Medical Image Analysis 74 (2021) 102215

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Comparison of domain adaptation techniques for white matter hyperintensity segmentation in brain MR images

Vaanathi Sundaresan[a,b,c,1,*], Giovanna Zamboni[a,d,e], Nicola K. Dinsdale[a,b], Peter M. Rothwell[d], Ludovica Griffanti[a,f,2], Mark Jenkinson[a,g,h,2]

[a] Wellcome Centre for Integrative Neuroimaging, Oxford Centre for Functional MRI of the Brain, Nuffield Department of Clinical Neurosciences, University of Oxford, UK
[b] Oxford-Nottingham Centre for Doctoral Training in Biomedical Imaging, University of Oxford, UK
[c] Oxford India Centre for Sustainable Development, Somerville College, University of Oxford, UK
[d] Centre for Prevention of Stroke and Dementia, Nuffield Department of Clinical Neurosciences, University of Oxford, UK
[e] Dipartimento di Scienze Biomediche, Metaboliche e Neuroscienze, Università di Modena e Reggio Emilia, Italy
[f] Wellcome Centre for Integrative Neuroimaging, Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford, Oxford, UK
[g] Australian Institute for Machine Learning (AIML), School of Computer Science, The University of Adelaide, Adelaide, Australia
[h] South Australian Health and Medical Research Institute (SAHMRI), Adelaide, Australia

ARTICLE INFO

Article history:
Received 15 March 2021
Revised 12 July 2021
Accepted 16 August 2021
Available online 17 August 2021

Keywords:
Deep learning
White matter hyperintensities
Domain adaptation
Segmentation

ABSTRACT

Robust automated segmentation of white matter hyperintensities (WMHs) in different datasets (domains) is highly challenging due to differences in acquisition (scanner, sequence), population (WMH amount and location) and limited availability of manual segmentations to train supervised algorithms. In this work we explore various domain adaptation techniques such as transfer learning and domain adversarial learning methods, including domain adversarial neural networks and domain unlearning, to improve the generalisability of our recently proposed triplanar ensemble network, which is our baseline model. We used datasets with variations in intensity profile, lesion characteristics and acquired using different scanners. For the source domain, we considered a dataset consisting of data acquired from 3 different scanners, while the target domain consisted of 2 datasets. We evaluated the domain adaptation techniques on the target domain datasets, and additionally evaluated the performance on the source domain test dataset for the adversarial techniques. For transfer learning, we also studied various training options such as minimal number of unfrozen layers and subjects required for fine-tuning in the target domain. On comparing the performance of different techniques on the target dataset, domain adversarial training of neural network gave the best performance, making the technique promising for robust WMH segmentation.

1. Introduction

White matter hyperintensities of presumed vascular origin (WMHs, also known as white matter lesions) are bright localised regions on T2-weighted and FLAIR images. They are commonly found in elderly subjects, however, they have also been related to various neurodegenerative (e.g. dementia, including Alzheimer's disease) and cerebrovascular diseases (e.g. stroke) (Wardlaw et al., 2013). Automated WMH segmentation is essential for further understanding the clinical impact of WMHs in a large population. Various methods using hand-crafted features have been used for WMH segmentation (Caligiuri et al., 2015), and in recent years, deep learning (DL) models are being increasingly used and have been shown to outperform traditional methods (Rachmadi et al., 2018; Kuijf et al., 2019). Many of the existing methods (using either hand-crafted features or DL models) were trained with a large amount of manual labels (Wang et al., 2012; Admiraal-Behloul et al., 2005; Ghafoorian et al., 2016) and/or evaluated on specific population group (Wang et al., 2012; Gibson et al., 2010; De Boer et al., 2009; Steenwijk et al., 2013; Jeon et al., 2011; Hong et al., 2020; Park et al., 2018), acquired with the same scanner/protocol or validated on isotropic or axial acquisition images (Ghafoorian et al., 2017a; Kuijf et al., 2019). However, in the real-world scenario, most of the clinical datasets are small in size, acquired using various protocols and scanners, and from people with diverse de-

* Corresponding author.
E-mail address: vaanathi.sundaresan@ndcn.ox.ac.uk (V. Sundaresan).
[1] https://www.ndcn.ox.ac.uk/team/vaanathi-sundaresan
[2] Contributed equally to this work.

https://doi.org/10.1016/j.media.2021.102215

mographic and pathological characteristics. In addition, specially in these datasets, limited amount or non-availability of manual segmentations constrains the training and segmentation performance of the model, especially due to the problem of overfitting. It is therefore very challenging to achieve robust performance metrics for segmentation of WMHs across datasets in the presence of such variations in image characteristics, lesion load, and availability of training data.

Several methods have been proposed for making models more adaptable to various '*domains*' (e.g. different scanners or acquisition protocols). These include reducing the variance in the image-level characteristics (Bordin et al., 2020) (induced by the scanner and acquisition protocol), estimating site effects to correct the measurements derived from the images (Fortin et al., 2018), by improving model generalisability (Ganin et al., 2016; Tzeng et al., 2015) (so that it is not affected by differences in intensity distributions or spatial resolution), or a combination of the above. Commonly used techniques to improve model generalisability include data augmentation (Shorten and Khoshgoftaar, 2019), and the use of ensemble networks (with different initialisations (Li et al., 2018) or planes (Prasoon et al., 2013)), which have been shown to be resistant to over-fitting (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Kamnitsas et al., 2017; Winzeck et al., 2019), which can occur with more complex models (Opitz and Maclin, 1999). However, these techniques cope mostly with minor variance in dataset characteristics within a domain and hence might not be sufficient for generalising across datasets obtained from different sources/domains.

Domain adaptation (DA) methods address the issue of discrepancies in the data distributions obtained from various domains that affect the robust performance of the model (Ben-David et al., 2010). DA methods, in general, aim to transfer the knowledge from a *source* domain to a *target* domain by leveraging the invariant features across different domains (Wilson and Cook, 2019; Pan and Yang, 2009). Various DA techniques used so far include minimising a distance metric of domain variance (Long et al., 2013; Pan et al., 2010; Wang and Schneider, 2014), using transferable features for creating intermediate feature representation between domains (Yosinski et al., 2014) and transfer learning (Pan and Yang, 2009; Yosinski et al., 2014). Within DA frameworks, the restriction posed by limited availability of manually labelled data for training has been addressed by proposing various semi-supervised (Cheng and Pan, 2014; Yao et al., 2015; Saito et al., 2019), self-labelling (Saito et al., 2017; Zou et al., 2019) and pseudo-labelling (Inoue et al., 2018) methods. Techniques such as self- and pseudo-labelling use small amount of labelled data along with large amount of unlabelled data to improve model performance (Lee et al., 2013). Hence, given the wide variations in lesion characteristics, contrast variations (e.g. GM voxels vs WMHs) and location priors (e.g. normal ventricle lining vs WMHs), these techniques could bias WMH segmentation results, especially in small non-representative datasets.

In *transfer learning* (TL) (Pan and Yang, 2009), one of the commonly used supervised DA techniques, the initial convolutional layers (domain invariant low-level features) are generally kept constant or *frozen*, while the final layers (task/domain specific high-level features) are fine-tuned on the target datasets. Fine-tuning of the pre-trained models has been shown to improve the performance on the target datasets (Tajbakhsh et al., 2016), especially when the intensity characteristics of images between domains are more similar (Yosinski et al., 2014; Wilson and Cook, 2019). TL has been applied for medical image segmentation tasks (Tajbakhsh et al., 2016), including lesion segmentation (Ghafoorian et al., 2017b). However, TL is limited by the fact that the training on different domains occurs separately and hence cannot combine features from both domains while training. Also, in addition to determining the layers to fine-tune, another crucial consideration often encountered while fine-tuning is the number of target training subjects required. This is because the performance of TL has been shown to rely (although to a less extent than training the model from scratch) on the amount of labelled target training data (Tajbakhsh et al., 2016).

Another successful DA approach, *unsupervised domain adversarial training* (Ganin et al., 2016; Tzeng et al., 2015; Lee et al., 2019; Fernando et al., 2013; Kouw and Loog, 2019; Wang and Deng, 2018) relies on domain invariant features to achieve good domain adaptation. Several adversarial training methods have been proposed, including the recent ones based on discriminator framework (Tzeng et al., 2017), partial transfer learning (Cao et al., 2018) (assuming that the target domain dataset is a subset of the source domain) and using associations between the source and target domains (e.g. increasing correlation/covariance, subspace alignment) (Haeusser et al., 2017; Fernando et al., 2013; Long et al., 2017a; Sun and Saenko, 2016). One of the earlier and commonly used adversarial training approaches, domain adversarial training of neural network (DANN) (Ganin and Lempitsky, 2015; Ganin et al., 2016) has been applied to several baseline architectures (Ganin et al., 2016; Schoenauer-Sebag et al., 2019) and is explored on multiple datasets (Gallego et al., 2020). The DANN technique has also been used in various comparative analyses (Tzeng et al., 2015; 2017; Zellinger et al., 2017) and has been proven to be one of the successful models for task-specific domain adaptation (Zellinger et al., 2017; Long et al., 2017b). The DANN model consists of a feature extractor network with a domain predictor and a label predictor. The adversarial training of the domain predictor is achieved using a *gradient-reversal layer*, placed between the feature extractor and the domain predictor, optimising the features and shared weights. This layer maximises the domain prediction loss, thus minimising the shift between the domains, while simultaneously making the model discriminative towards the main task of segmentation label prediction. While DANN was originally proposed in an unsupervised manner with respect to target domain (i.e. segmentation labels required only for source domain), it has also been shown to be beneficial under semi-supervised setting (i.e. using a fraction of target domain segmentation labels) (Ganin et al., 2016).

An alternative domain unlearning (DU) approach was proposed for domain and task adaptation using an iterative framework (Tzeng et al., 2015) (and recently adapted for unlearning scanner-related information between domains in (Dinsdale et al., 2020)). The method involved learning the domain prediction for a fixed feature representation and then minimising the domain shift between features resulting in a maximal domain confusion that is equally uninformative across domains.

In this work, we explore various domain adaptation techniques such as TL and adversarial adaptation methods including DANN and DU for obtaining a good WMH segmentation across various datasets, and to perform well irrespective of differences in data characteristics. We used a triplanar ensemble network (TrUE-Net), proposed in our recent work (Sundaresan et al., 2021) as a baseline model. Our objective is to adapt our baseline model to a different domain consisting of small dataset(s). In addition, when applying TL, we addressed two main issues while fine-tuning a model on a target dataset with limited training subjects: determining (1) the optimal layer of the model to start fine-tuning and (2) the minimal number of training subjects for reliable segmentation. We performed our experiments on 3 different datasets (including a publicly available dataset) with different acquisition and lesion characteristics, grouped into source and target domains. We experimented with several test strategies involving different DA techniques on the source-trained model and training the model directly on the target domain, to comprehensively study both innate and adapted performances of the model for WMH segmentation.

2. Materials and methods

2.1. Datasets used

Neurodegenerative cohort (NDGEN): The dataset, used in Zamboni et al. (2013), includes MRI data from 9 subjects with probable Alzheimer's Disease, 5 with amnestic mild cognitive impairment and 7 cognitively healthy control subjects (age range 63 - 86 years; mean age 77.1 ± 5.8 years; median age 77 years; F:M = 10:11). Total brain volume range: 1189282 - 1614799 mm^3, median: 1424669 mm^3. Manual segmentation was available for all datasets (WMH load range: 1878 - 89259 mm^3, median: 20772 mm^3). The images were acquired using a 3T Siemens Trio Scanner, with FLAIR (TR/TE = 9000/89 ms, flip angle 150°, FOV 220 mm, voxel size 1.1 × 0.9 × 3 mm, matrix size 256 × 256 × 35 voxels) and T1-weighted acquisitions (3D MP-RAGE sequence, TR/TE = 2040/4.7 ms, flip angle 8°, FOV 192 mm, voxel size 1 mm isotropic, matrix size 174 × 192 × 192 voxels).

Vascular cohort - Oxford Vascular Study (OXVASC): The dataset consists of 18 participants in the OXVASC study (Rothwell et al., 2004), who had recently experienced a minor non-disabling stroke or transient ischemic attack (age range 50 - 91 years; mean age 73.27 ± 12.32 years; median age 75.5 years; F:M = 7:11). Total brain volume range: 1290926 - 1918604 mm^3, median: 1568233 mm^3. Manual segmentation was available for all datasets (WMH load range: 3530 - 83391 mm^3, median: 16906 mm^3). The images were acquired using a 3T Siemens Trio Scanner, with FLAIR (TR/TE = 9000/88 ms, flip angle 150°, voxel size 1 × 3 × 1 mm, matrix size 174 × 52 × 192 voxels) and T1-weighted acquisitions (3D MP-RAGE sequence, TR/TE = 2000/1.94 ms, flip angle 8°, voxel size 1 mm isotropic, matrix size 208 × 256 × 256 voxels).

MICCAI WMH Segmentation Challenge training Dataset (MWSC): The dataset consists of 60 subjects from three different sources (20 subjects each) provided as training sets for the challenge (Kuijf et al., 2019) (http://wmh.isi.uu.nl/): UMC Utrecht, NUHS Singapore and VU Amsterdam. The brain volume ranges: 1257820 - 1844920 mm^3 (median 1473389 mm^3) for UMC Utrecht, 1147248 - 1532268 mm^3 (median: 1351325 mm^3) for NUHS Singapore and 1219614 - 1787321 mm^3 (median: 1441201 mm^3) for VU Amsterdam. Manual segmentations were available for all three datasets, with an additional exclusion label provided for other pathologies. We included these masks as parts of non-lesion tissue during both training and testing. The WMH volume ranges (excluding other pathologies) are 845 - 74991 mm^3 (median: 26240 mm^3) for UMC Utrecht, 786 - 61332 mm^3 (median: 17795 mm^3) for NUHS Singapore and 1522 - 43528 mm^3 (median: 6015 mm^3) for VU Amsterdam. FLAIR and T1-weighted images were available for this dataset (for more details regarding MRI acquisition parameters, refer to http://wmh.isi.uu.nl/). Even though preprocessed images were available, we used the original images and applied the preprocessing pipeline specified in Section 2.2 to maintain consistency and to avoid any biases due to the preprocessing method across domains in our experiments.

2.2. Data preprocessing

For all datasets, we reoriented FLAIR and T1-weighted images to the standard MNI space, performed skull-stripping with FSL BET (Smith, 2002) and bias field correction using FSL FAST (Zhang et al., 2001). We registered the T1-weighted image to the FLAIR using rigid-body registration using FSL FLIRT (Jenkinson and Smith, 2001) and cropped the field of view (FOV) close to the brain and applied Gaussian normalisation to the intensity values. For axial, sagittal and coronal slices, we resized the extracted slices to dimensions of 128 × 192, 192 × 120 and 128 × 80 voxels respectively, using bilinear interpolation.

2.3. Baseline method: triplanar U-Net Ensemble Network (TrUE-Net) architecture

As a baseline model, we used the triplanar ensemble architecture[3] proposed in Sundaresan et al. (2021). As shown in our prior work (table 5 in Sundaresan et al. (2021)), TrUE-Net provided results on par with the top performing methods of MWSC challenge (Kuijf et al., 2019) and with the method proposed in Ghafoorian et al. (2017a). Briefly, as shown in Fig. 1, the TrUE-Net architecture consists of three 2D U-Nets, one for each plane, taking FLAIR and T1 slices as input channels. We trimmed the depth of the classic U-Net (Ronneberger et al., 2015) in each plane to a depth of 3-layers (Fig. 2a), given the small size of lesions. In the ensemble model, we trained the U-Nets in each plane independently using 2D slices extracted in each plane. We used a combination of weighted cross-entropy (refer to Sundaresan et al. (2021) for more details) and Dice loss functions in order to overcome the effect of class imbalance between WMHs and healthy tissue. During testing, the predictions were obtained as 2D softmax output score maps for slices in each plane and were later assembled into 3D volumes and resized to the original dimensions. We then averaged the 3D volumes to get the final probability volume for the triplanar architecture.

2.4. Comparison of domain adaptation techniques

We studied the performance of various DA techniques using the following test strategies on the target test dataset (refer to Section 2.6) against the model that is trained directly on the target training dataset.

2.4.1. Strategy 1: train on the source domain and apply directly to the target domain

In order to determine the inherent generalisability of TrUE-Net, we trained the model on the source domain training datasets (training parameters in Section 2.5) and tested the model directly on the target domain test datasets.

2.4.2. Strategy 2: transfer the model trained on the source domain to the target domain with fine-tuning

We trained TrUE-Net (training details in Section 2.5) on the source domain datasets to get the source pre-trained model. We then fine-tuned the model by training it on the target dataset starting from the decoder end. For fine-tuning we used a smaller learning rate schedule (initially 1×10^{-4}, reduced by a factor 1×10^{-1} every 2 epochs, until it reaches 1×10^{-6}). Fig. 2a shows the layer numbers for the U-Net model. Given L layers in total, 'fine-tuning i layers' means that $L - i$ layers before i were frozen and the layers from i towards the decoder end were fine-tuned. The initial hyperparameter tuning (explained in Section 2.5) was performed using 18 subjects and fine-tuning layers starting from the end of encoder (3 layers from the end). Hence, we compared the results at this setting with other DA strategies. Additionally, for each fine-tuning, we increased the number of target training subjects, from 2 to 18 in steps of 2, and measured the performance of each fine-tuned model. Finally, we determined the best starting point for fine-tuning the model, and the optimal number of training data to obtain the best performance on the target dataset. Since TL involves both the domains, in addition to the existing DA strategies, we also compare the TL strategy with the case where the baseline

[3] TrUE-Net code available in https://git.fmrib.ox.ac.uk/vaanathi/truenet

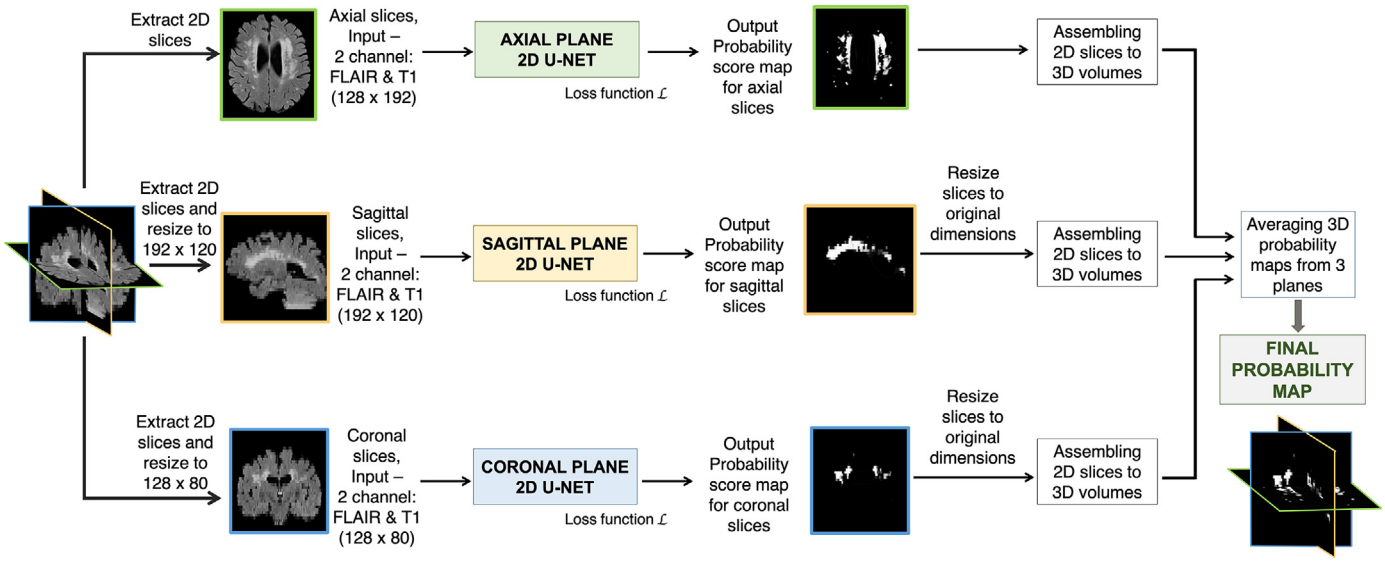


Fig. 1. Baseline architecture: Triplanar U-Net ensemble network (TrUE-Net) Sundaresan et al. (2021).

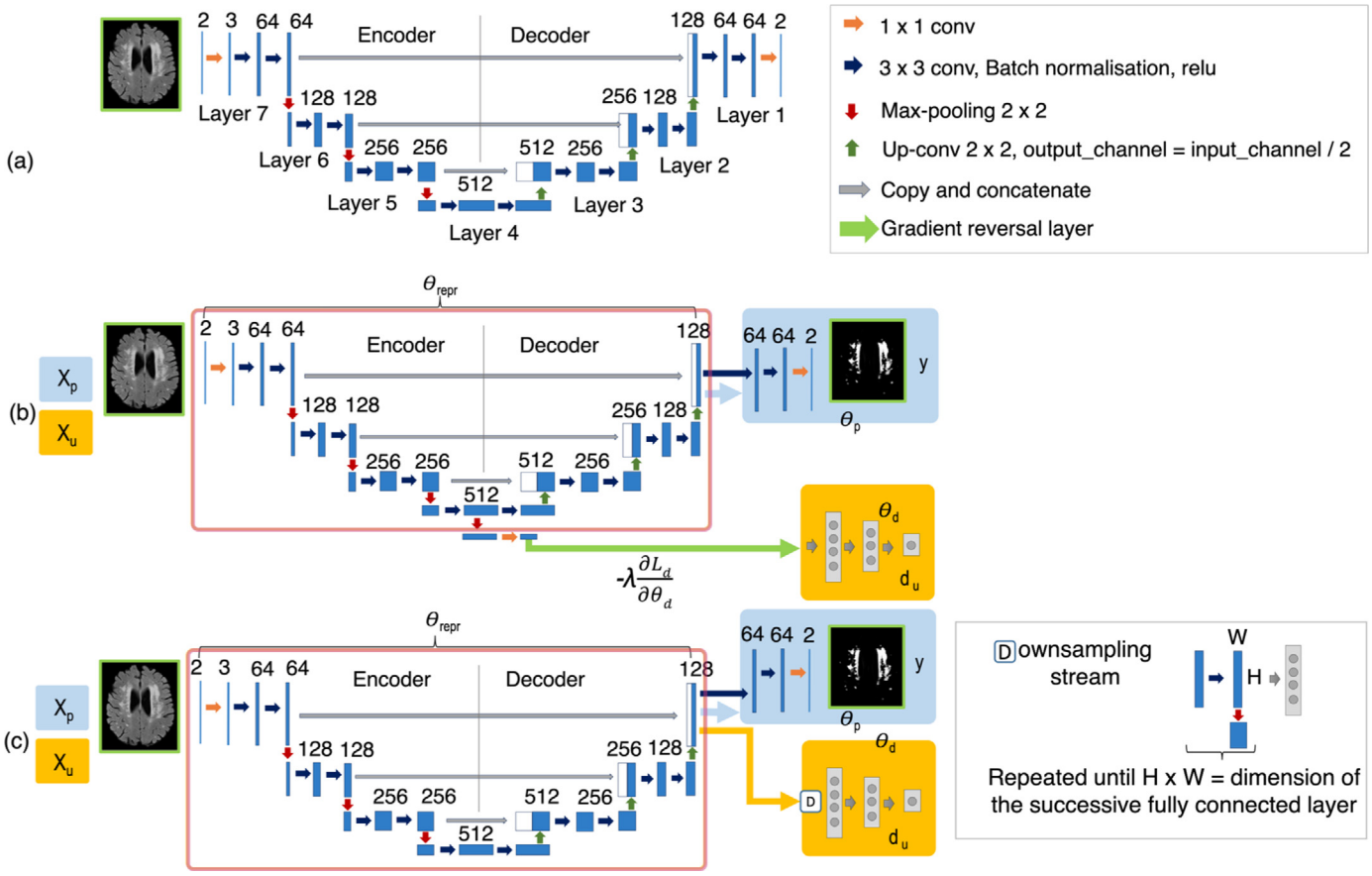
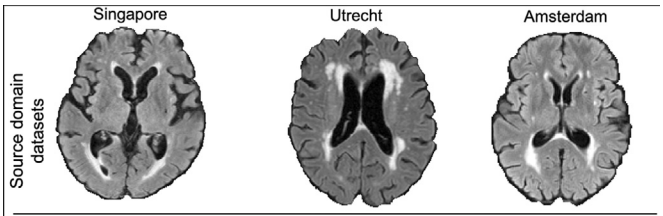


Fig. 2. Transfer learning (TL) framework, domain adversarial neural network (DANN) and domain unlearning (DU) architectures. (a) Layer numbers indicated on the baseline model for the TL strategy (numbered from decoder end indicating the order of fine-tuning), (b) DANN and (c) DU architectures, illustrating feature extractor (red box), lesion label predictor (blue) and domain predictor (orange) with corresponding training parameters θ_{repr}, θ_p and θ_d. The models take input features X_p and input domain information X_u and predicts output labels y, while unlearning output domains d_u. The DU model updates the label predictor, feature extractor and domain predictor in a sequential manner, while label prediction and domain unlearning occur simultaneously in DANN. For all the cases, only the axial U-Net is shown; note that sagittal and coronal models were modified in a similar manner. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



	Source	Target
Training dataset	Singapore (10) + Utrecht (10) + Amsterdam (10)	OXVASC (8) + NDGEN (10)
Validation set	Singapore (2) + Utrecht (2) + Amsterdam (2)	OXVASC (2) + NDGEN (2)
Test dataset	Singapore (8) + Utrecht (8) + Amsterdam (8)	OXVASC (8) + NDGEN (8)

Fig. 3. Sample axial slices shown from source (top row) and target (middle row) domains. The splits for training, validation and test datasets in source and target domains are provided in the table (bottom panel).

model is trained on the source and target training datasets combined together (refer to section 5 in the supplementary material).

2.4.3. Strategy 3: unsupervised domain adversarial training (DANN) on the source and target domains

We implemented DANN (Ganin et al., 2016) as shown in Fig. 2b by adding a domain predictor to the baseline TrUE-Net model. We added the domain predictor to the coarsest level after maximum levels of pooling (with 512 channels) at the end of the encoder, since it has high-level features with domain-specific information. We added a 2×2 max-pooling layer and 1×1 projection layers to 128 and 64 channels before the domain predictor. The domain predictor consists of three fully connected (FC) layers (with 1024, 512 and 32 nodes) alternating with two dropout layers ($P_{drop} = 0.2$), followed by a softmax layer. We added a gradient reversal layer between the feature extractor and the domain predictor, leading to adversarial training with respect to domain prediction. In this model, the domain predictor makes the model domain invariant by considering data from both domains, while the lesion label predictor optimises the model for accurate WMH segmentation. Hence, the domain predictor requires only domain labels for both source and target datasets, while the lesion label predictor requires manual segmentations from source datasets only, making the model unsupervised with respect to the target domain. We tested the DANN model on the test datasets from the source and target domains (table in Fig. 3), and measured model performance in each domain individually.

2.4.4. Strategy 4: semi-supervised domain adversarial training (semi-DANN) on source and target domains

We trained the DANN model (Fig. 2b) in a semi-supervised manner, wherein manual segmentations from a fraction of target training data were used in addition to source training datasets for training the lesion label predictor. The remaining target training data is used only for domain prediction. One of the main advantages of the unsupervised DANN model is that it does not require manually labelled data from the target dataset. Even then, our main aim for exploring the semi-DANN, despite the additional manual labelling effort, was to observe if there was any significant improvement over unsupervised DANN. Hence, we used a minimal proportion (25%, chosen empirically, which amounts to 4 subjects) of the labelled target training data, in addition to the source training dataset, for training the lesion label predictor. We tested the model on the source and target domain test datasets individually.

2.4.5. Strategy 5: iterative domain unlearning (DU) to remove scanner-bias between source and target domains

The DU model[4] (Dinsdale et al., 2020) is based on the iterative unlearning framework (Tzeng et al., 2015) for adversarial adaptation. The model, rather than using a gradient reversal layer, optimises two opposing loss functions in three sequential steps: (1) updating the feature representation and the lesion label predictor, (2) maximising the performance of a domain predictor given the fixed feature representation, and (3) updating the feature representation in order to maximally confuse the domain classifier. As in the unsupervised DANN (strategy 3), only domain labels are required for the target dataset, while the lesion label predictor uses manual segmentations of the source dataset only. As shown in Fig. 2c, we consider the final two 3×3 convolutional layers and the final softmax layer as our label predictor. The domain predictor, placed after the final decoder layer of the U-Net, consists of repetitive 2×2 max-pooling layers until the last layer dimensions match the first FC layer, followed by three FC layers (with 468, 96 and 32 nodes) alternating with two dropout layers ($P_{drop} = 0.2$), followed by a softmax layer. Note that while we added the domain predictor at the end of the encoder in DANN, we added the domain predictor at the end of decoder (the same point as label predictor) in DU (Fig. 2). This is because, in DANN, the domain unlearning happens simultaneously with shared weights between the predictors and hence we focus on the layer with coarsest features (showing overall WMH distribution) that is more domain specific. On the other hand, in DU, the training happens sequentially for each predictor (while freezing the other predictor) and hence we add the domain predictor directly at the point where the generalisability is most desirable for WMH segmentation.

2.4.6. Train on the target domain from scratch and apply to the target domain

We trained the TrUE-Net model on the target training dataset and tested the model on the target test dataset. This case is expected to perform better than source-trained and other DA strategies on the target test dataset (for the given training options and data) since it is not required to cope with domain variance. Hence, this represents the case of upper limit for the performance metrics and is included for reference purpose only, since it does not improve model generalisability across domains.

2.5. Implementation details

We implemented the networks in Python 3.6 using Pytorch 1.2.0. The baseline network (TrUE-Net), for source-trained, target-trained and TL strategies (pretraining and fine-tuning using 3-layers, 18 subjects), was trained on an NVIDIA Tesla V100, taking 5 mins (for 3 planes) per epoch for $\approx$ 15,000 samples with the training/validation split of 90/10%. We used the Adam Optimiser with $\epsilon = 10^{-4}$. We empirically chose a batch size of 8, and an initial learning rate of 1×10^{-3} and reducing it by a factor 1×10^{-1} every 2 epochs, until it reaches 1×10^{-5}, after which we maintain a fixed learning rate value (for more details, refer to Sundaresan et al. (2021)). Data augmentation was applied in an online manner using translation (x/y-offset $\in$ [-10, 10]), rotation ($\theta \in$

[4] DU code available at: https://github.com/nkdinsdale/Unlearning_for_MRI_harmonisation

[-10, 10]), random noise injection (Gaussian, $\mu = 0$, $\sigma^2 \in [0.01, 0.09]$) and Gaussian filtering ($\sigma \in [0.1, 0.3]$), increasing the dataset by a factor of 10 and 6 for axial and sagittal/coronal planes respectively. The hyperparameter values for the data augmentation transformations were randomly sampled from the closed intervals specified above using a uniform distribution. Additionally, for the domain predictor in DANN/semi-DANN and DU, we trained with the Momentum optimiser (momentum value of 0.9) and Adam optimiser respectively. We used a batch size of 8, with 50 epochs for pretraining and a criterion based on a patience value (number of epochs to wait for progress on validation set) of 25 epochs to determine model convergence for early stopping (converged at around 90 epochs for all cases). We used the learning rates of 10^{-3} and 10^{-4} for DANN/semi-DANN and DU respectively. These training hyper parameters were chosen empirically. In DU, we used a β value of 50 (a factor used for weighting the domain confusion (Dinsdale et al., 2020)). For determining the β value, we experimented with different values starting from 20 to 60 in steps of 10 and chose the value of 50, since it provided a domain accuracy value closer to 50% (indicating maximal confusion of domains) on the validation dataset. The DANN/semi-DANN and DU networks were trained on an NVIDIA Tesla V100, taking $\approx$ 10 and 7 mins per epoch respectively with the training/validation split of 90/10%.

2.6. Source and target domain datasets

The datasets used in this work were acquired using different scanners and sequences and therefore have different intensity characteristics and resolutions. For performing our domain adaptation (DA) experiments, we classified the available datasets into two domains. Rather than considering each dataset as an individual domain, we considered only two domains (source and target) for our experiments. This is because the datasets have varying degrees of similarity among them and also, given the limited number of subjects for each dataset (for training and testing), treating them as individual domains would be difficult and give unreliable results. For deciding the source and target datasets, we determined the homogeneity of image-level characteristics among the above 5 datasets, using a domain discriminator network. To this aim, we trained a domain discriminator model on the above datasets and determined the domain misclassifications among these datasets using a confusion matrix (for more details on this experiment and results, refer to Section 1 of supplementary material). Based on the results, we considered the MWSC dataset (3 cohorts) as our *source* domain datasets, and the combination of OXVASC and NDGEN as our *target* domain datasets. Examples of the source and target domain datasets, along with the training/validation/test data split for above test strategies is shown in Fig. 3.

2.7. Performance evaluation metrics

We used the following performance metrics:

- Dice Similarity Index (SI) = 2 × (true positive WMH voxels) / (true WMH voxels + positive WMH voxels).
- Voxel-wise true positive rate (TPR), the ratio of the number of true positive WMH voxels to the number of true WMH voxels.
- Voxel-wise false positive rate (FPR), the number of false positive WMH voxels divided by the number of non-WMH voxels.
- Cluster-wise TPR, the number of true positive WMH clusters (determined using 26-connected neighbourhood) divided by the total number of true WMH clusters.
- Absolute log-transformed volume difference (lAVD), which is defined by $lAVD = \left| log \frac{predicted\ segmentation\ volume}{manual\ segmentation\ volume} \right|$
- Cluster-wise F1-measure = 2 × (cluster-wise TPR × cluster-wise precision)/ (cluster-wise TPR + cluster-wise precision), where cluster-wise precision is the number of true positive WMH clusters divided by the total number of detected WMH clusters.

For each metric we determined the significant differences between the individual pairs of test strategies, correcting for multiple comparisons using Permutation Analysis of Linear Models (PALM) (Winkler et al., 2014).

3. Results

Figure 4 shows results for all strategies of the domain adaptation experiments for a sample high lesion load test subject from the OXVASC dataset (results on a low lesion load test subject are shown in figure S3 in the supplementary material). Figure 5 shows the boxplots of the performance metrics, while suppl. table S1 reports medians and interquartile ranges of performance metrics for different test strategies. For PALM results comparing the evaluation metrics between each pair of test strategies, refer to suppl. table S2. For the DANN, semi-DANN and DU models, Fig. 7 shows the visualisation of the feature representations using t-SNE plots (Van der Maaten and Hinton, 2008) at the layer before the label predictor with and without domain adaptation.

3.1. Strategy 1: train on the source domain and apply directly to the target domain

As shown in Fig. 4a, while the source-trained model detected most of the periventricular WMHs (PWMHs), it undersegmented their boundaries, and also missed some deep WMHs (DWMHs). From the boxplots in Fig. 5 (and suppl. table S1), strategy 1 showed the worst performance among all the cases. The segmentation was more precise in the low lesion load subjects (figure S3 in suppl. material) rather than high lesion load ones. This strategy showed significantly lower SI, cluster-wise TPR and F1 values compared to semi-DANN (significant even after correcting for multiple comparison across metrics). Also, cluster-wise TPR and F1-measures were significantly lower when compared to DANN and DU (strategies 3 and 5) respectively (suppl. table S2).

3.2. Strategy 2: transfer the model trained on the source domain to the target domain with fine-tuning

The TL strategy results are reported for the setting (using 18 subjects, starting from layer-3) that was used for tuning training hyperparameters (e.g. learning rate, optimiser parameters, batch size etc.). With this setting, TL provided better performance than strategy 1 for all the evaluation metrics. However, it gave lower cluster-wise F1 measure values and higher lAVD values when compared to other DA models. The difference in cluster-wise F1-measure was significant when compared to semi-DANN, as reported in suppl. table S2.

Later while determining the best setting for TL, the segmentation results improved with increased amounts of training data (Fig. 4). For instance, segmentation results with fine-tuning using 2 training subjects showed the worst results for both lesion loads, and improved with 12 and 18 training subjects. This is also evident from the heatmaps of the performance metrics shown for different numbers of training subjects and numbers of fine-tuned layers in Fig. 6. All the performance metrics showed best values for the training size of 18 subjects. The performance metrics were generally slightly lower when fewer layers in the decoder arm were fine-tuned, and also when there was less training data. However, the performance really started to increase

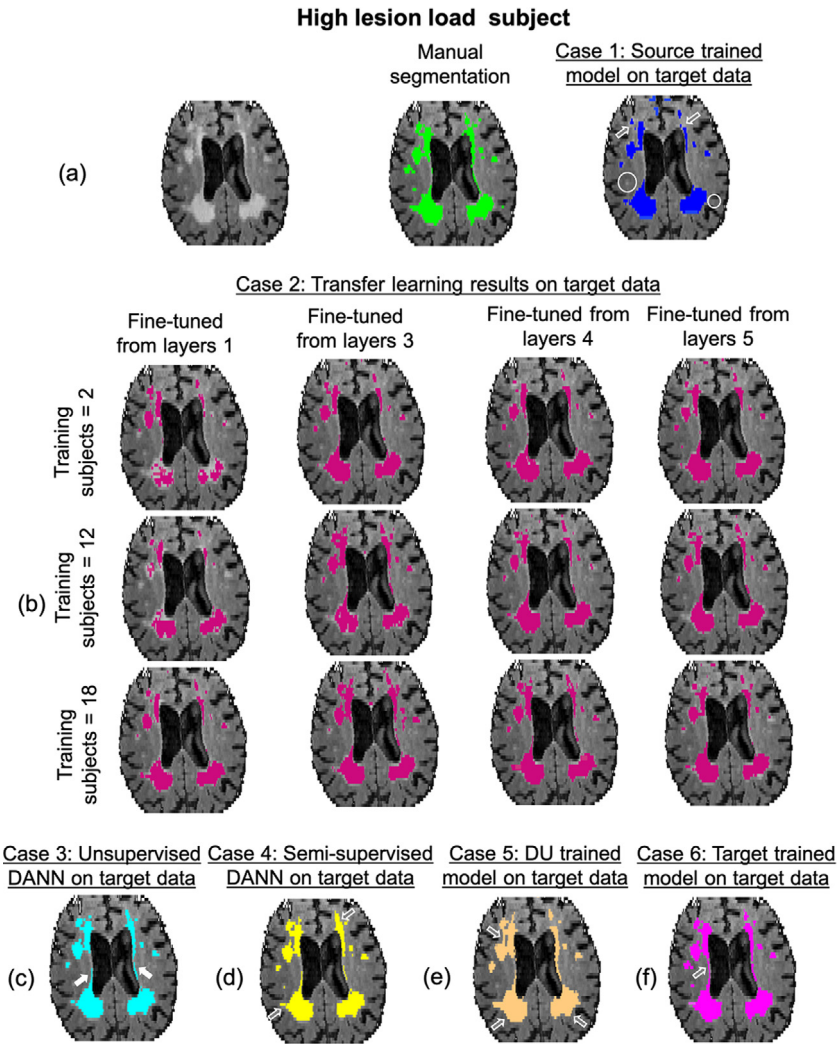


Fig. 4. Sample results of domain adaptation experiment test strategies: (a) Source-trained model, (b) TL models, (c) unsupervised DANN, (d) semi-supervised DANN (semi-DANN), (e) DU and (f) target-trained model on a high lesion load subject from the OXVASC dataset (target domain), along with the manual segmentation. The over/under-segmented regions of periventricular WMHs are indicated by hollow arrows, the correctly predicted regions by filled arrows and missed deep WMHs are shown in circles.

when fine-tuning was done in the intermediate coarser layers (layers 3 and 4), and was noticeably higher when encoder layer 4 was fine-tuned, with even a small amount of training data. This is due to the rich domain-specific information at the intermediate layers. Therefore the visual results were better for the middle two columns of Fig. 4b when compared to their corresponding results when fine-tuned from layer 1 (the first column). The best performance metrics were obtained when the pretrained model is fine-tuned starting from layer 4 with 18 training subjects (Fig. 6). In this case, TL provided better performance than strategy 1 and provided the highest median SI value among all strategies. The median performance metrics obtained at this setting are: SI value: 0.89, lAVD: 0.17, cluster-wise F1 measure: 0.62, cluster-wise TPR: 0.83, voxel-wise TPR: 0.84 and voxel-wise FPR: 1.6 × 10-4.

3.3. Strategy 3: unsupervised domain adversarial training on source and target domains

In the case of unsupervised DANN, the segmentation was better than both strategies 1 and 2, as shown in Fig. 4c on the target test dataset. The DANN model detected more lesion voxels along the ventricles and lesion edges when compared to the source trained model, and less false positives when compared to the TL strategy (especially in the low lesion load subjects). Even without using target labels for training, the model provided better delineation of PWMHs on the target test dataset, indicating the ability of the model to learn domain-invariant features. Unsupervised DANN gave the lowest voxel-wise FPR (significantly lower than DU), highest cluster-wise TPR and also the best voxel-wise TPR, on par with semi-DANN and target-trained models. On the source test dataset, the DANN model achieved median SI = 0.91, lAVD = 0.12, cluster-wise F1-measure = 0.82, cluster-wise TPR = 0.90, voxel-wise TPR = 0.89 and voxel-wise FPR = 0.9×10^{-4}. Also, DANN shows higher overlap between source and target feature representations at the layer before the label predictor, compared to the DU strategy, as shown in the t-SNE plot in Fig. 7b.

3.4. Strategy 4: semi-supervised domain adversarial training (semi-DANN) on the source and target domains

In the case of semi-DANN, the addition of a fraction of the labelled target data to label prediction provided improvement in the segmentation performance over source-trained, TL and unsupervised DANN (Fig. 5 and suppl. table S1). Semi-

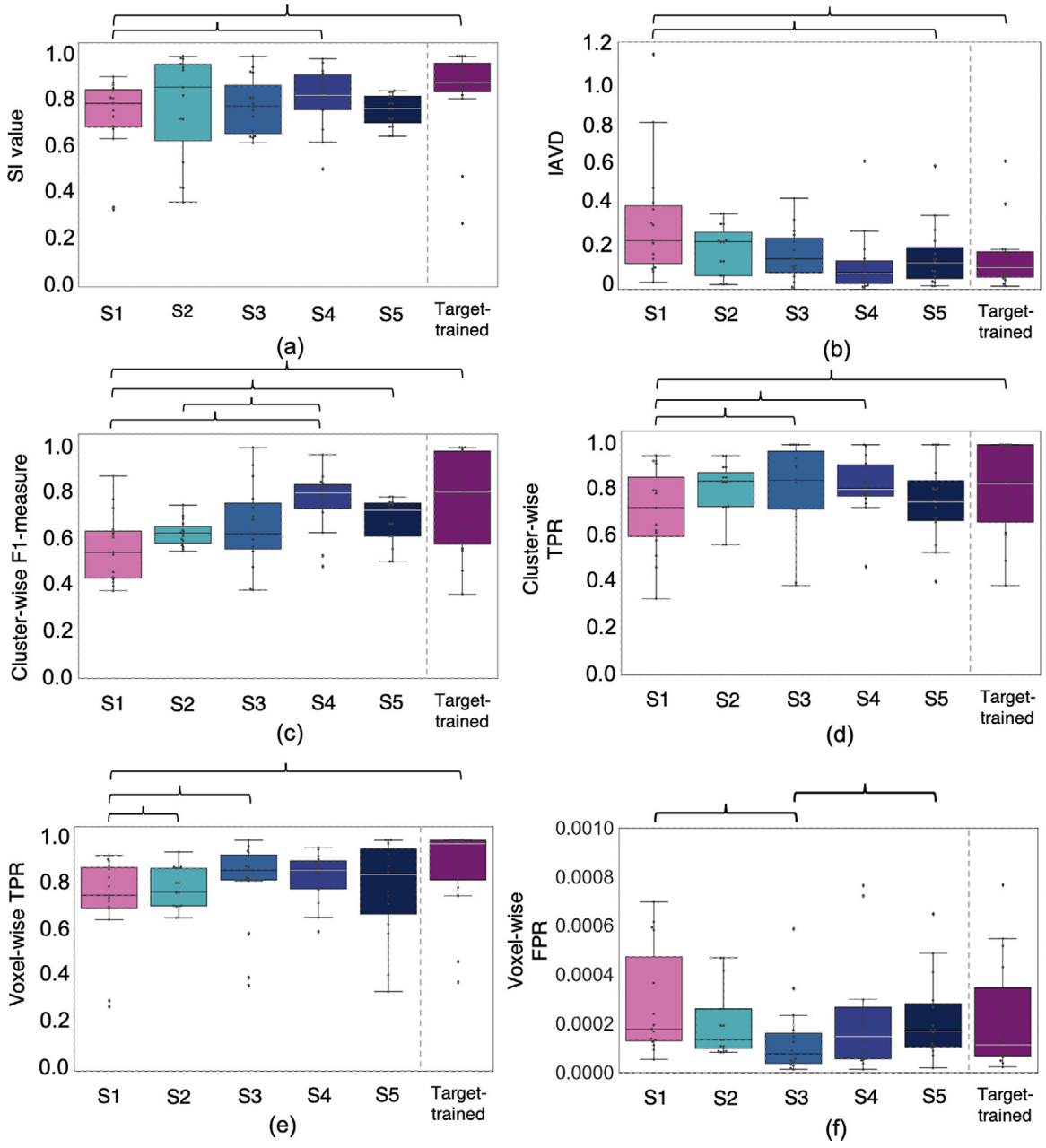


Fig. 5. Boxplots of performance metrics obtained for the 5 test strategies of the domain adaptation experiment, shown against the target-trained case, on the target test dataset (OXVASC + NDGEN) - (a) SI values, (b) lAVD, (c) cluster-wise F1-measure, (d) cluster-wise TPR, (e) voxel-wise TPR and (f) voxel-wise FPR values. For TL (strategy 2), we used the setting of 3 layers, 18 subjects for fine-tuning. The significant differences between the test strategies are indicated by brackets (after correcting for multiple comparisons across strategies).

DANN provided higher median cluster-wise F1-measure when compared to DANN, however with higher voxel-wise FPR as well. The improvement was subtle on visual assessment, and observable mainly along the boundaries of the PWMHs, as shown in Fig. 4d. The performance metrics achieved with semi-DANN method were not significantly different from those of target-trained model. On the source test dataset, the semi-DANN model achieved median SI = 0.86, lAVD = 0.09, cluster-wise F1-measure = 0.81, cluster-wise TPR = 0.87, voxel-wise TPR = 0.87 and voxel-wise FPR = 1.8×10^{-4}. We observed that the semi-DANN also brings the distribution of extracted features from the source and target domains together with greater overlap, when compared to the unsupervised DANN, as shown in Fig. 7c.

3.5. Strategy 5: iterative domain unlearning (DU) to remove scanner-bias between source and target domains

The DU model provided better performance than the source-trained model, but showed lower performance metrics compared to unsupervised DANN. However, none of the metrics were significantly different from strategy 3, except for higher voxel-wise FPR. On visual assessment, this strategy slightly oversegmented the PWMHs (Fig. 4e), mainly along the ventricles, which is also evident from the higher voxel-wise FPR values when compared to the unsupervised DANN. On the source domain test dataset, the DU model achieved SI = 0.83, lAVD = 0.10, cluster-wise F1-measure = 0.79, cluster-wise TPR = 0.84, voxel-wise TPR = 0.86 and voxel-wise FPR = 1.9×10^{-4}. From the visualisation of feature represen-

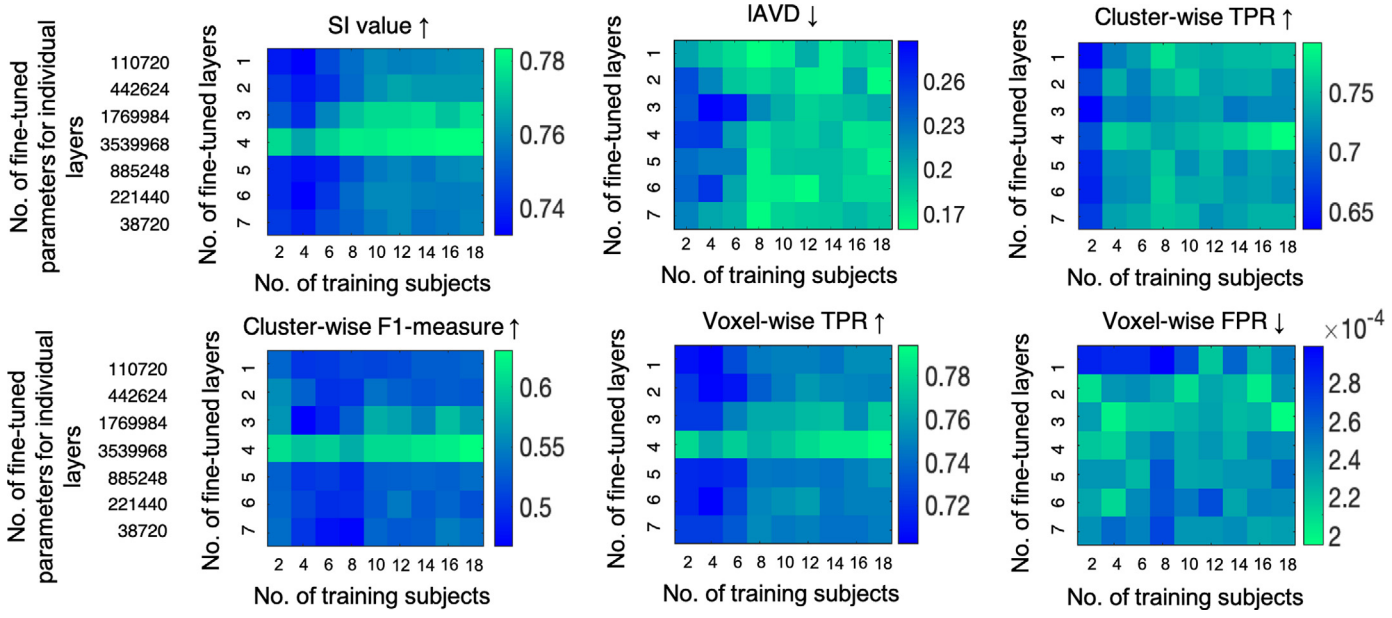


Fig. 6. Heatmaps of mean values of performance metrics for TL (strategy 2) on the target test dataset, corresponding to the number of training subjects and the number of fine-tuned layers. The maps are shown for (top row, left to right) SI values, lAVD, cluster-wise F1-measure, (bottom row, left to right) cluster-wise TPR, voxel-wise TPR and voxel-wise FPR values. The green end represents the best performance for all strategies, ↑ shows that higher values indicate better performance and ↓ shows vice versa. Note that given a number of fine-tuned layers, the layers prior to them in the encoder end were frozen, and the remaining layers towards the decoder end were fine-tuned. The number of parameters associated with individual layers has been reported (only for a single plane). For example, if the final 5 layers are fine-tuned, the sum of the top 4 values in the left column denotes the total number of parameters fine-tuned per planar U-Net. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tations in Fig. 7d, we can see that, for the DU model, while source and target features align as in the other DA strategies, they still show domain gaps with slightly less mixing of features from different domains.

3.6. Train on the target domain from scratch and apply to the target domain

The results from the model trained on the target training dataset showed the best segmentation performance on the target test dataset, as shown in Fig. 4f. But even in this case, the results showed a few false positive voxels in the high lesion load case, while the delineation of PWMHs was better than all the other strategies. The target-trained model achieved the best cluster-wise F1-measure, cluster- and voxel-wise TPR values with the lowest lAVD value as reported in suppl. table S1 and shown in Fig. 5, and significantly higher cluster-wise F1-measure and significantly better performance metrics (except the voxel-wise FPR) when compared to the source-trained model. Interestingly, the source-trained model fine-tuned on the target dataset with the best setting (18 subjects, starting from layer 4) provided better median SI value (although with wider interquartile range in SI values (0.63–0.97)) than the model trained from scratch on the target training dataset using same number of subjects. However, the other voxel- and cluster-wise metrics were better in target-trained case. Comparing with the adversarial training strategies, unsupervised DANN provided lower voxel-wise FPR values with on par cluster-wise TPR values.

4. Discussion and conclusions

In this work, we explored various domain adaptation techniques such as transfer learning, domain adversarial training and iterative domain unlearning for WMH segmentation using a triplanar ensemble model as the baseline method. Our baseline method provided better results than most of the ML methods using handcrafted features and provided results on par with the recently proposed DL methods (including the top-ranking methods of MWSC 2017) (Sundaresan et al., 2021). Also, on performing leave-one-out evaluation of TrUE-Net and the top-ranking method of MWSC 2017 (Li et al., 2018) on the datasets used as target domain in this study, TrUE-Net provided better performance metric values, especially on the OXVASC dataset (despite the lower resolution in the axial plane). In the case of TL, we also explored what would be the minimum number of subjects required for fine-tuning and which would be the best layers to fine-tune. We observed that domain adversarial training shows potential for better adaptation of the WMH segmentation task compared to other techniques on the given source and target datasets.

The source-trained model applied directly to the target test dataset achieved the worst performance out of all strategies due to the differences in image resolution, pathology, intensity characteristics and lesion distribution, as shown in Fig. 3 (also refer to Section 1 in supplementary material). The model trained on source and target domain datasets combined (section 5 in suppl. material) also performed better than the source-trained model, given that the model trained on the combined datasets learns the lesion characteristics of both domains.

The TL strategy provided better performance metrics compared to the source-trained model. Adding even a few subjects (2 - 4 subjects) from the target domain slightly improved the performance metrics when compared to the model trained on source-dataset only (refer to suppl. section 6 for more details), even though they were not on par with using >14 subjects for training or other adversarial training techniques. This is because the other strategies use more training data (either labelled or unlabelled) from the target domain which helps the model to learn the lesion characteristics of the target domain better. Although adding more representative training subjects could slightly improve the performance, we observed that in our case, a minimum of 14–16 subjects for fine-tuning might provide good results. However,

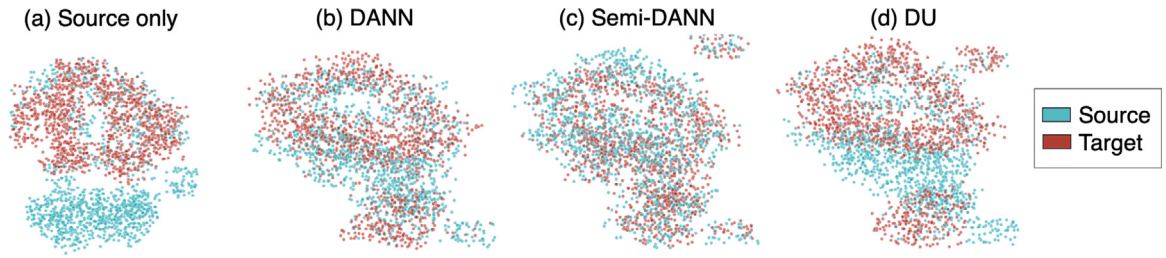


Fig. 7. The effect of domain adaptation on the extracted feature distributions for source (blue) and target domains (red). T-distributed Stochastic Neighbour Embedding (T-SNE) plots of the feature map values at the layer before the label predictor for (a) model trained on source dataset only, (b) unsupervised DANN, (c) semi-DANN and (d) DU. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the number of subjects required to improve WMH segmentation on the target domain also depends on the variation in features between domains. When we varied the number of layers to fine-tune, the performance improved when more intermediate features from both encoders and decoders (layers 3 and 4) were fine-tuned (heatmaps in Fig. 6). Generally, initial convolutional layers often contain low-level features (e.g. edges) that tend to be naturally domain invariant, and hence fine-tuning this layer does not improve the target performance much, as shown in (Ghafoorian et al., 2017b). On the other hand, the intermediate layers (Zeiler and Fergus, 2014; Girshick et al., 2014) and the layers with coarsest features at the encoder end (Ghafoorian et al., 2017b) contain higher-level information such as lesion pattern that are domain specific. Hence, fine-tuning the coarsest layers (e.g. layer 4) provided better performance on the target test dataset. Also, it has been shown that fine-tuning initial encoder layers adds more training parameters and requires a larger number of training samples (50–100 subjects) to avoid over-fitting (unavoidable even with 25 subjects in Ghafoorian et al. (2017b)). Given the encoder-decoder architecture of U-Net, while fine-tuning more decoder layers led to the steady improvement in the performance, fine-tuning the initial layers of encoder reduced the performance, possibly due to the shortage of representative training data required for training the initial layers, as observed in Ghafoorian et al. (2017b).

The performance of TL is better than training the baseline model on the combination of source and target dataset (refer to section 5 in supplementary material) on the target test dataset (even though the latter case shows better performance than strategy 1).

The unsupervised DANN performed better than TL and the source-trained model. The DANN model extracted domain invariant features (e.g. contrast between lesion and background, distribution of PWMHs) and provided better visual results with less noise and more precise segmentation of boundary voxels. The simultaneous label prediction and domain unlearning with shared weights provides regularisation in the DANN model, thus avoiding over-fitting to the training data.

The semi-DANN provided improvement over the DANN on the target test dataset. However, the DANN model detects less false positives when compared to the semi-DANN and provided comparable voxel-wise TPR values. Hence, while adding labelled target data might improve the performance of WMH segmentation in the target domain, it is necessary to weigh carefully the trade-off between improvement in the segmentation performance and the amount of manual effort involved, while choosing between unsupervised DANN and semi-DANN.

The DU model performed better than the source-trained model and provided performance metrics on par with the TL strategy with higher cluster-wise F1 measure. While training the DU model, we observed that the factor for weighting the domain confusion, β, plays a crucial role in achieving the domain invariance of the model. For the lower values of β, we found that the domain accuracy values were higher than 60% (where domain accuracy values closer to 50% are desirable indicating the maximal confusion of domains). We obtained the best results for the β value of 50 on the target dataset achieving a domain accuracy of 58%. Among the unsupervised models, DANN provided better performance than the DU model, with significant differences in voxel-wise FPR values. Also, DANN provided the better domain confusion with the domain accuracy of 47% at the layer before domain predictor (and a domain accuracy comparable to the DU model at the layer before the lesion label predictor as shown in the supplementary section 7). On visual assessment, the DU strategy oversegmented the PWMHs, while missing some DWMHs on the target test dataset, resulting in a lower cluster-wise TPR value.

The target domain features are aligned closer to the source features with a good overlap after domain adaptation (as shown in the t-SNE plots). The semi-DANN showed the maximum overlap of source and target feature representations at the layer before label predictor. The better overlap of domains with semi-DANN (compared to the unsupervised strategy) is expected due to the introduction of the labelled target data for training the label predictor in the semi-DANN. Among the unsupervised techniques, DANN showed better overlap of the features when compared to the DU model. It is worth noting the performance of the adversarial training techniques (such as DANN and DU) depends mainly on the variations in lesion distribution and the acquisition characteristics between the source and target datasets (since they do not require lesion labels from the target dataset), rather than uncertainties in the manual segmentations on the target dataset (as in the TL case).

The size of source and target domain dataset is an important factor that affects the performance of domain adaptation techniques. The model transferred from source to target domain in strategy 2, uses both source and target domain datasets for pre-training and fine-tuning respectively and hence learns the characteristics from both datasets. On the other hand, the difference in the source and target domain datasets' sizes and the inhomogeneity in inter-/intra-domain characteristics especially affect unsupervised methods like DANN. To better investigate this aspect, we chose the two best performing adversarial adaptation techniques from our test strategies, DANN and semi-DANN, and trained them after swapping the datasets used for source and target domains. We observed that while DANN model is susceptible to slight changes in the performance (however, none of them significant except voxel-wise FPR using Wilcoxon signed rank test), semi-DANN provided a consistent performance after domain swapping, without any significant difference in performance. For more details on the experiments and results, refer to suppl. Section 4.

We grouped the NDGEN and OXVASC datasets in the target domain according to their similarity (using a discriminator network), however these datasets still have differences in their characteristics (scanner, sequence, population). We therefore tested if the DA techniques performed differently in the two datasets and found no significant difference (refer to section 8 in suppl. material). This

demonstrates the ability of DA techniques to learn the domain invariance between the target datasets.

Concluding, we explored various DA techniques such as transfer learning and domain adversarial training techniques including DANN and DU. For the TL case, fine-tuning the intermediate layers towards the end of the encoder provided better results than fine-tuning the initial layers. The DANN models performed better than TL and the DU model on the target dataset. Particularly, the semi-DANN provided the best performance metrics with improvements over DU and TL cases. However, even without the addition of labelled target training data, the unsupervised DANN provided better cluster-wise and voxel-wise performance metrics compared to TL and DU, and results on par with the semi-DANN.

Declaration of Competing Interest

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

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

CRediT authorship contribution statement

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

Acknowledgements

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

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

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

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2021.102215.

References

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

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. Mach. Learn. 79 (1–2), 151–175.

Bordin, V., Bertani, I., Mattioli, I., Sundaresan, V., McCarthy, P., Suri, S., Zsoldos, E., Filippini, N., Mahmood, A., Melazzini, L., et al., 2020. Integrating large-scale neuroimaging research datasets: harmonisation of white matter hyperintensity measurements across Whitehall and UK Biobank datasets. bioRxiv.

Caligiuri, M.E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., Cherubini, A., 2015. Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review. Neuroinformatics 13 (3), 261–276.

Cao, Z., Long, M., Wang, J., Jordan, M.I., 2018. Partial transfer learning with selective adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2724–2732.

Cheng, L., Pan, S.J., 2014. Semi-supervised domain adaptation on manifolds. IEEE Trans. Neural Netw. Learn. Syst. 25 (12), 2240–2249.

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

Dinsdale, N.K., Jenkinson, M., Namburete, A.I., 2020. Unlearning scanner bias for MRI harmonisation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 369–378.

Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T., 2013. Unsupervised visual domain adaptation using subspace alignment. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2960–2967.

Fortin, J.-P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., et al., 2018. Harmonization of cortical thickness measurements across scanners and sites. NeuroImage 167, 104–120.

Gallego, A.-J., Calvo-Zaragoza, J., Fisher, R.B., 2020. Incremental unsupervised domain-adversarial training of neural networks. IEEE Trans. Neural Netw. Learn. Syst..

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning. PMLR, pp. 1180–1189.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17 (1). 2096–2030

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I.W., Sanchez, C.I., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., Platel, B., 2017. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Sci. Rep. 7 (1), 5110.

Ghafoorian, M., Karssemeijer, N., van Uden, I.W., de Leeuw, F.-E., Heskes, T., Marchiori, E., Platel, B., 2016. Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. Med. Phys. 43 (12), 6246–6258.

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R., de Leeuw, F.-E., Tempany, C.M., van Ginneken, B., et al., 2017. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 516–524.

Gibson, E., Gao, F., Black, S.E., Lobaugh, N.J., 2010. Automatic segmentation of white matter hyperintensities in the elderly using flair images at 3T. J. Magn. Reson. Imaging 31 (6), 1311–1322.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.

Haeusser, P., Frerix, T., Mordvintsev, A., Cremers, D., 2017. Associative domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2765–2773.

Hong, J., Park, B.-y., Lee, M.J., Chung, C.-S., Cha, J., Park, H., 2020. Two-step deep neural network for segmentation of deep white matter hyperintensities in migraineurs. Comput. Methods Programs Biomed. 183, 105065.
Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K., 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5001–5009.
Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.
Jeon, S., Yoon, U., Park, J.-S., Seo, S.W., Kim, J.-H., Kim, S.T., Kim, S.I., Na, D.L., Lee, J.-M., 2011. Fully automated pipeline for quantification and localization of white matter hyperintensity in brain magnetic resonance image. Int. J. Imaging Syst. Technol. 21 (2), 193–200.
Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.
Kouw, W.M., Loog, M., 2019. A review of domain adaptation without target labels. IEEE Trans. Pattern Anal. Mach. Intell..
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.
Kuijf, H.J., Biesbroek, J.M., de Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities; results of the WMH segmentation challenge. IEEE Trans. Med. Imaging.
Lee, D.-H., et al., 2013. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. Workshop on Challenges in Representation Learning, ICML, Vol. 3.
Lee, S., Kim, D., Kim, N., Jeong, S.-G., 2019. Drop to adapt: learning discriminative features for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 91–100.
Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.-S., Menze, B., 2018. Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. NeuroImage 183, 650–665.
Long, M., Cao, Z., Wang, J., Jordan, M. I., 2017a. Conditional adversarial domain adaptation. arXiv preprint arXiv:1705.10667.
Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S., 2013. Transfer feature learning with joint distribution adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2200–2207.
Long, M., Zhu, H., Wang, J., Jordan, M.I., 2017. Deep transfer learning with joint adaptation networks. In: International Conference on Machine Learning. PMLR, pp. 2208–2217.
Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE.. J. Mach. Learn. Res. 9 (11).
Opitz, D., Maclin, R., 1999. Popular ensemble methods: an empirical study. J. Artif. Intell. Res. 11, 169–198.
Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q., 2010. Domain adaptation via transfer component analysis. IEEE Trans. Neural Netw. 22 (2), 199–210.
Pan, S.J., Yang, Q., 2009. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.
Park, B.-y., Lee, M.J., Lee, S.-h., Cha, J., Chung, C.-S., Kim, S.T., Park, H., 2018. DEWS (DEep white matter hyperintensity segmentation framework): a fully automated pipeline for detecting small deep white matter hyperintensities in migraineurs. NeuroImage 18, 638–647.
Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M., 2013. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 246–253.
Rachmadi, M.F., Valdés-Hernández, M.d.C., Agan, M.L.F., Di Perri, C., Komura, T., Initiative, A.D.N., et al., 2018. Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology. Comput. Med. Imaging Graph. 66, 28–43.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Rothwell, P., Coull, A., Giles, M., Howard, S., Silver, L., Bull, L., Gutnikov, S., Edwards, P., Mant, D., Sackley, C., et al., 2004. Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004 (Oxford Vascular Study). Lancet 363 (9425), 1925–1933.
Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K., 2019. Semi-supervised domain adaptation via minimax entropy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8050–8058.
Saito, K., Ushiku, Y., Harada, T., 2017. Asymmetric tri-training for unsupervised domain adaptation. In: International Conference on Machine Learning. PMLR, pp. 2988–2997.
Schoenauer-Sebag, A., Heinrich, L., Schoenauer, M., Sebag, M., Wu, L. F., Altschuler, S. J., 2019. Multi-domain adversarial learning. arXiv preprint arXiv:1903.09239.
Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 60.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.
Steenwijk, M.D., Pouwels, P.J., Daams, M., van Dalen, J.W., Caan, M.W., Richard, E., Barkhof, F., Vrenken, H., 2013. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). NeuroImage 3, 462–469.
Sun, B., Saenko, K., 2016. Deep coral: correlation alignment for deep domain adaptation. In: European Conference on Computer Vision. Springer, pp. 443–450.
Sundaresan, V., Zamboni, G., Rothwell, P.M., Jenkinson, M., Griffanti, L., 2021. Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images. Med. Image Anal. p. 102184.
Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging 35 (5), 1299–1312.
Tzeng, E., Hoffman, J., Darrell, T., Saenko, K., 2015. Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4068–4076.
Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7167–7176.
Wang, M., Deng, W., 2018. Deep visual domain adaptation: a survey. Neurocomputing 312, 135–153.
Wang, X., Schneider, J., 2014. Flexible transfer learning under support and model shift. In: Advances in Neural Information Processing Systems, pp. 1898–1906.
Wang, Y., Catindig, J.A., Hilal, S., Soon, H.W., Ting, E., Wong, T.Y., Venketasubramanian, N., Chen, C., Qiu, A., 2012. Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts. NeuroImage 60 (4), 2379–2388.
Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., T O'Brien, J., Barkhof, F., Benavente, O.R., et al., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 12 (8), 822–838.
Wilson, G., Cook, D. J., 2019. A survey of unsupervised deep domain adaptation. arXiv preprint arXiv:1812.02849.
Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. NeuroImage 92, 381–397.
Winzeck, S., Mocking, S., Bezerra, R., Bouts, M., McIntosh, E., Diwan, I., Garg, P., Chutinet, A., Kimberly, W., Copen, W., et al., 2019. Ensemble of convolutional neural networks improves automated segmentation of acute ischemic lesions using multiparametric diffusion-weighted MRI. Am. J. Neuroradiol. 40 (6), 938–945.
Yao, T., Pan, Y., Ngo, C.-W., Li, H., Mei, T., 2015. Semi-supervised domain adaptation with subspace learning for visual recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2142–2150.
Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, pp. 3320–3328.
Zamboni, G., Wilcock, G.K., Douaud, G., Drazich, E., McCulloch, E., Filippini, N., Tracey, I., Brooks, J.C., Smith, S.M., Jenkinson, M., et al., 2013. Resting functional connectivity reveals residual functional activity in Alzheimer's disease. Biol. Psychiatry 74 (5), 375–383.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Springer, pp. 818–833.
Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., Saminger-Platz, S., 2017. Central moment discrepancy (CMD) for domain-invariant representation learning. arXiv preprint arXiv:1702.08811.
Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57.
Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J., 2019. Confidence regularized self-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5982–5991.