DAE-GCN: Identifying Disease-Related Features for Disease Prediction

Churan Wang[1,4], Xinwei Sun[2(✉)], Fandong Zhang[4], Yizhou Yu[4,5], and Yizhou Wang[3]

[1] Center for Data Science, Peking University, Beijing, China
[2] Peking University, Beijing, China
sxwxiaoxiaohehe@pku.edu.cn
[3] Department of Computer Science and Technology, Peking University, Beijing, China
[4] Deepwise AI Lab, Beijing, China
[5] The University of Hong Kong, Pokfulam, Hong Kong

Abstract. Learning disease-related representations plays a critical role in image-based cancer diagnosis, due to its trustworthy, interpretable and good generalization power. A good representation should not only be disentangled from the disease-irrelevant features, but also incorporate the information of lesion's attributes (*e.g.*, shape, margin) that are often identified first during cancer diagnosis clinically. To learn such a representation, we propose a **D**isentangle **A**uto-**E**ncoder with **G**raph **C**onvolutional **N**etwork (DAE-GCN), which adopts a *disentangling mechanism* with the guidance of a GCN model in the AE-based framework. Specifically, we explicitly separate the encoded features into disease-related features and others. Among such features that all participate in image reconstruction, we only employ the disease-related features for disease prediction. Besides, to account for lesions' attributes, we propose to leverage the attributes and adopt the GCN to learn them during training. Take mammogram mass benign/malignant classification as an example, our DAE-GCN helps improve the performance and the interpretability of cancer prediction, which can be verified by state-of-the-art performance on one public dataset DDSM and three in-house datasets.

Keywords: Disease prediction · Disentangle · GCN

1 Introduction

For image-based disease benign/malignant diagnosis, it is crucial to learn the disease-related representation for prediction, due to the necessity of trustworthy (to patients), explainable (to clinicians) and good generalization ability in healthcare. A good representation, should not only be disentangled from the

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-87240-3_5) contains supplementary material, which is available to authorized users.

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12905, pp. 43–52, 2021.
https://doi.org/10.1007/978-3-030-87240-3_5

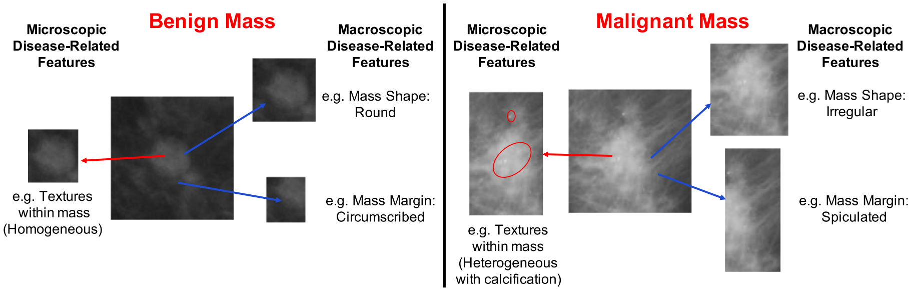
Benign Mass
Microscopic Disease-Related Features
e.g. Textures within mass (Homogeneous)
Macroscopic Disease-Related Features
e.g. Mass Shape: Round
e.g. Mass Margin: Circumscribed
Malignant Mass
Microscopic Disease-Related Features
e.g. Textures within mass (Heterogeneous with calcification)
Macroscopic Disease-Related Features
e.g. Mass Shape: Irregular
e.g. Mass Margin: Spiculated

Fig. 1. Illustrations of macroscopic/microscopic disease-related features in the mass. The left figure shows the benign case while the right shows the malignant case. Macroscopic-related features (*e.g.* margins, shapes) perform regular or clear in a benign mass while behave spiculated and irregular in a malignant mass. Microscopic-related features (*e.g.* the textures within the mass) perform differently between benign/malignant cases by taking a closer inspection. Textures within a malignant mass are heterogeneous (with calcification sometimes) but the textures in a benign mass are homogeneous.

disease-irrelevant features, but should also extract both macroscopic attributes and microscopic features of diseases. The macroscopic attributes mainly refer to the morphological attributes(*e.g.* shape, margin) [14] as summarized in American College of Radiology (ACR) [13], which are commonly adopted by doctors in benign/malignant diagnosis of many kinds of diseases. As all these attributes share the disease's information, they are correlated among each other [13]. In addition to these attributes, one can also extract some disease-related and microscopic features from the image, such as textures, curvatures of contour [5], *etc.*, which are helpful for prediction but may be beyond the observing ability of clinicians. Both examples of disease-related features, *i.e.*, macroscopic and microscopic ones are illustrated in Fig. 3.

Typically, in deep supervised learning, such a representation is regarded as the final hidden layer of the neural network [8,11]. However, it has been criticized [7] that these hidden layers can learn non-interpretable/semantic features, due to lack of prior knowledge which refers to clinical attributes in our scenario. Other representation learning works although leverage attributes [4,9] into learning, lack consideration of the microscopic features and empirical assurance to disentangle disease-related features from others. Besides, they do not model the correlation among attributes, which can make learning inefficient (Fig. 1).

For better representation learning, the disentanglement mechanism has been proved to be an effective way [1,3,12], since such a mechanism prompts different independent latent units to encode different independent ground truth generation factors that vary in the data [1]. Based on the above, to capture the disease-related features without mixing other irrelevant information, in this paper we propose a **D**isentangle **A**uto-**E**ncoder with **G**raph **C**onvolutional **N**etwork (DAE-GCN), which incorporates a *disentangling mechanism* into an AE framework, equipped with attribution data during training stage (the attributes are not provided during the test). Specifically, in our encoder network, we explicitly encode the image into three hidden factors: h_{ma}, h_{mi}, h_i which respectively correspond to macroscopic-related, microscopic-related and disease-irrelevant features. To achieve disentanglement, these hidden factors are fed into different constrains during the training phase. In details, among all h_{ma}, h_{mi}, h_i that participate in reconstruction of the whole image, we only use h_{ma}, h_{mi} in disease prediction and only h_{ma} to predict attributes, enforcing the disentanglement of h_{ma}, h_{mi}, h_i. To further leverage the correlation among attributes, we implement the GCN to facilitate learning.

To verify the utility of our method for diagnosing cancer-based on learning disease-related representations, we apply the DAE-GCN on Digital Database for Screening Mammography (DDSM) [2]) and three in-house datasets in mammogram mass benign/malignant diagnosis. It yields that our method successfully learns disease-related features and leads to a large classification improvement (**4%** AUC) over others on all datasets.

To summarize, our contributions are mainly three-fold: **a)** We propose a novel and general DAE-GCN framework, which helps disentangle the disease-related features from others to prompt image-based diagnosis; **b)** We leverage the GCN which accounts for correlations among attributes to facilitate learning; **c)** Our model can achieve state-of-the-art prediction performance for mass benign/malignant diagnosis on both the public and in-house datasets.

2 Methodology

Problem Setup. Our dataset contains $\{x_i, A_i, y_i\}_{i \in \{1,\ldots,n\}}$, in which x, A, y respectively denote the patch-level mass image, attributes (*e.g.*, circumscribed-margin, round-shape, irregular-shape), and the binary disease label. During test stage, only the image data is provided for feature extraction and prediction.

Figure 2 outlines the overall pipeline of our method, namely **D**isentangle **A**uto-**E**ncoder with **G**raph **C**onvolutional **N**etwork(DAE-GCN). As shown, our method is based on the auto-encoder framework, which has been empirically validated to extract effective features [10]. In the encoder, we separate encoded factors into three parts: macroscopic attributes h_{ma}, microscopic features h_{mi} and disease-irrelevant features h_i. During disentangle phase, we design the training with three constrains, which respectively reconstruct x, A, y via *image reconstruction*, *GCN learning* and *disease prediction*. In the following, we will introduce our encoder and disentangle training in Sect. 2.1, 2.2.

2.1 Encoder

We encode the original image into hidden factors h via an encoding network $f_{\theta_{\text{enc}}}$ parameterized by θ_{enc}: $h = f^{\theta}_{\theta_{\text{enc}}}(x)$. Such a hidden factor can contain many variations, which as a whole can be roughly categorized as three types: macroscopic disease-related attributes, microscopic disease-related features and other disease-irrelevant features. To only capture disease-related features into prediction, we explicitly separate h into three parts: h_{ma}, h_{mi}, h_i that aims to

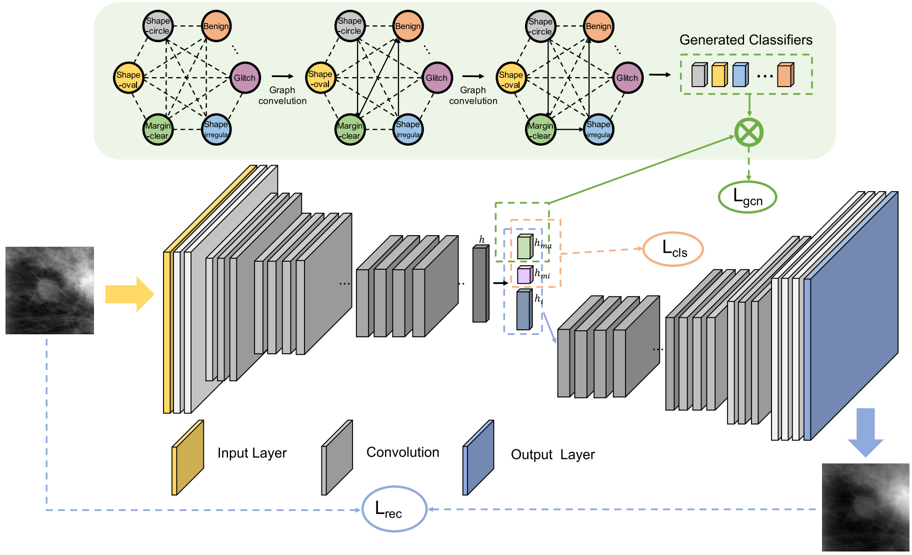

Fig. 2. Overview of the DAE-GCN, which is based on the auto-encoder framework. The encoded features h are separated into h_{ma}, h_{mi}, h_i marked by green, pink and blue. The decoder leads to three loss functions: the $\mathcal{L}_{\text{rec}}$ for reconstructing the original image using h as input; the $\mathcal{L}_{\text{gcn}}$ for learning attributes with h_{ma} as input; the $\mathcal{L}_{\text{cls}}$ for disease prediction with h_{ma}, h_{mi} as inputs. Specifically, we implement GCN for attribute learning to leverage the correlation among attributes. (Color figure online)

respectively capture the above three types of features. Accordingly, we append a three-constrain disentangle training, as elaborated in the subsequent section.

2.2 Disentangle Training

To achieve disentanglement of h_{ma}, h_{mi}, h_i, in particular the disease-related features (h_{ma}, h_{mi}) from others, we design a three-constrain training:*image reconstruction*, *GCN learning* and *disease prediction* with utilizing different factors.

Image Reconstruction. Since h_{ma}, h_{mi}, h_i describe different contents, they all participate in reconstructing the original image, via an decoding network $f_{\theta_{\text{dec}}}$: $\hat{x} = f_{\theta_{\text{dec}}}(f_{\theta_{\text{enc}}}(x))$. The reconstruction loss $\mathcal{L}_{\text{rec}}$ is defined as:

$$\mathcal{L}_{\text{rec}}(\theta_{\text{enc}}, \theta_{\text{dec}}) := \|x - \hat{x}(\theta_{\text{enc}}, \theta_{\text{dec}})\|_1 \tag{1}$$

Graph Convolutional Network Learning. To leverage correlation among attributes A into learning h_{ma}, we propose to implement the graph convolutional network (GCN), in which the computation rule is based on the graph $G := (V, E)$. Each node in V denotes the word embeddings of each attribute (details in supplementary). Each edge connecting V_i and V_j in E denotes the inner relevance between attributes. Specifically, we adopt [4] to define E as follows:

$$E_{i,j} = \begin{cases} p/\sum_{j=1}^{c} \tilde{E}_{i,j}, & \text{if } i \neq j \\ 1-p, & \text{if } i = j, \end{cases} \tag{2}$$

where $\tilde{E}_{i,j} = 0$ if $M_{i,j}/T_i < \tau$ and $= 1$ otherwise for some threshold hyperparameter $\tau \in (0,1)$, with $M_{i,j}$ denoting the number of concurrences of the i-th attribute and the j-th attribute, T_i denoting the number of occurrences of the i-th attribute. The $p \in (0,1)$ denotes the weight hyperparameter to avoid the over-smoothed problem, *i.e.*, the node from different clusters (shape-related vs margin-related) are indistinguishable [4]. Denote $H_l \in \mathbb{R}^{c \times d_l}$ as the output of the l-th layer, with c denoting the number of nodes. With $H_{l+1} = \sigma(EH_lW_l)$ for nonlinear activation σ and $W_l \in \mathbb{R}^{d_l \times d_{l+1}}$, each layer's output processes to the next, the last of which ($H_L \in \mathbb{R}^{c \times d_L}$ with $d_L = \dim(h_{ma})$) is fed into a softmax classifier for multi-label classification:

$$\hat{y}_i := \frac{\exp(\langle H_{i,L}, h_a \rangle)}{\left(\sum_{j \in \text{clus}(i)} \exp(\langle H_{j,L}, h_a \rangle)\right)}$$

where the clus(i) denoting the index set of attributes that belong to the same cluster (*e.g.*, shape, margin) with the i-th attribute. Denote θ_{gcn} as the whole set of parameters, *i.e.*, $\theta_{\text{gcn}} := \{W_l\}_{l \in \{0,\dots,L\}}$. The multi-label loss function for training such θ_{gcn} (and also θ_{enc}) $\mathcal{L}_{\text{gcn}}$ is denoted as the summation of cross-entropy loss for each cluster k:

$$\mathcal{L}_{\text{gcn}}(\theta_{\text{enc}}, \theta_{\text{gcn}}) := \frac{1}{N} \sum_n \sum_k \sum_{i \in \text{clus}_k} \left(\mathbb{1}(y_i = k) \log \hat{y}_i(H_0^n, \theta_{\text{gcn}})\right). \tag{3}$$

Disease Prediction. We feed disease-related features h_{ma}, h_{mi} into disease prediction via the classification network $f_{\theta_{\text{cls}}}$, with the θ_{cls} trained via the binary cross-entropy loss $\mathcal{L}_{\text{cls}}(\theta_{\text{enc}}, \theta_{\text{cls}})$. Combining $\mathcal{L}_{\text{cls}}$ with the $\mathcal{L}_{\text{rec}}$,$\mathcal{L}_{\text{gcn}}$ defined in Eq. (1), (3), the overall loss function $\mathcal{L}$ is:

$$\mathcal{L}(\theta_{\text{enc}}, \theta_{\text{dec}}, \theta_{\text{gcn}}, \theta_{\text{cls}}) = \mathcal{L}_{\text{rec}}(\theta_{\text{enc}}, \theta_{\text{dec}}) + \mathcal{L}_{\text{gcn}}(\theta_{\text{enc}}, \theta_{\text{gcn}}) + \mathcal{L}_{\text{cls}}(\theta_{\text{enc}}, \theta_{\text{cls}}).$$

3 Experiments

To evaluate the effectiveness of our DAE-GCN, we verify it on the patch-level mammogram mass benign/malignant classification. We consider both the public dataset DDSM [2] and three in-house datasets: Inhouse1, Inhouse2 and Inhouse3. For each dataset, the region of interests (ROIs) (malignant/benign masses) are cropped based on the annotations of radiologists the same as [9][1]. For all datasets, we randomly[2] divide the whole set into training, validation and testing

[1] We leave the number of ROIs and patients of each dataset and the description about the selection of attributes for DDSM in supplementary.

[2] Existing works about DDSM do not publish their splitting way and mention smaller count number of ROIs in DDSM compared with our statistics.

Table 1. AUC evaluation on public DDSM [2] and three in-house datasets.

Methodology	Inhouse1	Inhouse2	Inhouse3	DDSM [2]
Vanilla [8]	0.888	0.847	0.776	0.847
Chen *et al.* [4]	0.924	0.878	0.827	0.871
Guided-VAE [6]	0.921	0.867	0.809	0.869
ICADx [9]	0.911	0.871	0.816	0.879
Li *et al.* [11]	0.908	0.859	0.828	0.875
DAE-GCN **(Ours)**	**0.963**	**0.901**	**0.857**	**0.919**

as 8:1:1 in patient-wise. To provide convenience for latter works, we publish our spitted test set of DDSM [2] in supplementary.

3.1 Results

Compared Baselines. We compare our DAE-GCN the following methods: **a)** Li *et al.* [11] propose to reconstruct benign and malignant images separately via adversarial training; **b)** ICADx [9] also adopts the adversarial learning method and additionally introduces the attributes for reconstruction; **c)** Vanilla [8] directly trains the classifier via Resnet34; **d)** Guided-VAE [6] also implements disentangle network but lack the medical knowledge during learning; **e)** Chen *et al.* [4] only implements GCN to learning disease label and attributes.

Implementation Details. We implement Adam to train our model. For a fair comparison, all methods are conducted under the same setting; besides, they share the network structure of ResNet34 [8] as the encoder backbone. Area Under the Curve (AUC) is used as evaluation metrics in image-wise. More details are shown in the supplementary. For implementation of compared baselines, we directly load the published trained model of Vanilla [8], Chen *et al.* [4] during test; while for Guided-VAE [6], ICADx [9] and Li *et al.* [11] that without published source codes, we re-implement their methods.

Results and Analysis. As shown in Table 1, our methods can achieve state-of-the-art results on all datasets. Taking a closer look, the advantage of Guided-VAE [6] over Vanilla [8] may due to the disentangle learning in the former method. With further exploration of attributes via GCN, our method can outperform Guided VAE [6]. Although ICADx [9] incorporats the attributes during learning, they fail to model correlations among attributes, which limits their performance. Compared to Chen *et al.* [4] that also implement GCN to learn attributes, we additionally use disentangled learning which can help to identify disease-related features during prediction.

Table 2. Ablation Studies on (a) Inhouse1; (b) Inhouse2; (c) Inhouse3; (d) pubilic dataset DDSM [2]

Disentangle	Attribute Learning	$\mathbf{L}_{\mathbf{rec}}$	h_i	**AUC**(a)	**AUC**(b)	**AUC**(c)	**AUC**(d)
×	×	×	×	0.888	0.847	0.776	0.847
×	Multi-task	×	×	0.890	0.857	0.809	0.863
×	$\mathcal{L}_{\text{gcn}}$	×	×	0.924	0.878	0.827	0.871
×	×	✓	×	0.899	0.863	0.795	0.859
✓	×	✓	×	0.920	0.877	0.828	0.864
✓	×	✓	✓	0.941	0.882	0.835	0.876
✓	$\mathcal{L}_{\text{gcn}}$	✓	✓	**0.963**	**0.901**	**0.857**	**0.919**

Table 3. Overall prediction accuracy of multi attributes (mass shape,mass margin) on (a) Inhouse1; (b) Inhouse2; (c) Inhouse3; (d) pubilic dataset DDSM [2].

Methodology	ACC (Inhouse1)	ACC (Inhouse2)	ACC (Inhouse3)	ACC (DDSM [2])
Vanilla-multitask	0.715	0.625	0.630	0.736
Chen *et al.* [4]	0.855	0.829	0.784	0.875
ICADX [9]	0.810	0.699	0.671	0.796
Proposed Method	**0.916**	**0.880**	**0.862**	**0.937**

3.2 Ablation Study

To verify the effectiveness of each component in our model, we evaluate some variant models: *Disentangle* that denotes whether implement disentangled learning during reconstructing phase; and *multi-task* denotes using the multi-task model to learn attributes. The results are shown in Table 2.

As shown, deleting or changing any of the four components would lead to a descent of the classification performance. To be worthy of attention, using naive GCN also leads to a boosting of around 3%. Such a result can validate that the attributes data is quite helpful for the guidance of disease-related features learning. Meanwhile, the model with disentangle learning outperforms the one without it by a noticeable margin, which may be due to that the disease-related features can be identified without mixing information from others via disentangle learning. Moreover, with the guidance of exploring attributes, disease-related features can be disentangled better.

3.3 Interpretability

Attributes Prediction. To verify the prediction power of our learned representation h_{ma}, we report the accuracy of multi-label classification. Table 3 shows the results. As shown, our DAE-GCN outperforms other considered methods, which demonstrates that the learned h_{ma} can well capture the information to predict attributes.

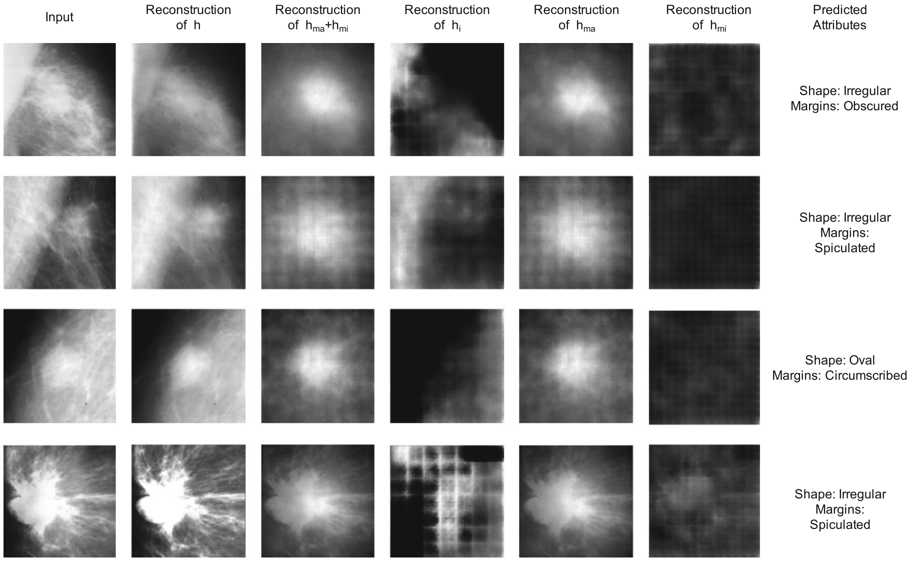
Input
Reconstruction of h
Reconstruction of h_ma+h_mi
Reconstruction of h_i
Reconstruction of h_ma
Reconstruction of h_mi
Predicted Attributes
Shape: Irregular Margins: Obscured
Shape: Irregular Margins: Spiculated
Shape: Oval Margins: Circumscribed
Shape: Irregular Margins: Spiculated

Fig. 3. Visualization of the disentanglement via reconstruction. From the left column to the right column: input; reconstruction using h ($h_{ma} + h_{mi} + h_i$); diseased-related features ($h_{ma} + h_{mi}$); disease-irrelevant features h_i; macroscopic attributes h_{ma}; microscopic features h_{mi} and predicted results of attributes.

Visualization. To further evaluate our learned representation, we visualize different parts in Fig. 3 via reconstruction effect. As we can see, the disease-related features ($h_{mi} + h_{ma}$) mainly reflect the lesion-related information since they only reconstruct the lesion regions without mixing others. The disease-irrelevant features h_i mainly learn features such as the contour of the breasts, pectoralis and other irrelevant grands without lesion information. The macroscopic attributes h_{ma} capture macroscopic attributes of the lesions; while the microscopic features h_{mi} learn features like global context, density or other invisible features but related to classification. For better validation of the success of disentangling, we use disease-related features ($h_{mi} + h_{ma}$) and disease-irrelevant features h_i encoded from our model while testing to train an SVM classifier respectively. The AUC of disease-related features ($h_{mi} + h_{ma}$) and disease-irrelevant features h_i for benign/malignant classification is separately 0.90 and 0.57 in DDSM [2]. This result further indicates the effectiveness and interpretability of our proposed DAE-GCN.

4 Conclusions and Disscusions

In this paper, we propose a novel approach called **D**isentangle **A**uto-**E**ncoder with **G**raph **C**onvolutional **N**etwork (DAE-GCN) to improve the mammogram classification performance. The proposed method performs disentangle learning by exploiting the attribute prior effectively. The promising results achieved on mammogram classification shows the potential of our method in benefiting the diagnosis of other types of diseases, *e.g.*, lung cancer, liver cancer and pancreatic cancer, which are left as our future work.

Acknowledgement. This work was supported by MOST-2018AAA0102004, NSFC-61625201 and ZheJiang Province Key Research & Development Program (No. 2020C03073).

References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013). https://doi.org/10.1109/TPAMI.2013.50
2. Bowyer, K., et al.: The digital database for screening mammography. In: Third International Workshop on Digital Mammography, vol. 58, p. 27 (1996)
3. Burgess, C.P., et al.: Understanding disentangling in β-vae. arXiv preprint arXiv:1804.03599 (2018)
4. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5177–5186 (2019)
5. Ding, J., et al.: Optimizing the peritumoral region size in radiomics analysis for sentinel lymph node status prediction in breast cancer. Acad. Radiol. (2020). https://doi.org/10.1016/j.acra.2020.10.015
6. Ding, Z., et al.: Guided variational autoencoder for disentanglement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7920–7929 (2020)
7. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=Bygh9j09KX
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Kim, S.T., Lee, H., Kim, H.G., Ro, Y.M.: Icadx: interpretable computer aided diagnosis of breast masses. In: Medical Imaging 2018: Computer-Aided Diagnosis, vol. 10575, p. 1057522. International Society for Optics and Photonics (2018). https://doi.org/10.1117/12.2293570
10. Klingler, S., Wampfler, R., Käser, T., Solenthaler, B., Gross, M.: Efficient feature embeddings for student classification with variational auto-encoders. International Educational Data Mining Society (2017)
11. Li, H., Chen, D., Nailon, W.H., Davies, M.E., Laurenson, D.I.: Signed laplacian deep learning with adversarial augmentation for improved mammography diagnosis. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 486–494. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_54
12. Ridgeway, K.: A survey of inductive biases for factorial representation-learning. arXiv preprint arXiv:1612.05299 (2016)

13. Sickles, E., D'Orsi, C., Bassett, L.: Acr bi-rads® mammography. acr bi-rads® atlas, breast imaging reporting and data system (2013)
14. Surendiran, B., Vadivel, A.: Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer. Int. J. Med. Eng. Inf. **4**(1), 36–54 (2012). https://doi.org/10.1504/IJMEI.2012.045302