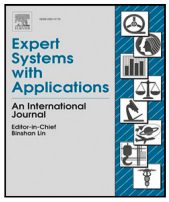
Expert Systems With Applications 243 (2024) 122897

Contents lists available at ScienceDirect

Expert Systems With Applications

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



A domain knowledge powered hybrid regularization strategy for semi-supervised breast cancer diagnosis

Xiaozheng Xie [a], Jianwei Niu [b,c,d], Xuefeng Liu [b,d,*], Yong Wang [e], Qingfeng Li [b], Shaojie Tang [f]

[a] School of Computer and Communication Engineering, University of Science and Technology Beijing, 100083 Beijing, China
[b] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, 100191 Beijing, China
[c] Beijing Advanced Innovation Center for Big Data and Brain Computing (BDBC), Beihang University, 310052 Hangzhou, China
[d] Zhongguancun Laboratory, 100190 Beijing, China
[e] Chinese Academy of Medical Sciences and Peking Union Medical College Beijing, 100021 Beijing, China
[f] Jindal School of Management, The University of Texas at Dallas, 75080 TX, USA

ARTICLE INFO

Keywords:
Consistency regularization
Domain knowledge
Virtual adversarial training
Breast cancer diagnosis

ABSTRACT

Semi-supervised learning has attracted much attention in medical image analysis as annotating medical images is substantially difficult. Existing semi-supervised learning methods mainly utilize some regularization strategies to encourage a network to make similar predictions for each sample under various perturbations (e.g. rotating and flipping). However, as these strategies are originally designed for natural images, they do not fully utilize the specific attributes in the medical domain. In this paper, we propose DK-HRS, a domain knowledge-powered hybrid regularization strategy for semi-supervised breast cancer diagnosis. More specifically, DK-HRS generates four types of perturbed samples, namely, domain knowledge-augmented samples, weakly and strongly-augmented samples, and virtual adversarial samples. These perturbed samples represent transformations of medical images from different aspects. DK-HRS utilizes the FixMatch, a popular consistency regularization method as the backbone, and integrates two additional modules: a virtual adversarial training module, and a domain knowledge-based contrastive learning module. Experimental results on four breast ultrasound datasets demonstrate that, by incorporating medical domain knowledge, DK-HRS achieves superior performance and outperforms some state-of-the-art semi-supervised methods by a large margin.

1. Introduction

Over recent years, emerging interest has occurred in adopting semi-supervised learning in medical image analysis. This is mainly due to the laborious and time-consuming labeling process of medical images. Semi-supervised learning, as a powerful approach to make full use of large amounts of unlabeled data, has demonstrated its effectiveness in various medical image analysis tasks, including pulmonary nodules diagnosis (Shi et al., 2022), multiple volumetric medical image segmentation (Wang & Li, 2023), breast cancer diagnosis and ophthalmic disease classification (Wang et al., 2021).

Most existing semi-supervised learning methods follow the regularization strategy, which in essence, utilizes unlabeled data by adding some extra constraints during the training process. The two popular approaches of the regularization strategy are consistency regularization (Bachman et al., 2014; Laine & Aila, 2017) and adversarial regularization (Miyato et al., 2019). In particular, consistency regularization devotes to enforce a network to produce consistent predictions for the same input sample under different perturbations, while adversarial regularization trains a network to assign similar labels to each input data and the neighbors in the adversarial direction. These two regularization methods demonstrate their effectiveness in many medical image analysis tasks including lung nodule classification (Xie et al., 2019), nucleus classification (Su et al., 2019), and the diagnosis of metastatic epidural spinal cord compression (Zhang et al., 2022).

However, as the regularization methods above are originally designed for tasks related to natural images, no specific attributes of medical images are directly utilized. We believe that a regularization strategy specially designed for medical images can achieve a better performance.

The most specific attribute of medical images lies the medical domain knowledge, which is mainly associated with how medical doctors browse images, the particular areas they focus on, and the features to which they give special attention. Nowadays, medical domain knowledge has been successfully applied to many supervised learning tasks

* Corresponding author.
E-mail addresses: xiexiaozheng@ustb.edu.cn (X. Xie), niujianwei@buaa.edu.cn (J. Niu), liu_xuefeng@buaa.edu.cn (X. Liu), wangyong@cicams.ac.cn (Y. Wang), liqingfeng@buaa.edu.cn (Q. Li), shaojie.tang@utdallas.edu (S. Tang).

https://doi.org/10.1016/j.eswa.2023.122897
Received 30 August 2023; Received in revised form 7 December 2023; Accepted 7 December 2023
Available online 9 December 2023

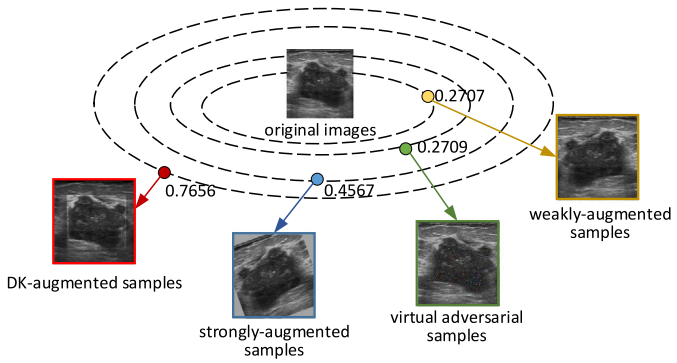


Fig. 1. Cosine distances between original images and their four corresponding perturbed samples. Different from the other three kinds of perturbed samples, DK-augmented samples are on average farthest away from the original images, which means thay may provide extra information that other perturbed samples do not contain.

and achieves large performance improvements (Xie, Niu, Liu, Chen et al., 2021).

Inspired by the success of applying medical domain knowledge to supervised learning tasks, we ask whether medical domain knowledge can also be helpful for semi-supervised learning. We believe the answer is positive. In the regularization strategy, perturbed samples of each unlabeled image are first generated via different methods (e.g., RandAugment Cubuk et al., 2019). Then the network is enforced to give similar predictions for all these samples, by which the most distinguishing features from each sample can be learned. Similarly, we can also generate perturbed samples by medical domain knowledge.

Then a key question is whether these 'domain knowledge augmented samples' (call as **DK-augmented samples** hereinafter) contain extra information than perturbed samples generated by other methods. We take breast cancer diagnosis as an example and carry out an experiment to illustrate this point. Here, the medical domain knowledge used in this paper is inspired by the fact that an experienced medical doctor generally focuses on the tumor area and the transition areas of the tumor margin. Correspondingly, the DK-augmented samples highlight more on these areas.

We randomly select about 3500 images from our breast ultrasound dataset and generate four kinds of perturbed samples (i.e., weakly-augmented samples, strongly-augmented samples, virtual adversarial samples and DK-augmented samples) for each image. Then, we compute the average cosine distances between the original images and their four corresponding perturbed samples, respectively, and the result is shown in Fig. 1. We can see that compared with the other three kinds of samples, DK-augmented samples, shown in the red box, are on average farthest away from the original images. We believe this is because these DK-augmented samples only contain the most important information by which radiologists make their decisions. This fact that DK-augmented samples are significantly different from other perturbed samples in feature space implies that they may provide extra information for the training of the model.

Based on the above observation, we present DK-HRS, a domain knowledge-powered hybrid regularization strategy for semi-supervised breast cancer diagnosis. DK-HRS utilizes labeled images to learn correct prediction, and for each unlabeled image, its four types of augmented samples are generated to help the network learn representative features. To illustrate the benefit of incorporating DK-augmented samples, we first test DK-HRS on a set consisting of 600 images. The classification results are shown in a two-dimensional t-SNE feature space (Van der Maaten & Hinton, 2008) in Fig. 2, where Fig. 2(a), (b) and (c) are results of using FixMatch, FixMatch+VAT and our method (DK-HRS), respectively. In particular, FixMatch only considers the weakly-augmented and strongly-augmented samples, FixMatch+VAT method also utilizes the virtual adversarial samples on the basis of FixMatch, and DK-HRS further incorporates the DK-augmented samples. We can see that with DK-augmented samples, two categories (i.e., benign and malignant) are better distinguished in Fig. 2(c), in which samples of the two categories are less overlapped with the highest diagnostic accuracy. Overall speaking, the contributions are summarized as follows:

- We find that samples augmented by medical domain knowledge are far from their original samples than other perturbed ones, and therefore potentially provide extra information during the training process of the model.
- We present a hybrid regularization strategy with medical domain knowledge for semi-supervised breast cancer diagnosis. Techniques like domain knowledge-based contrastive learning are adopted to let the network learn important features.
- Experimental results on four breast ultrasound datasets demonstrate the effectiveness of incorporating medical domain knowledge.

2. Related work

2.1. Semi-supervised learning for medical image analysis

Semi-supervised learning has been successfully used for many medical image analysis tasks (Basak & Yin, 2023; Ren et al., 2023; Wang et al., 2021; Xu et al., 2022; Yang et al., 2022). In general, most methods focus on the pseudo labeling (Huynh et al., 2022; Zeng et al., 2023) and regularization strategies (Bachman et al., 2014; Laine & Aila, 2017). The former trains the model in a supervised manner by generating pseudo labels for unlabeled images, while the latter forces the model to make similar predictions for samples under different perturbations.

The last few years have witnessed different variants of regularization methods for natural image tasks, such as Mean Teacher (Tarvainen & Valpola, 2017), virtual adversarial training (VAT) (Miyato et al., 2019), MixMatch (Berthelot et al., 2019) and FixMatch (Sohn et al., 2020). Perturbed samples in these methods are generally generated by adding some noises or using some data augmentation methods (e.g., RandAugment Cubuk et al., 2020 and CTAugment (Berthelot et al., 2020)). These regularization methods are soon applied to the medical domain. For example, on the basis of Mean Teacher, both local and global consistency are adopted to let the network learn better features for nuclei classification (Su et al., 2019). Region of interest (ROI) consistency is adopted in Xu et al. (2020) to guide the network to focus on the brain region for brain MRI quality assessment. In addition, by focusing on ambiguous regions in consistency target selection, a novel ambiguity-consensus mean-teacher model is proposed to drive the model to learn ambiguous-yet-effective information from unlabeled data. Experimental results on left atrium segmentation and brain tumor segmentation outperform many state-of-the-art semi-supervised methods (Xu et al., 2023).

Besides, the sample relation consistency paradigm is proposed in Liu et al. (2020) for skin lesion diagnosis and thorax disease classification. The consistency learning (i.e., FixMatch) and VAT are combined for tasks of breast cancer screening classification and multi-class ophthalmic disease classification (Wang et al., 2021). Furthermore, a novel cyclic prototype consistency learning is proposed for brain tumor and kidney segmentation. In this approach, label supervision is constructed for unlabeled data by using the consistency between labeled and unlabeled data (Xu et al., 2022). By combining consistency regularization with cluster assumption and active learning, the mutual collaboration of adaptive pseudo-annotation and information active annotation can be realized to boost medical image classification tasks (Zhang et al., 2022). However, all the above regularization methods are just designed for semi-supervised analysis tasks and do not incorporate the specific attributes in the medical domain.

Different from these methods, we incorporate medical domain knowledge in generating perturbed samples for the hybrid

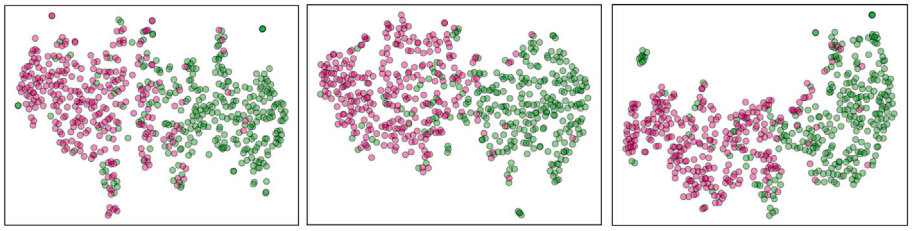


Fig. 2. T-SNE (Van der Maaten & Hinton, 2008) feature space representation of the classification results of 600 images. The pink and green points represent the malignant and benign tumors, respectively. (a), (b) and (c) are the results of using FixMatch (using weakly-augmented and strongly-augmented samples), FixMatch+VAT (using weakly-augmented, strongly-augmented and virtual adversarial samples) and our method (using weakly-augmented, strongly-augmented, virtual adversarial and DK-augmented samples), respectively. Results show that our method performs better with the help of medical domain knowledge.

regularization-based semi-supervised learning process. Compared with these methods, incorporating medical domain knowledge can help our model better learn distinguishing features and achieve superior performance.

2.2. Incorporating domain knowledge into deep learning for medical image analysis

Medical domain knowledge depicts the knowledge or experiences of medical doctors (especially radiologists) that they use during their daily diagnostic process. Recently, medical domain knowledge has been widely adopted for many supervised learning tasks and achieves superior performance, especially on small-scale medical datasets (Kong et al., 2022; Pham et al., 2022; Wang et al., 2023; Xie, Niu, Liu, Chen et al., 2021). For example, to mimic the from-easy-to-hard diagnostic pattern of doctors, a curriculum learning strategy is proposed for the fracture classification (Jimenez-Sanchez et al., 2019). In addition, to depict the two specific patterns of radiologists when reading breast contrast-enhanced ultrasound (CEUS) videos, the domain knowledge guided temporal attention module and channel attention module are designed and utilized in Chen et al. (2021).

More recently, domain knowledge is also introduced into the semi-supervised learning process, in which the generative adversarial network (GAN) and the domain adaptation method are widely used (Li et al., 2020; Yang et al., 2019). For example, eight kinds of domain features are utilized in the semi-supervised ultrasonography thyroid nodules classification (Yang et al., 2019). Additionally, to leverage the knowledge from another imaging modality (i.e., MRI), the semi-supervised domain adaptation approach (Dual-Teacher) is proposed for cardiac segmentation of CT images (Li et al., 2020).

In this paper, we introduce domain knowledge into the regularization-based semi-supervised learning method. On the basis of our previous work (Xie, Niu, Liu, Li et al., 2021), we incorporate domain knowledge into a hybrid regularization method, with a new representation of domain knowledge and a different regularization-based learning strategy. Specifically, the domain knowledge is represented as DK-augmented samples. These samples are combined with other samples generated by existing perturbation methods to enforce the network to learn important features.

2.3. Breast cancer diagnosis

Nowadays, lots of efforts are dedicated to the early detection of breast cancer by using different computer-aided diagnosis (CAD) models, and most of the recent studies directly adopt deep learning techniques (Ahmed & Muhammad, 2023; Li et al., 2022; Xu et al., 2019). For example, a Faster R-CNN based method is adopted to localize and classify masses from ultrasound images (Shin et al., 2019). In addition, in Xu et al. (2019), a deep selective attention approach is proposed to select valuable regions in histopathological images for breast cancer classification. Furthermore, the dual-path CNN-based mammogram analysis model is proposed to realize mask segmentation and classification simultaneously (Li et al., 2022).

Besides using deep models directly, some other methods also incorporate different kinds of information to improve the performance of detecting breast cancer. For example, margin information of tumors is utilized to let the network focus more on margin areas and achieve better results (Xie et al., 2022). Additionally, CNN features are generally fused with handcrafted features computed using conventional CAD methods (Antropova et al., 2017). Furthermore, using information of different modalities is also proven to be effective for breast cancer diagnosis (Samala et al., 2019).

More recently, some researchers focus more on the semi-supervised breast cancer diagnosis (Xie, Niu, Liu, Li et al., 2021; Zhang et al., 2020). For example, a method called BIRADS-SSDL integrates clinically approved breast lesion characteristics (BI-RADS features) into the semi-supervised learning process (Zhang et al., 2020), where the unsupervised BI-RADS feature map reconstruction and diagnosis-oriented image classification tasks are learned together in a multi-task learning framework. Different from the aforementioned methods, we focus on the regularization-based semi-supervised breast cancer diagnosis. In addition, domain knowledge is characterized as perturbed samples to enforce the model to learn the most representative and distinguishing features.

3. Method

The main structure of our DK-HRS is shown in Fig. 3. We can see that it mainly consists of two branches: the supervised learning branch (left) and the unsupervised one (right).

DK-HRS adopts FixMatch, a popular consistency regularization method, as the backbone, and also integrates two additional modules: the VAT module (highlighted as yellow), and the domain knowledge-based contrastive learning (DKCL) module (highlighted as lilac). Accordingly, four types of perturbed samples are utilized, namely, samples generated by domain knowledge, by weak and strong augmentation methods, and by virtual adversarial perturbation. These components collaborate to enforce the network to learn important features from different kinds of data transformations. It should be noted that ResNet18 is used as the encoder of the model, and followed by a multilayer perceptron (MLP) with one hidden layer to regularize the whole learning process.

3.1. The overview of DK-HRS

To better describe the learning process of our DK-HRS method, we first formulate the consistency regularization-based semi-supervised learning problem. In general, the training dataset mainly consists of the labeled set $D_L = \{(x_i, y_i)\}_{i=1}^{N_L}$ and the unlabeled set $D_U = \{(x_i)\}_{i=1}^{N_U}$, where x_i represents the input image, y_i is the one hot ground truth label for each labeled image x_i. Additionally, N_L and N_U are the size of labeled set and unlabeled set, respectively. In particular, N_U is assumed

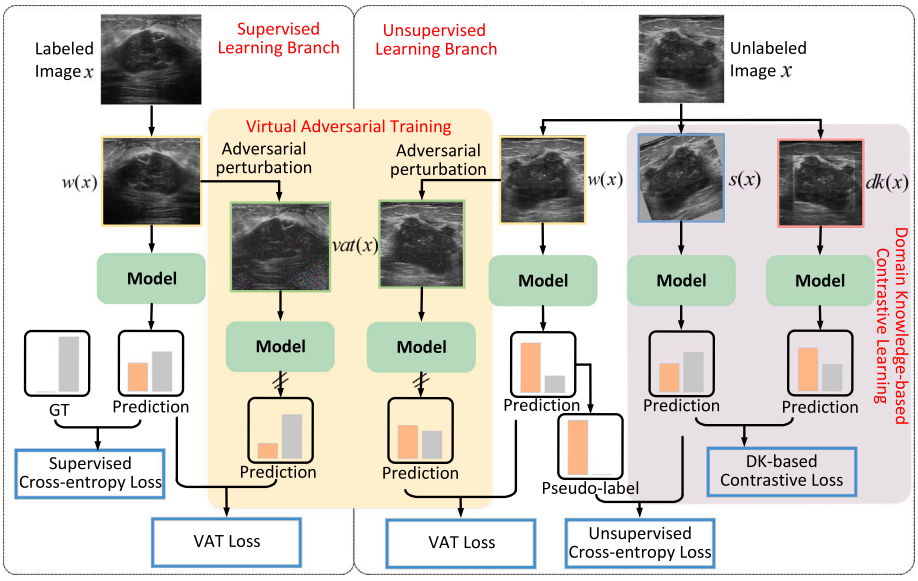


Fig. 3. Diagram of our DK-HRS method. It mainly consists of the supervised learning branch (the left part) and the unsupervised one (the right part). Based on the consistency regularization method FixMatch, the VAT module (highlighted as yellow) and the domain knowledge-based contrastive learning module (highlighted as lilac) are also integrated to boost the whole training framework. In particular, $w(x)$, $s(x)$, $vat(x)$ and $dk(x)$ are the weakly-augmented, strongly-augmented, virtual adversarial and DK-augmented samples for the input image x, respectively. Arrows with unequal signs represent the branch where the gradient is not back propagated.

to be much larger than N_L. During the training process, all labeled and unlabeled images are fed into the network concurrently to enforce the consistency learning process, and the overall optimization objective function can be defined as follows:

$$\min_{\theta} L_s\left(D_L,\theta\right)+\lambda L_c\left(D_L,D_U,\theta,\theta'\right), \tag{1}$$

where L_s is the supervised loss for evaluating the difference between network outputs and the ground truth labels for all labeled data. L_c represents the consistency loss for both labeled and unlabeled data. λ is a fixed scalar hyperparameter that is used to balance these two parts of losses. θ is a set of network parameters, and θ' is a fixed copy of θ indicting that the gradient is not propagated through θ'. In particular, the supervised loss is represented using the cross-entropy loss:

$$L_s=\frac{1}{N_L}\sum_{x_i,y_i\in D_L} H\left(y_i,f\left(x_i;\theta\right)\right), \tag{2}$$

where $H\left(p,q\right)$ represents the cross-entropy between distributions p and q, $f\left(\cdot\right)$ is the classification network.

Besides, the consistency loss L_c verifies the consistency between each input image and its perturbed samples. In particular, there are four different kinds of perturbed samples in DK-HRS (i.e., weakly-augmented, strongly-augmented, adversarial and DK-augmented samples), and the learning process of these perturbed samples are described in the following sections.

3.2. The backbone of DK-HRS

DK-HRS takes FixMatch as the backbone, in which the supervised and the unsupervised learning branches are constrained by their corresponding cross-entropy losses, respectively.

Concretely, in the supervised learning branch, all labeled images are first weakly-augmented (i.e., random horizontal flipping) and then fed into the network. The supervised cross-entropy loss is calculated to enforce the model to learn reliable discriminative patterns from labeled data. For the unsupervised learning branch, each image is augmented twice by using the weakly-augmented method and the strongly-augmented one (i.e., RandAugment Cubuk et al., 2019) separately. During the training process, each weakly-augmented image is first fed into the network to compute its pseudo-label, then the label is used to regularize the learning process from the strongly-augmented one. More specifically, the prediction class distribution of a weakly-augmented image is first computed as:

$$q_w=f\left(\widetilde{y}_i|w(x_i);\theta\right). \tag{3}$$

Then the largest class probability distribution (the maximum output of model) is defined as the pseudo-label by:

$$\hat{q}_w=\arg\max\left(q_w\right). \tag{4}$$

Finally, the pseudo-label is used to enforce the network to learn from the strongly-augmented image. The cross-entropy loss in this branch is given by:

$$L_u=\frac{1}{N_U}\sum_{i=1}^{N_U}\mathbb{1}\left(\max\left(q_w\right)\geqslant\gamma\right)H\left(\hat{q}_w,f\left(s(x_i);\theta\right)\right), \tag{5}$$

where $\mathbb{1}\left(\cdot\right)$ represents the indicator function, and γ is a scalar hyperparameter denoting the threshold to obtain the pseudo-label. $w(x_i)$ and $s(x_i)$ are the weakly-augmented and strongly-augmented images for the unlabeled image x_i, respectively.

3.3. The VAT module

Besides the weakly and strongly-augmented images, DK-HRS also incorporates the adversarial samples to enhance its robustness. In general, adversarial samples are formed by applying small but intentionally worst-case perturbations to examples from the dataset (Goodfellow et al., 2014). As an example, VAT (Miyato et al., 2019) adopts the adversarial samples for semi-supervised learning tasks. Specifically, the adversarial samples are generated from adversarial and virtual adversarial perturbations based on the labeled and unlabeled data, respectively. In this process, the loss function is defined using the negative measure of the local distribution smoothness (LDS) of the current model at each input data point, which can be formulated as:

$$\underset{x_i\in D_L\cup D_U}{\mathrm{LDS}}\left(x_i,\theta\right):=D_{\mathrm{KL}}\left[f\left(y|x_i;\theta'\right),f\left(y|x_i+r_{\mathrm{vadv}};\theta\right)\right], \tag{6}$$

$$r_{\mathrm{vadv}}:=\underset{r;\|r\|_2\leq\varepsilon}{\arg\max}\,D_{\mathrm{KL}}\left[f\left(y|x_i;\theta'\right),f\left(y|x_i+r\right);\theta'\right], \tag{7}$$

where $D_{\mathrm{KL}}[p, q]$ is the Kullback–Leibler (KL) divergence from distribution p to q. Here, as defined in Eq. (1), θ' is a fixed copy of θ and the gradient is not propagated through it. Additionally, r_{vadv} denotes the adversarial perturbation, which can be approximated with a linear approximation of D_{KL} with respect to r in Eq. (7), and the norm constraint $\varepsilon > 0$ represents the adversarial direction.

To incorporate the VAT into the FixMatch backbone, we apply it to weakly-augmented images from both labeled and unlabeled branches. In this process, the average of LDS on all training data is used as the regularization term in VAT and the loss function is defined as:

$$L_{vat} = \frac{1}{N_L + N_U} \sum_{x_i \in D_L \cup D_U} \mathrm{LDS}\left(w(x_i), \theta\right). \tag{8}$$

In this way, VAT can maximize the likelihood of the model while improving the LDS on each training example. Consequently, the semi-supervised learning process is robust against adversarial perturbations and different transformations.

3.4. The DKCL module

In the DKCL module, DK-augmented samples are designed and utilized to incorporate domain knowledge into the learning process of unlabeled data. In particular, we first introduce the generation method of these DK-augmented samples, and then describe how to incorporate them into the learning process.

The medical domain knowledge utilized when generating DK-augmented samples is based on the consensus of the BI-RADS (Breast Imaging Reporting and Data System) (Mendelson et al., 2013). BI-RADS gives advice to medical doctors about some important sonographic characteristics of tumors, including shape, margin and posterior features, to help them distinguish malignant tumors from benign ones. In particular, these features are generally closely related to some specific areas in images, such as the tumor area and its surrounding tissues, especially the transition areas of the tumor margin. Correspondingly, we design three kinds of images to represent the domain knowledge: the Region of Interest (ROI) image that contains tumor areas, the 1.5*ROI image that contains both tumor and its surrounding areas, and the margin-reinforced image which contains tumor area with reinforced pixels in margin areas.

As shown in Fig. 4, to generate these three kinds of images, we first adopt a fully automated segmentation method (Gomez-Flores & Ruiz-Ortega, 2016) to locate the breast tumor roughly in each image. It should be noted that no image segmentation annotation is needed in this process. Then the minimum bounding rectangle and 1.5 times bounding rectangle are calculated as the ROI area and 1.5*ROI area, respectively. Finally, pixels in ROI and 1.5*ROI areas are preserved, while pixels in other areas are set as the mean value of the dataset, by which the ROI image and the 1.5*ROI image are generated. For the margin-reinforced image, the tumor contour is first extracted from the segmentation result and is dilated to obtain a ribbon mask with a certain width along the contour line. Then, based on the ROI image, pixels within margin areas are set to twice as before to obtain the margin-reinforced image. To better combine the above information, these three images are mixed as the final DK-augmented image by alpha compositing, where the element-wise convex coefficients in compositing are randomly sampled from Dirichlet distribution as in Hendrycks et al. (2020).

Inspired by Li et al. (2021), we adopt contrastive learning to help the model learn more representative features. In general, contrastive learning typically pulls together the representations of a target image and its matching (positive) image in embedding space, while pushing apart the image from many non-matching (negative) ones. As contrastive learning requires a higher intensity of data augmentation (Chen et al., 2020), we define the contrastive prediction task between all DK-augmented samples with the strongly-augmented ones in each batch. Concretely, for a batch with N unlabeled images, the total number of DK-augmented and strongly-augmented samples is $2N$. We define the DK-augmented and the strongly-augmented samples generated from the same image as the positive pair, while the DK-augmented sample with all other strongly-augmented samples as negative pairs. Thus, the loss function for a positive pair of images is calculated by:

$$L_{con} = -\log \frac{\exp\left(\mathrm{sim}\left(z_i^{dk}, z_i^{s}\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\mathrm{sim}\left(z_i^{dk}, z_k^{s}\right)/\tau\right)}, \tag{9}$$

where $\mathrm{sim}(u, v) = u^T v / \|u\| \|v\|$ denotes the cosine similarity between l_2 normalized u and v. z_i^{dk} and z_i^{s} are the features extracted from the DK-augmented sample and strongly-augmented sample, respectively. $\mathbb{1}_{[k \neq i]} \epsilon \{0, 1\}$ is an indicator function evaluating to 1 if $k \neq i$, τ is a temperature parameter set as 0.07. The final loss is computed across all positive pairs from DK-augmented sample to strongly-augmented one and vice versa. Thus, the final loss in the DKCL module is defined as:

$$L_{dk} = \frac{1}{2N} \sum_{k=1}^{N} \left[L_{con}(2k-1, 2k) + L_{con}(2k, 2k-1)\right]. \tag{10}$$

3.5. The training framework of DK-HRS

In the whole DK-HRS training framework, the FixMatch, VAT and DKCL modules are performed concurrently to boost the final performance of the model. The detailed structure of the model used in DK-HRS is shown in Fig. 5, where ResNet18 is employed as the feature encoder, followed by a 2-layer MLP projection head to project the representation to the final 2-dimensional output space. Additionally, the ReLU activation function is adopted in the middle of two MLP layers to ensure the nonlinearity, and sizes of feature maps are listed near the corresponding maps.

During the training process, different kinds of perturbed samples (i.e., weakly-augmented, strongly-augmented, DK-augmented and virtual adversarial samples) are generated and injected into the model simultaneously. Meanwhile, the learning paths from the FixMatch, VAT and DKCL modules in return complement each other. The final loss function of DK-HRS is thus formulated as:

$$L = L_s + \lambda L_c, \text{ with } L_c = L_u + \alpha L_{vat} + \beta L_{dk}, \tag{11}$$

where λ, generally set as 1, is a fixed scalar hyperparameter to balance the losses denoting the relative weight of the consistency loss. α and β are weighting factors to balance these three kinds of loss terms, and α is also set to 1.

The weighting factor β is used to depict the effect of DKCL module, and hence can be regarded as a measure of the degree of incorporated domain knowledge. One simple approach is to set β to be a fixed value as in α. However, we find that it is more appropriate to set β as a ramp-down function.

The rationale behind is an analogy to how a medical doctor gets improved. In general, a trainee doctor generally relies more on the medical domain knowledge learned from text-books or other experienced doctors. Eventually, when he gains more and more hands-on experiences, he develops his own way of judgment and becomes less dependent on medical domain knowledge. Similarly, in the model training process, we believe a larger weight should be set for β at the beginning to ensure the model to learn more from medical domain knowledge. While when the learning process stabilizes, the model should learn more from unlabeled data without much guidance of medical domain knowledge. To emphasize the effect of domain knowledge, we apply a linear ramp-down function to control the value of β, the details will be elaborated in ablation studies in Section 4.4.

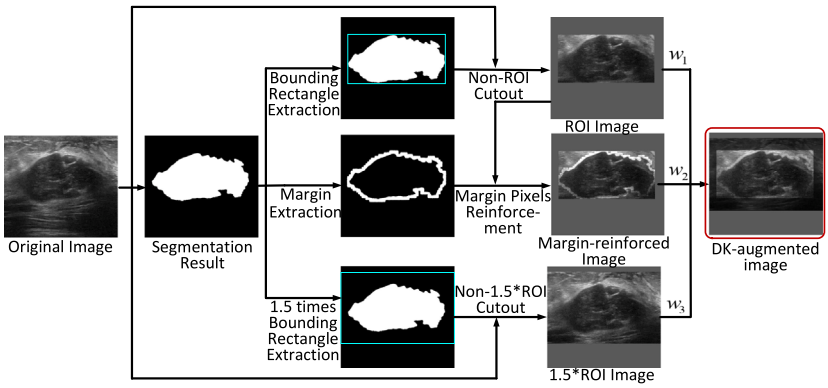


Fig. 4. Generating process of DK-augmented image.

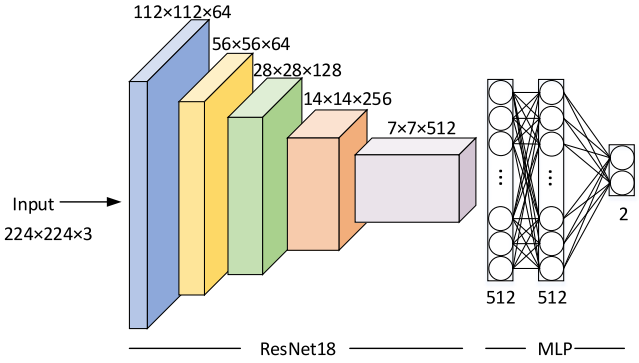


Fig. 5. The model structure used in DK-HRS, where ResNet18 is used as the encoder and followed by a MLP with one hidden layer.

4. Experiments

4.1. Datasets and experimental setup

Datasets: There are four breast ultrasound datasets used in our experiments, one is our private dataset BUS2019 and the other three are the public dataset BUSI (Al-Dhabyani et al., 2020), UDIAT (Yap et al., 2018) and the 3-class dataset (named as '3CUS' hereinafter) (Rodtook et al., 2018).

BUS2019: Our BUS2019 dataset consists of 7318 images from 3946 patients. These images are collected from the Cancer Hospital Chinese Academy of Medical Sciences from 2017 to 2019. In particular, 3596 images from 2039 patients are with benign tumors, while 3722 images from 1907 patients are with malignant ones. These images are collected from ultrasound systems with different vendors including PHILIPS, SIEMENS, HITACHI and so on. In addition, all images presented at least one tumor, and the benign or malignant label for each image is proved histopathology by biopsy.

The BUS2019 dataset is randomly divided into the training, validation and test sets according to the ratio 6:2:2. In detail, the training set consists of 4390 images (2158 benign images and 2232 malignant ones), while the validation set and test set are composed of 1464 images (719 benign images and 745 malignant ones), respectively. In the training set, different percentages of images are randomly selected as labeled ones, and the rest are treated as unlabeled ones. Note that images from the same patient are selected concurrently into the same set. In addition, all images for a certain patient are treated either as labeled or as unlabeled data in the training set.

BUSI: The public available dataset BUSI (Al-Dhabyani et al., 2020) is collected from Baheya Hospital in 2018, which contains 780 images from 600 patients (437 benign, 210 malignant and 133 normal). All images in BUSI are obtained from LOGIQ E9 and LOGIQ E9 Agile ultrasound systems, and the image resolution is 1280 × 1024. As the domain knowledge only focuses on benign and malignant images, normal images are not used in our experiments. In addition, as the ratio of benign and malignant images in BUSI is 2:1, we adopt data augmentation (i.e., rotating 10 degrees) on all malignant images to construct a balanced dataset. Meanwhile, 5-fold cross-validation is adopted to reduce the over-fitting problem on this dataset.

UDIAT: The UDIAT (Yap et al., 2018) dataset is collected in 2012 from the UDIAT Diagnostic Centre of the Parc Tauli Corporation, Sabadell (Spain). Obtained from a Siemens ACUSON Sequoia C512 system, the dataset mainly consists of 163 images from different women, in which 53 images are with cancerous masses and 110 are with benign lesions. The mean image size is 760 × 570 pixels, and each image presented one or more lesions. Similar to the BUSI dataset, the ratio of benign and malignant images in UDIAT is also about 2:1, we adopt data augmentation (i.e., left-to-right flipping) on all malignant images. Meanwhile, 5-fold cross-validation is also adopted on this dataset.

3CUS: Collected from the Department of Radiology of Thammasat Hospital University in Thailand, the dataset includes different types of ultrasound images (Rodtook et al., 2018). To test the performance of our method in multi-classification task, we selected 141 images with malignant tumors, 22 images with cysts, and 20 images with fibroadenomas from the dataset. The images are all obtained by using a Philips iU22 ultrasound machine, and the mean image size is 643 × 787. Due to the disparity in data number among these three categories, we adopt some data augmentation methods (i.e., left-to-right flipping, randomly rotating and randomly changing contrast) on all cyst and fibroadenoma images. Then, 5-fold cross-validation is performed on this dataset.

Evaluation metrics: We adopt five metrics to evaluate the performance of all experiments, including the Area Under Curve (AUC), accuracy, sensitivity, specificity and F1-score. It should be noted that for each metric, we compute the mean ± std over 5 trials.

Implementation details: To preserve the aspect ratio of the tumor in each image, all images are first scaled to make the short edge to 224 and then center-cropped to 224 × 224. For data augmentation, random horizontal flipping and the RandAugment method are used as the weak and strong augmentations, respectively. We train the whole method for 50 epochs, in which the batch sizes of labeled images and unlabeled images are set to 8 and 240, respectively. In addition, linear learning rate warm-up is adopted for the first 2 epochs until it reaches to the initial value of 0.0046875 (0.3 × BatchSize/512). Then the learning rate is decayed by multiplying 0.1 at 15, 30 and 45 epochs, respectively. We use the SGD optimizer with Nesterov momentum hyperparameter set as 0.9, and the weight decay is set as 5e–4. Moreover, the exponential moving average (EMA) technique is utilized with a decay 0.999 for all experiments. The confidence threshold γ is set as 0.9 and ϵ is set as 30. All experiments are implemented with PyTorch library, and trained on one GPU of NVIDIA Tesla V100-PCIE-16 GB. Our code can be accessed at https://github.com/xxzcs/DK_HRS.

Table 1
Quantitative evaluation of our method under different percentages of labeled data.

Method	Percentage		Metrics (%)				
	L	UL	AUC	Accuracy	Sensitivity	Specificity	F1
Upper Bound	100%	0	87.46 ± 0.17	81.41 ± 0.30	83.06 ± 1.03	79.69 ± 1.29	81.97 ± 0.29
Baseline	15%	0	79.34 ± 0.48	73.39 ± 0.17	76.56 ± 1.03	70.10 ± 1.39	74.54 ± 1.57
DK-HRS (ours)	15%	85%	87.46 ± 0.79	80.03 ± 1.33	78.34 ± 4.07	81.78 ± 1.61	79.91 ± 1.89
Baseline	20%	0	80.59 ± 1.00	74.74 ± 0.68	75.46 ± 1.69	73.99 ± 0.99	75.24 ± 0.88
DK-HRS (ours)	20%	80%	88.24 ± 0.42	81.49 ± 0.47	80.40 ± 2.01	82.61 ± 2.09	81.54 ± 0.59
Baseline	25%	0	81.36 ± 0.31	74.90 ± 0.46	76.16 ± 0.96	73.60 ± 1.68	75.54 ± 0.29
DK-HRS (ours)	25%	75%	88.51 ± 0.27	81.98 ± 0.45	82.79 ± 1.13	81.14 ± 1.54	82.38 ± 0.40
Baseline	30%	0	83.11 ± 0.50	75.92 ± 0.58	76.30 ± 0.68	75.52 ± 1.18	76.33 ± 0.50
DK-HRS (ours)	30%	70%	89.33 ± 0.24	82.36 ± 0.59	84.48 ± 1.42	80.17 ± 1.62	82.98 ± 0.59

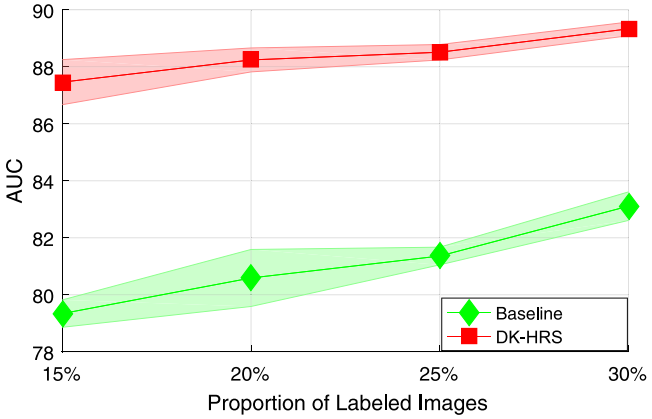


Fig. 6. The AUC values under different percentages of labeled images.

4.2. Effectiveness of DK-HRS

To prove the effectiveness of our DK-HRS method, we conduct experiments on different percentages of labeled images (i.e., 15%, 20%, 25% and 30%) on our BUS2019 dataset. We report the performance of all baselines (trained only with the corresponding labeled data) along with our DK-HRS method in Table 1. Note that the 'Upper Bound' represents the results trained using all images in the training set as labeled data and without introducing any information.

As we can see from Table 1, DK-HRS significantly improves the diagnostic performance of baseline methods under all percentages of labeled data settings. For example, when compared with the baseline method, the average of AUC is improved by 8.12%, 7.65%, 7.15% and 6.22% in 15%, 20%, 25% and 30% of labeled data settings, respectively. Meanwhile, the average of accuracy is improved by 6.64%, 6.75%, 7.08% and 6.44%, respectively in the four conditions. Moreover, DK-HRS trained with only 20% of labeled images achieves comparable or even better performance in some metrics compared with the Upper Bound method. In particular, it achieves higher AUC and specificity when compared with the Upper Bound method, and the improvements are 0.78% and 2.92%, respectively. Moreover, when using 30% of labeled images, DK-HRS achieves better performance on all metrics over the Upper Bound method, and the average of AUC and accuracy are 89.33% and 82.36%, respectively.

To better illustrate the performance of our DK-HRS method over the corresponding baseline method, we compare the AUC at different percentages of labeled data settings in Fig. 6. We can see the superior performance of DK-HRS over the baseline method at all proportions of labeled images. In addition, the performance improvements decrease with the increase of labeled data.

4.3. Comparison with other semi-supervised methods

We also compare the performance of our DK-HRS method with some state-of-the-art semi-supervised learning methods. The comparison methods mainly include (1) some general semi-supervised methods: Mean Teacher (Tarvainen & Valpola, 2017), VAT (Miyato et al., 2019), MixMatch (Berthelot et al., 2019), FixMatch (Sohn et al., 2020), UDA (Xie et al., 2020); and (2) three kinds of regularization based semi-supervised methods specifically designed for medical area: SRC-MT (Liu et al., 2020), the method proposed by Wang et al. (2021) and DK-Consistency (Xie, Niu, Liu, Li et al., 2021). In particular, built upon the Mean Teacher framework, SRC-MT further considers the consistency between the relationship among samples. The method of Wang et al. only combines FixMatch and VAT without introducing other information. DK-Consistency incorporates domain knowledge into the UDA model for breast cancer diagnosis. During the training process, the same network structure (i.e., ResNet18) and training protocol (i.e., data processing, training epochs, optimizer and learning rate schedule) as our method are adopted for all comparison methods. The experiments are performed on 20% of labeled data setting, and the results are listed in Table 2, where the best performance in each metric is shown in bold. In particular, the computation speed on the test set of each method is also evaluated by using the metric of FPS (frames per second).

We can see that DK-HRS substantially outperforms all other semi-supervised methods on almost all metrics. In particular, DK-HRS shows the highest average of AUC (88.24%), accuracy (81.49%), specificity (82.61%) and F1 (81.54%), and has 3.04%, 2.61%, 5.38% and 2.03% improvements when compared with the suboptimal method (DK-Consistency). In terms of sensitivity, DK-HRS also achieves a comparable performance compared with the best performance (80.40% and 81.21% on average). Moreover, we can also find that our DK-HRS method achieves comparable FPS with other methods. The phenomenon also proves that incorporating domain knowledge can improve performance with almost no additional model parameters. Overall speaking, DK-HRS can boost the performance on almost all metrics, and is prone to improve more on specificity than sensitivity. To clearly illustrate the comparison results, we show the AUC of these methods in Fig. 7 by error bars. We can see that on average, DK-HRS has the best performance among all comparison methods.

4.4. Ablation studies

To tease apart the importance of different components and hyper-parameters of DK-HRS, we design a set of ablation studies. Unless specified, all experiments are conducted under 20% of labeled data setting.

(1) **Effectiveness of different modules.**

To prove the effectiveness of different components in DK-HRS (i.e., FixMatch, VAT and DKCL), we analyze the performance of using each module alone and with different combinations of these three

Table 2
Comparison with other semi-supervised learning methods under 20% of labeled data setting.

Method	Percentage		Metrics (%)					
	L	UL	AUC	Accuracy	Sensitivity	Specificity	F1	FPS
Baseline	20%	0	80.59 ±1.00	74.74 ±0.68	75.46 ±1.69	73.99 ±0.99	75.24 ±0.88	**440 ±24**
VAT (Miyato et al., 2019)	20%	80%	81.60 ±0.55	75.67 ±0.39	74.60 ±1.85	76.77 ±1.72	75.72 ±0.63	401 ±12
MeanTeacher (Tarvainen & Valpola, 2017)	20%	80%	83.33 ±0.44	76.83 ±0.63	79.28 ±0.39	74.30 ±0.98	77.69 ±0.55	391 ±16
UDA (Xie et al., 2020)	20%	80%	83.30 ±0.76	76.75 ±1.12	79.65 ±1.62	73.74 ±1.45	77.71 ±1.12	303 ±4
MixMatch (Berthelot et al., 2019)	20%	80%	84.46 ±0.34	77.73 ±0.55	**81.21 ±1.35**	74.13 ±1.71	78.77 ±0.52	432 ±17
FixMatch (Sohn et al., 2020)	20%	80%	84.58 ±0.51	78.46 ±0.58	76.17 ±0.76	80.83 ±1.21	78.26 ±0.54	435 ±13
SRC-MT (Liu et al., 2020)	20%	80%	84.58 ±0.90	77.18 ±0.51	73.77 ±2.04	80.72 ±2.20	76.69 ±0.69	397 ±6
Wang et al. (2021)	20%	80%	84.95 ±0.19	78.95 ±0.39	77.54 ±1.29	80.40 ±1.73	78.94 ±0.36	414 ±13
DK-Consistency (Xie, Niu, Liu, Li et al., 2021)	20%	80%	85.20 ±0.53	78.88 ±0.45	80.49 ±1.19	77.23 ±1.66	79.51 ±0.39	341 ±45
DK-HRS (ours)	20%	80%	**88.24 ±0.42**	**81.49 ±0.47**	80.40 ±2.01	**82.61 ±2.09**	**81.54 ±0.59**	400 ±12

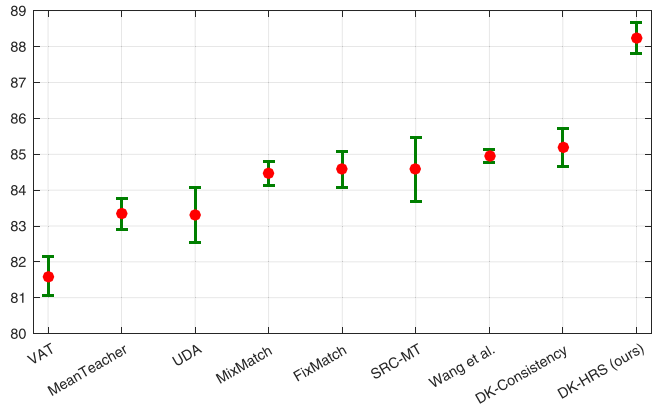


Fig. 7. AUC error bars of different methods.

modules. The quantitative results are listed in Table 3, and the best performance is shown in bold. We can clearly see that the best performance is achieved by combining all three modules simultaneously, which proves the superiority of our DK-HRS method. When using these three modules separately, DKCL achieves the best performance in most metrics when compared with FixMatch and VAT, and FixMatch performs better than VAT. This phenomenon shows the advantage of incorporating domain knowledge into the semi-supervised learning framework. In addition, the method incorporating domain knowledge into the regularization-based structure (combining FixMatch and DKCL) achieves better performance than the one only combining two kinds of regularization methods (i.e., FixMatch and VAT). The results also demonstrate the effectiveness of adopting domain knowledge rather than only using general regularization methods.

(2) **The effectiveness of using different kinds of DK-augmented images**.

In this section, we further analyze the diagnostic performance of incorporating different kinds of DK-augmented images. As mentioned in Section 3.4, the tumor area, tumor and its surrounding areas, and tumor margin areas are important for the diagnostic process. We design different kinds of DK-images to represent these areas and test them instead of the original DK-augmented images. Specifically, to depict the tumor area, the ROI image is selected. Some images with expanded ROI areas, i.e., the 1.5*ROI, 2*ROI, 2.5*ROI and 3*ROI images are used to incorporate the tumor and its surrounding areas. Additionally, to highlight the margin areas of tumor, the margin-alone image (only the pixels in margin areas are preserved, and the other pixels are set as the mean value of the dataset), margin-reinforced image (i.e., pixels within margin areas are set to twice as before) and margin-reinforced3 image (i.e., pixels within margin areas are set to three times as before) are designed. The quantitative results of these different conditions along with our method are listed in Table 4.

From Table 4, we can find that compared with the baseline method, the methods using these different kinds of DK images all perform well and improve the diagnostic performance in different extents. In particular, in all conditions of representing the tumor and its surrounding areas (i.e., 1.5*ROI, 2*ROI, 2.5*ROI and 3*ROI images), 1.5*ROI image performs best in almost all metrics. The reason can be attributed to that even with more tumor-surrounding information, larger ROI areas may include some disturbing information. Additionally, among all conditions of highlighting the margin areas of tumors, the method that uses margin-reinforced image achieves the best performance, outperforms the margin-alone and margin-reinforced3 images. The phenomenon shows that margin areas in each image need to be weighted appropriately, too little or too much retention of margin areas in the image may lead to information restriction or information loss. Consequently, these three DK images (i.e., the ROI, 1.5*ROI and margin-reinforced images) are chosen as the representations of domain knowledge in our method.

Based on the above results, some images that highlight all these important tumor areas are also designed and tested as the final DK-augmented images. In particular, as a straightforward representation, 1.5*ROI margin-reinforced images are first selected. Then, two kinds of fusion methods are also designed to combine the information from these three kinds of images, the DK-equal images (combine three kinds of images with equal weight) and our DK-augmented images (mix them up using random weights). The performance of these conditions is also listed in Table 4.

We can find that when compared with the other two conditions, our method achieves better performance with a relatively small standard deviation. We believe the reason behind this is that mixing-up operation generally combines the context and texture information

Table 3
Performance of using different components in DK-HRS under 20% of labeled data setting.

Method			Metrics (%)				
FixMatch	VAT	DKCL	AUC	Accuracy	Sensitivity	Specificity	F1
✓			84.58 ± 0.51	78.46 ± 0.58	76.17 ± 0.76	80.83 ± 1.21	78.26 ± 0.54
	✓		81.60 ± 0.55	75.67 ± 0.39	74.60 ± 1.85	76.77 ± 1.72	75.72 ± 0.63
		✓	86.58 ± 0.65	80.10 ± 0.25	**84.21 ± 1.16**	75.83 ± 1.12	81.15 ± 0.34
✓	✓		84.95 ± 0.19	78.95 ± 0.39	77.54 ± 1.29	80.40 ± 1.73	78.94 ± 0.36
✓		✓	86.06 ± 0.67	79.43 ± 0.80	77.34 ± 1.70	81.59 ± 2.19	79.28 ± 0.78
✓	✓	✓	**88.24 ± 0.42**	**81.49 ± 0.47**	80.40 ± 2.01	**82.61 ± 2.09**	**81.54 ± 0.59**

Table 4
The performance of using different kinds of DK-augmented samples.

DK-IMG	Metrics (%)				
	AUC	Accuracy	Sensitivity	Specificity	F1
Baseline	80.59 ± 1.00	74.74 ± 0.68	75.46 ± 1.69	73.99 ± 0.99	75.24 ± 0.88
ROI image	87.47 ± 0.37	80.33 ± 0.30	78.31 ± 2.53	82.42 ± 2.06	80.19 ± 0.74
1.5 ∗ ROI image	88.14 ± 0.72	**81.69 ± 1.18**	79.95 ± 4.24	83.50 ± 2.26	**81.58 ± 1.75**
2 ∗ ROI image	87.58 ± 0.38	81.13 ± 0.23	79.22 ± 1.50	83.12 ± 1.75	81.03 ± 0.30
2.5 ∗ ROI image	87.45 ± 0.45	80.44 ± 0.69	76.56 ± 3.04	84.45 ± 2.01	79.90 ± 1.19
3 ∗ ROI image	87.14 ± 0.23	79.99 ± 0.54	74.07 ± 1.65	**86.12 ± 1.22**	79.01 ± 0.76
Margin-reinforced image	87.19 ± 0.61	79.81 ± 0.76	75.11 ± 1.67	84.67 ± 0.91	79.10 ± 0.95
Margin-alone image	86.86 ± 0.43	78.35 ± 0.71	71.06 ± 3.07	85.90 ± 1.95	76.93 ± 1.30
Margin-reinforced3 image	84.48 ± 0.57	77.43 ± 1.38	71.68 ± 3.23	83.39 ± 0.98	76.34 ± 1.91
1.5 ∗ ROI margin-reinforced image	87.59 ± 0.50	80.61 ± 0.75	76.89 ± 3.24	84.48 ± 1.95	80.11 ± 1.30
DK-equal image	88.19 ± 0.56	81.03 ± 0.57	80.20 ± 2.27	82.48 ± 2.16	81.08 ± 0.73
DK-augmented image (ours)	**88.24 ± 0.42**	81.49 ± 0.47	**80.40 ± 2.01**	82.61 ± 2.09	81.54 ± 0.59

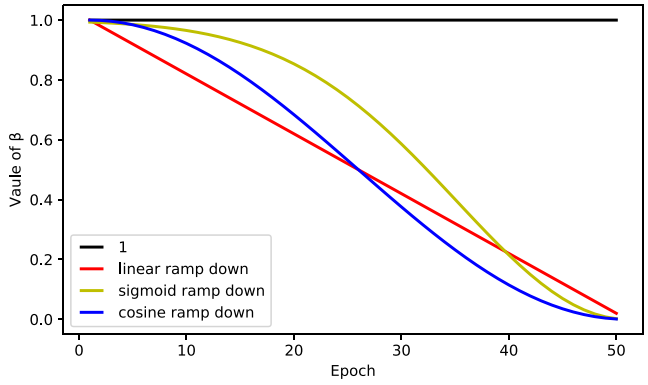


Fig. 8. Different ramp down functions of parameter β.

from different images by different proportions, which can promote the learning of distinguishing features and improve the robustness of the model. Additionally, when compared with the method of using 1.5*ROI margin-reinforced image, our method also shows better performance on almost all metrics. In particular, the former method is prone to improve specificity rather than sensitivity, while our method improves both of these two metrics simultaneously. The above results also demonstrate the effectiveness of designing and mixing up of these three kinds of DK images.

(3) **The impact of loss weight β.**

As mentioned in Section 3.5, it is more appropriate to set β as a ramp-down function. We test β on three kinds of ramp-down functions: the linear ramp down, sigmoid ramp down and cosine ramp down. Meanwhile, we also test the performance on a fixed value ($\beta = 1$). Fig. 8 shows the curves of these different weight functions, and the quantitative results of these different conditions are listed in Table 5.

We can see from Table 5 that all methods using ramp-down functions achieve higher AUC when compared with the one using constant weight 1. In addition, the linear ramp-down method performs best among all these three ramp-down conditions, which has the highest AUC, accuracy, specificity and F1, and achieves comparable sensitivity value compared with the highest one (80.40% vs. 80.43%). Besides, the method using linear ramp-down also has smaller standard deviations on most metrics.

(4) **The training manner of DKCL module.**

In DKCL, contrastive learning is carried out between all DK-augmented samples and strongly-augmented ones. It is interesting to know whether using weakly-augmented samples instead of strongly-augmented ones can achieve better results. To answer this question, we compare the performance of these two different conditions, and the results are shown in Table 6. We can find that using strongly augmented samples is more effective. The reason is that using strongly-augmented samples can force the model to learn the most distinguishing domain features from strong transformations, by which the model performs better learning ability.

(5) **Influence of Parameter ε.**

Parameter ε is adopted to compute the adversarial direction for each input image in the VAT module. To investigate the impact of using different values of ε on the final performance of DK-HRS, we conduct ablation studies by taking the value increments of 5 in the range of [5,50]. It should be noted that other settings remain unchanged in the training process. The average of AUC and accuracy on these different conditions are shown in Fig. 9. We can see that a better average of AUC is achieved when ε is set as 15 (88.31%) and 30 (88.24%). In addition, when we also consider the accuracy, $\varepsilon = 30$ has a higher accuracy than $\varepsilon = 15$ (81.49% vs. 81.26%). Thus we choose $\varepsilon = 30$ as it performs well on both metrics.

4.5. *Performance on public datasets*

We also evaluate the performance of our DK-HRS on three public breast ultrasound datasets, the BUSI (Al-Dhabyani et al., 2020), UDIAT (Yap et al., 2018), and 3CUS (Rodtook et al., 2018). The detailed experimental results on these three public datasets are as follows.

(1) Performance on public dataset BUSI

Firstly, we evaluate the performance of our DK-HRS method on the public dataset BUSI (Al-Dhabyani et al., 2020). During the training process, the same image pre-processing method used in our dataset is also applied to BUSI. The batch sizes of labeled and unlabeled data are set as 4 and 120, respectively, and the initial learning rate is set as

Table 5
The influence of β for the final performance.

DK-HRS (β)	Percentage		Metrics (%)				
	L	UL	AUC	Accuracy	Sensitivity	Specificity	F1
Baseline	20%	0	80.59 ± 1.00	74.74 ± 0.68	75.46 ± 1.69	73.99 ± 0.99	75.24 ± 0.88
1	20%	80%	87.43 ± 0.94	81.09 ± 0.92	79.75 ± 3.24	82.48 ± 1.95	81.07 ± 1.35
Sigmoid ramp down	20%	80%	87.88 ± 0.43	81.00 ± 0.80	**80.43 ± 2.92**	81.59 ± 1.83	81.14 ± 1.15
Cosine ramp down	20%	80%	88.07 ± 0.52	81.05 ± 0.99	79.87 ± 3.34	82.28 ± 1.86	81.06 ± 1.45
Linear ramp down	20%	80%	**88.24 ± 0.42**	**81.49 ± 0.47**	80.40 ± 2.01	**82.61 ± 2.09**	**81.54 ± 0.59**

Table 6
The influence of different training manner used in DKCL.

Metrics	Method	
	Weakly-augmented sample	Strongly-augmented sample
AUC	86.77 ± 0.37	**88.24 ± 0.42**
Accuracy	80.49 ± 0.81	**81.49 ± 0.47**
Sensitivity	78.30 ± 2.45	**80.40 ± 2.01**
Specificity	**82.77 ± 1.45**	82.61 ± 2.09
F1	80.32 ± 1.09	**81.54 ± 0.59**

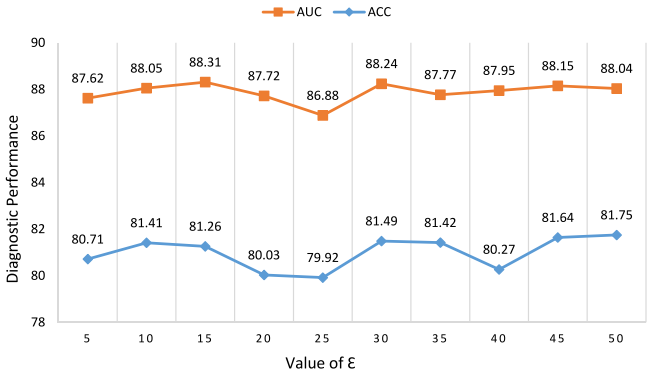


Fig. 9. The influence of ε on the performance of AUC and accuracy.

0.003. In particular, the EMA technique is not utilized for the training process, and all other parameters are set as the same as that in our BUS2019 dataset. Table 7 lists the quantitative results on BUSI, where the performance of the baseline method and our DK-HRS method on different percentages of labeled data settings are presented.

Additionally, to demonstrate the effectiveness of incorporating domain knowledge, we also show the performance comparison with the method proposed by Wang et al. (2021) and DK-Consistency (Xie, Niu, Liu, Li et al., 2021). Each value in the table represents the mean ± std result of three times of 5-fold cross-validation, and the best performance in each metric is also shown in bold under different labeled data settings.

From Table 7, we can see that DK-HRS shows better performance in most metrics (i.e., AUC, accuracy, sensitivity and F1) when compared with the corresponding baseline method and other comparison methods. For example, DK-HRS achieves the average AUC of 83.41%, 85.58% and 85.97% in 10%, 20% and 30% of labeled data settings, respectively. The performance is improved by 4.91%, 1.85% and 1.65% when compared with the corresponding baseline method, and improved by 1.29%, 1.57% and 1.18% when compared with the suboptimal method DK-Consistency. Similar to the performance on our dataset, the improvements on BUSI also show a trend of decline as the proportion of labeled data increases.

Besides, we can see that the method proposed by Wang et al. achieves the highest specificity at the cost of the lowest sensitivity. In contrast, our DK-HRS method along with the DK-Consistency method (both incorporating the information of domain knowledge) show better performance than the corresponding baseline method and the method of Wang et al. which can improve the sensitivity and specificity simultaneously. The phenomenon also demonstrates that incorporating domain knowledge into the semi-supervised learning process can greatly boost diagnostic performance than combining with two regularization-based methods.

(2) Performance on public dataset UDIAT

We also evaluate the performance of our DK-HRS method on the public dataset UDIAT. During the training process, the batch sizes of labeled and unlabeled images are 4 and 80, respectively, and the initial learning rate is set as 0.003. All other settings are set as the same as that in the BUS2019 dataset. In particular, EMA technique is not used in the training process. The Quantitative results of our method and some comparison methods on public dataset UDIAT are listed in Table 8, where the performance on different percentages of labeled data settings are presented (i.e., 10%, 20% and 30%). Each value in the table represents the mean ± std result of three times of 5-fold cross-validation, and the best performance in each metric is also shown in bold under different labeled data settings.

From Table 8, we can find that DK-HRS also shows better performance in UDIAT when compared with the baseline method and the other two comparison methods, especially in AUC and accuracy metrics. More specifically, DK-HRS obtains the average AUC of 55.92%, 60.65% and 61.71%, and the average accuracy of 62.01%, 64.25% and 65.33% in 10%, 20% and 30% labeled data settings, respectively. Under different labeled data settings, the AUC of DK-HRS is improved by 2.63%, 6.69% and 7.02% when compared with the corresponding baseline methods, and improved by 0.11%, 2.56% and 0.89% when compared with the suboptimal method DK-Consistency.

Additionally, in terms of sensitivity and specificity, the method of Wang et al. achieves the highest specificity but with the lowest sensitivity in different labeled data settings, while DK-Consistency achieves the highest sensitivity at the cost of reduced specificity. In contrast, our DK-HRS method can improve the diagnostic performance in both these two metrics in most conditions, and the improvements in specificity are also higher than that in sensitivity. This phenomenon also demonstrates the effectiveness of our DK-HRS method even in especially small-sized UDIAT dataset.

(3) Performance on the 3-class diagnostic task

Finally, we also verify the performance of our DK-HRS method on a 3-class diagnostic dataset 3CUS. In the training process, the batch sizes of labeled data and unlabeled data are 4 and 120, respectively, and the initial learning rate is set as 0.003. Similar to the BUSI and UDIAT, EMA is also not utilized in the training process of 3CUS. Additionally, all other parameters are set as the same as that in our BUS2019 dataset. In particular, all metrics are computed by using the macro-average manner for different categories.

Table 7
Quantitative evaluation on the public dataset BUSI.

Method	Percentage		Metrics (%)				
	L	UL	AUC	Accuracy	Sensitivity	Specificity	F1
Upper Bound	100%	0	92.29 ± 1.70	86.92 ± 1.90	79.37 ± 4.97	90.55 ± 2.43	79.72 ± 3.11
Baseline	10%	0	78.50 ± 2.34	74.18 ± 2.62	66.98 ± 7.11	77.61 ± 6.11	62.68 ± 2.73
Wang et al. (2021)	10%	90%	78.47 ± 5.54	77.57 ± 2.68	55.65 ± 4.11	**88.12 ± 4.65**	61.78 ± 3.26
DK-Consistency (Xie, Niu, Liu, Li et al., 2021)	10%	90%	82.12 ± 2.86	78.83 ± 2.76	66.19 ± 5.97	84.89 ± 3.74	66.95 ± 4.54
DK-HRS (ours)	10%	90%	**83.41 ± 3.38**	**79.02 ± 2.52**	**69.22 ± 7.93**	83.74 ± 3.54	**67.82 ± 3.65**
Baseline	20%	0	83.82 ± 4.18	78.98 ± 2.63	70.00 ± 5.75	83.30 ± 3.36	68.34 ± 3.87
Wang et al. (2021)	20%	80%	83.39 ± 3.92	79.65 ± 4.23	63.10 ± 3.24	**87.63 ± 6.97**	67.15 ± 4.44
DK-Consistency (Xie, Niu, Liu, Li et al., 2021)	20%	80%	84.10 ± 3.67	80.23 ± 3.32	69.05 ± 3.98	85.61 ± 4.38	69.49 ± 4.33
DK-HRS (ours)	20%	80%	**85.58 ± 3.36**	**81.34 ± 3.03**	**73.05 ± 6.99**	84.05 ± 6.79	**70.98 ± 3.16**
Baseline	30%	0	84.32 ± 2.83	79.55 ± 3.26	71.27 ± 4.80	83.53 ± 4.89	69.43 ± 4.06
Wang et al. (2021)	30%	70%	84.07 ± 2.28	80.47 ± 2.18	61.90 ± 3.95	**89.42 ± 4.37**	67.37 ± 2.31
DK-Consistency (Xie, Niu, Liu, Li et al., 2021)	30%	70%	84.79 ± 2.63	81.10 ± 2.70	69.52 ± 4.93	86.65 ± 4.78	70.53 ± 3.34
DK-HRS (ours)	30%	70%	**85.97 ± 2.23**	**81.98 ± 2.06**	**73.43 ± 4.60**	84.96 ± 4.61	**71.78 ± 3.31**

Table 8
Quantitative evaluation on the public dataset UDIAT.

Method	Percentage		Metrics (%)				
	L	UL	AUC	Accuracy	Sensitivity	Specificity	F1
Upper Bound	100%	0	77.54 ± 7.65	75.31 ± 5.99	62.08 ± 10.17	81.84 ± 6.80	62.44 ± 8.80
Baseline	10%	0	53.29 ± 12.57	57.11 ± 9.01	28.69 ± 13.87	71.21 ± 10.20	30.38 ± 12.76
Wang et al. (2021)	10%	90%	55.34 ± 12.93	61.27 ± 10.14	23.67 ± 20.65	**79.87 ± 16.87**	25.53 ± 18.40
DK-Consistency (Xie, Niu, Liu, Li et al., 2021)	10%	90%	55.81 ± 5.62	61.62 ± 5.20	**47.33 ± 13.56**	68.61 ± 12.70	**44.29 ± 5.83**
DK-HRS (ours)	10%	90%	**55.92 ± 14.44**	**62.01 ± 10.90**	42.09 ± 18.04	76.31 ± 18.08	44.03 ± 17.24
Baseline	20%	0	53.96 ± 7.00	60.05 ± 4.91	41.26 ± 13.81	69.42 ± 8.28	39.73 ± 9.75
Wang et al. (2021)	20%	80%	57.29 ± 9.73	64.05 ± 7.82	26.59 ± 16.98	**82.74 ± 15.06**	30.18 ± 15.83
DK-Consistency (Xie, Niu, Liu, Li et al., 2021)	20%	80%	58.09 ± 13.58	61.02 ± 10.05	**48.60 ± 16.37**	67.25 ± 12.07	**44.97 ± 12.57**
DK-HRS (ours)	20%	80%	**60.65 ± 7.56**	**64.25 ± 5.97**	46.08 ± 13.19	78.21 ± 10.21	44.61 ± 8.99
Baseline	30%	0	54.69 ± 6.88	60.58 ± 4.86	44.26 ± 18.97	67.23 ± 11.78	40.18 ± 12.35
Wang et al. (2021)	30%	70%	59.30 ± 9.72	63.47 ± 4.94	31.21 ± 27.66	**79.38 ± 19.29**	30.47 ± 15.85
DK-Consistency (Xie, Niu, Liu, Li et al., 2021)	30%	70%	60.82 ± 11.49	62.89 ± 9.36	**53.88 ± 12.32**	67.36 ± 13.02	**49.13 ± 9.27**
DK-HRS (ours)	30%	70%	**61.71 ± 8.03**	**65.33 ± 5.25**	47.91 ± 16.15	74.85 ± 11.25	48.42 ± 11.23

Table 9
Quantitative evaluation on the 3CUS dataset.

Method	Percentage		Metrics (%)				
	L	UL	AUC	Accuracy	Sensitivity	Specificity	F1
Upper Bound	100%	0	78.18 ± 6.95	77.78 ± 5.51	56.01 ± 9.74	80.89 ± 4.33	55.32 ± 10.27
Baseline	10%	0	60.65 ± 10.11	62.95 ± 11.23	41.95 ± 11.90	72.20 ± 5.65	39.40 ± 11.05
Wang et al. (2021)	10%	90%	62.75 ± 9.44	62.38 ± 5.65	40.15 ± 10.50	70.78 ± 5.06	40.20 ± 13.21
DK-Consistency (Xie, Niu, Liu, Li et al., 2021)	10%	90%	68.46 ± 8.95	63.93 ± 5.88	**49.07 ± 8.26**	76.36 ± 3.68	45.55 ± 6.22
DK-HRS (ours)	10%	90%	**69.73 ± 10.36**	**69.17 ± 7.27**	48.55 ± 9.59	**76.57 ± 3.58**	**45.62 ± 7.64**
Baseline	20%	0	66.70 ± 10.54	71.86 ± 9.35	49.71 ± 10.28	77.08 ± 5.18	47.05 ± 11.78
Wang et al. (2021)	20%	80%	70.10 ± 8.45	71.39 ± 7.28	45.02 ± 10.39	74.19 ± 6.44	42.94 ± 11.13
DK-Consistency (Xie, Niu, Liu, Li et al., 2021)	20%	80%	72.08 ± 9.48	71.47 ± 6.10	**54.39 ± 11.19**	79.02 ± 6.45	**50.80 ± 10.28**
DK-HRS (ours)	20%	80%	**73.07 ± 5.39**	**72.63 ± 5.03**	52.03 ± 13.63	**79.13 ± 5.01**	50.75 ± 11.45
Baseline	30%	0	68.80 ± 10.16	73.37 ± 7.47	46.88 ± 8.12	76.10 ± 5.93	48.65 ± 10.16
Wang et al. (2021)	30%	70%	71.28 ± 10.48	73.30 ± 7.16	46.28 ± 8.37	75.96 ± 6.16	46.36 ± 9.07
DK-Consistency (Xie, Niu, Liu, Li et al., 2021)	30%	70%	73.13 ± 8.40	71.93 ± 6.31	**55.11 ± 12.19**	78.99 ± 5.95	**52.78 ± 10.78**
DK-HRS (ours)	30%	70%	**74.02 ± 5.73**	**73.85 ± 5.35**	53.43 ± 13.76	**79.41 ± 5.74**	52.60 ± 12.32

The Quantitative results on the dataset 3CUS are listed in Table 9, where each value in the table represents the mean ± std result of three times of 5-fold cross-validation, and the best performance in each metric is also shown in bold under different labeled data settings.

We can see that from Table 9, DK-HRS shows better performance in most metrics on 3CUS dataset. For example, DK-HRS achieves the mean AUC of 69.73%, 73.07% and 74.02% in 10%, 20% and 30% labeled data settings, which is improved by 9.08%, 6.37% and 5.22% when compared with the corresponding baseline method, respectively. Meanwhile, DK-HRS also achieves the best accuracy among all these methods, the mean value is 69.17%, 72.63% and 73.85% in these three labeled data settings, respectively.

Besides, in terms of the sensitivity, specificity and F1 metrics, DK-HRS still performs comparable or better performance when compared with DK-Consistency, and has a big improvement over the corresponding baseline method and Wang's method. In particular, when compared with the method proposed by Wang et al. that only combining VAT with FixMatch, DK-HRS performs better in all metrics. The phenomenon also proves that incorporating domain knowledge is useful for the multi-class diagnostic task.

5. Conclusion and discussion

In this paper, we propose DK-HRS, a simple yet effective method to incorporate medical domain knowledge into the regularization strategy for semi-supervised breast cancer diagnosis. In DK-HRS, different kinds of perturbed samples are explored and adopted to boost the final performance. Extensive experiments on four datasets manifest that by incorporating domain knowledge, DK-HRS shows superior performance in comparison to some state-of-the-art methods.

There are a few interesting directions to explore in the future. Firstly, our current model is mainly tested on binary classification tasks (i.e., distinguishing benign lesions from malignant ones). Even if validated on a 3-class diagnostic task, the dataset is small-scale and unbalanced in categories. In the future, we will extend our work to more classification tasks on large-scale multi-class datasets. For these tasks, new kinds of domain knowledge may be required to generate DK-augmented images, and new incorporating approaches might also be designed to further improve the performance. Meanwhile, we will try to extend our approach of using DK-augmented images on more datasets of different kinds of diseases.

CRediT authorship contribution statement

Xiaozheng Xie: Conceptualization, Methodology, Software, Writing – original draft, Funding acquisition. **Jianwei Niu:** Resources, Supervision. **Xuefeng Liu:** Project administration, Writing – review & editing, Funding acquisition. **Yong Wang:** Resources, Data curation, Funding acquisition. **Qingfeng Li:** Investigation, Visualization, Validation. **Shaojie Tang:** Formal analysis, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by China Postdoctoral Science Foundation, China [grant number 2023M730224, 2017M620683], National Natural Science Foundation of China [grant numbers 61976012, 62372027, 62372028], and Fundamental Research Funds for the Central Universities, China [grant number FRF-TP-22-051A1].

References

Ahmed, I., & Muhammad, S. (2023). Bts-st: Swin transformer network for segmentation and classification of multimodality breast cancer images. *Knowledge-Based Systems*, *267*, Article 110393. http://dx.doi.org/10.1016/j.knosys.2023.110393.

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief*, *28*, Article 104863. http://dx.doi.org/10.1016/j.dib.2019.104863.

Antropova, N., Huynh, B. Q., & Giger, M. L. (2017). A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Medical Physics*, *44*, 5162–5171. http://dx.doi.org/10.1002/mp.12453.

Bachman, P., Alsharif, O., & Precup, D. (2014). Learning with pseudo-ensembles. In *Advances in neural information processing systems* (pp. 3365–3373).

Basak, H., & Yin, Z. (2023). Semi-supervised domain adaptive medical image segmentation through consistency regularized disentangled contrastive learning. In *Medical image computing and computer assisted intervention* (pp. 260–270). http://dx.doi.org/10.1007/978-3-031-43901-8_25.

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., & Raffel, C. (2020). Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International conference on learning representations*.

Berthelot, D., Goodfellow, I. nicholas. Carlini., Oliver, A., Papernot, N., & Raffel, C. (2019). Mixmatch: a holistic approach to semi-supervised learning. In *Advances in neural information processing systems* (pp. 5050–5060).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). PMLR.

Chen, C., Wang, Y., Niu, J., Liu, X., Li, Q., & Gong, X. (2021). Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. *IEEE Transactions on Medical Imaging*, *40*, 2439–2451. http://dx.doi.org/10.1109/TMI.2021.3078370.

Cubuk, E., Zoph, B., Shlens, J., & Le, Q. (2019). Randaugment: Practical data augmentation with no separate search. arXiv preprint arXiv:1909.13719.

Cubuk, E., Zoph, B., Shlens, J., & Le, Q. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 3008–3017). http://dx.doi.org/10.1109/CVPRW50498.2020.00359.

Gomez-Flores, W., & Ruiz-Ortega, B. (2016). New fully automated method for segmentation of breast lesions on ultrasound based on texture analysis. *Ultrasound in Medicine & Biology*, *42*, 1637–1650. http://dx.doi.org/10.1016/j.ultrasmedbio.2016.02.016.

Goodfellow, I., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Hendrycks, D., Mu, N., Cubuk, E., Zoph, B., Gilmer, J., & Lakshminarayanan, B. (2020). Augmix: A simple data processing method to improve robustness and uncertainty. In *Proceedings of the international conference on learning representations*.

Huynh, T., Nibali, A., & He, Z. (2022). Semi-supervised learning for medical image classification using imbalanced training data. *Computer Methods and Programs in Biomedicine*, *216*, Article 106628.

Jimenez-Sanchez, A., Mateus, D., Kirchhoff, S., Kirchhoff, C., Biberthaler, P., Navab, N., Ballester, M. Gonzalez., & Piella, G. (2019). Medical-based deep curriculum learning for improved fracture classification. In *Medical image computing and computer assisted intervention* (pp. 694–702). Springer.

Kong, J., He, Y., Zhu, X., Shao, P., Xu, Y., Chen, Y., Coatrieux, J. L., & Yang, G. (2022). Bkc-net: Bi-knowledge contrastive learning for renal tumor diagnosis on 3d ct images. *Knowledge-Based Systems*, *252*, Article 109369. http://dx.doi.org/10.1016/j.knosys.2022.109369.

Laine, S., & Aila, T. (2017). Temporal ensembling for semi-supervised learning. In *International conference on learning representation*.

Li, H., Chen, D., Nailon, W. H., Davies, M. E., & Laurenson, D. I. (2022). Dual convolutional neural networks for breast mass segmentation and diagnosis in mammography. *IEEE Transactions on Medical Imaging*, *41*, 3–13. http://dx.doi.org/10.1109/TMI.2021.3102622.

Li, K., Wang, S., Yu, L., & Heng, P. A. (2020). Dual-teacher: Integrating intra-domain and inter-domain teachers for annotation-efficient cardiac segmentation. In *Medical image computing and computer assisted intervention* (pp. 418–427).

Li, J., Zhou, P., Xiong, C., Socher, R., & Hoi, S. C. (2021). Prototypical contrastive learning of unsupervised representations. In *International conference on learning representations*.

Liu, Q., Yu, L., Luo, L., Dou, Q., & Heng, P. (2020). Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Transactions on Medical Imaging*, *39*, 3429–3440.

Mendelson, E., Bohm-Velez, M., Berg, W., Whitman, G., Feldman, I., & Madjar, H. (2013). *ACR BI-RADS ultrasound, ACR BI-RADS® Atlas, breast imaging reporting and data system*. Reston, VA: American College of Radiology.

Miyato, T., Maeda, S., Koyama, M., & Ishii, S. (2019). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*, 1979–1993. http://dx.doi.org/10.1109/TPAMI.2018.2858821.

Pham, T., Tao, X., Zhang, J., Yong, J., Li, Y., & Xie, H. (2022). Graph-based multi-label disease prediction model learning from medical data and domain knowledge. *Knowledge-Based Systems*, *235*, Article 107662. http://dx.doi.org/10.1016/j.knosys.2021.107662.

Ren, Z., Kong, X., Zhang, Y., & Wang, S. (2023). Ukssl: Underlying knowledge based semi-supervised learning for medical image classification. *IEEE Open Journal of Engineering in Medicine and Biology*, http://dx.doi.org/10.1109/OJEMB.2023.3305190.

Rodtook, A., Kirimasthong, K., Lohitvisate, W., & Makhanov, S. S. (2018). Automatic initialization of active contours and level set method in ultrasound images of breast abnormalities. *Pattern Recognition*, *79*, 172–182. http://dx.doi.org/10.1016/j.patcog.2018.01.032.

Samala, R. K., Chan, H. P., Hadjiiski, L., Helvie, M. A., Richter, C. D., & Cha, K. H. (2019). Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE Transactions on Medical Imaging*, *38*, 686–696.

Shi, F., Chen, B., Cao, Q., Wei, Y., Zhou, Q., Zhang, R., Zhou, Y., Yang, W., Wang, X., Fan, R., Yang, F., Chen, Y., Li, W., Gao, Y., & Shen, D. (2022). Semi-supervised deep transfer learning for benign-malignant diagnosis of pulmonary nodules in chest ct images. *IEEE Transactions on Medical Imaging*, *41*, 771–781. http://dx.doi.org/10.1109/TMI.2021.3123572.

Shin, S. Y., Lee, S., Yun, I. D., Kim, S. M., & Lee, K. M. (2019). Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. *IEEE Transactions on Medical Imaging*, *38*, 762–774.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., & Li, C. L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in neural information processing systems* (pp. 596–608).

Su, H., Shi, X., Cai, J., & Yang, L. (2019). Local and global consistency regularized mean teacher for semi-supervised nuclei classification. In *Medical image computing and computer assisted intervention* (pp. 559–567).

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems* (pp. 1195–1204).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wang, X., Chen, H., Xiang, H., Lin, H., Lin, X., & Heng, P. A. (2021). Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. *Medical Image Analysis*, *70*, Article 102010. http://dx.doi.org/10.1016/j.media.2021.102010.

Wang, H., & Li, X. (2023). Towards generic semi-supervised framework for volumetric medical image segmentation. In *Neural information processing systems*.

Wang, Y., Wang, Y., Cai, J., Lee, T. K., Miao, C., & Wang, Z. J. (2023). Ssd-kd: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. *Medical Image Analysis*, *84*, Article 102693.

Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. In *Advances in neural information processing systems* (pp. 6256–6268).

Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., & Yu, S. (2021). A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, *69*, Article 101985.

Xie, X. Z., Niu, J. W., Liu, X. F., Li, Q. F., Wang, Y., Han, J., & Tang, S. (2022). Dg-cnn: Introducing margin information into convolutional neural networks for breast cancer diagnosis in ultrasound images. *Journal of Computer Science and Technology*, *37*, 277–294.

Xie, X., Niu, J., Liu, X., Li, Q., Wang, Y., & Tang, S. (2021). Dk-consistency: A domain knowledge guided consistency regularization method for semi-supervised breast cancer diagnosis. In *IEEE international conference on bioinformatics and biomedicine* (pp. 3435–3442).

Xie, Y., Zhang, J., & Xia, Y. (2019). Semi-supervised adversarial model for benign malignant lung nodule classification on chest ct. *Medical Image Analysis*, *57*, 237–248. http://dx.doi.org/10.1016/j.media.2019.07.004.

Xu, J., Lala, S., Gagoski, B., Turk, E. Abaci., Grant, P. E., Golland, P., & Adalsteinsson, E. (2020). Semi-supervised learning for fetal brain mri quality assessment with roi consistency. In *Medical image computing and computer assisted intervention* (pp. 386–395).

Xu, B., Liu, J., Hou, X., Liu, B., Garibaldi, J., Ellis, I., Green, A., Shen, L., & Qiu, G. (2019). Attention by selection: A deep selective attention approach to breast cancer classification. *IEEE Transactions on Medical Imaging*, *39*, 1930–1941. http://dx.doi.org/10.1109/TMI.2019.2962013.

Xu, Z., Wang, Y., Lu, D., Luo, X., Yan, J., Zheng, Y., & yu Tong, R. K. (2023). Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. *Medical Image Analysis*, *88*, Article 102880. http://dx.doi.org/10.1016/j.media.2023.102880.

Xu, Z., Wang, Y., Lu, D., Yu, L., Yan, J., Luo, J., Ma, K., Zheng, Y., & Tong, R. K. Y. (2022). All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, *26*, 3174–3184.

Yang, Q., Liu, X., Chen, Z., Ibragimov, B., & Yuan, Y. (2022). Semi-supervised medical image classification with temporal knowledge-aware regularization. In *International conference on medical image computing and computer-assisted intervention* (pp. 119–129). Springer.

Yang, W., Zhao, J., Qiang, Y., Yang, X., Dong, Y., Du, Q., Shi, G., & Zia, M. B. (2019). Dscgans: Integrate domain knowledge in training dual-path semi-supervised conditional generative adversarial networks and s3vm for ultrasonography thyroid nodules classification. In *Medical image computing and computer assisted intervention* (pp. 558–566). Springer.

Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A. K., & Marti, R. (2018). Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, *22*, 1218–1226. http://dx.doi.org/10.1109/JBHI.2017.2731873.

Zeng, Q., Xie, Y., Lu, Z., & Xia, Y. (2023). Pefat: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In *2023 IEEE/CVF conference on computer vision and pattern recognition* (pp. 15671–15680). http://dx.doi.org/10.1109/CVPR52729.2023.01504.

Zhang, E., Seiler, S., Chen, M., Lu, W., & Gu, X. (2020). BIRADS features-oriented semi-supervised deep learning for breast ultrasound computer-aided diagnosis. *Physics in Medicine and Biology*, *65*, Article 125005. http://dx.doi.org/10.1088/1361-6560/ab7e7d.

Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q., & Ooi, B. C. (2022). Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20666–20676).