Supervised Domain Adaptation Applied to Heterogeneous, Multi-Center MR Imaging Datasets

Justin Park[a,b], Richard Frayne[a,b,c], and Mariana Bento[b,d]

[a]Radiology and Clinical Neuroscience, University of Calgary, Calgary, Canada
[b]Calgary Image Processing and Analysis Centre, Foothills Medical Center, Calgary, Canada
[c]Seaman Family MR Centre, Foothills Medical Center, Calgary, Canada
[d]Electrical and Software Engineering, University of Calgary, Calgary, Calgary, Canada

ABSTRACT

Deep learning neural networks are a common tool in medical imaging and frequently used to solve a variety of complex problems. Magnetic resonance (MR) images are frequently employed to develop these networks because of their high spatial resolution and user selectable image contrast between tissues. More advanced deep learning models are being developed, which when combined with improvements in MR image acquisition techniques, will allow for image analysis techniques that are more efficient yet able to solve increasingly challenging problems. A current significant disadvantage of deep learning networks is that they are extremely sensitive to the distribution of the data used for training, therefore, network implementation can be challenging in clinical applications with heterogeneous images. The main problem is that, in a clinical environment, data distributions of target datasets can vary from subject-to-subject due to differences in scanner vendor, magnetic field strength, and the setting of specific MR acquisition parameters. These variations create inherent scan variability that diversifies the data distributions of different datasets. This effect can result in the model becoming inaccurate and producing undesirable outcomes. Thus, to improve model generalizability, we explored a supervised domain-adaptation approach. To test this method, we created a convolutional neural network model that performed a classification task and was composed of three components: (1) a feature extractor, (2) a pathology classifier, and (3) a domain classifier. In a single, unified training process, the pathology classifier was trained by *minimizing* the pathology loss function and the domain classifier was trained by *maximizing* the domain loss function. This procedure allows the model to penalize learning of features specific to the domain, and thus attempts to produce a domain-invariant feature vector. The performance of this domain-adapted model was compared to the same model but without domain classification (*i.e.*, a baseline traditional model consisting of a feature extractor and a pathology classifier). We found that the domain-adapted model achieved a higher accuracy rate in predicting images from both source and target datasets.

Keywords: Domain adaptation, Image classification, Brain imaging, MR imaging

1. INTRODUCTION

Deep learning neural networks are commonly used in the field of clinical research to solve complex problems, such as image segmentation and classification, due to its profound ability to extract information from images.[1] For instance, convolutional neural network models have accurately segmented a variety of brain structures, such as white matter, gray matter, and cerebrospinal fluid from magnetic resonance (MR) images.[2] Convolutional neural networks have also successfully performed various classification tasks, such as the Alzheimer's disease and mild cognitive impairment classification task.[3] Despite these and many other recent successes, a common problem that these models face is their effectiveness to segment or classify images from an out-of-sample dataset.

In clinical applications, target domain images will generally experience some sort of "domain shift" when compared to the training (source) dataset.[4] For MR imaging, this shift can be generally due to the differences in acquisition parameters, such as the scanner vendor and magnetic field strengths (Figure 1). Furthermore,

Further author information: (Send correspondence to Justin Park)
Justin Park: E-mail: justin.park@ucalgary.ca

patient-specific MR sequences add additional variability in data distributions during acquisition. These images present inherent scan variability characteristics that modify the data distribution (introducing a domain shift) and lead to a higher misclassification rate and/or a poor generalization by the convolutional neural network model.

To increase the generalizability of convolutional neural network models, domain adaptation and data augmentation techniques have been explored. One of many methods that successfully achieved domain adaptation was to use adversarial discriminative networks.[5] This approach showed great results in classification tasks, however, it requires retraining on the target domain images, which is inconvenient in a clinical setting where acquisition parameters change frequently. Furthermore, to minimize the sensitivity of convolutional neural network models to data distribution, many data augmentation techniques have been developed, including: traditional methods (translation, zoom in/out, rotation, reflection, distortion, and introducing noise/artifacts) and generative adversarial networks (GANs).[6] The traditional method is more commonly applied to datasets because of its simplistic implementation. This method creates diversity in the dataset to improve the generalizability and the accuracy of convolutional neural network models. GANs are a much more complicated method of data augmentation where a neural network is used to create an image that contains a vastly different data distribution compared to the original image.[7] This method allows developers to create a domain-diverse dataset to train convolutional neural network models, which ultimately increases the generalizability.

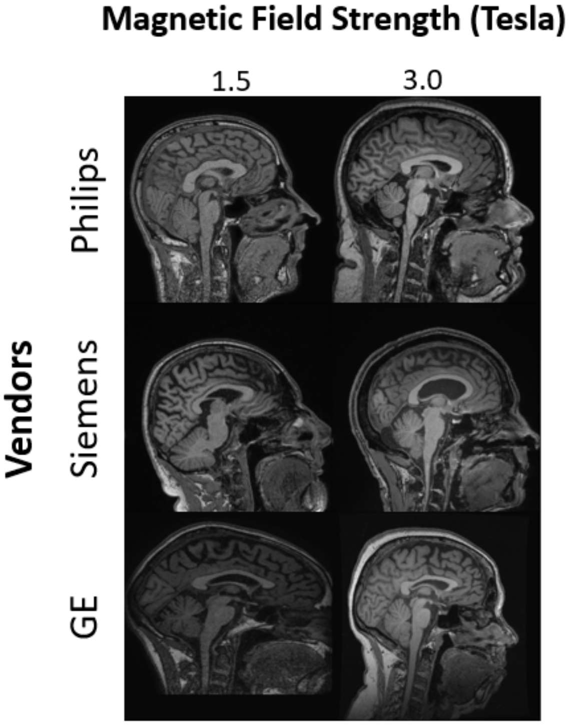

Figure 1. Domain shift between the various vendors and magnetic field strengths. To a human eye, inherent scan variability characteristics can be difficult to distinguish. However, convolutional neural network models are extremely sensitive to the data distribution, therefore, the model is able to detect the subtle changes in the image due to vendors and magnetic field strengths.

Despite these efforts to increase the generalizability through domain adaptation and data augmentation, the influences of these techniques are not enough to significantly increase the generalizability of convolutional neural network models for clinical applications. Therefore, we propose a supervised domain adaptation approach on heterogeneous, multi-center MR datasets to create a convolutional neural network model that can be compatible with rapidly changing clinical environments. This convolutional neural network model will have three main components: (1) a feature extractor, (2) a pathology classifier, and (3) a domain classifier. The feature extractor takes in a MR image and will condense it to a feature vector. This vector is then used by the pathology and domain classifier for prediction. Using this model, domain adaptation is achieved by simply implementing a gradient reversal layer to maximize the loss produced by the domain classifier while minimizing the loss of the pathology classifier during training.[8] Our hypothesis is that this approach should create a domain-invariant feature vector that isolates the information required by the pathology classifier to accurately predict the label of the input image. As a case of study, we will train a convolutional neural network model that will perform a binary classification task between a normal control participants and patients with Alzheimer's disease.

2. METHODS

2.1 Datasets

The experimental dataset consists of T1-weighted volumetric MR images from two publicly available datasets: the Alzheimer's Disease Neuroimaging Initiative (ADNI[9]) and the *Calgary-Campinas-359* (*CC-359*[10]). From the ADNI (source) dataset, we used data from two main classes: normal control subjects and Alzheimer's disease patients. Data was selected across multiple sites from three major scanner vendors (General Electric, Philips, and Siemens) and at two different field strengths (1.5 T and 3 T). The *CC-359* (target) dataset contains 359 normal control image volumes acquired considering the same scanner vendors and magnetic field strengths as the ADNI dataset.

A portion of the data used in this study was obtained from the ADNI database (adni.loni.usc.edu). ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W Weiner, MD. The primary goal of ADNI has been to test whether serial MR imaging, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. For up-to-date information, see www.adni-info.org.

Table 1. Number of T1-weighted volumetric imaging samples across different datasets and different acquisition parameters (by vendor and magnetic field strength). ADNI = Alzheimer's Disease Neuroimaging Initiative.

Datasets	Label/ Class	Siemens		Philips		GE		
		1.5 T	3 T	1.5 T	3 T	1.5 T	3 T	Total
ADNI (source)	Alzheimer's disease	141	143	34	82	92	91	583
	normal control	115	477	40	134	168	197	1,131
CC-359 (target)	normal control	60	60	59	60	60	60	359
		316	680	133	276	320	348	2,073

2.2 Model

To test the supervised domain adaptation approach, we created a convolutional neural network model that performs a simple binary classification task of distinguishing images between an Alzheimer's disease patient and a normal control subject. This model was composed of (1) a feature extractor, (2) a pathology classifier, and (3) a domain classifier (Figure 2).

The feature extractor was a neural network that condensed information about the image at the input. The most common form of a feature extractor is a convolutional neural network because it exceeds other neural networks in extracting topological information from an image. In a deep convolutional neural network model, the first few layers detect edges and extract the fundamental shapes that make up the image. However, in the deeper layers, the model will highlight features specific to the training dataset. To consolidate the feature extractor with the two classifiers, a transfer learning approach was used. Transfer learning is a common method that utilizes models developed for related or unrelated tasks that generally are trained with much larger quantities of training data. For our feature extractor, we chose to use a pre-trained VGG16 model. VGG16 is a convolutional neural network that has achieved high accuracy rates in ImageNet. It has been trained with roughly 1.2 million different images, therefore, it has very strong foundational layers for extracting elementary features. Furthermore, it has

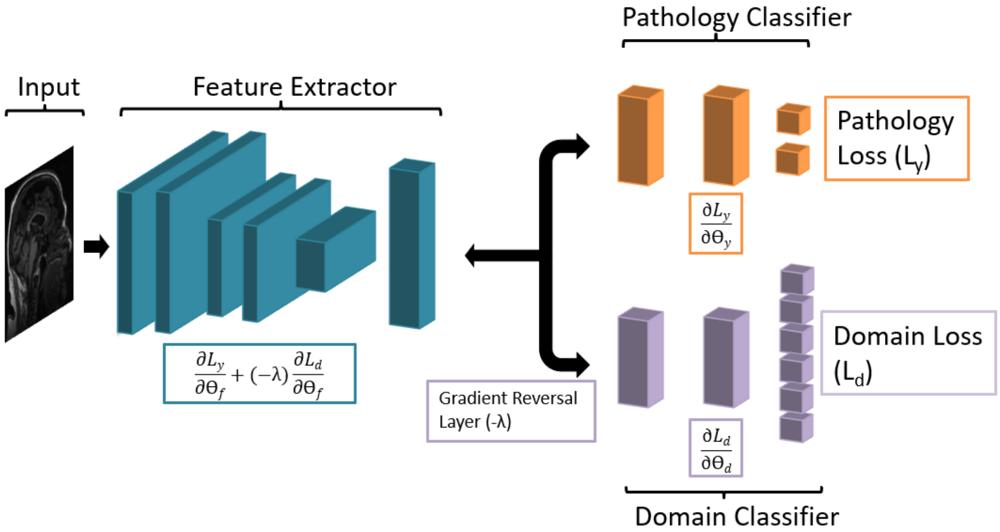
Pathology Classifier
Input
Feature Extractor
Pathology Loss (L_y)
$\frac{\partial L_y}{\partial \Theta_y}$
Domain Loss (L_d)
$\frac{\partial L_y}{\partial \Theta_f} + (-\lambda)\frac{\partial L_d}{\partial \Theta_f}$
Gradient Reversal Layer (-λ)
$\frac{\partial L_d}{\partial \Theta_d}$
Domain Classifier

Figure 2. Architecture of the proposed method. A transfer learning approach was used to consolidate the two classifiers with the feature extractor. In both classifiers, their respective loss was used to optimize the classifiers to minimize the loss. However, the feature extractor has a combined loss function that adjusts to minimize the loss of the pathology classifier while maximizing the loss for the domain classifier.

displayed robust feature extracting in transfer learning approaches through various classification studies.[11,12] To fine-tune the feature extractor, the last seven layers were trained through backpropagation with images from our training dataset.

The pathology classifier was a simple feed-forward neural network that produces a probabilistic binary outcome of whether the input feature vector is an Alzheimer's disease patient or a normal control subject. The loss function that was calculated by this classifier will be used in backpropagation to adjust the weights and biases of the pathology classifier and the feature extractor to minimize the loss. This step will incrementally improve the pathology classifier to learn the features that contribute to the Alzheimer's disease classification and optimize the VGG16 model to extract features more specific to the patient.

Similar to the pathology classifier, the domain classifier is also a feed-forward neural network that produced a probabilistic outcome of whether the input feature vector comes from a specific domain. This classifier has six different outcomes, which describes each combination of three vendors and two magnetic field strengths that were used to acquire our dataset (Figure 1). Using these predictions, a loss function was calculated and used to optimize the domain classifier to minimize the loss. However, when this loss function was backpropagated to the feature extractor, it went through a gradient reversal layer. During forward propagation, this layer acts as an identity transformation, however, during backpropagation, this layer reversed the gradient by multiplying a negative scalar value, $-\lambda$. This step causes all parameters in the feature extractor to be optimized to maximize the loss produced by the pathology classifier, which confuses the model about which domain the images are coming from. The result is a feature extractor that isolates for features that are only required to classify images by the pathology classifier.

To combine the domain adaption and pathology classification in one training step, a simple addition of loss functions can be performed[8] to modify the weights and biases of the feature extractor:

$$\frac{\partial L_y^i}{\partial \Theta_f} + (-\lambda)\frac{\partial L_d^i}{\partial \Theta_f} \tag{1}$$

where L_y^i and L_d^i represents the loss of pathology classifier and the loss of domain classifier both on the i-th training step, respectively. The variable Θ_f indicates all parameters to be adjusted in the feature extractor. Lastly, $-\lambda$ is the hyper-parameter used to reverse the gradient during backpropagation. Using this optimized loss function allowed us to create a single training step to minimize the loss of the pathology classifier while maximizing the loss of the domain classifier.

2.3 Experimental Design

To train and test our model, we first pre-processed the MR imaging data by taking two simple steps: resize the image and apply min-max intensity normalization. To resize the image to the specific shape for VGG16 feature extractor, we have used a bilinear interpolation to reshape the images to (224×224). Then applied min-max intensity normalization to each image to compress all intensity values between 0 and 1 (Figure 3).

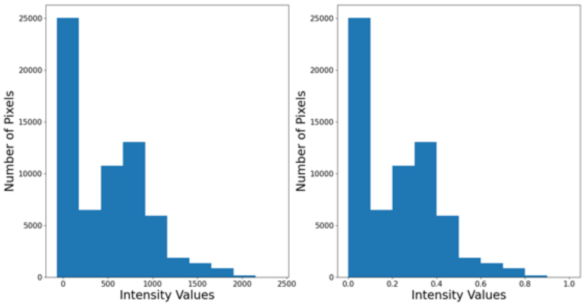

Figure 3. Comparison of intensity distribution of images by histograms from: (left) Raw data and (right) pre-processed data. Preserving the intensity distribution is critical for the model to learn the various domains the images come from.

To train our convolutional neural network model, we took 20 central image slices of three different views (incorporating volumetric information[13]) from a subject in a source dataset as shown in Figure 4 and aggregate them into a large pool of training data in a random order. We have chosen to use 20 central image slices because they contain the most relevant information about the brain. This step also increased the number of training samples by a factor of 20. In addition, we trained our model using three different views from each subject, allowing us to diversify the training images, which ultimately improves the generalization of our model.

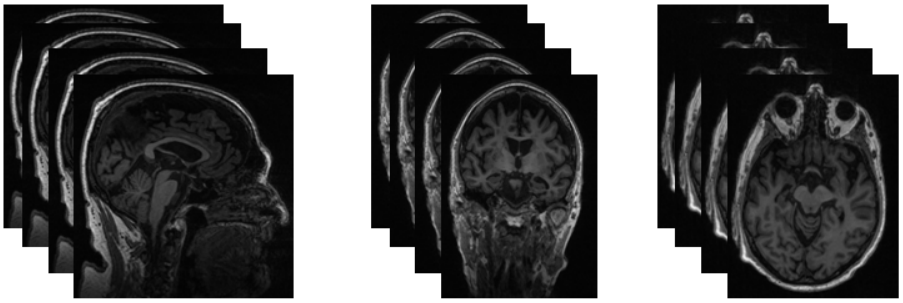

Figure 4. Sample images from the training dataset. For each subject, 20 central image slices of three different views were used to train the model: (left) sagittal view, (middle) coronal view and (right) axial view. This step significantly increased the number of training dataset and improved the generalizability of the model.

To compare our findings, a basic convolutional neural network model was created, which had the same feature extractor and pathology classifier as the domain adaptation model. However, the domain classification branch was absent in the basic model. This model was trained with the same training dataset and will serve as a baseline to compare the effects of our supervised domain adaptation approach on source and target datasets.

3. RESULTS

The domain adaptation model was only trained using the source dataset. Like all conventional models, the accuracy rate of pathology classifier increases as its loss decrease (Figure 5). However, the accuracy rate and the loss of the domain classifier after each training step remains constant. This sugests that the model is unable to learn the distinct qualities of images from various vendors, magnetic field strengths, and acquisition parameters by the use of the gradient reversal layer during backpropagation. This allows the model to predict on non-harmonious T1-weighted images with higher accuracy rate by minimizing the effects of domain-specific features on the outcomes of feature extractor and the pathology classifier.

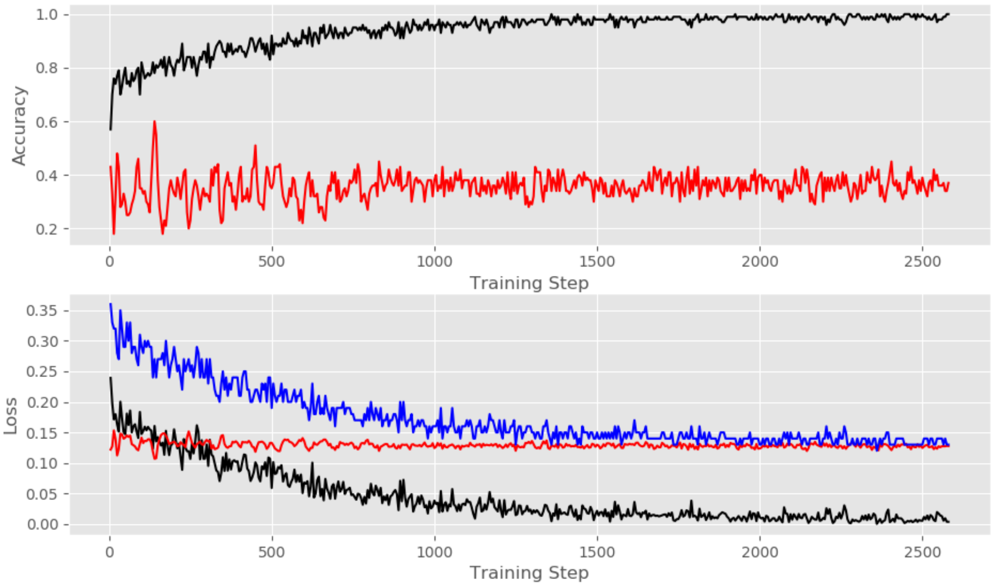

Figure 5. Accuracy rate (top) and loss (bottom) after each training step. Accuracy rates for the the pathology classifier (black) and domain classifier (red) are shown. Losses for the pathology classifier (black, $\frac{\partial L_y^i}{\partial \Theta_f}$) and domain classifier (red, $-\lambda\frac{\partial L_d^i}{\partial \Theta_f}$), respectively, are shown. The total loss (blue) is the sum of the pathology and domain classifier losses.

To further validate our approach, the predictions on source and target datasets of baseline model and domain adaptation model were compared. To predict the image classes, a similar method from training was used. The model predicted on all 20 central image slices of three different views for each subject in source and target dataset. These predictions were then averaged to give the final prediction of whether the sample is Alzheimer's disease or normal. The results in Table 2 show that, not only did the domain adaptation model surpass the accuracy rate of the baseline model in the target dataset, as expected, it also outperformed on the source dataset as well. Therefore, it is evident that the gradient reversal layer influences the model to have a greater generalizability to achieve higher accuracy rate on predictions of heterogeneous, multi-center MR datasets.

Table 2. Accuracy of convolutional neural network models classifying test dataset images from source and target domains. Overall, the domain adaptation model achieved a higher accuracy rate compared to the baseline model in both source and target datasets.

Models	Datasets	
	ADNI (source)	CC359 (target)
Baseline Model	91.3%	91.1%
Domain Adaptation Model	92.7%	97.8%

In this study, increasing the number of training samples and allowing margin of error had an impact on the results. To gather non-harmonious T1-weighted images, we utilized two large public datasets where images are gathered from multiple centers with various acquisition parameters and scanners. However, the experiments were limited by the amount of images available for specific cases, such as Alzheimer's disease patients. The initial idea was to use a 3D image to train our model because there are advantages of extracting features from multiple directions, however, the small dataset presented complications that outweighed the advantages of using 3D images. Therefore, we increased the number of training samples by using the 20 central image slices of three different views for each subject. This increased the training dataset by a factor of 60 and greatly improved the results of our model. Also, by using a similar method, we were able to incorporate volumetric predictions for testing our model.

4. CONCLUSION

In a clinical setting where images are acquired using various vendors and magnetic field strengths, as well as specific acquisitions parameters, it is difficult to implement computer assisted tools, such as those using deep learning neural networks. The difficulty is due to an inability of the tools to adapt to inherent scan variability characteristics and continuously changing data distributions. To increase the generalizability of deep learning models, we proposed a supervised domain adaptation approach. Our proposal, based on a domain classification with a gradient reversal layer combined to a classification model, was able to desensitize the model to various data distributions. We were able to mitigate the learning ability of domain adaptation model in domain classification by using the gradient reversal layer while maximizing the accuracy rate of the pathology classification. In future work, we intend to further enhance the generalizability of the convolutional neural network model, by using strategies such as data augmentation techniques and regularization applied in conjunction with the proposed approach, and also include other datasets in our experimental design.

ACKNOWLEDGMENTS

We thank the staff at the Calgary Image Processing and Analysis Centre (CIPAC) including Marina Salluzzi, PhD. The financial support of the Canada Foundation for Innovation (CFI), the the Canadian Institutes for Health Research are acknowledged. MB was supported by a fellowship from the Canadian Open Neuroscience Program (CONP). RF was supported by the Hopewell Professorship in Brain Imaging.

Some data collection and sharing was funded by ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

[1] Lundervold, A. S. and Lundervold, A., "An Overview of Deep Learning in Medical Imaging Focusing on MRI," *Journal of Medical Imaging* **29**(2), 102–127 (2019).

[2] Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., and Erickson, B. J., "Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions," *Journal of Digital Imaging* **30**, 449–459 (2017).

[3] Suk, H.-I. and Shen, D., "Deep Learning-Based Feature Representation for AD/MCI Classification," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*], **16**, 583–90 (2013).

[4] Wang, M. and Deng, W., "Deep Visual Domain Adaptation: A Survey," *Neurocomputing* **312**, 135–153 (2018).

[5] Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T., "Adversarial Discriminative Domain Adaptation," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2017).

[6] Shorten, C. and Khoshgoftaar, T. M., "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data* **6**(60) (2019).

[7] Creswell, A., White, T., Dumoulin, V., and et al., "Generative Adversarial Networks: An Overview.," *IEEE Signal Processing Magazine* **35**, 53–65 (2018).

[8] Ganin, Y. and Lempitsky, V., "Unsupervised Domain Adaptation by Backpropagation," in [*Proceedings of the 32nd International Conference on Machine Learning*], 1180–1189 (2015).

[9] S. Mueller, M. Weiner, L. J. T. and et al., """The Alzheimer's Disease Neuroimaging Initiative"," *Neuroimaging Clinics of North America* **15**, 869–877 (2001).

[10] R. Souza, O. Lucena, J. G. and et al., """An Open, Multi-vendor, Multifield-Strength Brain MR Dataset and Analysis of Publicly Available Skull Stripping Methods Agreement"," *NeuroImage* **170**, 482–494 (2018).

[11] Kaur, T. and Gandhi, T. K., "Automated Brain Image Classification Based on VGG-16 and Transfer Learning," in [*2019 International Conference on Information Technology (ICIT)*], 94–98 (2019).

[12] Tammina, S., "Transfer Learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images," *International Journal of Scientific and Research Publications (IJSRP)* **9**(10) (2019).

[13] Leite, M., Gobbi, D., Salluzi, M., Frayne, R., Lotufo, R., and Rittner, L., "3D Texture-based Classification Applied on Brain White Matter Lesions on MR Images," in [*Medical Imaging 2016: Computer-Aided Diagnosis*], **9785**, 682 – 687 (2016).