Learning Transferable 3D-CNN for MRI-Based Brain Disorder Classification from Scratch: An Empirical Study

Hao Guan[1], Li Wang[1], Dongren Yao[1], Andrea Bozoki[2], and Mingxia Liu[1(✉)]

[1] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
mxliu@med.unc.edu

[2] Department of Neurology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Abstract. Reliable and efficient transferability of 3D convolutional neural networks (3D-CNNs) is an important but extremely challenging issue in medical image analysis, due to small-sized samples and the domain shift problem (*e.g.*, caused by the use of different scanners, protocols and/or subject populations in different sites/datasets). Although previous studies proposed to pretrain CNNs on ImageNet, models' transferability is usually limited due to semantic gap between natural and medical images. In this work, we try to answer a key question: *how to learn transferable 3D-CNNs from scratch based on a small (*e.g.*, tens or hundreds) medical image dataset*? We focus on the case of structural MRI-based brain disorder classification using four benchmark datasets (*i.e.*, ADNI-1, ADNI-2, ADNI-3 and AIBL) to address this problem. (1) We explore the *influence of different network architectures* on model transferability, and find that appropriately deepening or widening a network can increase the transferability (*e.g.*, with improved sensitivity). (2) We analyze the *contributions of different parts* of 3D-CNNs to the transferability, and verify that fine-tuning CNNs can significantly enhance the transferability. This is different from the previous finding that fine-tuning CNNs (pretrained on ImageNet) cannot improve the model transferability in 2D medical image analysis. (3) We also study the *between-task transferability* when a model is trained on a source task from scratch and applied to a related target task. Experimental results show that, compared to directly training CNN on related target tasks, CNN pretrained on a source task can yield significantly better performance.

Keywords: Transferability · 3D-CNN · Deep learning · Brain MRI

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-87589-3_2) contains supplementary material, which is available to authorized users.

C. Lian et al. (Eds.): MLMI 2021, LNCS 12966, pp. 10–19, 2021.
https://doi.org/10.1007/978-3-030-87589-3_2

1 Introduction

Deep learning (*e.g.*, with convolutional neural networks, CNNs) has been extensively applied in medical image analysis [1–3], but usually suffers from the small-sample-size problem. To address this issue, recent studies proposed to fine-tune pretrained CNNs on ImageNet for medical imaging analysis [4–6]. However, direct transferring pretrained CNNs usually yields sub-optimal performance, due to the fundamental differences between natural and medical images. For instance, medical images (*e.g.*, T1-weighted MRIs) are typically 3-dimensional, whereas CNNs pretrained on ImageNet usually treat all the 2D slices within a subject scan independently. This will lead to loss of information [7,8]. In addition, significant differences in category distributions may also make pretrained CNNs on ImageNet (with 1,000 classes) unsuitable for analyzing medical images.

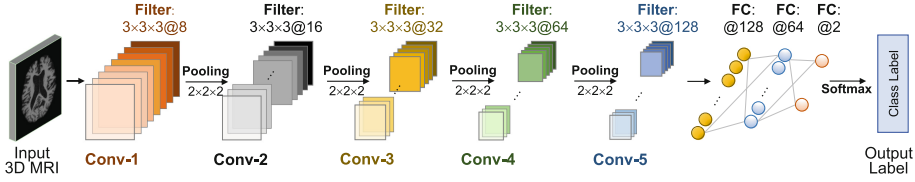


Fig. 1. Architecture of the baseline CNN model for 3D MRI classification, consisting of five convolutional (Conv) layers (with each followed by a pooling layer) and three fully-connected (FC) layers (followed by softmax activation). The term "$3 \times 3 \times 3$@8" denotes a convolutional layer with 8 filters (kernel size: $3 \times 3 \times 3$).

A reasonable alternative is to train 3D-CNNs from scratch on a source medical image dataset/site, and then to fine-tune them on a to-be-analyzed target dataset/site [9,10]. However, training 3D-CNNs from scratch has to face two challenges. *First*, the sample size is often limited (*e.g.*, tens or hundreds). Utilizing inappropriate network architecture will lead to severe over-fitting, which may greatly affect the transferability. *Second*, different sites/datasets may have significant domain shifts in terms of data distribution, caused by different scanners, protocols and populations [11–14]. Thus, it is highly desirable to boost the transferability of 3D-CNNs for medical image analysis.

In this paper, we aim to answer a key question: *how to train transferable 3D-CNNs from scratch with relatively small-sized medical images*? We study the case of structural MRI-based brain disorder classification on 4 benchmark datasets via two tasks: (1) Alzheimer's disease (AD) detection; and (2) mild cognitive impairment (MCI) conversion prediction. We build a baseline 3D-CNN model (see Fig. 1), and evaluate the performance of several variants of the baseline CNN model with different architectures. We then analyze the transferability of the baseline network with transfer learning on different target domains. We also explore the between-task transferability, with the CNN model trained on a source task (*i.e.*, AD detection) and applied to a related task (*i.e.*, MCI conversion prediction). Our empirical findings will provide some insights on how to improve the transferability of CNNs in 3D medical image analysis.

2 Materials and Methodology

2.1 Studied Subjects and MR Image Pre-processing

Four benchmark datasets are used in this work, *i.e.*, Alzheimer's Disease Neuroimaging Initiative (ADNI-1) [15], ADNI-2 [16], ADNI-3 [17] and Australian Imaging Biomarkers and Lifestyle Study of Aging database (AIBL) [18]. Note that the subjects that simultaneously appear in ADNI-1, ADNI-2 and ADNI-3 are removed from ADNI-2 and ADNI-3 to avoid data leakage and guarantee comparison fairness as suggested in [7]. Their domain heterogeneity mainly comes from the use of different scanning parameters (*e.g.*, 1.5T or 3T) and updated scanners. ADNI-1 contains 748 subjects with 1.5T T1-weighted structural MRIs, including 205 AD, 231 cognitively normal (CN), 165 progressive MCI (pMCI) and 147 stable MCI (sMCI) subjects. ADNI-2 has 708 subjects with 3T T1-weighted structural MRIs (*i.e.*, 162 AD, 205 CN, 88 pMCI and 253 sMCI). Note that those subjects with pMCI would convert to MCI within 36 months and sMCI remains stable. ADNI-3 involves 389 subjects with 3T T1-weighted structural MRIs (*i.e.*, 60 AD and 329 CN). AIBL consists of MRIs acquired from 549 subjects (*i.e.*, 71 AD, 447 CN, 11 pMCI and 20 sMCI). The demographic and clinical information of the studied subjects is shown in *Supplementary Materials*.

All brain MRIs are pre-processed through a standard pipeline, including skull stripping, intensity inhomogeneity correction, image re-sampling, and spatial normalization to the Automated Anatomical Labeling (AAL) template.

2.2 Methodology

Architecture of Baseline CNN. Figure 1 illustrates the architecture of the baseline CNN used in this work. This network consists of 5 convolutional (Conv) layers with $3 \times 3 \times 3$ filters, and 3 fully-connected layers with 128, 64 and 2 neurons, respectively. A softmax layer is used for classification. Each Conv layer is composed of a sequence of 3D convolutional filters, followed by batch normalization and ReLU activation function. To reduce the risk of over-fitting, a $2 \times 2 \times 2$ max pooling operation (stride: $2 \times 2 \times 2$) is added after each Conv layer. This is a very basic CNN model that can be flexibly extended in different applications.

Network Training. Considering the relatively larger number of subjects in ADNI-1, we use ADNI-1 as training/source data for model training and validation, whereas the other three datasets are treated as test/target data for transferability evaluation. A 5-fold cross-validation strategy is used. That is, subjects in ADNI-1 are randomly partitioned into 5 folds. For parameter selection, each fold is treated as the validation set in turn, with the rest as the training set.

The trained model is finally applied to target/test data. Such process is repeated five times to avoid bias caused by random partition. For network training, the Adam algorithm is used as the optimizer, with a learning rate of 0.0001. The batch size is set to 2. A drop-out rate of 0.5 is used to avoid over-fitting. We train the network for 30 to 50 epochs, and employ an early stopping strategy when the training or the validation loss continuously changes little for 10 epochs.

Table 1. Results of AD detection achieved by 3D-CNNs with different network depths on three target domains (with models trained on ADNI-1).

Target Domain	Method	AUC (%)	ACC (%)	BAC (%)	SEN (%)	SPE (%)
ADNI-2	CNN-5	93.69 ± 1.34	84.25 ± 2.69	82.65 ± 3.16	69.01 ± 7.23	96.29 ± 1.45
	CNN-7	91.81 ± 0.92	84.69 ± 1.58	83.51 ± 1.65	73.46 ± 2.66	93.56 ± 1.70
	CNN-10	91.38 ± 0.95	83.87 ± 0.70	82.27 ± 0.76	68.64 ± 2.56	95.90 ± 2.03
	CNN-12	91.35 ± 2.48	82.88 ± 2.24	81.42 ± 2.54	68.89 ± 5.53	93.95 ± 1.97
ADNI-3	CNN-5	94.57 ± 0.96	91.00 ± 0.92	78.87 ± 3.68	61.33 ± 8.85	96.41 ± 1.91
	CNN-7	93.51 ± 1.17	90.59 ± 1.91	81.76 ± 3.19	69.00 ± 5.73	94.53 ± 1.78
	CNN-10	91.53 ± 1.05	91.67 ± 1.24	78.59 ± 1.70	59.67 ± 3.61	97.51 ± 1.56
	CNN-12	92.25 ± 2.18	90.54 ± 1.15	80.10 ± 4.56	65.00 ± 9.34	95.20 ± 1.80
AIBL	CNN-5	92.55 ± 0.99	91.00 ± 1.21	82.94 ± 2.99	71.83 ± 7.65	94.05 ± 2.21
	CNN-7	89.25 ± 0.90	89.94 ± 0.76	83.06 ± 1.04	75.77 ± 4.61	90.34 ± 2.73
	CNN-10	90.84 ± 1.29	91.62 ± 1.42	80.33 ± 2.51	64.79 ± 5.63	95.88 ± 1.86
	CNN-12	89.60 ± 1.75	90.50 ± 1.56	81.94 ± 2.33	70.14 ± 5.92	93.74 ± 2.32

Evaluation Metric. Two classification tasks are included: (1) AD detection (*i.e.*, AD vs. CN classification), and (2) MCI conversion prediction (*i.e.*, pMCI vs. sMCI classification). Five metrics are used for performance evaluation, including (1) area under the ROC curve (AUC), (2) classification accuracy (ACC), (3) balanced accuracy (BAC), (4) sensitivity (SEN), and (5) specificity (SPE).

3 Transferability Vs. Different Network Architectures

Due to issues of small-sample-size and high feature dimension of medical images, it is often challenging to design a suitable network architecture to obtain good transferability. We now explore the influence of network capacity (in terms of *network depth* and *network width*) on the transferability of 3D-CNNs.

Influence of Network Depth. To explore the influence of network depth on transferability, we develop three variants of the baseline CNN with different number of convolutional (Conv) layers, including (1) **CNN-7** that contains 7 Conv layers with 8, 16, 32, 64, 64, 128 and 128 filters, respectively; (2) **CNN-10** that consists of 10 Conv layers with 8, 8, 16, 16, 32, 32, 64, 64, 128, and 128, respectively (with detailed architecture shown in *Supplementary Materials*); and (3) **CNN-12** that contains 12 Conv layers with 8, 8, 16, 16, 32, 32, 64, 64, 64, 128, 128 and 128 filters, respectively. The baseline model with 5 Conv layers is called **CNN-5** (see Fig. 1). The FC layers and the training strategy of these variants remain the same as the baseline CNN-5.

The performance of these models on three target/test domains is listed in Table 1. From the results, we can derive the following empirical findings. *First*, adding more layers does not necessarily lead to better AUC and ACC performance. For example, the AUC values of CNN-5 are generally better than those achieved by three deeper CNNs on three test datasets, and there are no significant differences in terms of ACC of four networks (with p-values = 0.5278,

Table 2. Results of AD detection achieved by 3D-CNNs with different network widths on three target domains (with models trained on ADNI-1).

Target Domain	Method	AUC (%)	ACC (%)	BAC (%)	SEN (%)	SPE (%)
ADNI-2	CNN-w1.0	93.69 ± 1.34	84.25 ± 2.69	82.65 ± 3.16	69.01 ± 7.23	96.29 ± 1.45
	CNN-w2.0	94.23 ± 0.55	86.87 ± 2.17	85.59 ± 2.58	74.69 ± 6.13	96.49 ± 1.22
	CNN-w4.0	93.52 ± 0.86	87.63 ± 1.61	86.94 ± 1.80	81.11 ± 4.91	92.78 ± 3.11
	CNN-w5.0	90.71 ± 1.92	83.38 ± 1.49	82.03 ± 1.56	70.49 ± 3.01	93.56 ± 2.14
ADNI-3	CNN-w1.0	94.57 ± 0.96	91.00 ± 0.92	78.87 ± 3.68	61.33 ± 8.85	96.41 ± 1.91
	CNN-w2.0	96.36 ± 0.58	92.70 ± 0.72	83.83 ± 2.10	71.00 ± 5.35	96.66 ± 1.47
	CNN-w4.0	94.71 ± 1.39	90.80 ± 2.56	84.88 ± 1.29	76.33 ± 5.94	93.61 ± 4.10
	CNN-w5.0	91.51 ± 2.10	89.51 ± 1.86	82.76 ± 2.62	73.00 ± 7.11	92.52 ± 3.01
AIBL	CNN-w1.0	92.55 ± 0.99	91.00 ± 1.21	82.94 ± 2.99	71.83 ± 7.65	94.05 ± 2.21
	CNN-w2.0	92.37 ± 0.98	90.42 ± 1.80	85.21 ± 1.47	78.03 ± 4.29	92.39 ± 2.55
	CNN-w4.0	91.82 ± 0.56	88.38 ± 4.55	84.90 ± 1.58	80.00 ± 7.87	89.80 ± 6.50
	CNN-w5.0	88.96 ± 1.84	86.60 ± 2.99	81.41 ± 1.86	72.05 ± 3.75	90.77 ± 2.27

0.5431 and 0.2348 on 3 target domains). *Second*, deeper CNNs tend to produce better sensitivity to precisely detect AD subjects. For example, CNN-7 achieves obvious improvement compared with CNN-5. However, adding too many layers (*e.g.*, > 10) does not necessarily benefit the detection rate of AD. *Besides*, all four networks show fluctuating performance (with larger standard deviations) in terms of sensitivity than other metrics. In contrast, all CNNs achieve relatively stable detection rate of cognitively normal samples. This implies that the cross-domain heterogeneity of AD subjects has more significant influence on the transferability of CNNs. It is interesting to pay more attention to patients when designing cross-domain transfer learning models.

Influence of Network Width. In this work, we refer to the number of filters in each convolutional layer as the network width. To explore the influence of network width on transferability, we develop three variants of the baseline CNN, by increasing the number of filters in each layer using a widening factor w. When $w = 2.0$, for example, the number of filters in each layer is doubled, and we name this variant as **CNN-w2.0**.

The results of CNNs with different network widths on three test domains are shown in Table 2. From Table 2, we can derive the following empirical findings. *First*, the network width does influence the transferability, and using very small or very large width will not benefit the transferability of CNN models. When the widening factor w> 1.0, the network tends to achieve better overall performance in terms of most metrics. But when $w = 5.0$, there is a significant performance degradation. This implies that there is a boundary when widening the network to achieve better transferability. *Besides*, widening the network helps increase the sensitivity for AD detection. Compared with the baseline CNN (*i.e.*, CNN-w1.0), wider networks (*e.g.*, CNN-w4.0) tend to have a significant better detection rate (SEN) of AD subjects, and its effect is more pronounced compared to increasing network depth. The possible reason is that using more filters helps capture more local-to-global discriminative features in brain MRIs to identify AD subjects.

4 Transferability Vs. Different Network Components

Influence of Fine-Tuning. A previous study proposed to fine-tune CNNs pretrained on ImageNet for 2D medical image analysis [19]. Although CNNs are pretrained with massive natural images, it is reported that when they are used for medical image analysis, fine-tuning cannot effectively improve their transferability. To analyze the behavior of the proposed 3D-CNN trained from scratch, we use some labeled target data to fine-tune the pretrained baseline CNN and evaluate its performance on the target domain. Specifically, we use ADNI-1 as the source domain to train the baseline CNN, and then fine-tune and test the network on ADNI-2. 20% of the samples in ADNI-2 are randomly selected and used for fine-tuning, while the remaining samples are used as test data for evaluating the transferability. Since the baseline CNN is trained in a 5-fold cross-validation manner, we can obtain five pretrained CNNs. Accordingly, the fine-tuning and evaluation procedures are carried out five times.

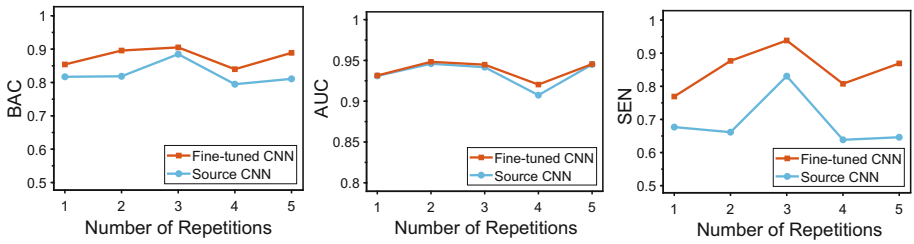


Fig. 2. Performance of fine-tuned and source CNNs in AD detection.

The evaluation results on ADNI-2 in terms of three key metrics are shown in Fig. 2. For comparison, we also report the results achieved by the pretrained CNNs (denoted as source CNN) without fine-tuning. From this figure, we have the following observations. *First*, the overall performance of the fine-tuned CNN is better than the source CNN pretrained on ADNI-1. The underlying reason may be that unlike the pre-trained 2D-CNN on ImageNet, the 3D-CNN trained from scratch using 3D MRIs can learn more disease-related information patterns. This provides a better initialization for the network, which lays the foundation for further network optimization through fine-tuning. *In addition*, fine-tuning makes the network achieve a more balanced accuracy (BAC). As shown in the right of Fig. 2, the most significant improvement comes from the detection of AD subjects (with much higher SEN). These results further support the conclusion in Sect. 3 that AD subjects are more informative in enhancing the transferability of 3D-CNNs, compared with cognitively normal subjects.

Contribution of Different Layers to Mitigating Domain Shift. When we train a 3D-CNN from scratch and apply it to a different domain, an interesting question is to determine the contribution of different layers on mitigating cross-domain differences in data distribution. That is, we'd like to investigate which

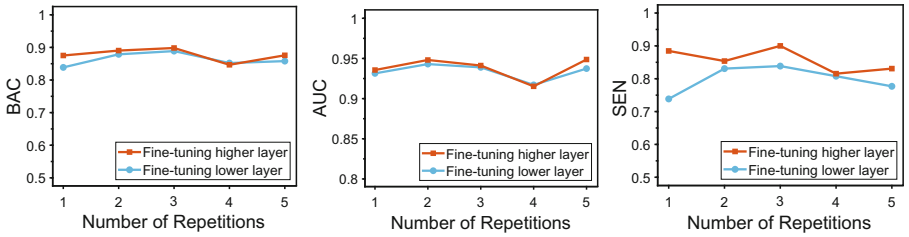
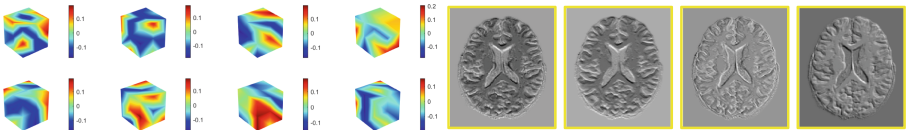


Fig. 3. Performance comparison of fine-tuning higher layer of the pretrained CNN and fine-tuning the first layer of the CNN in AD detection.



(a) Several learned filters at Conv1 (b) Feature maps generated by Conv1

Fig. 4. Visualization of learned convolution filters and feature maps at the first convolution layer of the fine-tuned CNN in AD detection.

part of a CNN is more easily influenced by the cross-domain data heterogeneity. We address this problem by fine-tuning certain layers while keeping the weights of other layers frozen. If fine-tuning a certain layer can achieve more performance improvement, it is assumed that this layer contributes more to alleviating the domain shift problem. Here, we focus on the low-level part (Conv-1) and the high-level part (Conv-5 and three fully-connected layers) in the baseline CNN. We fine-tune these two parts separately. That is, when fine-tuning the low-level part, the remaining layers are frozen; and vice versa.

Figure 3 reports the results in terms of three key metrics. It can be seen that fine-tuning the high-level part of the CNN produces better SEN and BAC results in most cases, compared with fine-tuning its low-level part. This implies that the high-level part with finer scales contributes more to mitigating the cross-domain data heterogeneity. The low-level part may help extract more domain-independent information. To verify this assumption, we visualize several Conv filters and feature maps in Conv-1 of the fine-tuned CNN in Fig. 4, from which we can see that Conv-1 mainly extracts low-level features (with enhanced edges).

5 Transferability to Related Task

We further investigate the between-task transferability when a model is trained on a source task from scratch and applied to a related target task. Considering that MCI is the prodromal stage of AD [20,21], we develop two CNNs which are trained with AD and CN subjects, and then evaluate their transferability for MCI conversion prediction. 1) CNN-AD1: Baseline CNN and a widened CNN (with w

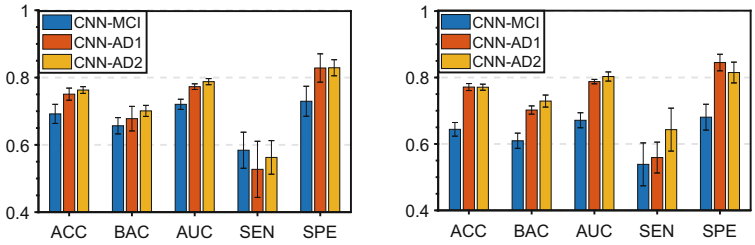


Fig. 5. Results of the baseline network (left) and its widened variant CNN-w4.0 (right), with models tested on ADNI-2 for MCI conversion prediction.

= 4.0) trained with AD and CN samples from ADNI-1. 2) CNN-AD2: Baseline CNN and a widened CNN (with w = 4.0) trained with AD and CN samples from ADNI-1 and ADNI-2. For comparison, we also train a baseline CNN and a widened CNN (with w = 4.0) with MCI samples in ADNI-1. All these networks are tested on ADNI-2 for MCI conversion prediction. The experiment is repeated five times and the results are shown in Fig. 5.

From Fig. 5, one can observe that the networks trained with only AD and CN samples can achieve comparable or even better performance in MCI conversion prediction, compared with CNN-MCI trained on MCI samples. The widened CNNs trained on AD and CN samples achieve greater transferability for MCI conversion prediction, compared with the baseline CNN. Especially, when using more training samples, CNN-AD2 produces much higher SEN values, indicating its superiority in detecting pMCI patients. We also observe that the widened CNN trained on MCI samples suffers some performance decrease which may be attributed to the limited number of training samples. The widened CNN trained on samples from ADNI-1 and ADNI-2 achieves the overall best performance. This may be due to the relatively large amount of training samples and the potential relationship between MCI and AD populations. These results show that it is beneficial to train a CNN model on one task and transfer it to related tasks.

6 Conclusion and Future Work

In this paper, we have conducted an experimental study on how to train a transferable CNN for 3D medical image analysis. Based on the results for brain MRI-based brain disorder classification, we have made some empirical findings. (1) Appropriately adding more layers and widening the network width (*i.e.*, using more filters at each Conv layer) is helpful to improve the transferability, especially for the enhancement of sensitivity. (2) Transfer learning via fine-tuning is beneficial to increase the transferability of the 3D CNN (trained from scratch on a source domain). Fine-tuning high-level layers is helpful to alleviate the domain shift issue. (3) A CNN trained on a specific task (*i.e.*, AD detection) with more training samples can be successfully transferred to a different but related

task (*i.e.*, MCI conversion prediction) with small-sized samples. The empirical findings may provide the community with some useful references and techniques on how to leverage 3D-CNNs for various medical image analysis tasks.

In the future, we will study the generalizability of our trained models in identifying other brain disorders such as Parkinson's disease and autism. In addition, we will explore neural architecture search for the analysis of architectures. Since different diseases may affect different brain regions, it is interesting to integrate an attention detection module to CNN, which will also be our future work.

Acknowledgements. This work was supported in part by NIH grants (Nos. AG041721, MH109773 and MH117943).

References

1. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
2. Mazurowski, M.A., Buda, M., Saha, A., Bashir, M.R.: Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. J. Magn. Reson. Imaging **49**, 939–954 (2019)
3. Guan, H., Liu, Y., Yang, E., Yap, P.T., Shen, D., Liu, M.: Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. Med. Image Anal. **71**, 102076 (2021)
4. Morid, M.A., Borjali, A., Del Fiol, G.: A scoping review of transfer learning research on medical image analysis using ImageNet. Computers in Biology and Medicine (2020)
5. Cuingnet, R., Gerardin, E., Tessieras, J., et al.: Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage **56**(2), 766–781 (2011)
6. Tajbakhsh, N., et al.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging **35**, 1299–1312 (2016)
7. Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., et al.: Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med. Image Anal. **63**, 1–19 (2020)
8. Guan, Z., Kumar, R., Fung, Y.R., Wu, Y., Fiterau, M.: A comprehensive study of Alzheimer's disease classification using convolutional neural networks. arXiv:1904.07950 (2019)
9. Liu, M., Zhang, J., Adeli, E., Shen, D.: Landmark-based deep multi-instance learning for brain disease diagnosis. Med. Image Anal. **43**, 157–168 (2018)
10. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. arXiv:2102.09508 (2021)
11. AlBadawy, E.A., Saha, A., Mazurowski, M.A.: Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. Med. Phys. **45**(3), 1150–1158 (2018)
12. Pooch, E.H., Ballester, P.L., Barros, R.C.: Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. arXiv:1909.01940 (2019)
13. Stacke, K., Eilertsen, G., Unger, J., Lundström, C.: A closer look at domain shift for deep learning in histopathology. arXiv:1909.11575 (2019)

14. Wang, M., Zhang, D., Huang, J., Yap, P.T., Shen, D., Liu, M.: Identifying autism spectrum disorder with multi-site fMRI via low-rank domain adaptation. IEEE Trans. Med. Imaging **39**(3), 644–655 (2019)
15. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging **27**(4), 685–691 (2008)
16. Jack Jr, C.R., et al.: Magnetic resonance imaging in Alzheimer's Disease Neuroimaging Initiative 2. Alzheimer's Dementia **11**(7), 740–756 (2015)
17. Weiner, M.W., et al.: The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement. Alzheimer's Dementia **13**(5), 561–571 (2017)
18. Ellis, K.A., Bush, A.I., Darby, D., et al.: The Australian Imaging, Biomarkers and lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int. Psychogeriatrics **21**(4), 672–687 (2009)
19. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 3347–3357 (2019)
20. Gauthier, S., et al.: Mild cognitive impairment. Lancet **367**(9518), 1262–1270 (2006)
21. Sabbagh, M.N., et al.: Early detection of mild cognitive impairment (MCI) in primary care. J. Prev. Alzheimer's Disease **7**, 165–170 (2020)