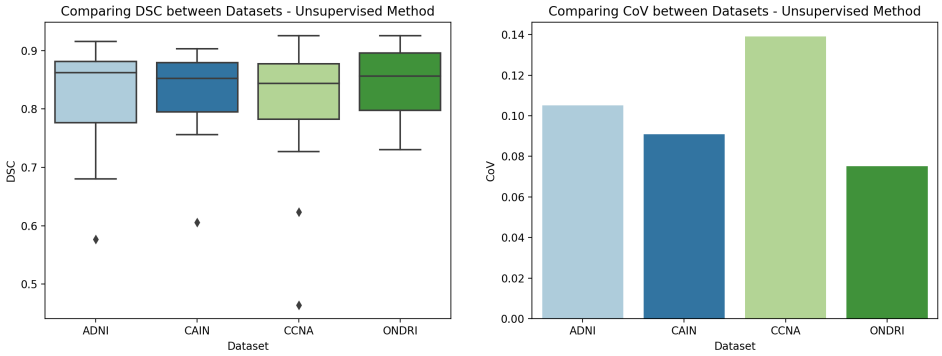

Figure 2: Unsupervised ventricle segmentation (SS masks) on 120 heldout imaging volumes.

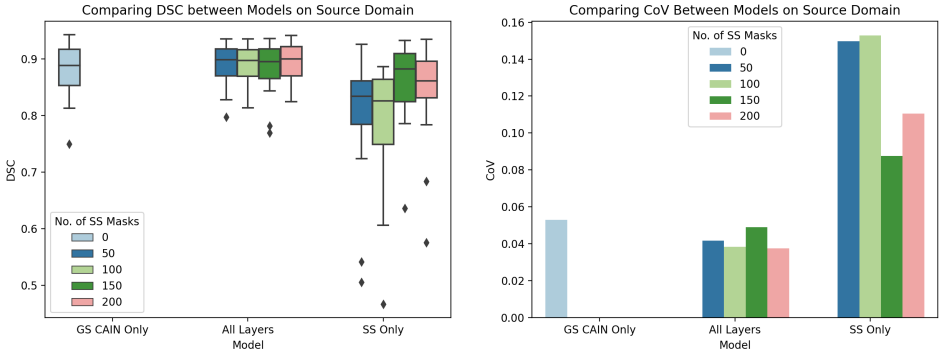

Figure 3: Source domain performance on 30 held-out CAIN imaging volumes.

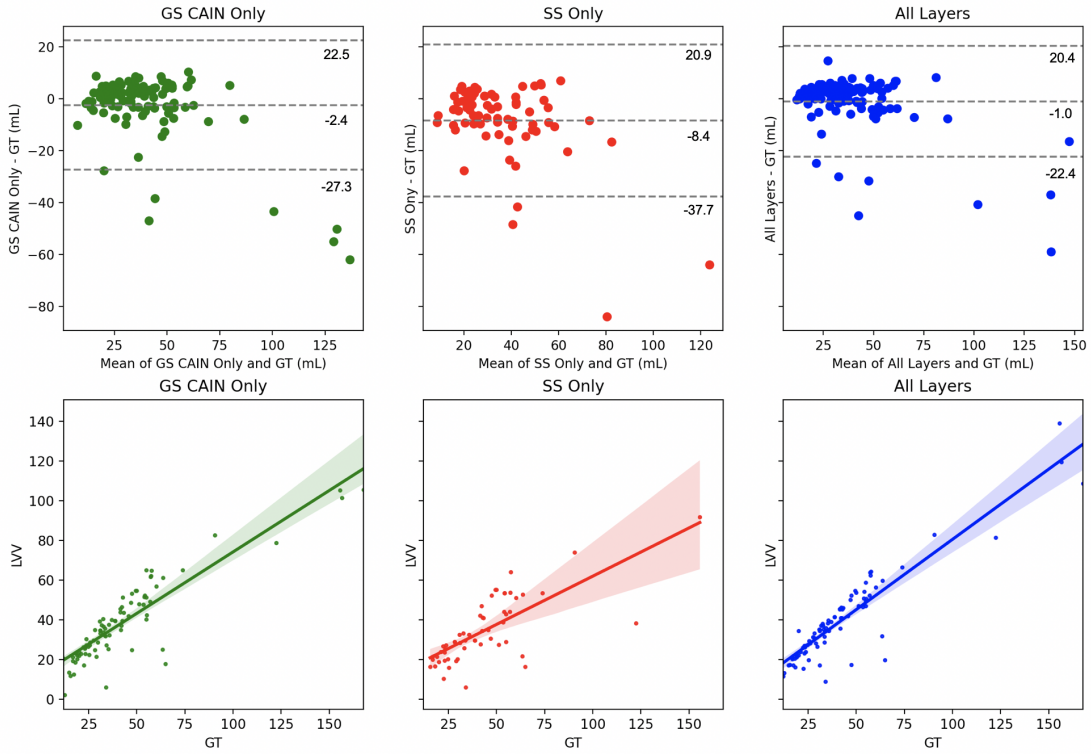

Figure 4: Linear regression and Bland Altman plots between ground truth and predicted LVVs. SS only and All Layers are from models with 100 SS masks.

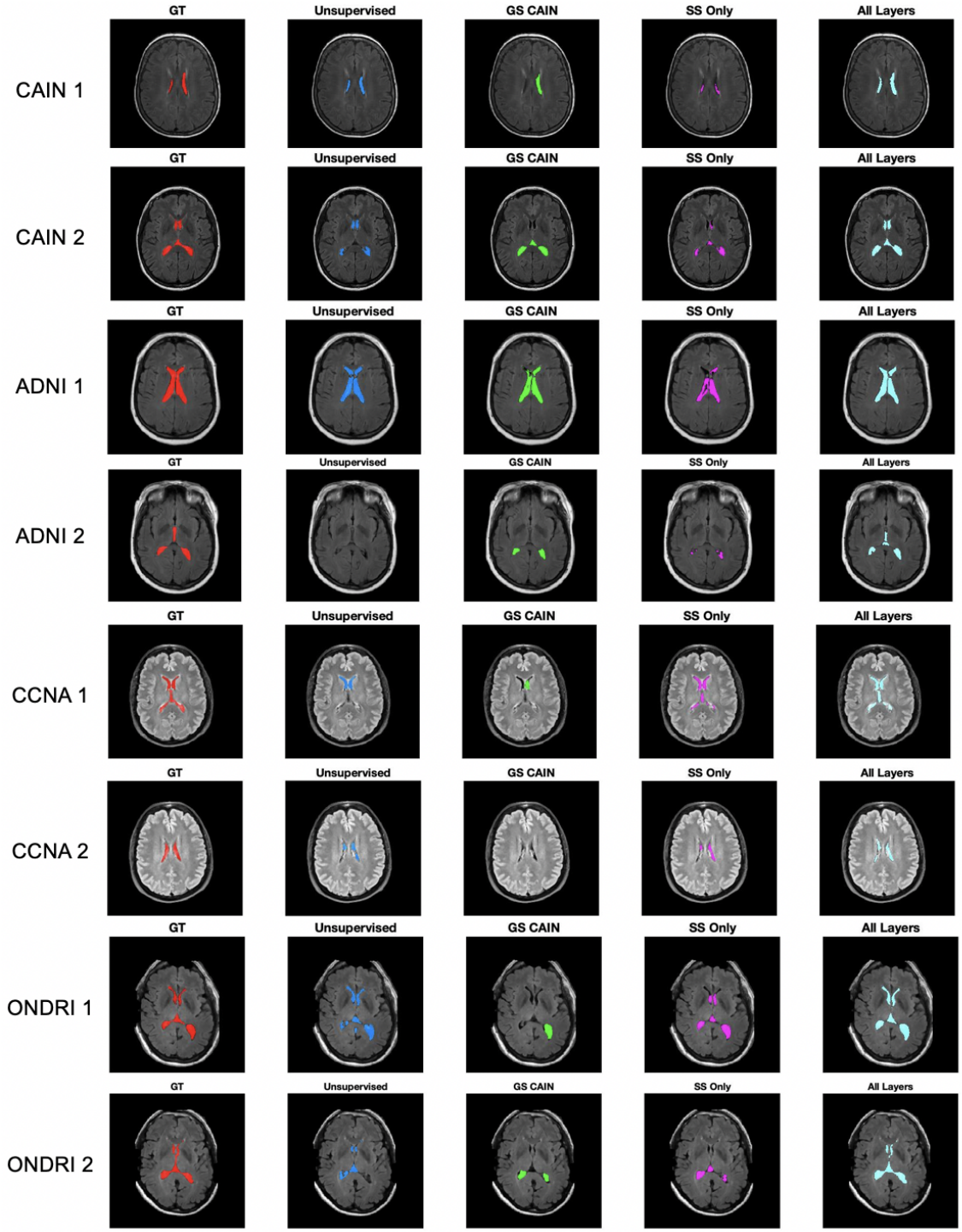

Figure 5: Prediction examples. SS Only and All Layers models trained on 100 SS masks.

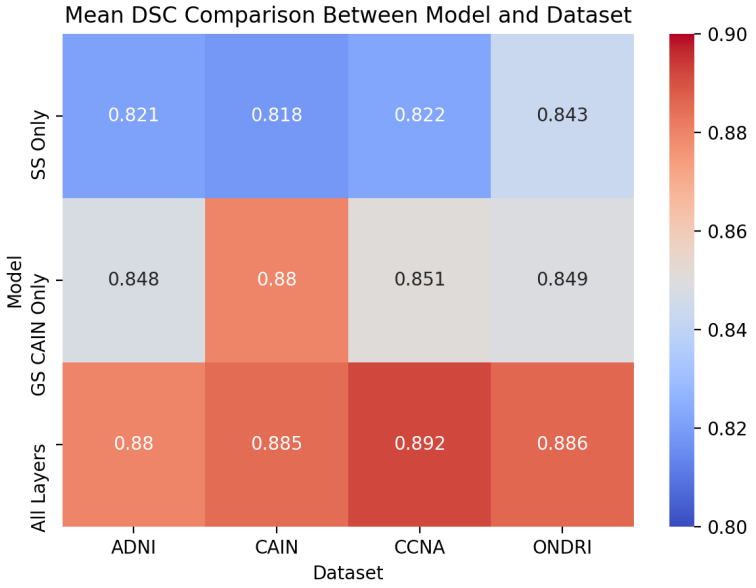

Figure 6: Mean DSC heatmap between models by dataset.