Journal of Big Data

RESEARCH Open Access

Supervised contrastive pre-training models for mammography screening

Zhenjie Cao[1,2], Zhuo Deng[1], Zhicheng Yang[3], Jie Ma[4] and Lan Ma[1*]

*Correspondence:
malan@sz.tsinghua.edu.cn

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[2] PingAn Tech, Shenzhen, China
[3] PAII Inc., Palo Alto, USA
[4] Shenzhen People's Hospital, Shenzhen, China

Abstract

Breast cancer is now the most deadly cancer worldwide. Mammography screening is the most effective method for early detection and diagnosis of breast cancer. Due to the lack of labeled mammograms, building an AI system for mammography screening often relies heavily on human-designed data augmentation, which doesn't always perform robustly when applied to clinical scenarios. This paper presents a novel framework of Supervised Contrastive Pre-training followed by Supervised Fine-tuning (SCP+SF) for mammography screening. Unlike the previous approaches, the proposed supervised contrastive pre-training does not need a data augmentation module. We apply the SCP+SF framework to two challenging and important mammography screening tasks for breast cancer: mammographic abnormality screening and mammographic malignancy screening. Our extensive experiments on a large-scale dataset show that the supervised contrastive pre-training (SCP) can substantially improve the final model performance compared with the traditional direct supervised training approach. Superior results of AUC and specificity/sensitivity have been achieved on two clinically significant mammographic screening tasks in comparison with previously reported State-Of-The-Art approaches. We believe this work is the first to show that supervised contrastive pre-training (SCP) followed by supervised fine-tuning (SF) can outperform the supervised counterpart on these two critical medical imaging tasks.

Keywords: Breast cancer, Mammography screening, Supervised contrastive pre-training

Introduction

According to WHO, breast cancer has surpassed lung cancer as the world's number one cancer in terms of morbidity and mortality in 2020 [1]. Mammography screening is the most cost-effective method for early detection of breast cancer, with approximately 48 million mammograms performed annually in the US. It has been reported that the U.S. radiologists ranged from 66.7% to 98.6% for sensitivity and from 71.2% to 96.9% for specificity in mammogram-based breast cancer diagnosis [2]. Despite the recent study [3] showing that AI systems have the potential to surpass human experts in breast

cancer interpretation, further improvement of AI systems' accuracy and robustness is demanded before their large-scale adoption.

Previous deep neural nets (DNN)-based mammographic screening systems are trained directly by supervised learning [3–9]. The lack of labeled data causes these methods to normally rely heavily on human-designed data augmentation during training. To solve this problem, we propose a Supervised Contrastive Pre-training + Supervised Fine-tuning (SCP+SF) framework. Unlike the previous approaches, the proposed supervised contrastive pre-training does not need a data augmentation module. It first performs the SCP pre-training through a carefully designed Siamese contrastive learning module, searching for a better clustered embedding space, then transfers the pre-trained encoder to the SF module for the supervised fine-tuning phase.

Contrastive learning has been applied to self-supervised visual representation learning [10–14], exemplified by the success of *SimCLR/SimCLR-v2* [13, 14], which shows that self-supervised pre-training on ImageNet with a simple contrastive learning framework can generate competitive results on downstream image classification tasks when compared with fully supervised learning. Some follow-up work [15] demonstrates that contrastive pre-training can also be applied to fully supervised settings and further improve the SOTA performance on ImageNet.

Fundamentally, contrastive pre-training is a clustering process, with the objective of learning an embedding space to minimize the samples' intra-class variances while maximizing their inter-class variances. The rationale behind the supervised learning approach subsequent to contrastive pre-training is that supervised fine-tuning can be carried out more effectively in an embedding space where the training samples have been better set apart.

In this paper, we demonstrate the effectiveness of the SCP+SF framework through two important and challenging mammography screening tasks, namely (1) identifying normal mammograms with high confidence and (2) identifying malignant mammograms with high specificity (please see Sec. Task 1: mammographic abnormality screening and Task 2: mammographic malignancy screening for more detailed explanation). Our results show that for both tasks, the proposed SCP+SF framework significantly outperforms their counterparts. In particular, when trained on our in-house dataset of 134,488 images from 30,487 patients and tested on 2,538 images from 640 patients with biopsy ground truth, both screening models trained with SCP+SF surpass the previously reported SOTA approaches [4–6] by a large margin, in terms of AUC and specificity/sensitivity.

Figures 1a and b (best viewed in color) visualize the sample projections from the proposed contrastive learning module before and after the SCP phase, clearly illustrating the improvement in the separability of the two clusters representing the healthy and at-risk populations. Figures 1c and d are the sample projections from our proposed dual-view model, with direct supervised learning, and with the proposed SCP+SF training framework. They further demonstrate that the SCP+SF results in better clustering quality.

The main thrust of this paper is summarized below:

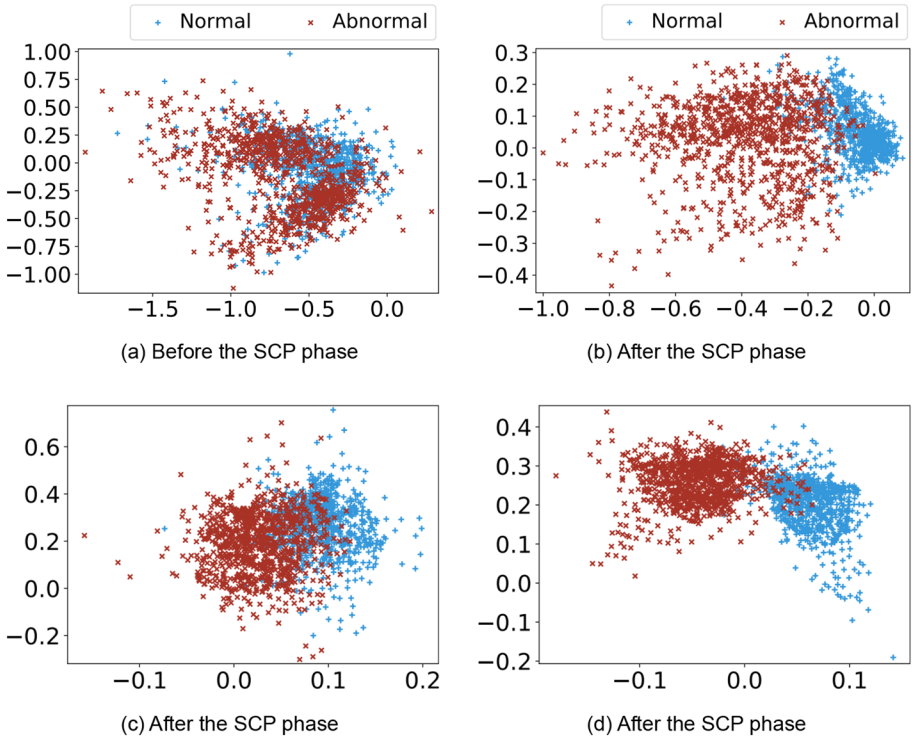

Fig. 1 Visualization of the normal and abnormal sample projections. **a**, **b** are from the Siamese contrastive learning module, and **c**, **d** are from the final dual-view model by t-SNE [16]. x,y axises indicating the projection plane before the Sigmoid function

1. We show that on two clinically meaningful mammography screening tasks, abnormality screening and malignancy screening, our models trained with the framework of Supervised Contrastive Pre-training followed by Supervised Fine-tuning (SCP+SF) consistently achieve superior performance over previously reported SOTA approaches.
2. The SCP+SF framework does *not* require data augmentation module and uses the regular L2 loss. This novel framework benefits from a carefully designed Siamese contrastive learning module and a patient-constrained sample selection scheme.

Related work

Mammographic screening

Previous works on deep learning based mammographic screening have focused on triage tasks of 1) identifying the normal mammograms with no lesions at all [6, 17–19], 2) identifying the malignant mammograms [3–5, 7, 20, 21]. We experiment with both tasks using our proposed framework. It is worth pointing out that most of the previous works on task No.1 use biopsy results as the ground truth to classify the patient data, while in our experiments, we treat the BI-RADS 1[1] category mammograms as normal for task No.1, which is a more conservative standard.

[1] We will explain the BI-RADS categories in detail in Sec. BI-RADS assessment categories

Most of the previous methods take the traditional direct supervised learning approach. McKinney et al. [3] detected lesions from mammograms' each view and then accumulated lesion-level cancer risk scores to produce image/breast/case-level cancer risk scores. Wu et al. [21] pre-trained the screening model on a large amount of data with BI-RADS labels before fine-tuning it with biopsy ground truth. However, its pre-training phase is *not* based on the contrastive learning principle and, therefore, different from our approach.

In particular, Yala et al. [6] is a classic multi-view-based method and achieved superior results on its own data, while the approaches in [4, 5] represent the latest SOTA mammography screening methods. All these three approaches' effectiveness has been proven on large datasets. Hence, all three approaches have been tested on our large-scale datasets and are selected as the competing methods as presented in Sec. Experimental results and discussion.

Contrastive pre-training

Most contrastive learning works have been conducted within the realm of self-supervised learning on ImageNet data [10–14], involving certain forms of contrastive loss [22]. The recent work of *SimCLR-v2* [14] shows that self-supervised contrastive pre-training can compete with its fully supervised counterpart after supervised fine-tuning on downstream tasks. By encouraging the similarity of same-class data and distancing different-class data in the feature domain, the supervised contrastive learning approaches from *SupCon* [15] achieve superior representation to that of trivial supervised learning.

Contrastive pre-training in medical imaging

Model pre-training leveraging contrastive learning has also emerged in medical imaging domain [11, 23–30]. In [29], a library of models called "Model Genesis" is obtained through self-supervised pre-training by recovering anatomical patterns from transformed medical images. The recent work of [27] proposes to carry out contrastive pre-training to assist the model in learning a better embedding space across different modalities. Contrastive learning improves representation for image classifications of chest x-rays and CT images [31, 32]. Self-supervised learning fashion-based mammography screening approaches focus on leveraging the unlabeled data [33–35], which is different from our problem setup. Medical image segmentation can also benefit from the pre-training using contrastive loss in semi-supervised settings [24, 30]. Because it is expensive to obtain a large number of labeled medical images, most of them apply contrastive learning in an unsupervised fashion. To the best of our knowledge, our work is the first to leverage supervised contrastive pre-training for mammography screening.

Methodology

SCP+SF framework

The overall architecture of the SCP+SF framework is illustrated in Fig. 2. The *Siamese contrastive learning module* is designed to carry out the SCP phase. The resulting Siamese encoders are then transferred to the *single-view learning module* and the *dual-view*

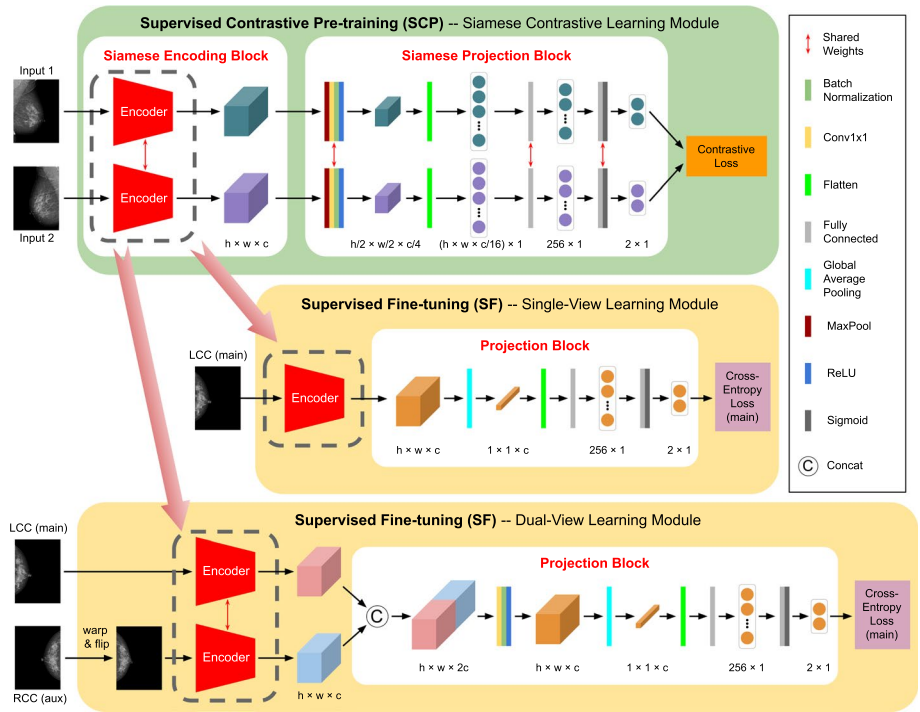

Fig. 2 The SCP+SF framework, consisting of the Siamese contrastive learning module and SF modules for both the single-view and the dual-view model

learning module to continue the SF phase, respectively, as shown by the maroon arrows in Fig. 2. We further elaborate on both phases in the following subsections.

SCP phase

The SCP phase is carried out by the Siamese contrastive learning module, which consists of a Siamese encoding block and a Siamese projection block.

Siamese contrastive learning is a technique used in computer vision to improve the feature representation of images through a unique architecture and training method. The approach utilizes a Siamese network, which consists of two identical subnetworks that share the same weights and parameters. Each subnetwork processes a different input image. The goal is to learn a feature space where similar inputs are pulled closer together while dissimilar inputs are pushed apart. The learned representations through Siamese contrastive learning can be transferred to various downstream tasks.

In the encoding block, one pair of the input mammographic images are simultaneously fed into the shared-weight encoders. The encoded features are then projected into a lower dimensional space by max-pooling and 1×1 conv operations before being flattened into two 1-dimensional vectors. The 1-D vectors are further reduced to 2×1 output vectors through fully connected layers and sigmoid operation, representing the likelihood for each class.

The contrastive loss[22] is designed to draw the samples from the same class closer and separate the samples from different classes farther apart in the projected space. Given a pair of input images (I, I'), we use the regular L2 distance in the loss function and set *margin* as $\sqrt{2}$:

$$L(I, I') = \begin{cases} D^2 & \text{if } l_I = l_{I'} \\ \max\left(0, margin - D\right)^2 & \text{if } l_I \neq l_{I'}, \end{cases} \tag{1}$$

where

$$D = \left\| \mathcal{P}_{\text{sia}}(\mathcal{E}_{\text{sia}}(I)) - \mathcal{P}_{\text{sia}}(\mathcal{E}_{\text{sia}}(I')) \right\|_{L2}, \tag{2}$$

and $\mathcal{E}_{\text{sia}}(\cdot)$ and $\mathcal{P}_{\text{sia}}(\cdot)$ denote the Siamese encoder block and nonlinear projection block, respectively; l_I and $l_{I'}$ indicate the corresponding BI-RADS labels. The loss for a batch of N image pairs can be simply defined as $\mathcal{L}_{\text{batch}} = \sum_{i=1}^{N} L(I_i, I_i')$.

Other types of the loss function (e.g., inner product based) as in [13, 15] are also experimented with and showed no significant difference. The SCP phase is completed once the training of this module ends.

SF phase

Each mammogram typically includes four views, the left and right craniocaudal (LCC/RCC), and mediolateral-oblique (LMLO/RMLO), and the triage screening model can take one or multiple views as input.

Single-view model

The *Single-view learning module* in Fig. 2 illustrates the network architecture of the single-view model. During the SF phase, its encoder block is directly transferred from the Siamese encoding block trained in the SCP phase and kept intact, while the projection block after the encoder is fine-tuned based on the regular cross-entropy loss.

Dual-view model

In practice, radiologists routinely identify the abnormalities through bilateral analysis of mammography image pairs (i.e., LCC/RCC, or LMLO/RMLO). Therefore, we also experimented with the bilateral views as the input for a dual-view model in addition to the single-view model.

The *Dual-view learning module* in the SF phase comprises a dual-view-based input structure, a Siamese encoder, and a projection block, as shown in Fig. 2. Since our screening model output is for each image, we designate one image of the bilateral pair as the main input, and the other image serves as the auxiliary input. For the example shown in Fig. 2, the LCC view is the main input, and the RCC view from the same patient serves as the auxiliary input. The RCC input will first be registered and warped according to the LCC view before being fed into the shared-weight pre-trained encoder in tandem with the LCC view. The output encoded features are then concatenated before being projected into a lower dimension and further reduced to a 2×1 vector. Similar to the single-view model, the encoder block of the dual-view model is directly transferred from the SCP phase and fixed during the SF phase.

Sample selection strategy

During supervised contrastive learning, a batch of images is first randomly selected from the training set. Then, the positive and negative pairs are identified according to the sample labels within this selected batch [15]. Limited by the affordable batch size,

we experimented with two slightly different sampling strategies. One method is random sampling, where the training batch includes N pairs of positive and negative image pairs directly sampled from *the entire training set* with the corresponding labels. Since our dual-view model takes input from a pair of images from the same patient, we also experimented with a patient-constrained sampling method, where each randomly sampled positive or negative pair must come from *the same patient*. The training label comes from the main input as in Fig. 2. The auxiliary input will be another image from this patient.

Experiment design

In this section, we first provide preliminary knowledge about BI-RADS assessment categories. We then present the details of our dataset and describe our two clinically meaningful screening tasks. The training parameter settings and evaluation metrics are also provided.

BI-RADS assessment categories

BI-RADS is a widely adopted standard for risk assessment of breast cancer using unified terminology and reporting schema [36], applicable to mammography and other breast imaging modalities. It has seven numeric categories: 0 – Incomplete (need additional imaging evaluation); 1 – Normal; 2 – Benign; 3 – Probably benign; 4 – Suspicious for malignancy; 5 – Highly suggestive of malignancy; 6 – Known biopsy-proven malignancy. BI-RADS 0 is a special category, indicating that the subject needs to have further imaging evaluation, such as additional mammographic views or other breast imaging modalities. BI-RADS 0 is often regarded as between BI-RADS 3 and 4 by radiologists [37, 38]. In the literature [3, 21, 39], mammographic screening models have been developed and tested only with large-scale private datasets conforming to the realistic patient distribution. Since no similar public dataset is available, we follow the convention and perform experiments on our own large-scale datasets. BI-RADS categories can be assigned to a lesion, an image, a breast, or a patient. Our screening tasks use the image-level BI-RADS annotation.

Datasets

Our data for training and validation is collected from three collaborative hospitals at distinct geographical locations using three different vendors' equipment and dated from 2011 to 2018. Screening equipment is MAMMOMAT Fusion from Siemens Healthineers, Senographe Crystal from GE Healthcare, and Hologic. For this study, we collect Full-Field Digital Mammography (FFDM) only, as FFDM reduces overlapping tissue issues that can obscure small lesions and improves detection rates in dense breasts. All the data are initially stored in hospitals in PACS (Picture Archiving and Communication Systems). They have pixel values of 500–2000 HU. The original captured images are in DICOM format. To process DICOM images, we first apply Pydicom to extract its metadata and pixel data. Then, with normalization, windowing, and leveling, the image data turns into Numpy arrays that the deep learning models can process. Most of the dimensions are 3328 x 4096 pixels with very few exceptions, on which we use zero-padding to

resize into the same dimension. We downsample them three times before sending them into models.

In our study, the dataset covers 30,487 patients, of which 13,931 patients have at least one side breast diagnosed as abnormal (other than BI-RADS 1), and 16,556 patients' both breasts are diagnosed as normal/healthy (BI-RADS 1). Our 640-patient test set is collected within 31 consecutive days (March 2019) from one of those three hospitals, in which 405 patients have at least one side breast diagnosed as abnormal (other than BI-RADS 1), and 235 patients' both breasts are diagnosed as normal/healthy (BI-RADS 1). Mammograms in the test set come with biopsy-proven malignancy results. The image-level details of our datasets are listed in Table 1. For abnormal cases, the BI-RADS categories are listed by increasing risk level, where BI-RADS 0 is often regarded as between BI-RADS 3 and 4 by radiologists [37, 38].

Public datasets like DDSM [40] have BI-RADS information on patients' data. However, BI-RADS 1 data only accounts for less than 30% of the complete dataset, while it should be over 70% in real clinical cases. Thus, the dataset's distribution is not consistent with screening mammography scenarios. Other public mammography datasets either contain no BI-RADS 1 patients' data or only have diagnostic mammograms (when a screening mammogram does show an abnormality, a diagnostic mammogram may be needed). Besides, in the literature [3, 21, 39], screening models have also been developed and tested only with large-scale private datasets conforming to the realistic patient distribution. Since no similar public dataset is available, we follow the convention and perform experiments on our own large-scale datasets.

Our entire data collection has been approved by the ethics committee with the IRB-number LL-XJS-2020011.

Experimental settings

The dataset is split into the training and validation sets by an 8:1 ratio. All input images are resized to 1008×800 and retain the original aspect ratio. All implementations use Python 3.11 and Pytorch 2.0.0. The SCP and SF phases share the following training parameter settings. The initial learning rate is 1×10^{-5} with 4 warming-up steps and reduced to 1×10^{-6} after 100 epochs. Adam is used [41], with a weight decay of 5×10^{-4}. Two NVIDIA A100 GPUs (40 G memory each) are used, and the batch size for contrastive learning is set to 12 due to the computation limit. The model training normally completes within 300 epochs. Our final model's runtime is less than 3 s on our machine.

Evaluation metric

Since our goal is to screen out a portion of mammograms confidently, for Task 1 as in Sec. Task 1: mammographic abnormality screening, BI-RADS 1 mammograms are considered as *positives*, and other BI-RADS images are *negative* samples. While for Task 2, as in Sec. Task 2: mammographic malignancy screening, non-malignant mammograms are considered *positives*.

We set the sensitivity (recall rate of normal/non-malignant images) at 20%, which is commonly used in clinical studies for mammogram triage screening [42], and compare the specificity rate (percentage of correctly classified abnormal images) of different

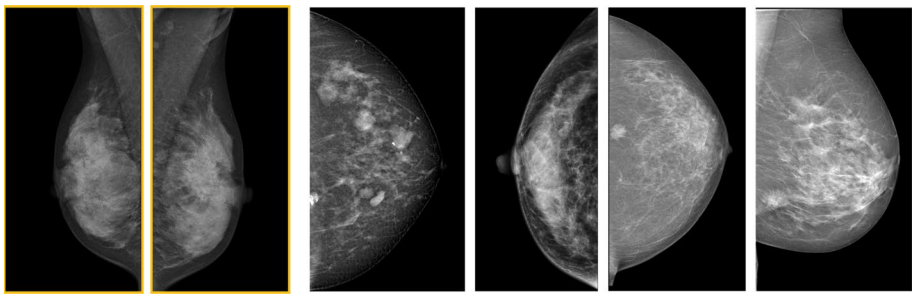

Fig. 3 Examples of mammograms in our dataset. The left two are normal cases and the right four are all abnormal mammograms, with mass, architectural distortion, and calcification in them

approaches. In addition, AUC is used to compare the overall performance of the classification models. We then investigate how many abnormal images are incorrectly screened out as false positives. We use specificity (*i.e.*, true negative rate) as the metric to indicate how many abnormal images are correctly retained. Table 2 has illustrated the above details.

Task 1: mammographic abnormality screening

This triage screening task aims to identify normal mammograms (BI-RADS 1) with near-perfect accuracy in physical screening scenarios. Patients with any suspicious regions in the breasts should not be screened out as normal patients. This task can be further defined as a binary classification problem of BI-RADS 1 (normal/healthy) vs. other BI-RADS categories (abnormal). The majority of screening mammograms belong to BI-RADS 1; thus, screening them off can assist radiologists in reducing their workload. Our collaborative radiologists recommend this screening task because safely screening out a portion of BI-RADS 1 mammograms while retaining almost all the mammograms that are in the abnormal BI-RADS categories has huge potential to reduce the radiologists' workload. Figure 3 shows examples of mammograms from normal and abnormal categories.

Task 2: mammographic malignancy screening

A malignancy screening aims to identify mammograms with malignant findings with high specificity so that we can confidently screen out some non-malignant patients. The malignancy screening has been well studied in the literature. In order to properly compare the performance of some recent methods with ours, we reformat our datasets into malignant and non-malignant portions. Since there is no hard boundary to divide the BI-RADS categories regarding malignancy, biopsy is regarded as the gold standard of malignancy determination. However, a biopsy is not applied to every patient, depending on the follow-up clinical recommendation after the screening. We simplify this task by gathering BI-RADS 1∼3 into the non-malignant class and BI-RADS 4∼6 into the malignant class. BI-RADS 0 is excluded due to its uncertainty of malignancy. Several recently representative methods [3, 6, 20], which have been tested on their own large-scale datasets, are selected and re-implemented. In particular, Yala *et al.* [6] proposed a classic multi-view-based method; Shen *et al.* [4] and Truong *et al.*

[5] represented the latest SOTA multi-view screening methods with mammography. All three approaches have been tested on large-scale datasets. We compare their performance with our proposed method in Sec. Task 2: mammographic malignancy screening task. The training parameter settings are as in Sec. Experimental settings.

Experimental results and discussion

Task 1: Mammographic abnormality screening task

We first present our experimental results of the mammographic abnormality screening task (BI-RADS 1 vs. other BI-RADS categories) on the test set using methods listed in Table 2 according to the evaluation metrics described in Sec. Evaluation metric. Table 2 shows the performance comparison of the triage screening task with different approaches, where Yala et al. [6], Shen et al. [4] and Truong et al. [5] are all previous SOTA models, alongside our proposed single-view and dual-view models with different SCP+SF training strategies. We also scrutinize the number of abnormal images that are incorrectly screened out from the 1192 abnormal images in the test set and further break it down according to the BI-RADS level (in the order of 2, 3, 0, 4, 5, 6). Therefore, a non-zero number appearing at a later position in the last column of Table 2 has a more harmful impact on performance. To demonstrate the efficacy of our design, we next interpret Table 2 from three aspects as follows.

Comparisions with previous SOTA methods

We re-implemented the previous SOTA approaches for this comparison. Our single-view model with SCP+SF and random sample selection generates the best single-view performance. Our dual-view model trained with SCP+SF and patient-constrained sample selection (Row 6) generates the best result overall, outperforming Shen [4] and Truong [5], which are both previous SOTA screening models. At higher sensitivity, such as 0.8, our SCP+SF with patient-constrained sample selection method improves the specificity from 0.806 to 0.888 compared to our vanilla dual view model. We have also included three SOTA models for natural image classification for comparision in Table 2, ViT [43], Swin-Transformer [44], and ConvNexT-Base [45].

Effectiveness of SCP+SF

As shown in Table 2, for both single-view and dual-view models, the SCP+SF framework effectively improves the overall performance of the models, including the AUC and the specificity at given sensitivity (20%). In turn, the number of total incorrectly screened abnormal images is reduced. In addition, for the dual-view model, the SCP+SF framework can completely remove the error made for BI-RADS 4,5,6 images, which is critical in practice since those images often correspond to higher cancer risk. Such results would enable a confident deployment of such AI system as it confidently screens out no-risk cases. In comparison, existing methods [4–6] still have misclassified BI-RADS 4,5,6 cases in Table 2, which, in physical examination scenarios, would be risky. For the single-view model, the SCP+SF framework can also reduce the error for BI-RADS 5,6 images to near zero. We further confirm from the separate

biopsy reports that the incorrectly screened images from the dual-view methods with SCP+SF do not include any malignant findings.

Table 4 shows the ablation study results of the backbone in SCP. We compare a set of ResNets [46] and ConvNexTs [45] an see no significant difference between them. To run the system efficiently, ConvNext-Small has been selected to serve our encoder.

To evaluate the impact of data augmentation, we train our state-of-the-art (SOTA) model from Table 2 using standard data augmentation techniques, including flipping, rotation, and Gaussian noise addition. Despite employing the same training recipes, the AUC and specificity on the test set are 0.8806 and 0.9801, respectively, both falling short of the current SOTA results.

Effectiveness of patient-constrained training strategy

To compare our proposed patient-constrained training strategy, we develop another training approach that selects two *random* images from the entire dataset (Row 5 and 8). Patient-constrained sampling improves the dual-view model over the random sampling method, while random sampling is slightly better than the patient-constrained sampling for the single-view model, and both are consistent with our expectations.

Mammograms from the same patients have many similarities in terms of texture, density, etc. Our patient-constrained training strategy in the SCP phase enables the encoder to learn the patient-level similarities and dissimilarities of the same patient. This strategy is able to narrow down the solution space when a bilateral dual-view input structure is given in the SF Phase. However, such characteristics don't exist for single-view models. As a result, this strategy only improves the performance of dual-view-based models in Table 2.

Task 2: mammographic malignancy screening task

The performance comparison among various methods is shown in Table 3, where the AUC values are for the non-malignant class, and a sensitivity of 20% means that 20% non-malignant mammograms are confidently screened out from all non-malignant mammograms. In our dual-view-based methods, the effectiveness of the SCP phase is again demonstrated, and the efficacy of the patient-constrained training strategy is also validated over the random-selection-based one. Even though the SOTA multi-view-based method (Row 3) is slightly better than our dual-view-based SF phase without SCP (Row 4), our complete method of dual-view-based SF with patient-constrained SCP (Row 6) surpasses all previous SOTA methods by reporting the best AUC value of 0.9270. The specificity value of 100% means that there isn't any malignant image screened out as a false positive non-malignant one. We thus confidently state that our method achieves the SOTA performance on the mammographic malignancy screening task. Figure 5 shows that the AUC values are in line with the training epochs on our validation set, using our best method (Row 6 in Table 3). We see that the AUC value is rising up significantly while the training epoch is increasing. Similar to the previous section, we have also experimented with three SOTA image classification approaches, ViT [43], Swin-Transformer [44], and ConvNexT-Base [45]. However, the results are far from competitive, with their AUCs reaching 0.8241, 0.8298, and 0.8369.

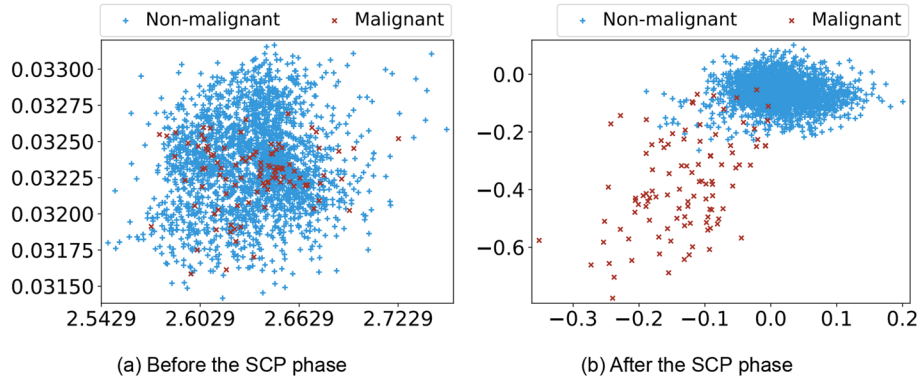

Fig. 4 Visualization of the malignant and non-malignant outputs of the Siamese contrastive learning module on the test set data before and after the SCP phase by t-SNE. (**a**) Distribution before the SCP phase. (**b**) Distribution after the SCP phase. x,y axes indicating the projection plane before the Sigmoid function

Table 1 Number of mammograms in each BI-RADS category

Type	BI-RADS	Dataset	
		Train & Val	Test
Abnormal	2	35290	728
	3	14508	226
	0	1570	136
	4	3288	88
	5	732	6
	6	174	8
	Total	55562	1192
Normal (Healthy)	1	78926	1346
Total		134488	2538

Figure 4 also shows t-SNE visualization of the malignant and non-malignant outputs of the Siamese contrastive learning module on the test set data before and after the SCP phase. We observe that the SCP phase significantly pulls the diffuse malignant samples out from the non-malignant samples and effectively forms two clusters in an embedding space. The effectiveness of SCP is again demonstrated.

Discussion of the results

Technical aspect

All prior methods for building reliable AI systems in medical domains face the common challenge of limited labeled data. These methods typically rely on direct supervised training, which heavily depends on hand-crafted data augmentation techniques to compensate for the scarcity of labeled examples. While contrastive pre-training with large amounts of unlabeled data has been successfully adopted in other fields to mitigate data shortages-owing to the ease of acquiring such data-this approach is impractical in mammography screening due to the difficulty of collecting extensive datasets for unsupervised contrastive pre-training.

To address these limitations, our method introduces Supervised Contrastive Pre-training (SCP), leveraging limited labeled data without requiring additional data

Table 2 Performance comparison of different approaches on the abnormality triage screening task (task 1). *SCP+SF(R)* refers to SCP with random sample selection and *SCP+SF(PC)* indicates SCP with patient-constrained sample selection as in Sec. Sample selection strategy.The last column shows the number of incorrectly screened abnormal images with a breakdown according to their BI-RADS levels (in the order of 2, 3, 0, 4, 5, 6). The 95% confidence intervals (CI) are shown in the square brackets. Bold values represent the achieved best results

Method	**AUC**	**Sensitivity = 20%**	
		Specificity	**# of *incorrectly* screened out images out of 1192 abnormal images**
Yala *et al.* [6] 2021)	0.8452 [0.8432, 0.8476]	0.9748 [0.9712, 0.9778]	30 (10, 8, 9, 2, 0, 1)
Shen *et al.* [4] (2021)	0.8602 [0.8529, 0.8681]	0.9765 [0.9737, 0.9793]	28 (8,11, 8, 1, 0, 0)
Truong *et al.* [5] (2023)	0.8733 [0.8707, 0.8765]	0.9790 [0.9766, 0.9814]	25 (8, 7, 9, 1, 0, 0)
ViT [43] (2020)	0.8555 [0.8551, 0.859]	0.9706 [0.9701, 0.9709]	35 (11, 12, 9, 3, 0, 0)
Swin-T [44] (2022)	0.8688 [0.8685, 0.8691]	0.9732 [0.9725, 0.9739]	32 (10, 10, 9, 3, 0, 0)
ConvNexT [45] (2022)	0.8342 [0.8328, 0.8356]	0.9673 [0.9665, 0.9680]	39 (13, 13, 7, 5, 1, 0)
Single-view	0.8348 [0.8327, 0.8372]	0.9713 [0.9688, 0.9739]	34 (7, 9, 14, 2, 1, 1)
Single-view SCP+SF (R)	0.8512 [0.8486, 0.8544]	0.9740 [0.9736, 0.9754]	31 (8, 9,10, 2, 0, 0)
Single-view SCP+SF (PC)	0.8441 [0.8380, 0.8502]	0.9757 [0.9721, 0.9793]	29 (10, 6,11, 1, 0, 1)
Dual-view	0.8566 [0.8540, 0.8592]	0.9782 [0.9764, 0.9800]	26 (7, 8, 10, 1, 0, 0)
Dual-view SCP+SF (R)	0.8888 [0.8829, 0.8947]	0.9815 [0.9796, 0.9834]	22 (7, 5, 10, 0, 0, 0)
Dual-view SCP+SF (PC)	**0.9046 [0.9007, 0.9085]**	**0.9841 [0.9825, 0.9859]**	**19 (7, 5, 7, 0, 0, 0)**

Table 3 Results of the mammographic malignancy screening task (task 2). The 95% confidence intervals (CI) are shown in the square brackets. Bold values represent the achieved best results

Method	**AUC**	**Spe. (Sen. = 20%)**
Yala *et al.* [6] (2021)	0.8671 [0.8667, 0.8671]	0.9808 [0.9804, 0.9811]
Shen *et al.* [4] (2021)	0.8946 [0.8940, 0.8952]	0.9902 [0.9899, 0.9905]
Trong *et al.* [5] (2023)	0.9071 [0.9069, 0.9074]	0.9902 [0.9901, 0.9903]
Dual-view	0.8929 [0.8922, 0.8936]	0.9804 [0.9797, 0.9801]
Dual-view SCP+SF (random)	0.9070 [0.9065, 0.9075]	**1.0 [1.0, 1.0]**
Dual-view SCP+SF (patient-constrained)	**0.9270 [0.9264, 0.9276]**	**1.0 [1.0, 1.0]**

Table 4 Ablation study on different encoders in SCP. Bold values represent the achieved best results

Encoder	**AUC**	**Specificity (Sen.=20%)**
ResNet-22	0.9040	98.32%
ResNet-34	0.9013	98.15%
ResNet-50	0.8993	98.15%
ConvNexT-Tiny	0.9025	98.29%
ConvNexT-Small	**0.9046**	**98.41%**
ConvNexT-Base	0.8975	97.99%

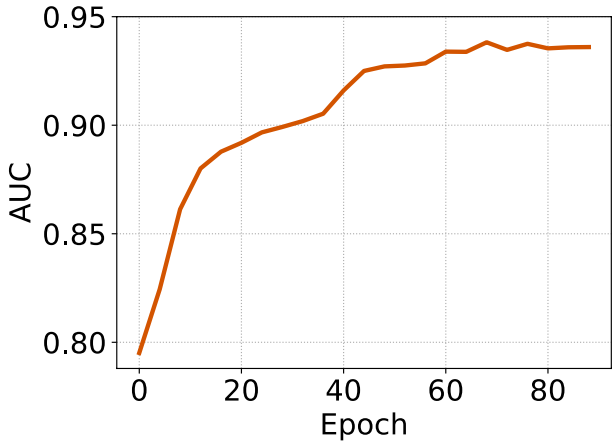

Fig. 5 AUC vs. training epochs for the malignancy screening task

augmentation. This innovative design not only harnesses the benefits of contrastive pre-training to enhance model generalizability but also eliminates the need for handcrafted data augmentation strategies. Subsequent Supervised Fine-tuning (SF) further refines the model's representation capabilities, leading to superior performance. Table 2 and 3 highlight the state-of-the-art (SOTA) results achieved by our SCP+SF framework, beating existing approaches. Moreover, our dual-view input paradigm capitalizes on the inherent symmetry of mammograms, improving model performance. The patient-constrained sample selection strategy effectively narrows the solution space, complementing the dual-view approach. Both strategies have been validated through ablation studies, as shown in Table 2.

Clinical aspect

Previous methods in mammography screening have primarily focused on identifying abnormal or malignant cases. However, due to limitations in achieving near-perfect sensitivity and specificity, these methods often require significant human intervention, making their deployment in real clinical settings challenging. In contrast, our system targets a different aspect of clinical need. In physical examination scenarios, the majority of screening mammograms fall under BI-RADS category 1, indicating no apparent abnormalities. Our AI system aims to confidently recognize a portion of healthy (BI-RADS 1) cases while ensuring that nearly all abnormal mammograms are retained for further review. Specifically, our experimental results demonstrate that when screening out 20% of normal cases, no instances from BI-RADS categories 4, 5, or 6 were misclassified by our AI system. Consequently, this AI system can be effectively deployed to assist radiologists in reducing their workload by confidently screening out certain BI-RADS 1 cases. By identifying abnormalities and malignancies that human radiologists may overlook, AI can contribute to earlier diagnosis, which is crucial for effective treatment and better outcomes. With improved accuracy and less overdiagnosis, patients may experience less anxiety and uncertainty while awaiting results. This capability significantly reduces the workload for radiologists by allowing them to focus on more complex cases and enhance their efficiency without compromising diagnostic accuracy.

Limitations and future prospects

There are still a few limitations of this work that need improving in the future. First, it has yet to be tested on an open-world dataset to prove its universal practicability. The lack of diversity could lead to dataset bias. Hence, we are working on agreements with other medical centers in Hong Kong and Macau that contain mammograms screened by vendors different from our current ones. By doing so, we will enlarge our dataset to foster its generality. Second, the runtime needs to accelerate through better parallel processing. Embedding ZeRO [47] is one solution we believe could benefit the overall runtime, and we will work on one next up. Besides, a great portion of incorrect predictions are caused by occlusion at different viewing angles. We believe including modalities other than mammography, like hispathological data and MRI, can not only alleviate the impact of this problem but also has high clinical value, as they all count as clinically meaningful approaches to screening breast cancer. Lastly, inspired by the recent success of large foundation models, we will work to build them into our current pipeline.

Conclusion

Mammography screening is the most effective and widely used approach for early breast cancer detection. Due to the heavy workload of radiologists, an AI system serving as their daily assistant would greatly ease their pressure. This paper presents a novel framework of Supervised Contrastive Pre-training followed by Supervised Fine-tuning (SCP+SF) for two critical mammography screening tasks. Different from the previous supervised learning methods, the proposed framework leverages the limited medical data well without any data augmentation operation. Our extensive experiments demonstrate that the SCP+SF framework can substantially improve the final model performance. When compared with previous approaches, the proposed framework achieves SOTA performance on both mammography screening tasks.

Such an AI system has great potential to boost medical service efficiency and eventually improve the breast cancer survival rate. The AI system can minimize errors due to fatigue or cognitive overload experienced by radiologists, particularly during high-volume screening. In real medical and clinical scenarios, especially in physical examination departments or physical examination centers, radiologists can quickly filter out confidently normal or malignant cases on the PACS (Picture Archiving and Communication Systems) device deployed with this AI software. This can significantly reduce the workload of radiologists in the remaining physical examination scenarios, freeing up their energy and time to focus more on patients with more ambiguous and controversial conditions.

Acknowledgements

This work is supported by Ping An Technology (Shenzhen) Co., Ltd. and PAII Inc.

Author contributions

Zhenjie Cao is the primary author of this work and contributes to the core novelties and design of this work. Zhuo Deng is responsible for the design of the experiment and ablation studies. Zhicheng Yang assists in conducting the experiments with workstations from PAII Inc. Jie Ma is the director of the Radiology Department of our collaborative hospital and advises the whole data collection and clinical applicability and value. Lan Ma is the supervisor of Zhenjie Cao and instructs on the whole project.

Funding

This work is funded by Shenzhen Science and Technology Innovation Bureau under GJHZ20220913142613025.

Availability of data and materials
The data that support the findings of this study are available from our collaborating hospitals, but restrictions apply to the availability of these data, which were used under license for the current study and so are not publicly available. We will continue to work with our partners to meet certain criteria before releasing these data. No datasets were generated or analysed during the current study.

Code availability
As declared in previous sections, this work is supported by two institutes. The code, considered intellectual property by both institutes, has been put on certain restrictions before being made public. Next, we will apply for patent protection before releasing them.

Declarations

Ethics approval and consent to participate
This project is approved by the IRB number LL-XJS-2020011.

Competing interests
The authors declare that they have no Competing interests

Received: 12 June 2024 Accepted: 17 January 2025
Published online: 05 February 2025

References

1. Organization WH. Breast Cancer. http://www.who.int/news-room/fact-sheets/detail/breast-cancer Accessed 15 July 2023.
2. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DS, Kerlikowske K, Henderson LM, Onega T, Tosteson AN, Rauscher GH, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. Radiology. 2017;283(1):49–58.
3. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GC, Darzi A, et al. International evaluation of an ai system for breast cancer screening. Nature. 2020;577(7788):89–94.
4. Shen Y, Wu N, Phang J, Park J, Liu K, Tyagi S, Heacock L, Kim SG, Moy L, Cho K, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Med Image Anal. 2021;68: 101908.
5. Truong Vu YN, Guo D, Taha A, Su J, Matthews TP. M &m: Tackling false positives in mammography with a multi-view and multi-instance learning sparse detector. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. 2023; pp. 778–788
6. Yala A, Mikhael PG, Strand F, Lin G, Smith K, Wan Y-L, Lamb L, Hughes K, Lehman C, Barzilay R. Toward robust mammography-based models for breast cancer risk. Sci Transl Med. 2021;13(578):4373.
7. Cea MVS, Diedrich K, Bakalo R, Ness L, Richmond D. Multi-task learning for detection and classification of cancer in screening mammography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer; pp. 241–250.
8. Lehman CD. Artificial intelligence to support independent assessment of screening mammograms-the time has come. JAMA Oncol. 2020;6(10):1588–9.
9. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, Clauser P, Helbich TH, Chevalier M, Mertelmeier T. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol. 2019;29(9):4825–32.
10. Lee H, Seol J, Lee S-G, Park J, Shim J. Contrastive learning for unsupervised image-to-image translation. Appl Soft Comput. 2024;151: 111170.
11. Wu G, Jiang J, Liu X. A practical contrastive learning framework for single-image super-resolution. IEEE Transactions on Neural Networks and Learning Systems. 2023.
12. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; pp. 9729–9738.
13. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, PMLR. 2020; pp. 1597–1607.
14. Chen T, Kornblith S, Swersky K, Norouzi M, Hinton GE. Big self-supervised models are strong semi-supervised learners. Conf Neural Inf Proc Syst. 2020;33:22243–55.
15. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D. Supervised contrastive learning. Adv Neural Inf Process Syst. 2020;33:18661–73.
16. Lvd Maaten, Hinton G. Visualizing data using t-sne. J Mach Learn Res. 2008;9:2579–605.
17. Beuque MP, Lobbes MB, Wijk Y, Widaatalla Y, Primakov S, Majer M, Balleyguier C, Woodruff HC, Lambin P. Combining deep learning and handcrafted radiomics for classification of suspicious lesions on contrast-enhanced mammograms. Radiology. 2023;307(5): 221843.
18. Trieu PD, Lewis SJ, Li T, Ho K, Wong DJ, Tran OT, Puslednik L, Black D, Brennan PC. Improving radiologist's ability in identifying particular abnormal lesions on mammograms through training test set with immediate feedback. Sci Rep. 2021;11(1):9899.
19. Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. Eur Radiol. 2020;31:1–6.

20. Taha A, Truong Vu YN, Mombourquette B, Matthews TP, Su J, Singh S. Deep is a luxury we don't have. In: International conference on medical image computing and computer-assisted intervention, Springer, 2022; pp. 25–35.
21. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, Jastrzębski S, Févry T, Katsnelson J, Kim E, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging. 2019;39(4):1184–94.
22. Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE. 2006; vol. **2**, pp. 1735–1742.
23. Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, Loh A, Karthikesalingam A, Kornblith S, Chen T, et al. Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp. 3478–3488.
24. Li H, Yang X, Liang J, Shi W, Chen C, Dou H, Li R, Gao R, Zhou G, Fang J. Contrastive Rendering for Ultrasound Image Segmentation. In: International conference on medical image computing and computer-assisted intervention, 2020; pp. 563–572.
25. Luo Y, Liu W, Fang T, Song Q, Min X, Wang M, Li A. Carl: Cross-aligned representation learning for multi-view lung cancer histology classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023; pp. 358–367.
26. Jin Q, Zou C, Cui H, Sun C, Huang S-W, Kuo Y-J, Xuan P, Cao L, Su R, Wei L, et al. Multi-modality contrastive learning for sarcopenia screening from hip x-rays and clinical information. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023; pp. 85–94.
27. Zhong Y, Xu M, Liang K, Chen K, Wu M. Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023; pp. 724–733.
28. Basak H, Yin Z. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023; pp. 19786–19797.
29. Zhou Z, Sodha V, Siddiquee MMR, Feng R, Tajbakhsh N, Gotway MB, Liang J. Models genesis: Generic autodidactic models for 3d medical image analysis. In: International conference on medical image computing and computer-assisted intervention, 2019; pp. 384–393.
30. Karani KCEEN, Konukoglu E. Contrastive learning of global and local features for medical image segmentation with limited annotations. In: Advances in Neural Information Processing Systems, 2020; 33.
31. Zhang J, Xie Y, Li Y, Shen C, Xia Y. Covid-19 screening on chest x-ray images using deep learning based anomaly detection. 2020. arXiv preprint arXiv:2003.12338.
32. Wang Z, Liu Q, Dou Q. Contrastive cross-site learning with redesigned net for COVID-19 CT classification. IEEE J Biomed Health Inform. 2020;24(10):2806–13.
33. Azizi S, Culp L, Freyberg J, Mustafa B, Baur S, Kornblith S, Chen T, MacWilliams P, Mahdavi SS, Wulczyn E, et al. Robust and efficient medical imaging with self-supervision. 2022. arXiv preprint arXiv:2205.09723.
34. Wantlin K, Wu C, Huang S-C, Banerjee O, Dadabhoy F, Mehta VV, Han RW, Cao F, Narayan RR, Colak E, et al. Benchmd: A benchmark for modality-agnostic learning on medical images and sensors. 2023. arXiv preprint arXiv:2304.08486.
35. Miller JD, Arasu VA, Pu AX, Margolies LR, Sieh W, Shen L. Self-supervised deep learning to enhance breast cancer detection on screening mammography. 2022. arXiv preprint arXiv:2203.08812.
36. Sickles EA. Acr bi-rads® atlas, breast imaging reporting and data system. American College of Radiology. 2013, 39.
37. Nelson HD, O'Meara ES, Kerlikowske K, Balch S, Miglioretti D. Factors associated with rates of false-positive and false-negative results from digital mammography screening: an analysis of registry data. Ann Intern Med. 2016;164(4):226–35.
38. Castells X, Torá-Rocamora I, Posso M, Román M, Vernet-Tomas M, Rodríguez-Arana A, Domingo L, Vidal C, Baré M, Ferrer J, et al. Risk of breast cancer in women with false-positive results according to mammographic features. Radiology. 2016;280(2):379–86.
39. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. Radiology. 2019;293(1):38–46.
40. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data. 2017;4: 170177.
41. Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014. arXiv preprint arXiv:1412.6980.
42. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 2019;292(1):60–6.
43. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. arXiv preprint arXiv:2010.11929.
44. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; pp. 10012–10022.
45. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022; pp. 11976–11986.
46. 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, 2016; pp. 770–778.
47. Rajbhandari S, Rasley J, Ruwase O, He Y. Zero: Memory optimizations toward training trillion parameter models. In: SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE. 2020; pp. 1–16.

Publisher's Note