Learning with Domain-Knowledge for Generalizable Prediction of Alzheimer's Disease from Multi-site Structural MRI

Yanjie Zhou, Youhao Li, Feng Zhou, Yong Liu, and Liyun Tu(✉)

School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
tuliyun@bupt.edu.cn

Abstract. Construct a generalizable model for the diagnosis of Alzheimer's disease (AD) is an important task in medical imaging. While deep neural networks have recently advanced classification performance for various diseases using structural magnetic resonance imaging (sMRI), existing methods often provide suboptimal and untrustworthy results because they do not incorporate domain-knowledge and global context information. Additionally, most state-of-the-art deep learning methods rely on multi-stage preprocessing pipelines, which are inefficient and prone to errors. In this paper, we propose a novel domain-knowledge-constrained neural network for automatic diagnosis of AD using multi-center sMRI. Specifically, we incorporate domain-knowledge into a ResNet-like architecture. We explicitly enforce the network to learn domain invariant and domain specific features by jointly training multiple weighted classifiers, so that pixel-wise predictive performance generalizes to unseen images. In addition, the network directly takes segmentation-free and patch-free images in original resolution as input, which offers accurate inference with global context information and accurate individualized abnormalities to further refines reproducible predictions. The framework was evaluated on a set of sMRI collected from 7 independent centers. The proposed approach identifies important discriminative brain abnormalities associated with AD. Experimental results demonstrate superior performance of our method compared to state-of-the-art methods.

Keywords: Domain-knowledge encoding · Patch-free · Structural magnetic resonance imaging (sMRI) · Alzheimer's disease

1 Introduction

Alzheimer's disease (AD) is one of the most pervasive neurodegenerative disorders, causing an increasing morbidity burden that may outstrip diagnosis and management capacity with the population ages. The assessment of AD usually involves the acquisition of structural magnetic resonance imaging (sMRI) images, since it offers accurate visualization of the anatomy and pathology of the brain. Brain abnormalities (e.g., atrophy, enlargement, malformation) are known to be

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 452–461, 2023.
https://doi.org/10.1007/978-3-031-43904-9_44

the most discriminative and reliable biomarkers [1] of AD that can be observed and analyzed through sMRI. However, automatic and reproducible identification of AD remains challenging due to heterogeneous of sMRI collected from different centers.

Recently, convolutional neural networks (CNN) have been used for automatic classification of AD from sMRI. Many methods [2,3] use a bag of patches selected from the skull-stripped brain region, which ignores global context information that can play a significant role in identifying lesions for accurate inference [4]. Many studies [5–8] proposed to characterize AD using segmented anatomies (e.g., gray matter or hippcampus). These methods rely on the accurate segmentation of the anatomies which is usually performed in a multi-stage data processing pipeline with the help of third-party softwares (e.g., FreeSurfer [9]) driven by a prior template. However, template-driven methods depend on variable image registration accuracy and highly affected by the anatomical variability between subjects, introducing errors to the characterization of individualized abnormalities. Similarly, methods (e.g., [10]) use detected landmarks also depend on a template-driven pipeline. Taking advantage of attention mechanism, some methods [5] proposed to diagnose AD using sMRI images from multiple centers. However, the classification performance is either hardly reproducible or difficult to compare across studies. One of the major reasons is that existing methods are often trained with samples from the same training (source) domain, while testing samples come from an independent new (target) domain with a different feature distribution. In the literature, this situation relates to domain adaptation [11–16] or domain generalization [17–19]. A widely used solution for the problem is to learn a domain-invariant latent feature space [20]. Unfortunately, there is no guarantee that the target samples' features will fall into the shared source domain-invariant representation, and in practice it is that new domains typically do not.

In this paper, we propose a novel domain-knowledge-constrained neural network for the diagnosis of AD using sMRI from multiple source domains. We designed a new domain-knowledge encoding module into a ResNet-like architecture for feature learning that yields a latent feature space with domain specific and domain shared information. In addition, we propose to use segmentation-free, resampling-free, patch-free 3D sub-images, which offers global context information and subject-level abnormalities to further refines generalizable and reproducible predictions.

2 Methods

We propose to design and implement an end-to-end neural network (Fig. 1) for automatic, robust, and reproducible diagnosis of AD using sMRI images, with the hope to identify and understand the most discriminative anatomical regions associate with AD. The model operates in 3 major steps: a) crop the input sMRI image to keep a sub-region (red rectangle), containing relevant anatomy structures (e.g., hippocampus, caudate, ventricles) associate with AD; b) extract

features shared by all training sources based on ResNet [21]; c) design a domain-knowledge encoding module and a set of label predictors to constrain the feature learning process for better generalization.

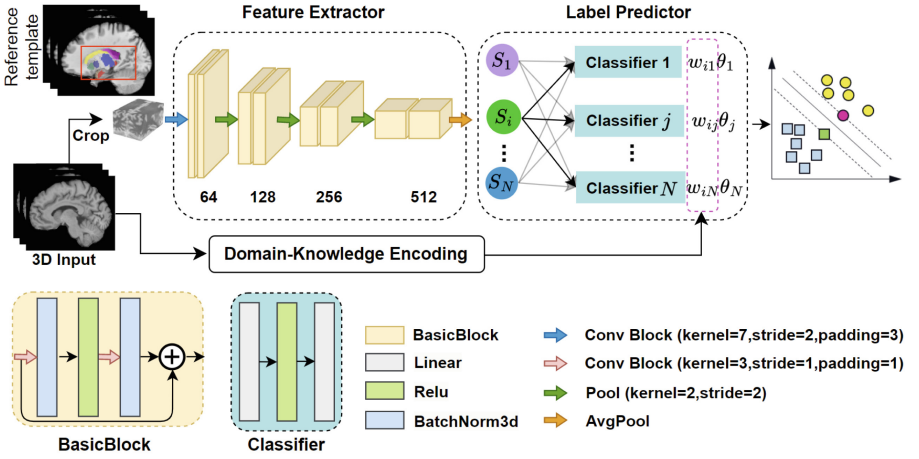

Fig. 1. Schematic of the proposed generalizable classification model. **Feature extractor** is a ResNet18-like 3D network that extracts high-dimensional features from MRI images for classification using 3D convolution and residual connection. **Basic block** is the basic component of the feature extractor and consists of two 3D convolutional layers, two BatchNorm layers, a ReLu layer and residual connection. **Classifier** is a multilayer perceptron (MLP), consisting of two linear layers and a ReLu layer. **Domain-Knowledge Encoding** captures domain invariant features and domain-specific features and generates weights for classifiers based on domain similarity. **Label Predictor** specifies that our model has multiple mutually independent classifiers, and the predictions of all classifiers are weighted and summed to obtain the final output. (Color figure online)

2.1 Patch-Free 3D Feature Extractor

We first estimate a bounding box around relevant anatomical objects in the input sMRI. The objects are automatically identified by affine registration, which transforms the reference template to each image in the dataset to estimate label for the image. We note that, the estimated labels are only used to locate the bounding box, it has no effect on the individual's atrophy since we pad extra space to ensure the cropped image contain all interested objects with respect to registration errors. Then, we crop the input image using the located bounding box to obtain the sub-image as input to our network. It need to be clarified that the cropping size is a fixed tuple determined by the maximum bounding box containing informative anatomical objects associated with AD.

To encode global context information, we propose a patch-free 3D feature extractor for different source domains, which is expected to learn domain-invariant features while not eliminating domain-specific features. Each domain has a unique label classifier, allowing adjustments for domain differences. Based on ResNet, we design our feature extractor as shown in Fig. 1. Each basic block consists of two convolutional layers. Each convolutional layer is followed by a batch normalization and a nonlinear activation function LeakyReLU. The basic block can be wrote as:

$$X_{l+1} = F(W_i, X_l) + W_s X_l, \tag{1}$$

where X_l and X_{l+1} are the input and output of the basic block and $F(W_i, X_l)$ denotes the nonlinear mapping in the basic block. Since the dimensions of X and $F(W_i, X)$ must be the same for summation, we use the linear mapping W_s to adjust the dimensions of X in the shortcut connection.

In the proposed method, we use global average pooling function which is more suitable for disease classification, because the global average pooling operation reflects the information of gray matter volume in brain regions and preserves the relative position relationship between different channels of the feature map.

In the output layer, we use a softmax classifier based on cross-entropy loss to calculate the loss between the predicted and true labels.

$$\mathcal{L} = \textit{cross-entropy}(\widehat{Y}_i(X_i \in D_s; \omega), Y_i) \tag{2}$$

2.2 Global Average Pooling

Global average pooling solves the problem of excessive image feature dimensions. If the feature maps of 3D images are directly expanded for classification, it will significantly increase the number of classifier parameters and increase the time and space complexity of training. Global average pooling averages the 3D feature maps in the channel dimension, preserving the relative position relationship between channels and reducing the resources required for model training.

The dimension change in the global average pooling is $[B, C, D, H, W] \rightarrow [B, C, 1, 1, 1]$, where B denotes the batch-size and C denotes the channel number.

$$GAP(\delta) = \frac{1}{D \times H \times W} \sum_{i=1}^{D} \sum_{j=1}^{H} \sum_{k=1}^{W} \delta_{i,j,k} \tag{3}$$

where δ denotes the image feature extracted by ResNet, and D, H, W denote the three dimensions of the feature.

Since global average pooling has fewer parameters, it can prevent over-fitting to some extent, further more, global average pooling sums out the spatial information, thus it is more robust to spatial translation of the input.

2.3 Domain-Knowledge Encoding

The domain-knowledge encoding module is designed to give relative similarity weights to source domains from a new sample. The weights reflect the similarity between the testing sample and source domains, allowing the module to share strength only between similar domains.

Our model uses multiple classifiers for prediction from the features extracted by the feature extractor. The classifiers are independent from each other. We feed the image features to different classifiers and generate weights to each classifier, summing the predictions of each classifier according to the weights as the final output.

$$\widehat{Y} = \sum_{j=1}^{c\_num} \omega_{ij} \cdot classifier_j(\delta(X \in D_i), \theta_j) \tag{4}$$

where $\widehat{Y}$ denotes the prediction result of X, c_num denotes the number of classifiers, D_i denotes the center which X belongs, δ denotes the extracted feature from X, $classifier_j$ denotes one classifier and θ_j are the parameters in $classifier_j$.

Multiple classifiers can capture the invariant and specific feature distributions between different domains, comparing the similarity of feature distributions between training source and unseen target domains by a joint training of the admixture classifiers, generating weights to integrate the feature distributions of known domains to fit the unknown domain feature distributions.

3 Experiments and Results

3.1 Data Description

Structural T1-weighted brain MRI data of 809 subjects (468 male, 341 female, age 68.16 ± 8.12 years, range 42–89 year) were acquired from 7 in-house independent multiple centers as detailed in [5,22]. In total, 552 subjects (295 of normal control (NC), 257 of AD) were used for leave-center-out training. The rest 257 subjects with mild cognitive impairment (MCI) were used as an independent dataset for evaluation and compared with clinical diagnosis metrics.

3.2 Implementation Details

We first evaluated the model using leave-center-out cross-validation, where one center was selected for testing at a time and all remaining centers were used for training. Then, we applied the trained model on an independent validation set of unseen images for subjects with MCI. All images were cropped to have the same size of $[80, 128, 72]$. Image features were extracted with $3 \times 3 \times 3$ convolution in the network and $2 \times 2 \times 2$ convolution with a stride of 2 replacing the maximum pooling. The extracted features were passed through a global average pooling layer (Sect. 2.1). $N = 6$ independent classifiers were used.

During training, we sorted all training centers and feed the image features from $site_i$ to all classifiers, and set the weight of $classifier_{j(j=i)}$ to 1 and the weight of the rest classifiers to 0. We used cross-entropy to calculate the prediction error and update the parameters of the feature extractor and $classifier_j$ by backpropagation. In testing stage, we feed the image features from the test center to all classifiers, and the final prediction was used the weighted average of predicted probability over all classifiers as the final prediction.

We used SGD algorithm to optimize the model coefficients, and set the initial learning rate to 0.001 and reduce the learning rate to one-tenth of the previous value every 50 epochs. The method was implemented using PyTorch 1.1 with Python 3.7. The experiments were run on an Intel Xeon CPU with 16 cores, 43 GB. RAM and a NVIDIA A5000 GPU with 24 GB memory. The code and model are available at https://github.com/Yanjie-Z/DomainKnowledge4AD.

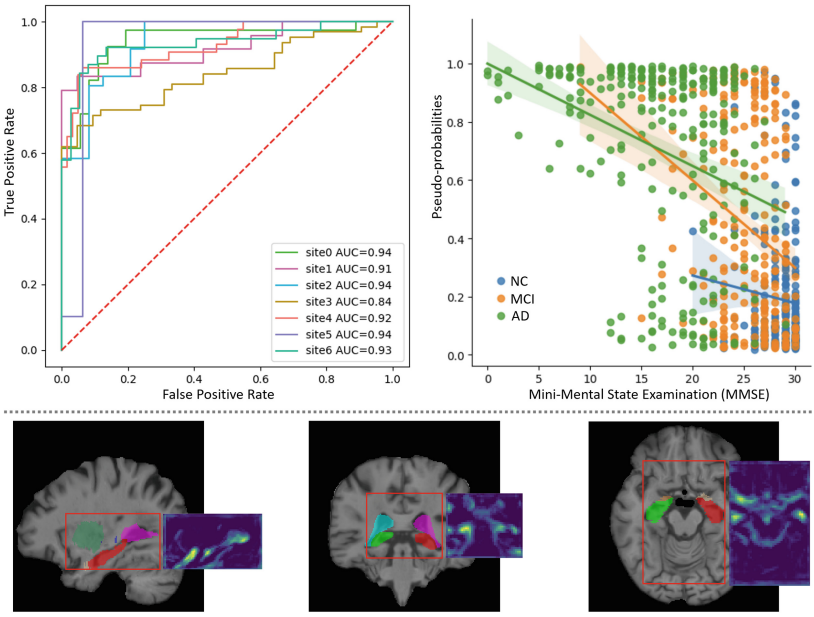

Fig. 2. First row: the left panel evaluates the AUC-ROC curve for each domain through leave-center-out cross validation, and the right panel investigates the association between the predicted probabilities and clinical measure (MMSE) in subjects with Alzheimer's disease (AD), mild cognitive impairment (MCI), and healthy controls (NC). Second row: attention map for an arbitrary example sMRI of a subject with AD, illustrating the most discriminative features learnt from the proposed approach.

3.3 Performance Evaluation

To evaluate the proposed approach, we feed 2 different types of input to the conventional 3D-ResNet [21] and each obtains a models: 1) ResNet, which use the

original image as input, and 2) Baseline, which use the bounding box cropping strategy as proposed in Sect. 2.1. In addition, we incorporated the patch-free cropping strategy inspired by [4] to crop the middle-half sub-region of the original input sMRI image of the brain, and feed to ResNet, which we denote as ResNet-PF. The prediction performance are compared in Table 1.

Table 1. Comparisons among different methods with leave-center-out cross-validation. Abbreviations: ACC = accuracy, AUC = area under the curve of the receiver operating characteristic, AVG = average performance over centers. ACC in percentage.

		S0	S1	S2	S3	S4	S5	S6	AVG
ResNet	LOSS	0.90	0.53	0.35	1.34	0.53	0.38	0.56	0.66
	ACC	87.05	88.31	**90.83**	72.14	79.39	87.57	87.71	84.71
	AUC	0.91	0.91	0.96	0.83	0.86	0.94	0.94	0.91
ResNet-PF	LOSS	0.79	0.53	0.39	1.76	0.64	0.38	0.45	0.71
	ACC	84.00	86.67	88.89	72.63	82.86	95.56	89.78	85.77
	AUC	0.91	0.90	0.94	0.83	0.86	0.95	0.94	0.90
Baseline	LOSS	0.47	0.50	0.37	1.37	0.72	0.24	0.42	0.58
	ACC	87.66	87.03	88.36	71.40	82.85	95.40	89.00	**85.95**
	AUC	0.93	0.86	0.95	0.83	0.92	0.95	0.93	**0.91**
Proposed	LOSS	0.32	0.39	0.34	0.85	0.34	0.20	0.33	0.39
	ACC	**90.79**	**88.88**	88.33	**74.28**	**91.42**	**97.77**	**93.33**	**89.25**
	AUC	0.94	0.92	0.94	0.84	0.93	0.94	0.93	**0.92**

Our model achieves an average classification accuracy of 89.25% on all test centers during cross-validation, compared to the average classification accuracy of 85.95% with baseline (without the use of domain knowledge encoding module).

We used AUC-ROC curves to evaluate the classification effectiveness [13,17,23] of the model on the test centers, and we counted the AUC-ROC curves for seven centers and compared them accordingly in Fig. 2.

To evaluate the interpretability of the model, we used Grad-CAM [24] to analyze the sensitive regions of the model in discriminating AD. We found that the model focused on the hippocampus in the images during prediction, which confirms that AD and the hippocampus have a significant correlation. We also find that the model pays more attention to the hippocampus in discriminating AD than healthy controls. Figure 3 compares the 3D attention map between a subject with AD and a healthy subject who never has AD, demonstrating obvious higher values in hippocampus region.

4 Discussion

We proposed a novel reproducible and generalizable neural network to assist the automatically diagnosis of AD that benefits from domain knowledge and

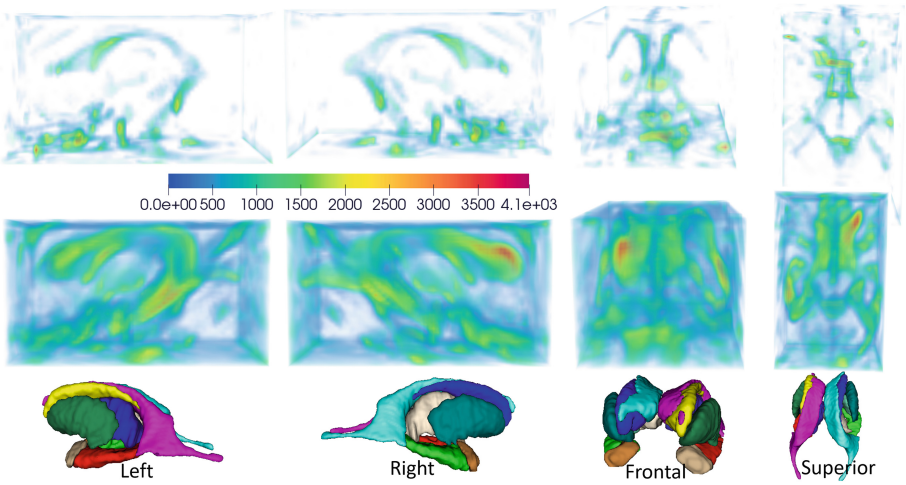

Fig. 3. 3D attention maps for a healthy subject (first row) and a subject with AD (second row) in 4 different views (column). The bottom row shows a visual navigator.

global contextual information with the help of segmentation-free, resampling-free, patch-free sub-image. The model was evaluated with leave-center-out cross-validation and with an independent set of unseen images for subjects with MCI (Fig. 2). It obtains an average accuracy of 89.25%, loss of 0.39 and AUC of 0.92 comparing with 85.95%, 0.58 and 0.91 using ResNet. We apply the proposed model to images from a new domain (never used during training), demonstrating promising results.

We did ablation studies to evaluate the proposed method (Table 1), unsurprisingly, the cropped images obtain the best performance. Figures 2 and 3 evaluated the explainability of the proposed neural network. The results suggest that the hippocampus and ventricles regions suffer the most in AD, which are consistent with multi-stage segmentation-based methods [5], and clinical measures (in terms of MMSE) on an independent dataset (Fig. 2).

Our results and all comparative frameworks tend to predict worse for center 3, probably because it has some subjects with AD who have higher MMSE (Fig. 2) making the diagnosis challenging. As opposite, all models provide the best accuracy for center 5. We will further explore possible reasons of this center imbalance in future work. Another limitation of the presented study is the empirical estimation of early stop strategy during leave-center-out cross validation based on the observed loss ranges. In future work, we will also explore a more automated mechanism to increase model robustness for images from more center.

5 Conclusion

We proposed a novel end-to-end domain-knowledge constrained neural network for automatic and reproducible diagnosis of AD using sMRI images. We proposed a new domain-knowledge encoding module that learn simultaneously with a ResNet-like feature extractor for domain specific and domain shared representations. The network directly takes the segmentation-free, patch-free images in original resolution as input, which is able to learn with global contextual information for subject-level pathological brain dysmorphologies features to further refines reproducible predictions. Our experiments demonstrate superior performance and generalize well to completely unseen domain.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62201091, the Startup Funds at Beijing University of Posts and Telecommunications (BUPT), and the BUPT innovation and entrepreneurship support program under 2023-YC-A208. We are grateful to the Multicenter Alzheimer Disease Imaging Consortium (PI: Prof. Xi Zhang, Prof. Yuying Zhou, Prof. Ying Han, and Prof. Qing Wang). The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies or sponsors.

References

1. Guptha, S.H., Holroyd, E., Campbell, G.: Progressive lateral ventricular enlargement as a clue to Alzheimer's disease. Lancet **359**(9322), 2040 (2002). https://doi.org/10.1016/S0140-6736(02)08806-2
2. Zhu, W., Sun, L., Huang, J., Han, L., Zhang, D.: Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. IEEE Trans. Med. Imaging **40**(9), 2354–2366 (2021)
3. Wen, J., et al.: Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med. Image Anal. **63**, 101694 (2020). https://www.sciencedirect.com/science/article/pii/S1361841520300591
4. Wang, H., et al.: Super-resolution based patch-free 3D medical image segmentation with self-supervised guidance (2022). https://arxiv.org/abs/2210.14645
5. Jin, D., et al.: Generalizable, reproducible, and neuroscientifically interpretable imaging biomarkers for Alzheimer's disease. Adv. Sci. **7**(14), 2000675 (2020)
6. Goenka, N., Tiwari, S.: Deep learning for Alzheimer prediction using brain biomarkers. Artif. Intell. Rev. **54**(7), 4827–4871 (2021)
7. Gutiérrez-Becker, B., Wachinger, C.: Deep multi-structural shape analysis: application to neuroanatomy. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 523–531. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_60
8. Nguyen, H.-D., Clément, M., Mansencal, B., Coupé, P.: Interpretable differential diagnosis for Alzheimer's disease and Frontotemporal dementia. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part I, pp. 55–65. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-16431-6_6

9. Hedges, E.P., et al.: Reliability of structural MRI measurements: the effects of scan session, head tilt, inter-scan interval, acquisition sequence, freesurfer version and processing stream. NeuroImage **246**, 118751 (2022). https://www.sciencedirect.com/science/article/pii/S1053811921010235
10. Zhang, J., Gao, Y., Gao, Y., Munsell, B.C., Shen, D.: Detecting anatomical landmarks for fast Alzheimer's disease diagnosis. IEEE Trans. Med. Imaging **35**(12), 2524–2533 (2016)
11. Danig, S., Orsborn, A.L., Moorman, H.G., Carmena, J.M.: Design and analysis of closed-loop decoder adaptation algorithms for brain-machine interfaces. Technical report 7 (2013)
12. Li, Y., Murias, M., Major, S., Dawson, G., Carlson, D.E.: On target shift in adversarial domain adaptation. In: AISTATS, March 2019
13. Hoffman, J., et al.: CyCADA: cycle-consistent adversarial domain adaptation. Int. Conf. Mach. Learn. **5**(11), 3162–3174 (2018). http://arxiv.org/abs/1711.03213
14. Sun, S., Shi, H., Wu, Y.: A survey of multi-source domain adaptation. Inf. Fusion **24**, 84–92 (2015)
15. Dozat, T.: Incorporating Nesterov momentum into Adam. In: ICLR Workshop, vol. 1, pp. 2013–2016 (2016)
16. Jiang, J.: A literature survey on domain adaptation of statistical Classifiers. UIUC Technical report, pp. 1–12, March 2008
17. Balaji, Y., Sankaranarayanan, S., Chellappa, R.: MetaReg: towards domain generalization using meta-regularization. In: NeurIPS, vol. 2018-Decem, pp. 998–1008 (2018). http://papers.nips.cc/paper/7378-metareg-towards-domain-generalization-using-meta-regularization
18. Li, D., Yang, Y., Song, Y.-Z., Hospedales, T.M.: Learning to generalize: meta-learning for domain generalization. In: Thirty-Second AAAI Conference on Artificial Intelligence, vol. 4 (2018). https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16067
19. Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.-Z., Hospedales, T.M.: Episodic training for domain generalization. In: IEEE International Conference on Computer Vision (2019). https://arxiv.org/pdf/1902.00113.pdf
20. Johansson, F.D., Sontag, D., Ranganath, R.: Support and invertibility in domain-invariant representations. In: Chaudhuri, K., Sugiyama, M. (eds.) Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, ser. Proceedings of Machine Learning Research, vol. 89, pp. 527–536. PMLR, 16–18 April 2019
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv e-prints, arXiv:1512.03385, December 2015
22. Zhao, K., et al.: Independent and reproducible hippocampal radiomic biomarkers for multisite Alzheimer's disease: diagnosis, longitudinal progress and biological basis. Sci. Bull. **65**(13), 1103–1113 (2020). https://www.sciencedirect.com/science/article/pii/S2095927320302140
23. Tu, L., Talbot, A., Gallagher, N.M., Carlson, D.E.: Supervising the decoder of variational autoencoders to improve scientific utility. IEEE Trans. Signal Process. **70**, 5954–5966 (2022)
24. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017)