



Deep learning on digital mammography for expert-level diagnosis accuracy in breast cancer detection

Jinrong Qu¹ · Xuran Zhao^{2,3} · Peng Chen² · Zhaoqi Wang¹ · Zhenzhen Liu⁴ · Bailin Yang² · Hailiang Li¹

Received: 13 December 2020 / Accepted: 9 June 2021 / Published online: 28 June 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Recently, computer-aided diagnosis (CAD) systems powered by deep learning (DL) algorithms have shown excellent performance in the evaluation of digital mammography for breast cancer diagnosis. However, such systems typically require pixel-level annotations by expert radiologists which is prohibitively time-consuming and expensive. Medical institutes would wonder if a high-performance breast cancer CAD system can be trained by exploring their own huge amount of historical imaging data and corresponding diagnosis reports, without additional annotations workload of their radiologists. In this study, we show that a DL classification model trained on historical mammograms with only image-level pathology labels (which can be automatically extracted from medical reports) can achieve surprisingly good diagnostic performance on newly incoming exams compared with experienced radiologists. A DL model called DenseNet was trained and cross-validated with 5979 historical exams acquired before September 2017 with biopsy-verified pathology and tested with 1194 newly obtained cases after that. For both cross-validation and test sets, the ROCs generated by DL predictions were above the ROCs generated by ratings from radiologists. For the suspicious cases which radiologists suggest biopsy (BI-RADS category 4 and 5), the DL model can reject 60% of false biopsies on benign breasts while keeping 95% sensitivity. For the mammograms based on which radiologists were not able to make a diagnosis (BI-RADS 0), the DL model still achieved an AUC score of 79%. Moreover, the model is able to localize lesions on mammograms although such information was not provided in the training phase. Finally, the impact of input image resolution and different DL model architectures on the diagnostic accuracy were also presented and analyzed.

Keywords Deep learning · Digital mammography · Breast cancer · Diagnostic accuracy · Artificial intelligence

1 Introduction

Breast cancer is the most frequently diagnosed cancer and accounts for 23% of all cancer cases and 14% of cancer-related deaths among women in the world [1]. Digital mammography, a 2-D X-ray examination of the breast, is the most commonly used method for breast cancer screening and diagnosis and can effectively reduce breast cancer mortality [2]. A typical mammography exam is composed of four images: for each breast, two images are taken from two different projection angles, namely Cranial–Caudal (CC) taken from above and mediolateral-oblique (MLO) taken horizontally. A radiologist interprets the four images based on the detection and classification of breast lesions such as breast masses and micro-calcifications, and gives a BI-RADS score to each breast as a risk assessment of breast cancer. Patients with BI-RADS categories 4 and 5 will be suggested to receive a biopsy operation, and the presence of

Jinrong Qu, Xuran Zhao, and Peng Chen contributed equally.

✉ Xuran Zhao
zxr@zjgsu.edu.cn

¹ Department of Radiology, Affiliated Cancer Hospital of Zhengzhou University, Henan Cancer Hospital, Zhengzhou 450008, China

² School of Computer and Information Engineering, Zhejiang Gongshang University, No. 18, Xuezheng Road, Hangzhou 310018, China

³ Shenzhen Xpectvision Technology Co., Ltd, Shenzhen, China

⁴ Department of Breast Surgery, Affiliated Cancer Hospital of Zhengzhou University, Henan Cancer Hospital, Zhengzhou 450008, China

cancer will be confirmed by the pathology report. The breast cancer diagnosis procedure is illustrated in Fig. 1. Although effective, radiologists' diagnostic accuracy to identify cancer from mammograms was only moderate. In the United States, 74.4% of patients who received biopsies were proven non-cancer, while 15% of true cancers will not be identified by radiologists [3]. The interpretation of mammograms is also subjective and the performances are highly dependent on the practicing experience of individual radiologists [4]. As a result, computer-aided diagnosis (CAD) has been introduced to provide an objective view to radiologists.

In recent years, deep learning (DL) models have achieved a notable set of accomplishments, and state-of-the-art DL models based on convolutional neural networks (CNN) have demonstrated excellent performance in various computer vision tasks. A number of deep learning models have also been developed for automatic lesion detection and classification on mammographic images, and reported performances were comparable or even surpassed that of human radiologists [5–9]. However, the superior performance of these models, which are trained and tested on a particular crafted data set, may not be achieved on other dataset acquired by different devices by different medical institutions. A recent study [10] have shown that two state-of-the-art breast cancer classification models which have reported super-expert performances on public datasets achieved only near-random classification results when evaluated on an external real-world clinical dataset, caused by the data distribution mismatch between the training and testing data. As a result, instead of deploying a DL model trained with external datasets and taking the risk of significant performance decrease, major medical institutes would wonder if a high-performance breast cancer detection model can be trained internally by exploring their own huge amount of historical imaging data and corresponding diagnosis reports. In our opinion, such model should have two desirable properties to

make real contribution to clinical practices: (1) Image level labels form model training. The model should be trained based on image-level ground truth labels which can be automatically extracted from medical report to exploit the huge sum of existing records while preventing the need for pixel-level annotation. However, most existing approaches require pixel-level annotations of abnormality regions, which will cause additional workload of radiologists; (2) Model interpretability. The model should be able to locate the abnormality region to alert radiologists, even if this information is not provided in the training phase.

The purpose of our study is to show that an expert-level breast cancer diagnosis model based on DL can be obtained by training with large-scale historical data resources of a medical institute, and to investigate its potential applications in real-world clinical applications. The methodology of the study is illustrated as a flowchart in Fig. 2. First, we constructed a large real-world clinical mammographic dataset with solid cancer/non-cancer ground truth labels. 7108 patients (more than 3000 with proven cancer) who received mammographic exams and a following biopsy operation were identified from our institutional Clinical Information System (CIS) and their corresponding mammographic images, imaging report and pathology report were extracted and anonymized. Malignant/benign labels, and well as BI-RADS scores given by radiologists are automatically extracted from textual reports. Second, the whole database is divided into a train-val set (historical data) and a test set (recent exams) according to a temporal order. Different DL classification models are trained and validated, and the best performing model, namely DenseNet-121, is selected as the final model. Third, the performance of the selected model is evaluated with different subset of the test set, simulating different clinical application settings. Finally, we present further analysis about the influences of different DL

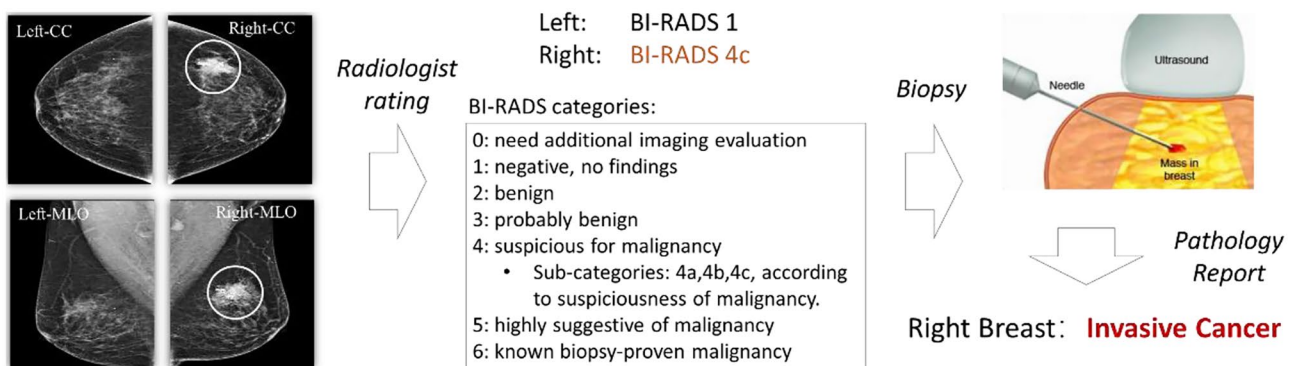
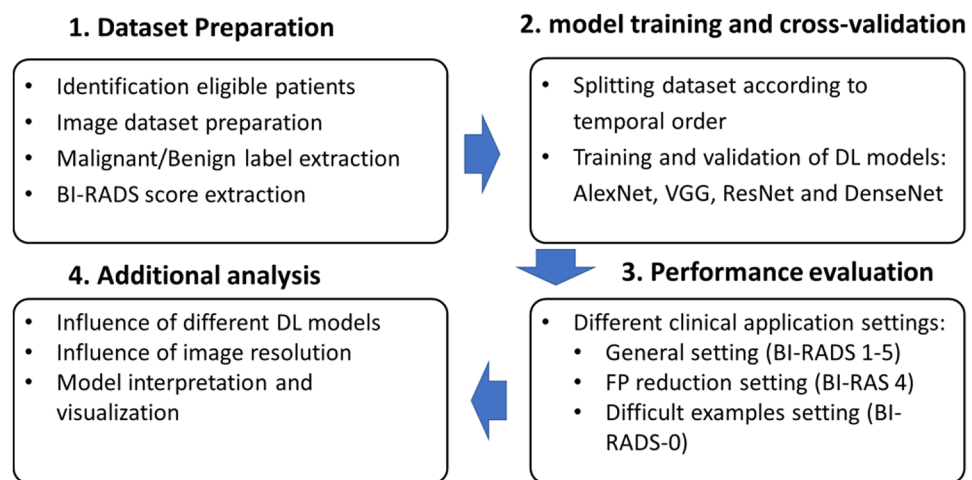


Fig. 1 An illustration of the breast cancer examination and diagnosis procedure. Left: example of a mammographic exam, a highly suspicious tumor is identified in both CC and MLO projections of the right breast; middle: BI-RADS scores extracted from the corresponding

medical report and an explanation of different BI-RADS categories. BI-RADS 4c leads to a biopsy suggestion; right: an illustration of breast biopsy which confirms the presence of cancer

Fig. 2 Methodology of the study as a flowchart



models, image resolutions, and show the capability of our model of locating cancer-related regions with neural network visualization techniques.

The main contributions of this paper are summarized as follows:

- As a leading cancer-specific hospital of the most populated province in China, we constructed a large-scale clinical digital mammographic dataset. The dataset includes more than 7000 biopsy-verified mammogram exams and more than 3000 them are cancer cases. The number of cancer cases in our dataset is among the largest reported in related literature.
- We show that a DL breast cancer classification model trained on historical mammograms with only image-level pathology label can achieve comparable (if not superior) diagnostic performance on newly incoming exams experienced radiologists.
- We show that our model can achieve surprisingly good classification performance (AUC 79%) with difficult samples which radiologists regarded as “information incomplete” (BI-RADS category 0). Which can potentially reduce the number of unnecessary additional examinations.
- We show that the DL model is able to locate lesions on mammograms, although this information was not provided during the training phase.

The rest of the article is organized as following. Section 2 includes materials and methods, which describes the details of our dataset and experimental methods. In Sect. 3, experimental results and feature visualization are presented. In Sect. 4, we discuss related work and limitations of its current implementation. Finally, a conclusion of this paper is presented in Sect. 5.

2 Materials and methods

2.1 Dataset collection

Our study was performed with anonymized, retrospectively collected digital mammographic images and their corresponding pathology report in our institute. The study was approved by the local Ethics Committee, and informed consent from patients was waived after the review of the institutional review board due to the retrospective nature. We extract from PACS of our institute all mammogram images (recorded using a Hologic Selenia digital mammography system) and the corresponding pathological records of all 7108 patients who underwent both a mammographic exam and biopsy operation from August 2014 to March 2018. Among them, 887 patients had biopsies on both breasts and other patients had biopsy on a single breast, giving a total of 7173 breast cases (14,346 mammogram images including CC and MLO views) along with their ground truth pathology labels (3938 benign and 3235 malignant). Note that our dataset only consisted of breast cases with biopsy-confirmed malignancy to make sure the ground truth malignancy was solid and non-subjective. For the comparison of radiologists’ performance against the proposed deep learning model, we also extracted Breast Imaging-Reporting and Data System (BI-RADS) categories of each breast from the mammographic reports. In all mammogram exams of our institute, the BI-RADS categories were given to each single breast by a reporting radiologist of at least 3-year mammogram reading experiences and double checked (and modified if necessary) by a chief radiologist with more than 10 years of mammogram reading experience. Finally, each breast case in our dataset was comprised of 2 mammogram images from CC and MLO

views, a ground truth pathology (malignant or benign) and a BI-RADS category in [0, 1, 2, 3, 4, 4A, 4B, 4C, 5, 6].

An independent test set containing 1194 breast cases from 1183 patients whose mammography was performed during the last 6 months (September 2017 to March 2018) of our data collection period were first split out to evaluate the DL model's ability of diagnosing new exams by learning from historical data. The rest 5979 breast cases from 5925 patients which was acquired before September 2017 were used for fivefold cross-validation, and was denoted as train-val set. This train-val set was randomly split into five subsets on patient level, and in each fold of the cross-validation, one subset was used as validation set and the 4 others were used for training. There was no overlap of patients among the training, validation and test set in each fold of cross-validation. The summary of the train-val and test dataset, including the distributions of benign and malignant cases in each BI-RADS category, was presented in Table 1. Note that a significant number of biopsies were also performed in breasts with BI-RADS 1 (no finding) or 2 (benign finding). This is due to the fact that the majority of mammographic exams performed in our institute (Henan Cancer Hospital, the leading cancer-specialized hospital in the most populated province in China) were diagnostic, and often went along with breast ultrasound and MRI. Breast lesions non-identifiable in mammograms might be detected by other imaging modalities which lead to biopsy suggestions. All cases of BI-RADS category 4 were merged to category 4B since the two categories have very similar probability of malignancy. Nevertheless, the BI-RADS in the test set may better reflect the diagnostic performance of radiologists.

2.2 Deep learning model

In this study, we used deep CNN classification models trained on ground truth pathology to classify a mammogram image into malignant or benign.

Since the success AlexNet [11] in the general image recognition task, consistent breakthroughs in CNN model development were made year after year. In the annual Large Scale Visual Recognition Challenge (ILSVRC), the top 5 classification error of the winning model each year dropped annually from 16% (AlexNet, 2012) to 2.25% (SE-ResNet, 2017), which was well below the average human error rate of 5% on the same task. It is thus meaningful to check the influence of the ever-evolving DL models on mammographic diagnostic performance. Meanwhile, mammogram image resolution is pivotal to radiologists' diagnosis, helping to better distinguish small or nuanced features of lesions potentially leading to cancer. As a result, a number of representative deep learning models including AlexNet (2012) [11], VGGNet (2014) [12], ResNet (2016) [13] and DenseNet (2017) [14] operating on different image resolutions were trained and evaluated. Finally, the configuration of DenseNet-121 model (DenseNet model with 121 convolutional layers) with an image resolution of 1024×832 achieved the best cross-validation accuracy. The main results reported in following sections were based on this configuration, while the impact of the CNN architectures and image resolution on classification performance is also reported and analyzed.

All DICOM files were first converted to PNG images and resized to 1024×832 in resolution, roughly keeping the original aspect ratio. To attenuate overfitting, data augmentation including random vertical/horizontal flipping

Table 1 Distribution overview of the train-val and test set

BI-RADS	Train-val set			Test set		
	Benign	Malignant	Sum (M%)	Benign	Malignant	Sum (M%)
0	935	329	1264 (26.0%)	187	65	252 (25.8%)
1	236	27	263 (10.3%)	37	6	43 (14.0%)
2	1052	75	1127 (6.7%)	206	18	224 (8.0%)
3	381	65	446 (14.6%)	85	14	99 (14.1%)
4	193	365	558 (65.4%)	43	70	113 (61.9%)
4A	266	71	337 (21.1%)	57	17	74 (23.0%)
4B	144	258	402 (64.2%)	28	47	75 (62.7%)
4C	69	1079	1148 (94.0%)	13	212	225 (94.2%)
5	6	361	367 (98.4%)	0	72	72 (100.0%)
6	0	67	67 (100.0%)	0	17	17 (100.0%)
Sum	3282	2697	5979 (45.1%)	656	538	1194 (45.1%)

Fivefold cross-validation was performed on the train-val set, and the resulting model performances are evaluated on the test set. The numbers in table denote breast cases, each containing 2 images of MLO and CC views, a BI-RADS category, and its ground truth pathology. The percentages in parenthesis indicate the portion of malignant cases in a specific BI-RADS category

and random rotation in $(-10,10)$ degrees were applied to each input image. Adam optimizer [15] was employed for model optimization. The CNN model was pretrained with ImageNet dataset [16] which contains 1.3 million non-medical images and then fine-tuned with our dataset. During fine-tuning, the initial learning rate was set to 10^{-4} and was reduced by a factor of 10 when the validation loss did not reach a new minimum for 3 consecutive epochs. The training process ended when the learning rate fell below 10^{-6} . Although the training process was performed on single images, the predicted malignancy scores of MLO and CC views were averaged to generate a single prediction for each breast. The deep learning network was implemented using PyTorch [17] and ran on a workstation equipped with 3 Nvidia GTX 2080Ti GPUs (11 GB memory each), which can fit a maximum batch size of 12 under the configuration of DenseNet-121 and 1024×832 input resolution. However, such small batch size will severely damage model convergence and accuracy due to the batch normalization layers in most recent CNN architectures [18]. As a result, a batch sub-division approach was adopted: the model was updated once only after the gradients were accumulated for 3 batches, making an effective batch size of 36. In our experiment, the fivefold training and cross-validation finished in 15 h; and in test phase, the inference of 1193 exams took less than 5 min.

2.3 Performance of human readers

To compare the performance of DL models and the radiologists' evaluation, ROCs were generated from both the prediction scores of DL models and the BI-RADS categories extracted from the mammogram reports, and the area under curve (AUC) was used as a metric of prediction accuracy. Although using BI-RADS to generate ROC curves was a common practice in many existing literatures [19, 20], there were disputations. Jiang et al. [21] argued that BI-RADS should not be used to estimate ROC curves mainly because BI-RADS category 0, 1 and 2 are not ordinal with respect to the estimated probability of cancer. But [21] also suggested that for diagnostic mammography in particular, BI-RADS can be made ordinal by (1) Removing all cases with BI-RADS category 0; (2) combining BI-RADS 1 and 2 into one category, and then can be used to generate ROC curves. Following these suggestions, when the DL model performances had to be compared with radiologists, we only use breast cases with BI-RADS 1/2, 3, 4A, 4B, 4C and 5 as six ordinal points with respective to malignancy probability estimated by the reporting radiologist, then use each as a threshold to calculate the true positive rate (TPR) and false positive rate (FPR) from which an ROC can be generated to represent radiologists' performance.

2.4 Evaluation setting

To simulate different clinical application scenarios, we evaluated the performance of the proposed DL model in three different settings, depending on the different samples used in the validation and test set. These evaluation settings include:

- (1) Setting BI-RADS 1–5: In this setting, cases with BI-RADS 0 were removed from the validation set and test set, and ROCs of both the DL model and the radiologists were generated with the rest cases. The evaluation in this setting reflected the general performance of DL model compared to the reporting radiologists on mammograms which radiologists regarded as “contains complete information to make decision”.
- (2) Setting BIRADS 4–5: In this setting, evaluation was performed on samples belonging to BI-RADS category 4A, 4B, 4C and 5, which was a subset of cases which radiologists would recommend for biopsy. The evaluation in this setting answers the question: Among breast cases which a radiologist would suggest biopsy, can DL effectively reduce the number of unnecessary biopsies without losing sensitivity?
- (3) Setting BI-RADS 0: In this setting, the DL performance was evaluated only on cases with BI-RADS category 0. This experiment dealt with the DL performance to distinguish benign and malignant cases which radiologists regarded to as “information incomplete, need other imaging modalities”.

Although the validation and test sets were different, noted that we did NOT train different models for different settings. The training process was performed only once for each fold of cross-validation on all BI-RADS categories, because the training process did not require BI-RADS but need more samples.

2.5 Interpreting DL model's decision with visualization

A major drawback of DL-based classification models in medical diagnosis is that it often operates as a black box, which generates prediction without explaining. Doctors do not know if such prediction should be trusted if it contradicts with their own decision. However, a number of recent progresses in neural network visualization may improve the interpretability of DL models. Zhou et al. [22] proposed class activation mappings (CAM) method to localize objects in image from a CNN model trained on image-level labels. Rajpurka et al. [23] used CAM to localize Pneumonia-related areas in chest X-Ray images. In this study, we also use CAM to visualize the area most indicative of malignancy in mammograms, making deep learning model's prediction

more interpretable. Moreover, such visualization also gave us a cue about the behavior of DL models to know their advantages and disadvantages, which is helpful for the development of new models.

3 Results

In each fold of the fivefold cross-validation, a single DL model was obtained. For each model, their ROC curve was calculated on their validation set as well as the test set and compared to the ROC curve generated with BI-RADS scores from radiologist ratings. Area under the ROC Curve (AUC) is used as the evaluating metric of both the DL model and radiologists performances. We report below the performance of DL models and radiologist rating in each of the testing scenario introduced in the previous section.

3.1 Setting BI-RADS 1–5

In this evaluation setting, the performances of DL models and radiologists were evaluated on samples within BI-RADS categories 1/2, 3, 4A, 4B, 4C and 5. The ROC curves generated for cross-validation and test were shown in Fig. 3. For fold 1 to 5 in cross-validation, the AUCs of DL models were 0.947, 0.953, 0.931, 0.931, and 0.940 respectively (mean = 0.939, std = 0.009), while the AUCs generated from radiologist BI-RADS were 0.912, 0.922, 0.906, 0.921 and 0.909 respectively (mean = 0.913, std = 0.006). For the test set, the 5 DL models obtained

from fivefold cross-validation achieved a mean AUC of 0.940 with standard deviation of 0.003, while the AUC of radiologists rating was 0.906. It can be observed in ROC curves of both validation and test that, the DL models' performance was similar (or slightly better) to the compared radiologists at high sensitivity and high specificity regions where radiologists were relatively sure about their decisions (BI-RADS $\geq 4C$ or BI-RADS < 3); whereas DL models showed notable better performance in middle specificity regions where the radiologists' decisions were obscure (BI-RADS 3, 4A and 4B). It is also worth noted that in the test set, the DL model was able to achieve sensitivity 0.9 at specificity of 0.85. On the other hand, the radiologists ROC only reached same sensitivity of 0.9 at specificity of 0.7, which means that the DL model can prevent nearly half of false positives in this scenario.

3.2 Setting BI-RADS 4–5

In this task, the performances of DL models were evaluated only on samples within BI-RADS categories 4A, 4B, 4C and 5, which represented a subset of exams in which radiologists would suggest a biopsy based on the mammogram alone. The ROC curves generated for cross-validation and test is shown in Fig. 4. In cross-validation and test set respectively, the DL model achieved a mean AUC of 0.932 and 0.927, while the mean AUCs of radiologists were 0.837 and 0.847, respectively.

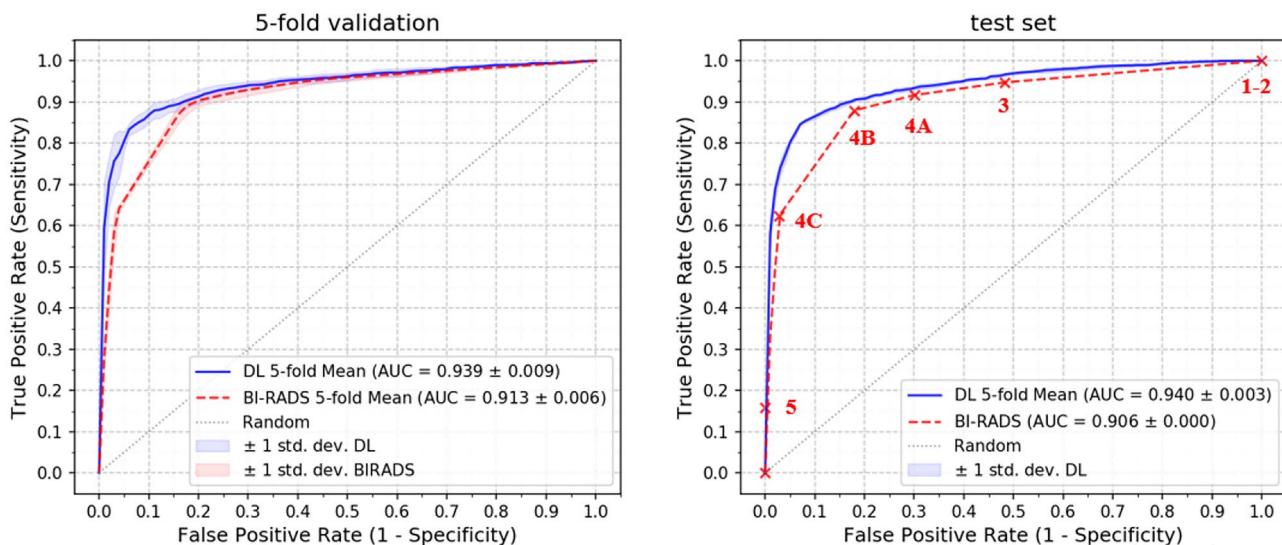


Fig. 3 ROC curves of DL model and radiologists with samples with BI-RADS 1–5. In cross-validation (left), the mean and standard deviation of the ROC curves of DL model and radiologists was shown in blue and red respectively. In test set (right), Red crosses

were the operating point of human reader by setting cases with BI-RADS higher than a certain threshold as positive prediction. It can be noticed that on each operation point, the DL model performs better than human radiologists

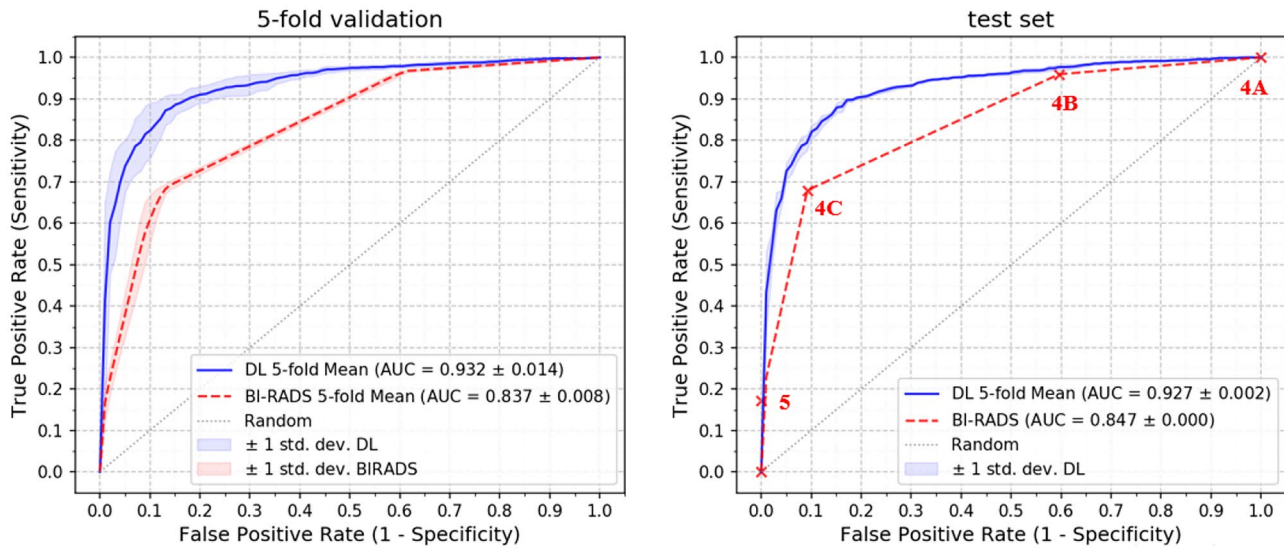


Fig. 4 ROC curves of DL model and radiologists with samples with BI-RADS 4A, 4B, 4C and 5

3.3 Setting BI-RADS 0

Finally, we evaluated the performance of the DL model to distinguish the malignancy of BI-RADS 0 samples, which radiologists regarded to as “information incomplete, need other imaging modalities”. The ROC curves were shown in Fig. 5. The mean AUCs of cross-validation set and test set were 0.758 and 0.787, respectively. It was shown that even for samples which radiologists regarded to as information incomplete, the proposed DL model can still well distinguish the malignant and benign classes, which can potentially reduce the number of unnecessary additional examinations.

3.4 Influence of DL models and image resolution

We trained and validated different CNN model architectures with different image resolutions to check the influence of these two factors on the final diagnostic accuracy. Due to the limited computational resources, instead of performing fivefold cross-validation, all models were only trained and evaluated once using training-validation split of the first fold. Moreover, all data regardless of BI-RADS (0–6) were used for validation to compute ROCs, since comparison with human reader performance was no longer needed.

We first compared performances of a series of representative deep learning models including AlexNet (2012),

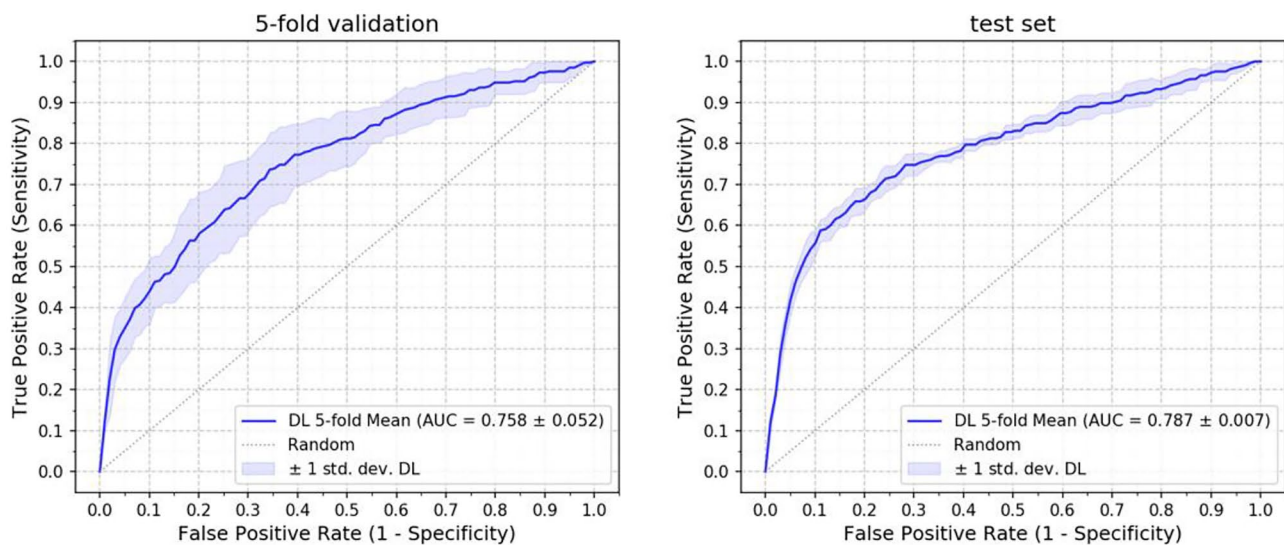


Fig. 5 ROC curves on samples with BI-RADS 0 in cross-validation (left) and test set (right)

VGGNet (2014), ResNet (2016) and DenseNet (2017). Since AlexNet and VGGNet only accept a fixed image resolution at 224×224 pixels, other models were also trained and evaluated at this resolution. The performances of different models were summarized in Table 2(a). The earliest CNN model architecture AlexNet performed worst (AUC = 0.822) and the most recent DenseNet-121 model achieved the best performance (AUC = 0.892), showing a direct impact of the development in deep learning technology on diagnostic accuracy. However, as can be noticed in Table 2(a), the models' classification accuracy in mammograms was not always correlated with their classification accuracy in the general image recognition task. The DenseNet-121 model had lower classification accuracy in ImageNet contest than both ResNet-50 and ResNet-101 models, but it achieved the best performance in the mammogram classification task. In our opinion, the reasons are twofold: (1) Although DenseNet-121 has more convolutional layers, it has much less trainable model parameters than VGGNet, ResNet-50 and ResNet-101. Since the mammogram training set in this study was much smaller than the ImageNet training set (> 1 Million images), DenseNet models were less prone to overfitting. (2) The low-level features of mammograms were very different from non-medical images, and DenseNet models are especially good at learning low-level features, resulting in more effective transfer learning.

To evaluate the influence of the image resolution, the DenseNet-121 model was trained and validated with different image resolutions, and the result was presented in Table 2(b). The AUC first improves along with the increased resolution from 0.882 for 224×224 to 0.921 for 1024×832 , then decreased when the resolution moved further to 1536×1248 . This might be because DenseNet (and most other off-the-shelf CNN classification architectures) assumes

input resolution of 224×224 , thus has a limited receptive field. When the image resolution increases, the CNN acquires more and more detailed but local information, but failing to capture the global information. As a result, there is a trade-off between the importance of learning detailed (at high resolution) and global (at low resolution) features.

3.5 Model interpretation by visualization

To interpret the DL model's predictions, the heat maps were produced to visualize the areas of the image which is most indicative of the malignancy using class activation mappings (CAM) method proposed in [23]. For most of malignant mammograms, the CAM correctly identified the regions containing the malignant lesions. Figure 6a shows several examples DL visualizations and the radiologists' annotations of on the same mammogram. Figure 6b shows a mammogram with invasive cancer which was given a BI-RADS category 4B by radiologist and a malignant probability of 0.96 by the DL model. The DL model visualization (left) correctly identified the mass region as highly indicative of cancer, as well as auxiliary signs such as the thickened skin and the axillary nodule, which was not annotated by the radiologist. Visualization also helps us to understand the behavior on CNN models trained with different image resolutions and provides hints for further improvement of model structure. Figure 6c shows the activation patterns in on a spiculated mass that the radiologist rated as BI-RADS 4C. It can be observed that in lower resolution (512×416), the whole mass region was activated, while in higher resolution (1024×832) more local regions like the spikes around the mass was considered most indicative of malignancy. Although under both resolutions, the model correctly predicted the malignancy (score > 0.99), the decision was made

Table 2 Summary of influence of the choice of DL models and image resolutions on classification performance

(a) Performance different CNN models on 224×224 image resolution measured by AUC on the validation set of the first fold. The ImageNet Errors were measured by the top 5 classification error rates, which were extracted from papers in which the models were first proposed

CNN models	Year	# of conv layers	ImageNet Error	AUC
AlexNet	2012	7	15.3%	0.822
VGGNet	2014	16	7.1%	0.875
ResNet-50	2015	50	5.3%	0.872
ResNet-101	2015	101	4.6%	0.883
DenseNet-121	2017	121	6.6%	0.892

(b) Performance of DenseNet-121 model on different image resolution

Resolution	AUC
224×224	0.892
512×416	0.916
1024×832	0.921
1536×1248	0.908

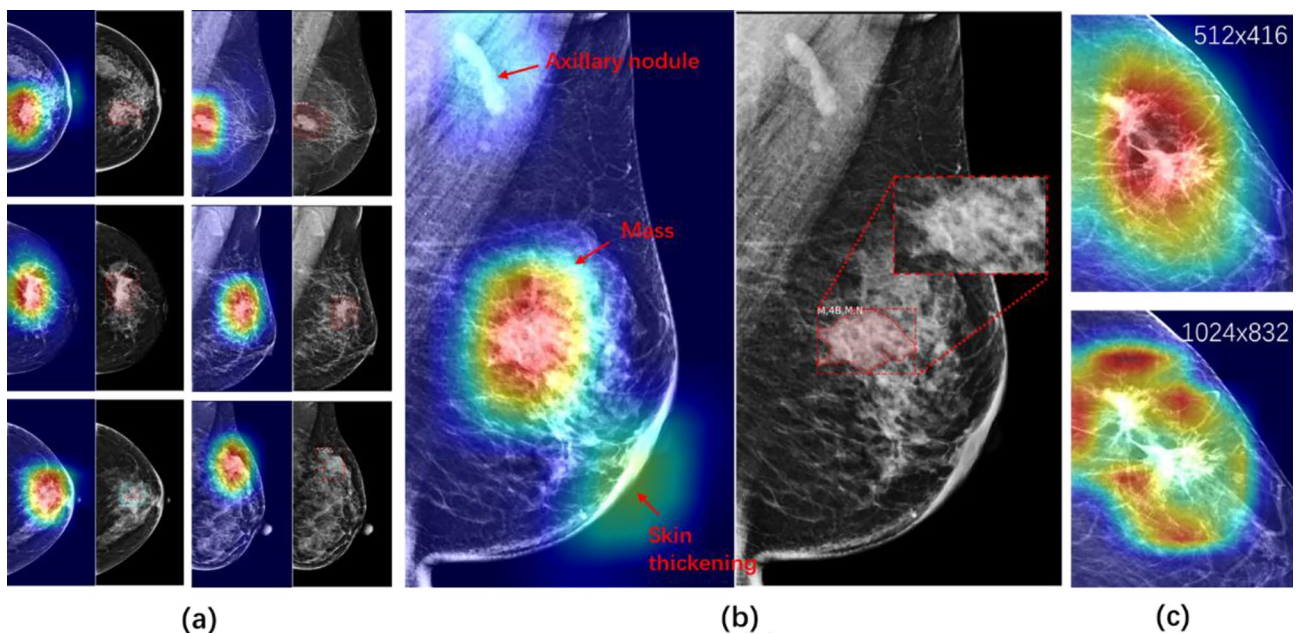


Fig. 6 **a** Heatmap visualization of regions most indicative of malignancy on random samples of malignant images in test set (left), and annotations of the most significant lesions annotated by radiologists (right). **b** A mammogram with invasive cancer which was given a BI-RADS category 4B by radiologist and a malignant probability of 0.96 by the DL model. The DL model visualization (left) not only

correctly identified the mass region as highly indicative of cancer, but also showed auxiliary signs such as the thickened skin and the axillary nodule as weakly indicative of cancer. **c** Malignancy Heatmap on a spiculated mass image generated by two DenseNet121 models trained with different image resolutions. Global features were learnt at low resolution while local features were learnt at high resolution

based on different image features. As a result, a new CNN architecture able to learn both global and local features may further improve the performance.

4 Discussions

In this study, we present a novel investigation showing that deep learning methods trained on a large set (over 5000 patients) of historical mammographic exams with ground truth pathology can out-perform experienced radiologists in distinguishing malignant case from benign ones, showing the potential to reduce the number of biopsies performed on benign breasts in clinical applications. Even on cases which human readers regarded to as “information incomplete” (rated as BI-RADS 0), the DL model can still achieve a decent AUC around 80%.

4.1 Relation to previous literature

The main objective of this paper is to explore the possibilities that if a major medical institute can train a high-performance breast cancer classification model by exploring their huge amount of historical imaging data to assist radiologists to perform more accurate diagnosis, but without increasing their workload to perform pixel-level annotations for model

training. For this purpose, we discuss previous literature in two aspects: datasets and algorithms.

4.1.1 Datasets

In most technical oriented literature, proposed algorithms were trained and evaluated with public datasets. The most widely used public mammographic datasets include the DDSM dataset [24] and the InBreast dataset [25], which hardly meet clinical need. The DDSM dataset is the largest publicly available mammogram dataset which contains screen film mammography images of 2620 patients. Screen film mammography devices are nowadays obsolete and the image style of the DDSM dataset is vastly different from that of modern digital mammograms. A recent study [10] have shown that two state-of-the-art breast cancer classification models which have been trained and evaluated with the DDSM dataset achieved near-random classification results when applied to real-world digital mammogram images. The InBreast dataset consists full-field digital mammogram images acquired by a modern digital mammography machine, but only contains 115 patients, which is unlikely to train and validate a clinically reliable deep learning model for breast cancer diagnosis. To fill this gap, other research group constructed their own digital mammogram datasets which are generally not publicly available, such as the NYU

[5] dataset. The NYU dataset consists of mammographic exams from 141,473 patient in total, but only 985 of them have biopsy-proven cancer. In comparison, the dataset constructed and used in this study is much smaller in total, but contain more than 3000 biopsy-proven cancer cases, which provides the deep learning model more general cancer-related features.

4.1.2 Algorithms

Deep learning algorithms in mammography have been previously studied by a number of researchers and can be roughly categorized into two classes, namely detection-based and classification-based approaches, depending on if manual annotations of abnormal regions is required for model training. For detection-based approaches, Kooi et al. [26] trained a customized CNN model with extracted image patches from lesion-annotated mammograms and show that such model out-performed traditional CAD system in screening mammography; Ribli et al. [6] used Faster-RCNN detector to detect lesions in mammograms and estimate their malignancy and achieve good performance in the public INBreast database [25]; Shen et al. [7] proposed a two-stage training strategy where the abnormality annotation is required in the first stage, and the image-level label is used to fine-tune the model in the second stage [8] and [27] showed that a commercialized deep-learning-based CAD system can achieve a cancer detection accuracy non-inferior to an average breast radiologist. However, all those detection-based methods need extra manual annotations of lesion locations by radiologists which is impractical for major medical institutions where radiologists are already under high workload. Moreover, these methods or systems can only detect masses and micro-calcifications while ignoring other axillary signs of cancer.

The method used in this study fall in to the category of classification approaches and can be trained with image-level annotations only, which can be directly extracted from medical reports with no additional efforts required; hence, enables large-scale data collection and training. In [28] the authors employed AlexNet to distinguish recalled but benign mammograms. The mammogram images were downscaled to a very low resolution of 224×224 pixels which has limited value for real-world clinical application. Zhu et al. [29] proposed a multi-instance learning model for the detection of breast cancer without the need to annotate the training data. However, their method was trained and tested on a small dataset of 115 patients. On the other hand, our study perform training and validation with a large real-world dataset of more than 7000 patients with biopsy-proven pathology, which provides more solid proof for such a model to aid radiologists in making more accurate diagnosis.

4.2 Limitations

Our study has limitations that need to be acknowledged. The first is about the data collection. Although we collected a large digital mammogram dataset with reliable ground truth pathology, all samples were from the same institute and recorded by the same mammogram machine. There was no guarantee that the trained model can perform equally well on mammograms shot by devices from other manufactures. Moreover, the dataset used in this study contained only the breast cases which finally underwent biopsies. Although we did this intentionally to guarantee a solid ground truth of malignancy in both training and validation, a fraction of both true negative and false negative cases in terms of radiologist decision were screened out, causing that the ROCs generated did not fully represent the performance of both radiologists and DL models in real clinical diagnosis scenarios. Future improvements of the study include training and evaluating DL models on a more comprehensive datasets including mammo-gram cases both with and without biopsies, whereas those considered-benign cases are verified by patient tracking. The second limitation of our study is about the radiologist performance evaluation. Because of the large size of the dataset, we did not have enough resources to asking extra radiologists to re-interpret those historical mammograms, and the radiologist performance evaluation was only based on historical BI-RADS. Though every mammo-gram exam had underwent double-reading by a radiologist with no-less than 3-year mammogram reading experience and a senior radiologist with more than 10-year experience, we cannot readily claim the performance reported is representative of breast radiologists in general. As a result, we do not claim a “super-radiologist” performance of deep learning in general diagnostic mammogram setting, although experiments in our study showed that the proposed DL model can effectively reject a large fraction of unnecessary biopsies suggested by our radiologists with little loss in sensitivity.

In the algorithm aspect, although different DL classification models such as VGGNet, ResNet and DensNet were trained and evaluated, we do not claim our model performance to be the state-of-the-art, which is not the key point of this paper. The main contribution of our paper is a proof-of-concept that a major medical institute is able to train a high-performance breast cancer classification model by exploring their huge amount of historical imaging data, and analyze the potential clinical use cases of the model. Recent progresses in deep learning such as fine-grained classification [30, 31], multi-view learning [32], 3D learning [33] can be integrated into the current model to further boost the system performance.

5 Conclusion

To conclude, we showed that under certain conditions (sufficient training sample, homogeneous data source, etc.), a well-designed deep neural network model can be trained on historical mammogram-pathology pairs to make accurate diagnosis on breast cancer, showing comparable or better accuracy than experienced radiologists. Our study also showed the potential of further optimization of the neural network architecture specific to mammography problem, and the development of more effective interpretation tools to enable radiologists to discover knowledge from neural networks instead of only providing knowledge to train them. Finally, we are optimistic that the incorporation of deep-learning-based artificial intelligence into clinical workflow of breast cancer diagnosis may lead to better working efficiency and diagnostic accuracy.

Funding This work is supported by National Natural Science Foundation of China (Grand No. 61702453), Natural Science Foundation of Zhejiang Province (Grand No. Q17F030003).

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

- Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., Thun, M.J.: Cancer statistics, 2008. *CA Cancer J. Clin.* **58**(2), 71–96 (2008)
- DeSantis, C., et al.: Breast cancer statistics, 2013. *CA Cancer J. Clin.* **64**(1), 52–62 (2014)
- Lehman, C.D., Arao, R.F., Sprague, B.L., Lee, J.M., Buist, D.S., Kerlikowske, K., et al.: National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology* **283**(1), 49–58 (2016)
- Sickles, E.A., Wolverton, D.E., Dee, K.E.: Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology* **224**(3), 861–869 (2002)
- Wu, N., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans. Med. Imaging* **39**(4), 1184–1194 (2019)
- Ribli, D., et al.: Detecting and classifying lesions in mammograms with deep learning. *Sci. Rep.* **8**(1), 1–7 (2018)
- Shen, L., et al.: Deep learning to improve breast cancer detection on screening mammography. *Sci. Rep.* **9**(1), 1–12 (2019)
- Rodríguez-Ruiz, A., et al.: Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* **290**(2), 305–314 (2019)
- Gardezi, S.J.S., et al.: Breast cancer detection and diagnosis using mammographic data: Systematic review. *J. Med. Internet Res.* **21**(7), e14464 (2019)
- Wang, X., et al.: Inconsistent performance of deep learning models on mammogram classification. *J. Am. Coll. Radiol.* **17**(6), 796–803 (2020)
- Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Adv. Neural. Inf. Process. Syst.* **25**(2), 1097–1105 (2012)
- Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Preprint at [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
- He, K., Zhang, X., Ren, S., & Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
- Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708 (2017)
- Kingma, D.P., Ba, J.A.: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255 (2009)
- Paszke, A., Gross, V., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: NIPS 2017 Workshop on Autodiff (2017)
- Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
- Becker, A.S., Marcon, M., Ghafoor, S., Wurnig, M.C., Frauenfelder, T., Boss, A.: Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Invest. Radiol.* **52**(7), 434–440 (2017)
- Rodríguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., et al.: Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J. Natl. Cancer Inst.* **111**(9), 916–922 (2019)
- Jiang, Y., Metz, C.E.: BI-RADS data should not be used to estimate ROC curves. *Radiology* **256**(1), 29–31 (2010)
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929 (2016)
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., et al.: Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint [arXiv:1711.05225](https://arxiv.org/abs/1711.05225) (2017)
- Lee, R.S., et al.: A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **4**(1), 1–9 (2017)
- Moreira, I.C., et al.: Inbreast: toward a full-field digital mammographic database. *Acad. Radiol.* **19**(2), 236–248 (2012)
- Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R., et al.: Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017)
- Rodríguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., et al.: Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J. Natl. Cancer Inst.* **111**(9), 916–922 (2019)
- Aboutalib, S.S., Mohamed, A.A., Berg, W.A., Zuley, M.L., Sumkin, J.H., Wu, S.: Deep learning to distinguish recalled but benign mammography images in breast cancer screening. *Clin. Cancer Res.* **24**(23), 5902–5909 (2018)
- Zhu, W., et al.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham (2017)

30. Liang, Y., et al.: Urbanfm: inferring fine-grained urban flows. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (2019)
31. Ouyang, K., et al.: Fine-grained urban flow inference. *IEEE Trans. Knowl. Data Eng.* (2020).
32. Yan, C., et al.: Deep multi-view enhancement hashing for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* (2020).
33. Yan, C., et al.: 3D room layout estimation from a single RGB image. *IEEE Trans. Multimedia* **22**(11), 3014–3024 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.