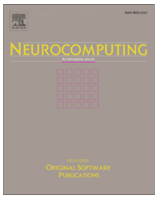
Neurocomputing 472 (2022) 152–165

Contents lists available at ScienceDirect

Neurocomputing

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



Computer-aided diagnosis of breast cancer in ultrasonography images by deep learning

Xiaofeng Qi [a,1], Fasheng Yi [b,1], Lei Zhang [a], Yao Chen [c], Yong Pi [a], Yuanyuan Chen [a], Jixiang Guo [a], Jianyong Wang [a], Quan Guo [a], Jilan Li [a], Yi Chen [a], Qing Lv [c,*], Zhang Yi [a,*]

[a] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, PR China
[b] College of Computer Science, Chengdu University, Chengdu 610106, PR China
[c] Department of Galactophore Surgery, West China Hospital, Sichuan University, Chengdu 610041, PR China

ARTICLE INFO

Article history:
Received 12 April 2021
Revised 7 October 2021
Accepted 14 November 2021
Available online 25 November 2021

Keywords:
Breast cancer
Ultrasonography
Deep neural networks
Initial screening
Computer-aided diagnosis

ABSTRACT

Ultrasonography of the breast mass is an important imaging technology for diagnosing breast cancer. In China, ultrasound equipment is widely used in medical institutions. Patients obtain a report with highlighted ultrasonography images through an initial clinical screening. However, analyzing these images manually is highly subjective for the variation in the clinical competence of doctors, resulting in poor consistency and low sensitivity. In this study, an automated breast cancer diagnosis system is developed to increase diagnostic accuracy. The system is deployed on mobile phones, takes a photo of the ultrasound report as input and performs diagnosis on each image. The developed system consists of three subsystems. The first subsystem is to reduce noise in the taken photos, reconstructing high-quality images. We develop the first subsystem based on the frameworks of stacked denoising autoencoders and generative adversarial networks. The second subsystem is to classify images into malignant and non-malignant; to extract high-level features from the images, deep convolutional neural networks are employed. The third subsystem is to detect anomalies in model performances, reducing false negative rates. Generative adversarial networks are utilized to distinguish false negative samples from true negative samples. 18,225 breast ultrasonography images and 2416 ultrasound reports are collected to train and evaluate the system. Experimental results show that the performance of our system is comparable to that of human experts. It is believed that this is the first system for breast cancer diagnosis deployed on mobile phones. The developed system is integrated with a cloud computing platform and accessible online to aid in the initial screening and diagnosis of breast cancer, thereby promoting earlier treatment, reducing the morbidity and mortality.

© 2021 Published by Elsevier B.V.

1. Introduction

Breast cancer is a major public health problem among women worldwide. It is the most common invasive cancer and the second leading cause of death for women around the world [1]. The incidence of breast cancer is the highest among cancers for women in 154 countries, and still rising sharply. This is likely caused by a variety of factors, such as a lack of awareness about breast cancer and limitations regarding treatment. Breast cancer can be cured if treated early, so accurate early detection is the key to control this disease and reduce its fatality [2]. Diagnosis by pathological examination is the gold standard for breast cancer. However, pathological examinations are time-consuming, resource-intensive, and invasive [3,4]. Since the breasts are superficial organs of the human body, anomalies in breast mass can be detected by imaging modalities. In clinical applications, imaging techniques are widely used in the detection of breast cancer. One of the most popular imaging modalities for breast cancer detection is mammography. However, the sensitivity of mammography is reduced in younger women with dense breasts; In China, 70% of women are with dense breasts, the false negative rate of mammography is relatively high. In this case, other modalities are often suggested. Ultrasonography is capable of detecting and classifying nodules in the breast mass and is commonly used for physical examination in middle-income countries; it is more suitable for use in daily clinical practice due to its non-radiation, low invasiveness, and low cost. In China, patients usually go to different levels of hospitals for primary screening.

---

* Corresponding authors.
E-mail addresses: lqlq1963@163.com (Q. Lv), zhangyi@scu.edu.cn (Z. Yi).

[1] These authors are co-first authors.

https://doi.org/10.1016/j.neucom.2021.11.047
0925-2312/© 2021 Published by Elsevier B.V.

In an ultrasound examination, the sonographer examines breast tissues from different angles, using various modes and pressures of the probe, and captures images encompassing the main characteristics of the tissues of interest. The sonographer then summarizes the entire examination, providing descriptions of the breast tissues along with diagnostic suggestions. The patient receives an inspection report, which contains highlighted images and a summary. A sample ultrasound inspection report is presented in Fig. 1. In most cases, the breast surgeon performs diagnoses and provides clinical advice based on the ultrasound inspection report. If there are uncertainties in the analysis of this report, the patient is recommended to undergo further examinations such as mammography and biopsy. Therefore, the analysis of ultrasonography images plays an important role in the diagnosis of breast cancer.

However, the analysis of ultrasonography images is challenging for sonographers and breast surgeons. The results of ultrasound examinations are easily influenced by inherent noise and the operations of sonographers. It is difficult to judge whether a lesion in mammary tissue is malignant or benign when there are some similar features, and human doctors from different hospitals are with poor consistency. Thus, experienced breast surgeons and well-trained sonographers are important for breast diagnosis. Although ultrasound equipment is widely used in medical institutions, there are far from enough experienced breast surgeons and sonographers, and huge numbers of patients have difficulty receiving an advanced diagnosis. Manual analysis of breast ultrasonography images is highly subjective, human doctors may have different diagnosis for the same image. The specificity and sensitivity of manual diagnoses are 91% and 84% in the classification of breast cancer [5]. This suggests that there may be numerous missed diagnoses and unnecessary biopsy operations.

Automated breast cancer diagnosis is helpful to reduce the cost of time and human resources [5–8]. For breast ultrasonography images, a convenient and accurate intelligent diagnosis system can assist breast surgeons in making diagnoses more reliable and efficient. Several computer-aided diagnosis approaches have been studied regarding breast ultrasonography image detection, segmentation and classification [9–15]. These traditional machine learning methods such as linear discriminant analysis [16,17] and support vector machine [18–20] have made great contributions to aiding diagnosis based on breast ultrasonography images. However, these methods require substantial feature engineering. The deep learning methods that are trained end-to-end can extract features with no need for manual feature selection. Some studies on deep learning [21] have also been performed, using approaches like the restricted Boltzmann machine (RBM) [22,23] and multilayer perceptron (MLP) [24]. However, the lack of data in these experiments leads to the problems of overfitting and weak generalization ability. Recently, with the breakthroughs in research on neural networks, a deep convolutional neural network (CNN) [25–30] model has been developed that can acquire very high accuracy in complex image classification tasks such as ImageNet Challenge [31,32], driving advances in classification and detection. A convolutional layer in CNNs has much fewer connections and is easier to train than a fully connected layer. Compared with RBM and MLP, deep CNNs learn hierarchical abstract features from pixels and perform better in large datasets.

In this study, a fully automated diagnostic system for breast ultrasonography images is proposed. The system focuses on a practical application scenario, in which the users are patients: usually, the collected digital ultrasound images are stored in the Picture Archiving and Communication Systems(PACS) in the local area network in hospital, and are not accessible to patients and wide area network. After an ultrasound examination, the patient receives a report as presented in Fig. 1. Therefore, the users need to use mobile phones to take photos of ultrasound reports, which is feasible in the typical application scenario in today's medical imaging workflow. In this case, the system is deployed on mobile phones, providing a convenient analysis service for patients irrespective of the timing and location. The diagnosis is accomplished through the *R-S-A* procedure: (1) *Report*: As described above, an ultrasound inspection report is provided after an ultrasound examination; this report is usually printed on paper for breast surgeons and patients. In primary hospitals such as community hospitals and county-level hospitals, there is a shortage of experienced surgeons and sonographers to analyze these reports. (2) *Shoot*: The only procedure that the patients need to perform is to take a photo of the ultrasound inspection report with a smart phone, feeding the photo into out system. (3) *Analysis*: The system takes the photo as input, detects each image in the report, and performs diagnosis on the ultrasonography images.

Following the *R-S-A* procedure: *Report*, *Shoot*, and *Analysis*, the proposed system consists of three parts. The main component of our system is the DeepCls subsystem, which is an image classifier based on the Inception-v3 architecture [33]. Each ultrasonography image in the report is classified by the deep convolutional neural network into "malignant" and "non-malignant", which is the most important issue in breast ailments.

Here, we propose a three-stage procedure: input standardization, classification, missed diagnosis detection, for CAD system development. The ultrasound inspection report is printed on paper and shot by various models of smartphone, so there is substantial noise during the *Shoot* process. To overcome the influence of noise, another subsystem is added before the DeepCls subsystem, which we call DeepRec. The DeepRec subsystem is capable of reconstructing high-quality images from low-quality photos. The reconstructed images are then fed into the DeepCls subsystem to perform diagnosis. The DeepRec subsystem consisted of a deep convolutional generator with fractional strided convolutions [34,35] and two discriminators based on the GANs framework. To train the generator, low-level losses such as mean square error and total variance loss as well as high-level losses like perceptual loss and adversarial loss were employed. In breast cancer examinations, the top priority is reducing the number of missed diagnoses. As such, after the DeepCls subsystem, an extra subsystem called DeepAti is proposed. The DeepAti subsystem aims at recognizing missed diagnoses, reducing the false negative rate. To deal with the diversity of read-world data, these three stages are indispensable for clinical application. All three subsystems work in a cascade manner, as shown in Fig. 2.

2. Materials and methods

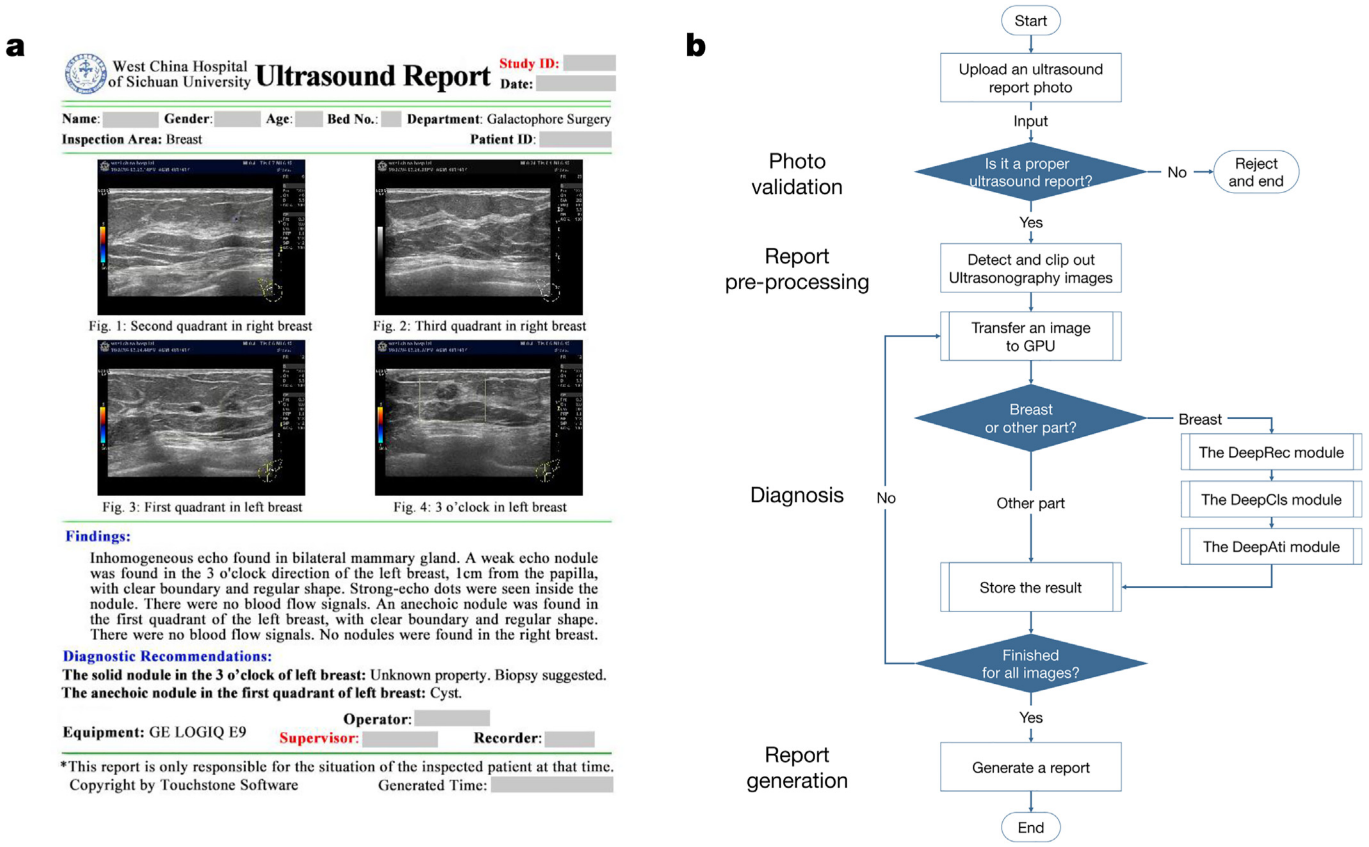
2.1. Patient and image characteristics

To develop the three subsystems, we collected 18125 breast ultrasonography images from three hospitals. 7328 images were obtained from West China Hospital, Sichuan University (WCH), 9014 images were from Chengdu Military General Hospital (CMGH), and 1784 were from People's Hospital of Deyang City (PHDY).

The images were used to construct the Cls-Set, which was used to train and evaluate the DeepCls subsystem. In the Cls-Set, there were 8146 malignant images and 9979 non-malignant ones.

The Cls-Set was split into a training set, a validation set and a testing set, with 12,083, 3020, and 3022 images, respectively. The training set was used to train the DeepCls subsystem, the model with the highest performance on the validation set was used for testing.

To develop and evaluate the DeepRec subsystem, we constructed the Rec-Set. Each sample in the Rec-Set was a pair includ-



West China Hospital of Sichuan University **Ultrasound Report**

Study ID: [redacted]
Date: [redacted]

Name: [redacted] **Gender:** [redacted] **Age:** [redacted] **Bed No.:** [redacted] **Department:** Galactophore Surgery
Inspection Area: Breast **Patient ID:** [redacted]

Fig. 1: Second quadrant in right breast

Fig. 2: Third quadrant in right breast

Fig. 3: First quadrant in left breast

Fig. 4: 3 o'clock in left breast

Findings:

Inhomogeneous echo found in bilateral mammary gland. A weak echo nodule was found in the 3 o'clock direction of the left breast, 1cm from the papilla, with clear boundary and regular shape. Strong-echo dots were seen inside the nodule. There were no blood flow signals. An anechoic nodule was found in the first quadrant of the left breast, with clear boundary and regular shape. There were no blood flow signals. No nodules were found in the right breast.

Diagnostic Recommendations:

The solid nodule in the 3 o'clock of left breast: Unknown property. Biopsy suggested.
The anechoic nodule in the first quadrant of left breast: Cyst.

Operator: [redacted]
Equipment: GE LOGIQ E9 **Supervisor:** [redacted] **Recorder:** [redacted]

*This report is only responsible for the situation of the inspected patient at that time.
Copyright by Touchstone Software Generated Time: [redacted]

Fig. 1. An example of ultrasound inspection report and the analysis process of the proposed system. a. At the top of the report is some information of the patient. The body of the reports is composed of several ultrasonography images and descriptions of the tissues. At the bottom of the report is the diagnostic suggestions from the sonographer. b. Analysis process of the proposed system.

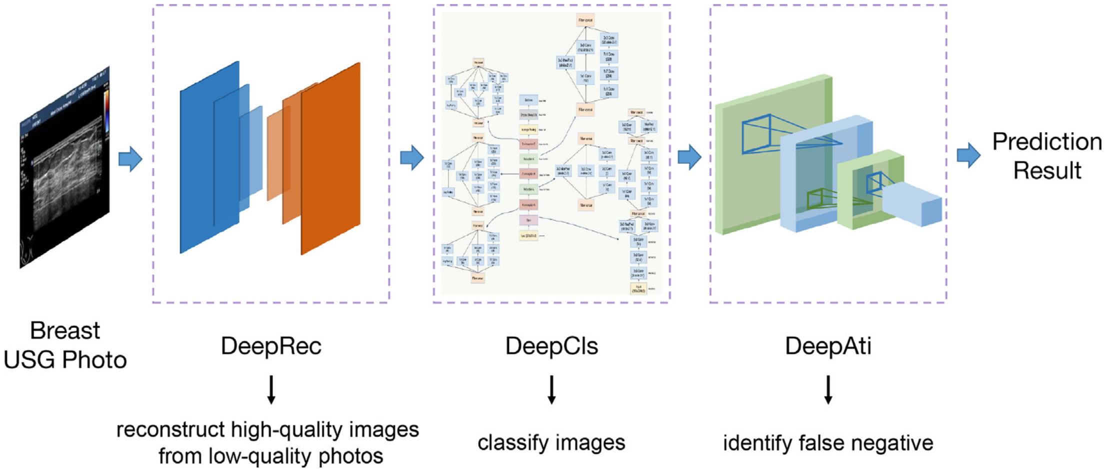


Fig. 2. Main framework of the proposed system. The input breast image is processed by the DeepRec subsystem, the DeepCls subsystem and the DeepAti subsystem subsequently.

ing a manually shot low-quality photo and the corresponding original high-quality image.

We collected 2416 ultrasound reports from West China Hospital, Sichuan University, and used smartphones to take photos of the reports. The reports contain 15,514 ultrasonography images, covering all of the images from West China Hospital included in the Cls-Set.

Since the ultrasound reports from Chengdu Military General Hospital and People's Hospital of Deyang City were not collected, to maintain consistency of the validation set and testing set, the images from these hospitals in the validation set and testing set were organized and printed on A4 paper to generate ultrasound reports. Therefore, the validation set and testing set of the Rec-Set include 3020 and 3022 pairs of images, respectively, which is the same as the Cls-Set. The left images were used to construct the training set, which includes 13,071 pairs. We used 25 models of smart phones to take pictures of the reports, to mirror what would typically occur in a real-world scenario in the R-S-A procedure. The process for constructing the datasets is presented in Fig. 3.

The fully automated diagnostic system proposed in this study follows the *R-S-A* procedure: *Report*, *Shoot*, *Analysis*, and there are three subsystems proposed in it. In the following subsections, we describe these three subsystems, namely, the DeepRec subsystem, the DeepCls subsystem and the DeepAti subsystem in detail.

2.2. The DeepRec subsystem

2.2.1. Data preparation

To reduce the side effects caused by the *Shoot* process, we proposed the DeepRec subsystem. The DeepRec subsystem is capable of reconstructing high-quality images from low-quality images with noise, such as a shadow, reflection, or elastic deformation. To train the DeepRec subsystem, we constructed a large dataset of image pairs, which we called the Rec-Set. Each sample in the Rec-Set was a pair including a manually shot low-quality photo of the ultrasonography image and the corresponding original high-quality image.

In this study, we used 25 models of smart phones to take pictures of the ultrasound reports included in the Cls-Set. There were 2416 reports and 15514 images collected from West China Hospital, Sichuan University. All reports were shot by different people, from various directions, and in various illumination conditions. Since the ultrasound reports from Chengdu Military General Hospital and People's Hospital of Deyang City were not collected, to maintain consistency of the validation set and testing set, the images from these hospitals in the validation set and testing set were organized and printed on A4 paper to generate ultrasound reports. We then use smart phones to take photos of these reports.

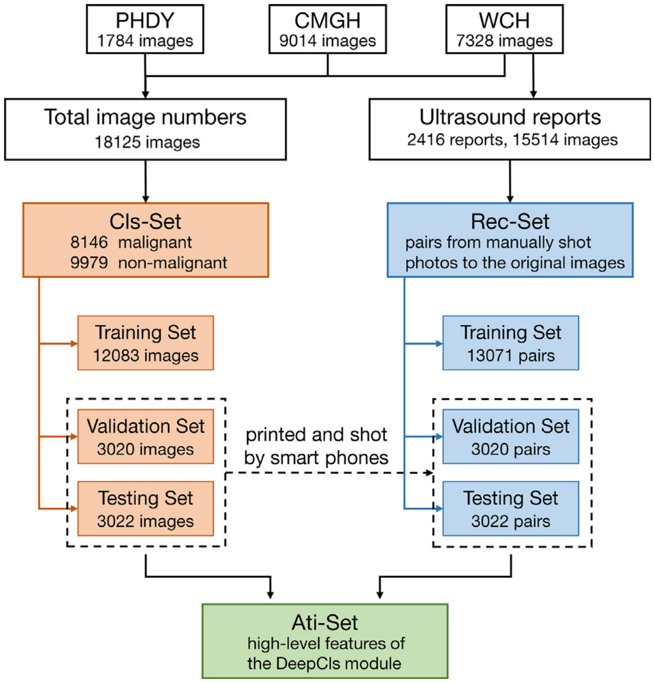


Fig. 3. Construction process of the datasets. Images collected from West China Hospital, Sichuan University (WCH), Chengdu Military General Hospital (CMGH), and People's Hospital of Deyang City (PHDY) were used to construct the Cls-Set. Ultrasound reports obtained from West China Hospital, Sichuan University were used to construct the Rec-Set, along with all images in the validation set and testing set of the Cls-Set. The Ati-Set was formed using the high-level features from the Cls-Set and the Rec-Set.

2.2.2. Model evaluation

To evaluate the performance of the DeepRec subsystem, the DeepCls subsystem was employed. After the training of the DeepRec subsystem, low-quality photos of the images from the validation and testing set of the Cls-Set were used as input of the DeepRec subsystem, and the reconstructed images were evaluated using the DeepCls subsystem. We used the same criteria as described in Section 2.3.4 to evaluate the reconstructed images.

2.2.3. Image reconstruction network

The DeepRec subsystem was designed based on the frameworks of stacked denoising autoencoders (SDAEs) and generative adversarial networks (GANs). Stacked denoising autoencoders is an image-to-image model introduced by Vincent et al. [36]. Taking an image as input, the autoencoder is trained to reconstruct the image in an unsupervised manner. However, unlike the standard autoencoders, our method takes a low-quality image as input and attempts to reconstruct a high-quality image. The idea of generative adversarial networks was introduced by Goodfellow et al. based on game theory [37]. The goal of GANs is to define a game between two networks: a generator network $G(z;\theta^{(G)})$ that generates samples from vectors of noise z, and a discriminator $D(x;\theta^{(D)})$ that is trained to discriminate fake samples generated by $G(z;\theta^{(G)})$ from real samples. Training GANs requires finding a Nash equilibrium of the competition between the generator and the discriminator; the generator attempts to fool the discriminator by generating convincing samples, and the discriminator attempts to recognize fake sample accurately.

In this study, we propose an image-to-image model based on SDAEs and GANs. The main framework of the DeepRec subsystem is presented in Fig. 4. The subsystem contains a 16-layer generator and two discriminators. Since the DeepRec subsystem is used for image reconstruction, to maintain input features, residual blocks are employed in the generator architecture. The generator contains two parts, an encoder of 3 downsampling layers and 5 residual blocks [29] to extract features from the input images, and a decoder composed of 2 fractionally strided convolutional layers and 1 convolutional layers to reconstruct high-quality images, as presented in Fig. 4. The first discriminator $D_1(x)$ takes reconstructed and original images as input, which follows the framework of standard GANs. The second discriminator $D_2(x)$ is one of the unique parts of our approach, it takes the latent representations of the generator and $D_1(x)$ as input. Experimental results showed that the extra discriminator boosted the reconstruction performance of the generator.

To train the DeepRec subsystem, we formulate our objective function by combining four loss functions:

$$\mathcal{L} = w_{adv}\mathcal{L}_{adv} + w_{con}\mathcal{L}_{con} + w_{tv}\mathcal{L}_{tv} + w_{per}\mathcal{L}_{per} \tag{1}$$

where $w_{adv}, w_{con}, w_{per}$ and w_{tv} are the weighting parameters of individual losses.

The first part of $\mathcal{L}$ is the adversarial loss $\mathcal{L}_{adv}$ of two discriminators $D_1(x)$ and $D_2(x)$. $\mathcal{L}_{adv}$ consists of two parts as presented in Eqs. 2 and 3. Following the definition of [37], the target to train the generator and the discriminators is the minimax value function $V(G, D_1)$ and $V(G, D_2)$:

$$\begin{aligned} \min_{G}\max_{D_1} \quad & \mathbb{E}_{x\sim p_d}[\log D_1(x)] + \mathbb{E}_{\hat{x}\sim p_s}[\log(1 - D_1(G(\hat{x})))] \\ \min_{G}\max_{D_2} \quad & \mathbb{E}_{x\sim p_d}[\log D_2(G_E(x))] + \mathbb{E}_{\hat{x}\sim p_s}[\log(1 - D_2(G_E(\hat{x})))] \end{aligned} \tag{2}$$

where p_d and p_s indicate the distribution of original images and shot images, respectively. G_E represents the *encoder* part of the generator, as shown in Fig. 4.

Proposed by Salimans et al. [38], feature matching has been shown to reduce the training instability of GANs. In this study, we also employed feature matching as an extra loss. We used acti-

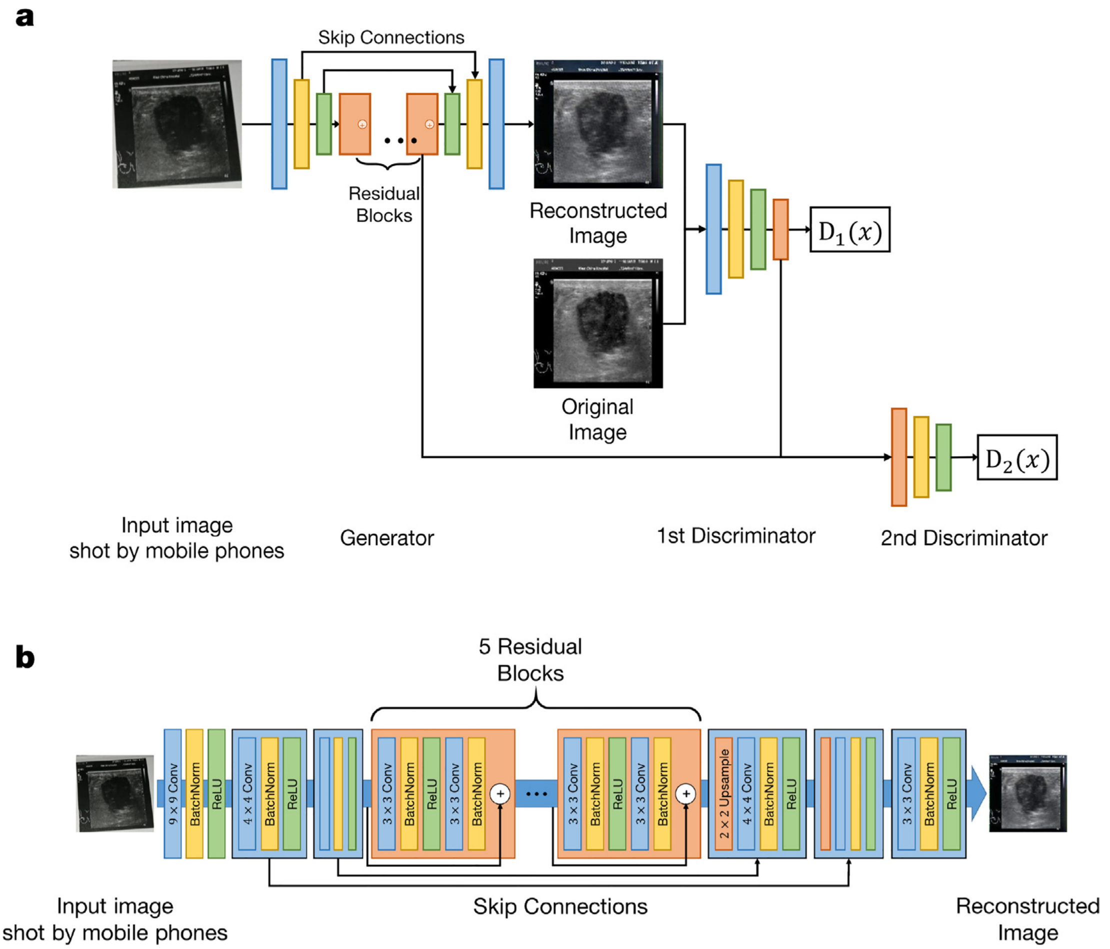


Fig. 4. The framework of the DeepRec subsystem. a. The subsystem consists of a generator and two discriminators. The input image shot by a mobile phone is fed into the generator, and the reconstructed image along with the original image are used to compute losses for training. b. The generator uses 3 downsampling layers and 5 residual blocks to encode features from the input image. The decoder is composed of 2 fractionally strided convolutional layers and 1 convolutional layers.

vations of the third and second layers of $D_1(x)$ and $D_2(x)$ to compute the feature matching loss. Formally, let f_1 and f_2 denote the functions that output the internal activations of the discriminators, the loss is defined as:

$$\|f_1(x) - f_1(G(\hat{x}))\|_2 + \|f_2(G_E(x)) - f_2(G_E(\hat{x}))\|_2 \tag{3}$$

where x and $\hat{x}$ denote the original and shot images.

The second part of $\mathcal{L}$ is the contextual loss $\mathcal{L}_{con}$, which is widely used in autoencoders. Previous studies showed that measuring the distance between the target and the generated images enables the generator to learn contextual information [39]. $\mathcal{L}_{con}$ is defined using the Euclidean distance between the original images x and the generated images $G(\hat{x}) : \mathcal{L}_{con} = \|x - G(\hat{x})\|_2$. To encourage spatial smoothness in the generated images, we also made use of *total variation regularizer* $\mathcal{L}_{tv}$, following previous studies on generated models [40,41].

Moreover, to encourage the reconstructed images to have feature representations similar to the original images, we employed the *perceptual loss* $\mathcal{L}_{per}$. We use the perceptual loss to measure high-level feature distance between the original images x and the generated images $G(\hat{x})$. In this study, $\mathcal{L}_{per}$ is defined as the Euclidean distance between high-level features of x and $G(\hat{x})$, based on the 16-layer VGG network [42] pretrained on the ImageNet dataset. We computed the loss at layer **relu3_3** of the standard 16-layer VGG network architecture.

2.3. The DeepCls subsystem

2.3.1. Image collection

In the development of the DeepCls subsystem, a large dataset of breast ultrasonography images was constructed. The images of breast mass were captured by sonographers during ultrasound examinations, using different kinds of color Doppler instrument, including Philips iU22, ATL HDI5000 and GE LOGIQ E9. The collected images were then labeled according to the descriptions by sonographers and pathological records. For each ultrasound examination, the ultrasonography images captured in this examination and the pathological reports of the same lesion were collected. If the patient underwent other examinations or surgical operations, corresponding reports such as an operation note and an immunohistochemical analysis report were also collected. Cases without pathological reports were not used in this study.

In this study, 18,125 images from October 2014 to August 2017 were collected. All of the ultrasonography images were obtained from West China Hospital, Sichuan University (WCH), People's Hospital of Deyang City (PHDY), and Chengdu Military General Hospital (CMGH).

2.3.2. Data annotation

The images were labeled by three tiers of sonographers and breast surgeons. In the first tier, ultrasound examination reports were collected, and sonographers and medical students from the three hospitals mentioned above analyzed the images, providing morphological descriptions and an initial diagnosis of the tissues. Images containing severe artifacts such as aspiration needles or those captured after surgical operations were excluded. In the second tier, each image that had passed the first tier was labeled by two breast surgeons individually referring to pathological reports. Diagnosis from the sonographers and the results of pathological examinations were recorded. In the third tier, if there were inconsistencies between the two annotators, the annotations were then evaluated by another professional and experienced breast cancer expert with over 20 years of experience, who excluded the images if the annotations were not accurate.

Each ultrasonography image was classified into two classes: malignant or non-malignant. Recognizing malignant changes is the most pressing concern in clinical applications. In most cases, only high-risk lesions are examined by pathological methods, many lesions are not included in pathological examination. For these images, the labels were determined referring to the Breast Imaging Reporting and Data System (BI-RADS), which is widely used in breast cancer examinations. In BI-RADS, ailments in breast mass are graded into 6 categories: BI-RADS 1 and 2 represent normal or benign tissues, while BI-RADS 3, 4, 5 and 6 indicate an increasing likelihood of malignancy. In this study, images with diagnoses of greater severity than BI-RADS 4B were labeled as malignant. The constructed dataset was named the Classification Set (Cls-Set).

2.3.3. Data partition and augmentation

In the Cls-Set, there were 18,125 images in total, 8146 images were malignant and 9979 were non-malignant. The Cls-Set showed a data imbalance where more images were non-malignant. This is consistent with clinical scenarios because there are fewer patients suffering from breast cancer than those with other ailments.

The dataset was split into a training set, a validation set and a testing set. 4/6 samples of the dataset were used for training, 1/6 for validation, and 1/6 for testing. Images of each class were uniformly distributed in each subset. In the Cls-Set, 12,083, 3020, and 3022 samples were used for training, validation, and testing, respectively. We also applied data augmentation to our datasets. To enhance the model stability in the proposed *R-S-A* procedure, in the training procedure, we processed the images through three steps. (1) We first selected a central crop from the original image while the crop size varied with a scale rate from 0.8 to 1.0 of the whole image. (2) The cropped image is resized to 256×256 pixels. (3) We applied a random horizontal flip and rotation from -15 degrees to +15 degrees to the image for training.

In the testing procedure, there was no central crop, flip and rotation. The original image was resized to 256×256 pixels for testing.

2.3.4. Algorithm evaluation

For the DeepCls subsystem, we used the accuracy and the area under the receiver operating characteristic curve (henceforth referred to as AUC) as evaluation criteria. We also used the sensitivity (also called the true positive rate or recall) and the specificity (also called the true negative rate). We define malignant as "positive" and non-malignant as "negative".

2.3.5. Image classification network

We utilized Google's Inception-v3 architecture [33] with an input size of 256×256 pixels to construct the network. The motivation behind utilizing this architecture was to use multi-scale convolutional kernels to learn features with different types, shapes, and sizes. The application of multi-scale convolution kernels is known as the Inception module, which was first introduced by the GoogLeNet. Feature maps from the previous layer are passed through several convolutional layers and pooling layers with different kernel sizes, such as 1×1, 3×3, and 5×5, the output feature maps are then concatenated. We replaced the final classification layer from the default network architecture with two fully-connected layers with 1024 and 2 neurons, respectively, equipping the neural network with a powerful classification capability and robustness. Because the size of our specific datasets was limited, we pre-trained the Inception-v3 network architecture on the well-known ImageNet dataset. Previous studies indicate that the parameters obtained by pre-training on a large dataset can be transferred to another application trained on a different dataset [43].

For the transfer learning, all networks were pre-trained on the ImageNet dataset which contains 1.2×10^6 training images, 5×10^4 validation images and 10^5 testing images. The images were classified into 1,000 classes. After pre-training, the networks were fine-tuned on our own datasets with ADADELTA [44] as the optimizer. Dropout [45] was applied to the first fully-connected layer of the network with a keep probability of 0.8. The mini-batch size was fixed at 12. The experiments were implemented using MXNet [46], an open-source scalable deep learning framework. It took around 30 h to fine-tune the network, using an Ubuntu 14.04 workstation with an Intel Xeon CPU, an NVIDIA Tesla K40m GPU and 64 GB available in RAM memory.

2.4. The DeepAti subsystem

2.4.1. Data preparation

The DeepAti subsystem was proposed to recognize missed diagnoses, reducing the false negative rate. To train and evaluate this subsystem, we constructed a large dataset Ati-Set, which was based on the Cls-Set and the Rec-Set. Each sample in the Ati-Set was a pair from the input features to the target.

We used the high-level features of the DeepCls subsystem as the input, and the target was one of "true negative" and "false negative". To construct the dataset, all images from the Cls-Set were fed into the DeepCls subsystem, samples that were classified as "negative" were picked out, and were further organized into two categories: true negative samples and false negative samples, in accordance with the classification labels. For each image, we used the features at the first fully-connected layer after global pooling as the input of the DeepAti subsystem. Formally, the input feature $\mathcal{X} \in \mathbb{R}^{1024}$ is a vector of length 1024, and the target $\mathcal{Y} \in \{0, 1\}$, where 0 represents false negative samples and 1 represents true negative samples.

2.4.2. Algorithm evaluation

To evaluate the performance of the DeepAti subsystem, we also used the criteria defined in Section 2.3.4. Each image from the Cls-Set passed through the DeepCls subsystem and the DeepAti subsystem successively, generating a final prediction. Then, the accuracy, the AUC value, the sensitivity, and the specificity were calculated. To evaluate the DeepAti subsystem on the manually shot photos, images from the Rec-Set were fed into the DeepRec subsystem, the DeepCls subsystem, and the DeepAti subsystem successively; the prediction results were used to calculate metrics.

In anomaly detection tasks [47,48], another criterion is widely used, which is FPR at 95% TPR. It measures the false positive rate (FPR) when the true positive rate (TPR) is equal to 95%. The false positive rate is calculated as FPR = FP/(FP + TN), while the true positive rate is calculated as TPR = TP/(TP + FN). In this study, a similar criterion was defined as the **FTNR at 95% TTNR**. Since only the images that are classified as "negative" are fed into the DeepAti subsystem, the **FTNR** can be interpreted as the probability that a false negative sample is misclassified as true negative, and the **TTNR** denotes the probability that a true negative sample is correctly classified as true negative. Detailed descriptions are presented in Fig. 5, and the **FTNR** and **TTNR** are calculated as follows:

$$\begin{aligned} \mathrm{FTNR} &= \tfrac{\mathrm{FTN}}{\mathrm{FTN}+\mathrm{TFN}} \\ \mathrm{TTNR} &= \tfrac{\mathrm{TTN}}{\mathrm{TTN}+\mathrm{FFN}} \end{aligned}. \quad (4)$$

2.4.3. False negative identification network

The objective of the DeepAti subsystem is to identify false negative samples; in this study, we used the high-level features of the DeepCls subsystem as input to perform the identification. The output of the DeepCls subsystem can be regarded as the confidence of the classifier, and the confidence can also be represented by intermediate layer features. In previous studies [47,49], most approaches used the output probabilities to recognize inaccurate classification. The motivation of this study followed a hypothesis that the extracted features of the classifier correspond to different distributions for True Negative and False Negative, and we built a GAN to identify the difference. The DeepAti subsystem was constructed based on the framework of standard GANs, as shown in Fig. 6.

The generator contained four fully-connected layers with 128, 256, 512 and 1024 neurons, and the discriminator contained three fully-connected layers with 512, 256, and 1 neurons. In the training process, only true negative samples were used. As proposed in [37], if the input x of the discriminator is a sample from real data, the output $D(x)$ should be close to 1; if the input $\hat{x}$ is from generated data, the output $D(\hat{x})$ should be close to 0. When the training converges, the discriminator can hardly recognize fake samples, and the outputs $D(x)$ and $D(\hat{x})$ are both close to 0.5. In this study, we used true negative samples to train the GAN. In the evaluation progress, the high-level features were fed into the discriminator. If $D(x)$ was close to 0.5, the input x was regarded as true negative; if $D(x)$ was not close to 0.5, the input x was regarded as false negative.

3. Results

3.1. Performance of the model

The results of different settings of our system are presented in Fig. 7 and Table 2. We evaluated our system with four settings: (1) DeepCls; (2) DeepRec + DeepCls; (3) DeepCls + DeepAti; and (4) DeepRec + DeepCls + DeepAti. (1) and (3) are evaluated on the high-quality Cls-Set, (2) and (4) are evaluated on the low-quality Rec-Set.

3.1.1. DeepCls

We evaluated our system for recognizing breast cancer based on ultrasonography images. We first evaluated the DeepCls subsystem on the testing set of the high-quality Cls-Set. In the classification of malignant and non-malignant samples, the DeepCls subsystem achieved an accuracy of 94.51% (95%CI, 93.63–95.29%), with a sensitivity of 94.04% (95%CI, 92.65–95.24%) and a specificity of 94.89% (95%CI, 93.72–95.90%). The ROC curve is presented in Fig. 7 and the area under the ROC curve was 0.982 (95%CI, 0.978–0.986).

Following the *R-S-A* procedure in a real-world scenario, the system would perform diagnosis based on low-quality photos. We thus next evaluated the DeepCls subsystem on the testing set of the Rec-Set, which was composed of low-quality photos of the images in the Cls-Set. Since the image quality was much lower than that of the Cls-Set, as shown in Fig. 8, and the DeepCls subsystem was not trained with such data, the performance droped markedly. The DeepCls subsystem achieved an accuracy of 84.48% (95%CI, 83.14–85.75%), with a sensitivity of 76.60% (95% CI, 74.26–78.83%) and a specificity of 90.92% (95%CI, 89.44–92.26%). The area under the ROC curve was 0.920 (95%CI, 0.910–0.929).

3.1.2. DeepRec + DeepCls

The second stage was to evaluate the DeepRec subsystem on the testing set of the Rec-Set together with the DeepCls subsystem. Employing the DeepRec subsystem, the performance of our system increased significantly. The system achieved an accuracy of 89.34% (95%CI, 88.91–90.42%), which was 4.43% higher than the DeepCls subsystem. The sensitivity was 87.71% (95%CI, 85.85–89.41%), which increased by more than 10%. The specificity was 90.68% (95%CI, 89.18–92.03%) and the AUC value was 0.953 (95%CI, 0.946–0.960).

We also compared our approach with several encoder-decoder-based image-to-image models and image denoising algorithms, the results are presented in Table 1,2. Our approach outperforms all other methods.

3.1.3. DeepCls + DeepAti

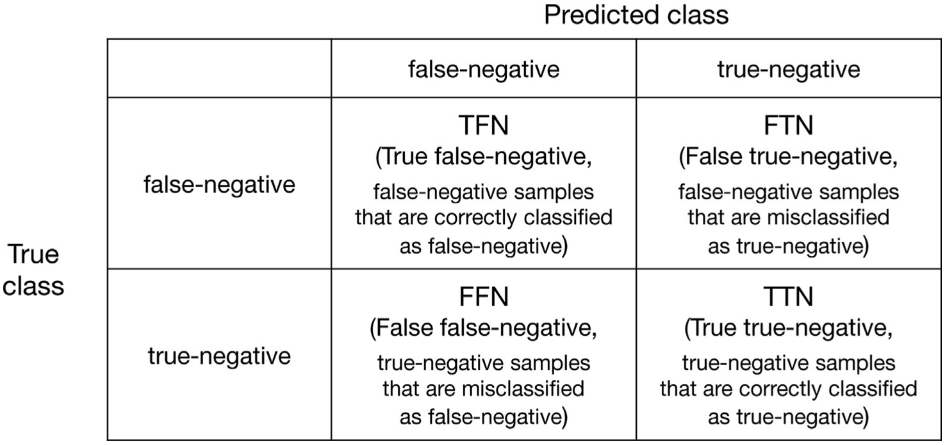
The third stage was to evaluate the DeepAti subsystem. We first demonstrate its performance together with the DeepCls subsystem on the testing set of the Cls-Set. The DeepAti subsystem aims at reducing the false negative rate, and can explicitly trade specificity for sensitivity according to different settings of the TTNR (defined in Section 2.4.2). Setting the TTNR at 95%, the sensitivity of our system increased up to 97.20% (95%CI, 96.18–98.01%). The specificity was 89.90% (95%CI, 88.35–91.30%) and the accuracy was 93.18% (95%CI, 92.23–94.06%). Although there were small drops in the specificity and accuracy, the sensitivity increased markedly, which could be valuable in real-world applications. The AUC value is 0.981 (95%CI, 0.977–0.986). If the TTNR were set to a higher threshold, the specificity and accuracy would be higher, and the sensitivity would be lower. With a TTNR of 98%, the sensitivity was 96.17% (95%CI, 95.01–97.13%), the specificity was 92.24% (95%CI, 90.85–93.48%), and the accuracy was 94.01% (95%CI, 93.10–94.83%). The AUC value was 0.982 (95%CI, 0.978–0.986). The DeepAti subsystem achieved an FTNR of 46.91% at 95% TTNR.

3.1.4. DeepRec + DeepCls + DeepAti

The final stage was to evaluate the entire system, consisting the DeepRec subsystem, the DeepCls subsystem, and the DeepAti subsystem. We demonstrate the performance on the testing set of the low-quality Rec-Set. Setting the TTNR at 95%, the sensitivity of our system was 92.42% (95%CI, 90.88–93.77%). The specificity was 85.45% (95%CI, 83.66–87.11%), the accuracy was 88.58% (95%CI, 87.40–89.70%), and the AUC value was 0.952 (95%CI, 0.949–0.959). Setting the TTNR at 98%, our system achieved a sensitivity of 90.58% (95%CI, 88.90–92.08%), with a specificity of 87.97% (95% CI, 86.31–89.50%) and an accuracy of 89.15% (95%CI, 87.98–90.23%). The AUC value was 0.952 (95%CI, 0.949–0.960) and the FTNR was 61.68% at 95% TTNR.

3.2. Comparison of the system with human experts

An independent test set of 100 breast ultrasonography images (50 malignant and 50 non-malignant) was used to compare our AI system with human experts. All images were printed on paper and 9 human experts were instructed to perform diagnosis based on the ultrasonography images. The 9 experts were composed of three professors with senior titles, three doctors with intermediate titles and three doctors with junior titles. In each group, the 3 experts came from central city, prefecture-level city and county-level city, respectively. The results of the proposed system and human experts on this test set are presented in Fig. 7 and Table 3.



		Predicted class	
		false-negative	true-negative
True class	false-negative	TFN (True false-negative, false-negative samples that are correctly classified as false-negative)	FTN (False true-negative, false-negative samples that are misclassified as true-negative)
	true-negative	FFN (False false-negative, true-negative samples that are misclassified as false-negative)	TTN (True true-negative, true-negative samples that are correctly classified as true-negative)

Fig. 5. The definition of our confusion matrix of the DeepAti subsystem. In our work, only the images that are classified as "negative" are fed into the DeepAti subsystem. Therefore, the DeepAti subsystem aims at distinguishing false negative sample from true negative samples. Here, the false-negative samples that are correctly classified by the subsystem as false-negative are defined as true false-negative (TFN) samples, the true-negative samples that are misclassified as false-negative are defined as false false-negative (FFN) samples, the false-negative samples that are misclassified as true-negative are defined as false true-negative (FTN) samples, and the true-negative samples that are correctly classified as true-negative are defined as true true-negative (TTN) samples.

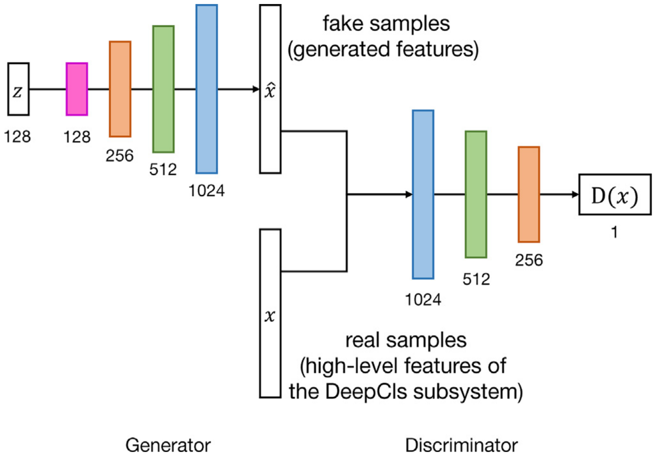


Fig. 6. The framework of the DeepAti subsystem. The subsystem is based on the GAN architecture. In the training process, the generator attempts to reconstruct features which are close to the high-level features of the DeepCls subsystem, and the discriminator attempts to recognize the reconstructed features. In the inference process, the discriminator is used to distinguish false negative samples from true negative samples.

Our system achieved an accuracy of 87%, outperforming six of the nine experts.

The sensitivities and specificities of the experts are plotted on the ROC curve of our system, as presented in Fig. 7. The performance of our system was comparable of human experts.

Furthermore, the averaged time per image used by the human experts was 30–50 s. In this case, only 100 samples were used for the comparison to avoid overwork for human experts. For the proposed system, it only took around 5 s to process each image. The analysis by the system was very efficient.

3.3. Application of the system in real-world scenarios

The proposed system for analyzing breast ultrasonography images was integrated with a cloud computing platform for automated breast cancer diagnosis and report generation. The system is accessible with smartphones, the user only needs to upload a photo of the ultrasound report to obtain diagnostic results. As presented in Fig. 1, the proposed diagnostic task includes four stages: photo validation, report pre-processing, diagnosis and report generation.

3.3.1. Photo validation

The uploaded photos are of the breast ultrasound reports of patients. If a user attempts to upload an unstandardized picture, the system would fail to analyze the ultrasonography images. Therefore, the system should only receive proper ultrasound reports to ensure the stability of the whole system and avoid unpredictable results. We employed the system with a classification model to reject fake or murky images, which is a fine-tuned CNN model that can distinguish ultrasound reports from other ordinary photos. The dataset used to build this model contains ultrasound reports and ordinary photos from the MSCOCO Dataset [52].

3.3.2. Report pre-processing

Each ultrasound report contains several ultrasonography images. In this stage, we used a well-designed canny edge detector to find contours and traverse each contour to obtain the bounding boxes of the images. Subsequently, all ultrasonography images are resized to 256×256 pixels. Normally, not all ultrasonography images in an ultrasound report are from breast mass, there may be images from other parts, such as thyroid and kidney. An extra classifier was thus used to recognize the images from breast mass.

3.3.3. Diagnosis

At this stage, each ultrasonography image passed through the DeepRec subsystem, the DeepCls subsystem and the DeepAti subsystem successively. For the DeepAti subsystem, we set the TTNR at 95%. We used an NVIDIA Tesla K40m GPU to perform the computation, speeding up the average response time for each report to less than 5 s.

3.3.4. Report generation

Finally, the proposed system creates a report based on the results. The classification results and probabilities are listed image-by-image. The ultrasonography images are organized the same as the original ultrasound report, diagnosis results and output probabilities are listed following the ultrasonography images.

The system was deployed online as a mobile application. Since July 2018, the web service API of our system has been open to third

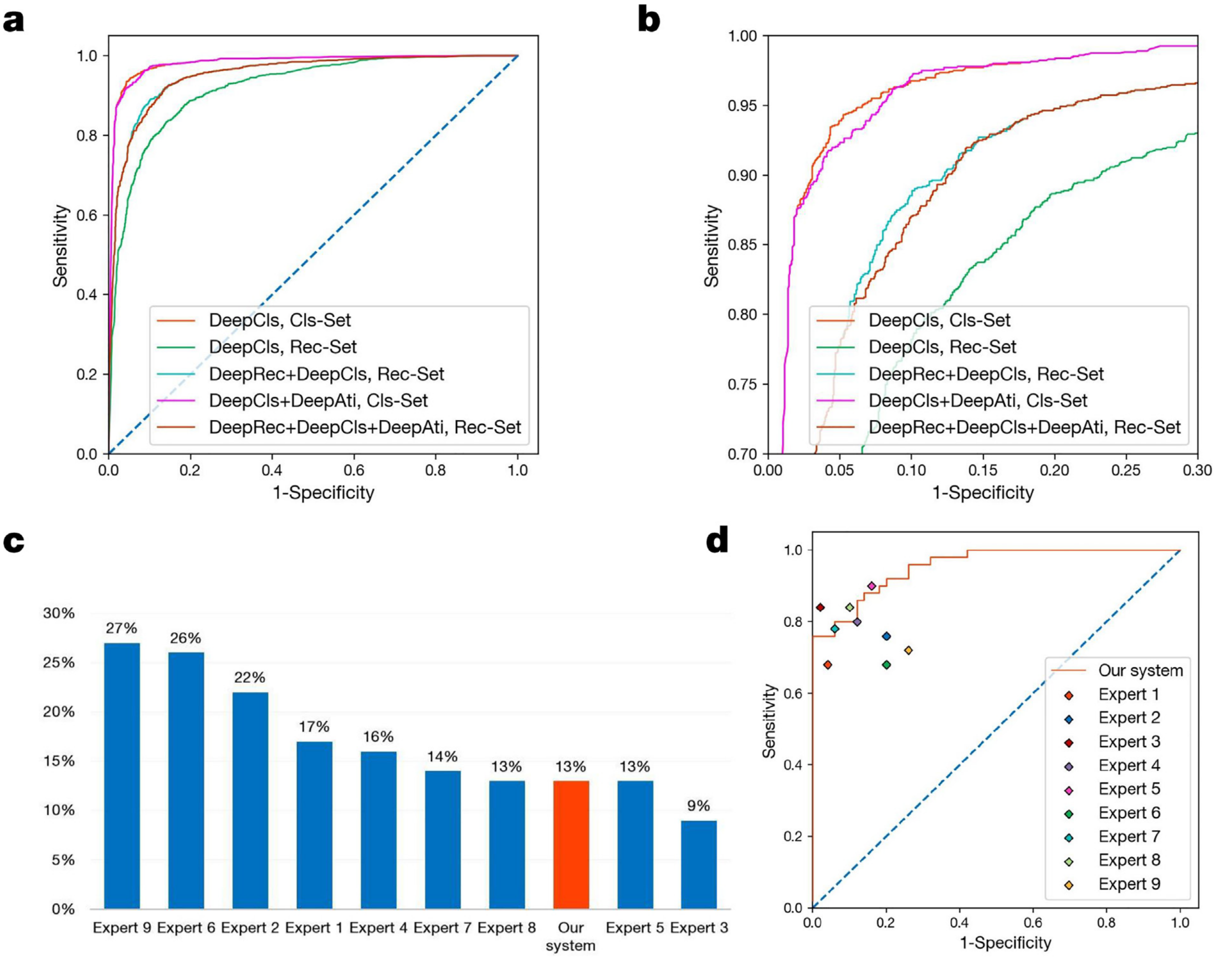


Fig. 7. Statistic results. a. The receiver operating characteristic curve (ROC) of our system, with different settings. b. Zoomed area of the ROC curve. c. Comparison of our system with human experts in error rates. d. Comparison of our system with human experts in the ROC curve.

parties. From July 2018 to October 2019, the number of external accesses was 21729, being accessed by over 10,000 users.

4. Discussion

Automated breast cancer analysis algorithms are desirable for breast cancer diagnosis and treatment. Unlike pathological examinations which require biopsies or immunohistochemical examinations, breast cancer is detectable by imaging modalities. Ultrasound is a widely used imaging technology, and is nonradiative and convenient. In a breast ultrasound examination, several images of the breast mass are captured, and sonographers and breast surgeons perform an initial diagnosis based on the ultrasonography images.

Owing to its low cost, ultrasound equipment is widely used in medical institutions, from community clinics to central hospitals. However, there are far from enough sonographers and breast surgeons with the ability to analyze ultrasonography images. The diagnosis by human doctors is highly subjective. To solve this problem, in this study, we propose an automated breast cancer diagnosis system for the recognition of malignant changes in breast ultrasonography images. To make the system convenient to use, it is deployed on mobile phones and is accessible for users everywhere. The user only needs to undergo an ultrasound examination at a medical institution, use a mobile phone to take a photo of the ultrasound report, and upload it into our system; diagnostic results will then be returned to the user within seconds. The results could be used as an initial screening or a second opinion of breast cancer diagnosis.

Following the *R-S-A* procedure: *Report*, *Shoot*, and *Analysis*, the proposed system is composed of three subsystems. In the first step, to reduce the side-effects caused by noise in the *Shoot* process, the DeepRec subsystem is employed to reconstruct high-quality images from low-quality photos. Samples reconstructed by the DeepRec subsystem are shown in Fig. 8. In the second step, the DeepCls subsystem takes the reconstructed image as input and classifies it as malignant or non-malignant. As the main component of our system, the DeepCls subsystem is based on convolutional neural networks and multi-scale convolutional kernels. In the third step, the DeepAti subsystem is employed to recognize missed diagnoses and reduce the false negative rate. All three subsystems work in a cascade manner. To train and evaluate the subsystems, three large datasets were constructed. To reduce the side effects of over-fitting and increase the generalization ability of our system, 18,125 breast ultrasonography images of a large number of patients were collected from three hospitals. Moreover, reliable annotation is crucial for the training of neural networks. To ensure the correctness of our datasets, all of the images were labeled by three tiers of sonographers and breast surgeons.

The results demonstrate that the system can be trained using large-scale datasets and achieve performance comparable to that of human doctors. With the *R-S-A* procedure, the system can perform initial screening and diagnosis of breast cancer based on ultrasound reports. This automated system could be used in some

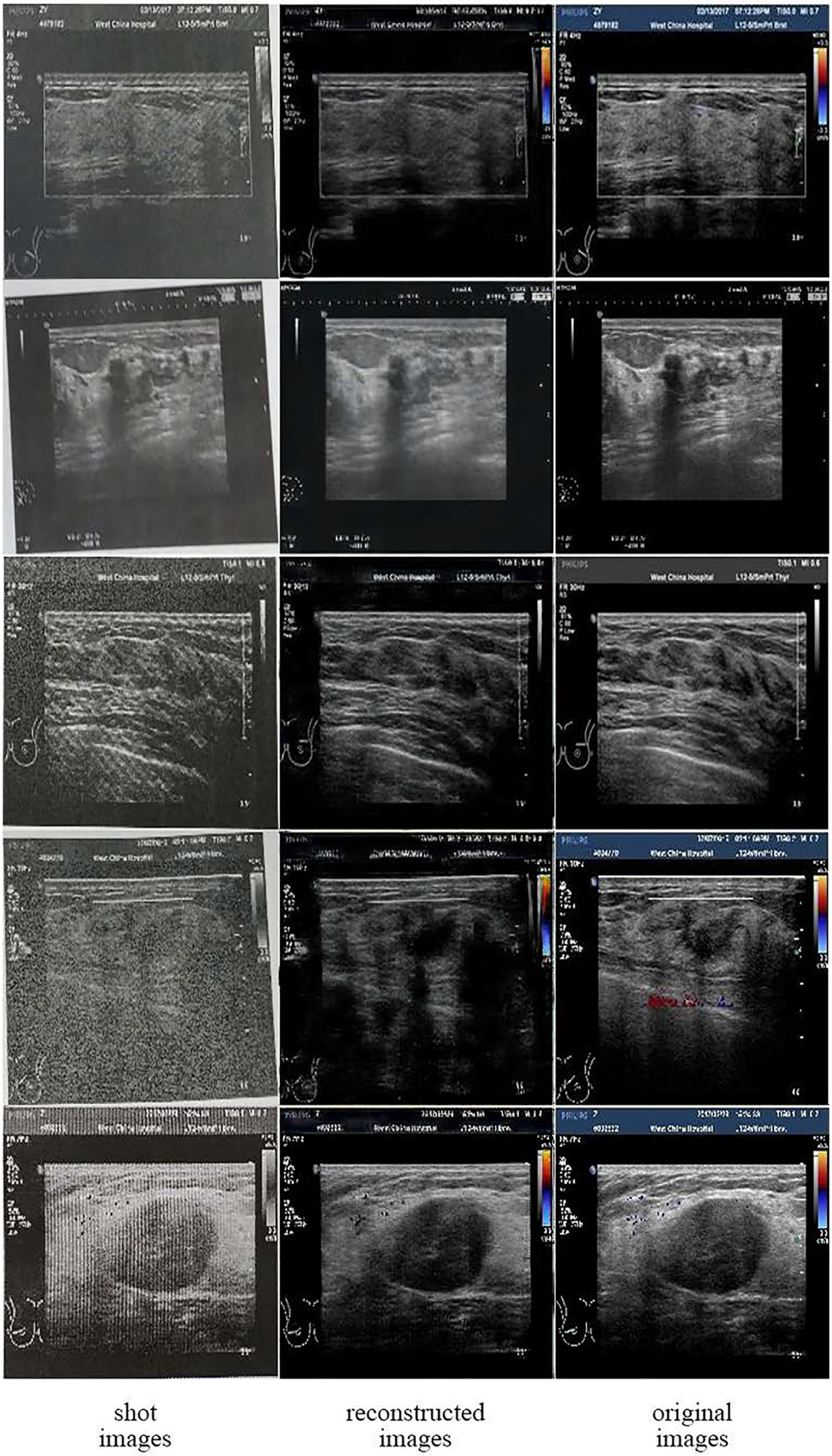


Fig. 8. Reconstructed samples of the DeepRec subsystem. First column: images shot by mobile phones. Second column: images reconstructed by the DeepRec subsystem. Third column: original images from the Cls-Set.

Table 1
The DeepRec subsystem compared with other methods.

Method	Accuracy	Sensitivity	Specificity
Encoder-decoder	79.95%	87.12%	74.08%
U-Net [50]	87.39%	87.49%	87.31%
EPLL [51]	77.96%	66.30%	87.49%
Ours	**89.34%**	**87.71%**	**90.68%**

Table 2
Results of the proposed system.

System settings	TTNR	Dataset	Accuracy	Sensitivity	Specificity	AUC
DeepCls	–	Cls-Set	**94.51%**	94.04%	**94.89%**	**0.982**
	–	Rec-Set	84.48%	76.60%	90.92%	0.920
DeepRec + DeepCls	-	Rec-Set	89.34%	87.71%	90.68%	0.953
DeepCls + DeepAti	0.95	Cls-Set	93.18%	**97.20%**	89.90%	0.981
	0.98		94.01%	96.17%	92.24%	**0.982**
DeepRec + DeepCls + DeepAti	0.95	Rec-Set	88.58%	92.42%	85.45%	0.952
	0.98		89.15%	90.58%	87.97%	0.952

Table 3
Results of the proposed system and human experts.

	Accuracy	Sensitivity	Specificity
Expert 1	83.00%	68.00%	98.00%
Expert 2	78.00%	76.00%	80.00%
Expert 3	91.00%	84.00%	98.00%
Expert 4	84.00%	80.00%	88.00%
Expert 5	87.00%	90.00%	84.00%
Expert 6	74.00%	68.00%	80.00%
Expert 7	86.00%	78.00%	94.00%
Expert 8	87.00%	84.00%	90.00%
Expert 9	73.00%	72.00%	74.00%
Averaged	82.56%	77.78%	87.33%
Our system	87.00%	86.00%	88.00%

medical institutions both for patients and doctors. In addition, because the system can have multiple operating points and different configurations of the subsystems, its performance can be tuned to match different clinical applications. In this study, the system is tuned to have a high sensitivity for a screening setting.

There are several limitations of the proposed system. Firstly, the system can only recognize malignant changes in the breast mass, while other tasks such as the detection of solid nodules are also very important in clinical settings. In the future, we plan to annotate the datasets with more fine-grained classes and equip the system with extended functions. Secondly, the neural networks learn features from data automatically, the features are not interpretable. However, human doctors are concerned about the features leading to the diagnosis. Understanding what are the features that a deep neural networks uses to make prediction is important both in the deep learning community and the medical image analysis community. Thirdly, the system is deployed on mobile phones. The user only needs to take a photo of the ultrasound report and upload it to our system to get diagnosis results. Although this is a convenient way for patients and doctors to use, it is infeasible in some countries. In China, the ultrasound report includes several representative ultrasonography images. However, in some countries, the ultrasound report only contains descriptions and conclusions of the sonographer, no ultrasonography images are included. We hope that the users in these countries could also use our system. For example, using only the DeepCls and the DeepAti subsystems, our system can be applied on the electronic images taken by ultrasound equipments. In this case, our system can be used in most countries. This could facilitate the screening of breast cancer and make the medical resources in developed areas available to remote and low-resource areas.

5. Conclusion

In this study, an automated computer-aided breast cancer diagnosis system for breast ultrasonography images is proposed to improve the efficiency and reliability of breast cancer screening. The system is deployed on mobile phones, takes a photo of the ultrasound report as input and recognizes malignant changes. Potentially, the system can be modified for usage in many other scenarios. Three subsystems working in a cascade manner are proposed based on deep neural networks for the diagnosis. To train and evaluate the system, three large-scale annotated ultrasonography breast image datasets are constructed. The performance of our system is comparable to that of human experts, making it applicable in real-world clinical scenarios. The main limitation of the proposed system is the lack of interpretability, therefore, the future work includes identification of ultrasonic characterizations and the classification of more types of diseases.

CRediT authorship contribution statement

Xiaofeng Qi: Methodology, Software, Data-curation, Visualization, Writing-original-draft, Writing-review-editing. **Fasheng Yi:** Methodology, Software. **Lei Zhang:** Methodology, Writing-review-editing. **Yao Chen:** Resources, Data-curation, Writing-original-draft. **Yong Pi:** Methodology, Software. **Yuanyuan Chen:** Methodology, Software. **Jixiang Guo:** Methodology, Software. **Jianyong Wang:** Methodology, Software. **Quan Guo:** Methodology, Software. **Jilan Li:** Methodology, Software, Writing-original-draft. **Yi Chen:** Resources. **Qing Lv:** Conceptualization, Methodology, Data-curation, Writing-review-editing, Supervision. **Zhang Yi:** Conceptualization, Methodology, Writing-review-editing, Supervision.

Declaration of Competing Interest

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

Acknowledgements

This work was supported by the National Major Science and Technology Projects of China [Grant No. 2018AAA0100201] and the National Natural Science Foundation of China [Grant No.61806136] and the National Natural Science Foundation of China [Grant No.61806136].

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (6) (2018) 394–424.

[2] C.H. Lee, Screening mammography: proven benefit, continued controversy, Radiol. Clin. North Am. 40 (3) (2002) 395–407, https://doi.org/10.1016/s0033-8389(01)00015-x.

[3] A.S.Y. Leong, Z. Zhuang, The changing role of pathology in breast cancer diagnosis and treatment, Pathobiology 78 (2) (2011) 99–114, https://doi.org/10.1159/000292644.

[4] M. Veta, J.P. Pluim, P.J. Van Diest, M.A. Viergever, Breast cancer histopathology image analysis: a review, IEEE Trans. Biomed. Eng. 61 (5) (2014) 1400–1411.

[5] M.L. Giger, N. Karssemeijer, J.A. Schnabel, Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer, Annu. Rev. Biomed. Eng. 15 (2013) 327–357, https://doi.org/10.1146/annurev-bioeng-071812-152416.

[6] G. Carneiro, J. Nascimento, A.P. Bradley, Automated analysis of unregistered multi-view mammograms with deep learning, IEEE Trans. Med. Imag. 36 (11) (2017) 2355–2365.

[7] T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Mérida, C.I. Sánchez, R. Mann, A. den Heeten, N. Karssemeijer, Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.

[8] Z. Zhuang, Y. Kang, A.N. Joseph Raj, Y. Yuan, W. Ding, S. Qiu, Breast ultrasound lesion classification based on image decomposition and transfer learning, Med. Phys. 47 (12) (2020) 6257–6269. doi:10.1002/mp.14510. .

[9] A. Jalalian, S.B. Mashohor, H.R. Mahmud, M.I.B. Saripan, A.R.B. Ramli, B. Karasfi, Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review, Clin. Imag. 37 (3) (2013) 420–426.

[10] S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W.-C. Bang, Y.-K. Seong, A deep learning framework for supporting the classification of breast lesions in ultrasound images, Phys. Med. Biol. 62 (19) (2017) 7714.

[11] Q. Huang, B. Hu, F. Zhang, Evolutionary optimized fuzzy reasoning with mined diagnostic patterns for classification of breast tumors in ultrasound, Inf. Sci. 502 (2019) 525–536, https://doi.org/10.1016/j.ins.2019.06.054.

[12] Q. Huang, Y. Chen, L. Liu, D. Tao, X. Li, On combining biclustering mining and adaboost for breast tumor classification, IEEE Trans. Knowl. Data Eng. 32 (4) (2020) 728–738, https://doi.org/10.1109/TKDE.2019.2891622.

[13] Y.-W. Chang, Y.-R. Chen, C.-C. Ko, W.-Y. Lin, K.-P. Lin, A novel computer-aided-diagnosis system for breast ultrasound images based on bi-rads categories, Appl. Sci. 10 (2020) 1830, https://doi.org/10.3390/app10051830.

[14] Q. Huang, Y. Huang, Y. Luo, F. Yuan, X. Li, Segmentation of breast ultrasound image with semantic classification of superpixels, Med. Image Anal. 61 (2020) , https://doi.org/10.1016/j.media.2020.101657 101657.

[15] J. Xing, C. Chen, Q. Lu, X. Cai, A. Yu, Y. Xu, X. Xia, Y. Sun, J. Xiao, L. Huang, Using bi-rads stratifications as auxiliary information for breast masses classification in ultrasound images, IEEE J. Biomed. Health Inf. 25 (6) (2021) 2058–2070, https://doi.org/10.1109/JBHI.2020.3034804.

[16] N.V. Boulgouris, Z.X. Chi, Gait recognition using radon transform and linear discriminant analysis, IEEE Trans. Image Process. 16 (3) (2007) 731–740.

[17] C.H. Park, H. Park, A comparison of generalized linear discriminant analysis algorithms, Pattern Recogn. 41 (3) (2008) 1083–1097.

[18] C.J. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discovery 2 (2) (1998) 121–167.

[19] J.C. Platt, Fast training of support vector machines using sequential minimal, Optimization (1999).

[20] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (10) (2000) 906–914.

[21] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[22] R. Salakhutdinov, A. Mnih, G. Hinton, Restricted boltzmann machines for collaborative filtering, in: Proceedings of the 24th International Conference on Machine Learning, ACM, 2007, pp. 791–798.

[23] H. Larochelle, Y. Bengio, Classification using discriminative restricted boltzmann machines, in: Proceedings of the 25th International Conference on Machine Learning, ACM, 2008, pp. 536–543.

[24] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control, Signals Syst. 2 (4) (1989) 303–314.

[25] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324, https://doi.org/10.1109/5.726791.

[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, International Conference on Neural Information Processing Systems (2012) 1097–1105, https://doi.org/10.1145/3065386.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, IEEE Conference on Computer Vision and Pattern Recognition 2015 (2015) 1–9, https://doi.org/10.1109/cvpr.2015.7298594.

[28] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456. .

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, https://doi.org/10.1109/cvpr.2016.90.

[30] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252, https://doi.org/10.1007/s11263-015-0816-y.

[32] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, IEEE Conference on Computer Vision and Pattern Recognition 2009 (2009) 248–255, https://doi.org/10.1109/cvprw.2009.5206848.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, https://doi.org/10.1109/cvpr.2016.308.

[34] M.D. Zeiler, D. Krishnan, G.W. Taylor, R. Fergus, Deconvolutional networks, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2528–2535.

[35] M.D. Zeiler, G.W. Taylor, R. Fergus, et al., Adaptive deconvolutional networks for mid and high level feature learning., in: ICCV, vol. 1, 2011, p. 6. .

[36] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408, URL:http://dl.acm.org/citation.cfm?id=1756006.1953039.

[37] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the 27th International Conference on Neural Information Processing Systems – Volume 2, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, pp. 2672–2680. URL:http://dl.acm.org/citation.cfm?id=2969033.2969125. .

[38] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, X. Chen, Improved techniques for training gans, in: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29, Curran Associates Inc., 2016, pp. 2234–2242. URL: http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf. .

[39] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017 (2017) 5967–5976.

[40] H.A. Aly, E. Dubois, Image up-sampling using total-variation regularization with a new observation model, IEEE Trans. Image Process. 14 (10) (2005) 1647–1659, https://doi.org/10.1109/TIP.2005.851684.

[41] A. Mahendran, A. Vedaldi, Understanding deep image representations by inverting them, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015 (2015) 5188–5196, https://doi.org/10.1109/CVPR.2015.7299155.

[42] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556. .

[43] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp 3320–3328.

[44] M.D. Zeiler, Adadelta: an adaptive learning rate method, Comput. Sci. .

[45] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, Comput. Sci. 3 (4) (2012) 212–223.

[46] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, Z. Zhang, Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems, arXiv preprint arXiv:1512.01274. .

[47] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, in: International Conference on Learning Representations, 2018. URL:https://openreview.net/forum?id=H1VGkIxRZ. .

[48] K. Lee, H. Lee, K. Lee, J. Shin, Training confidence-calibrated classifiers for detecting out-of-distribution samples, in: International Conference on Learning Representations, 2018. URL:https://openreview.net/forum?id=ryiAv2xAZ. .

[49] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, in: 5th International Conference on Learning Representations, 2017, URL:https://openreview.net/forum?id=Hkg4TI9xl.

[50] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28..

[51] D. Zoran, Y. Weiss, From learning models of natural image patches to whole image restoration, International Conference on Computer Vision (2011) 479–486, https://doi.org/10.1109/ICCV.2011.6126278.

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. Lawrence Zitnick, Microsoft coco: common objects in context, European Conference on Computer Vision (2014), https://doi.org/10.1007/978-3-319-10602-1_48.

Xiaofeng Qi is currently working toward the Ph.D degree with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. His current research interests include deep neural networks and medical image analysis.

Fasheng Yi received a Ph.D. degree in computer applications from the University of Electronic Science and Technology of China, Chengdu, China, in 2008. Currently, he is an associate professor at College of Computer Science, Chengdu University, Chengdu, China. His current research interests include Artificial Intelligence Systems Development and Big Data.

Lei Zhang received the B.S. and M.S. degrees in mathematics and the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2002, 2005, and 2008, respectively. She was a Post-Doctoral Research Fellow with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, from 2008 to 2009. She is currently a Professor with Sichuan University, Chengdu. Her current research interests include theory and applications of neural networks based on neocortex computing and big data analysis methods by very deep neural networks. She is currently serving as an Associate Editor for the IEEE Transactions on Cognitive and Developmental Systems (2017-) and the IEEE Transactions on Neural Networks and Learning Systems (2019-). She serves as the Chair of the Chengdu Chapter of the IEEE Computational Intelligence Society.

Yao Chen received the Master's Degree from the Department of Galactophore Surgery, West China Hospital, Sichuan University, in 2020. She is currently a resident doctor with the Department of Galactophore Surgery, West China Hospital, Sichuan University, Chengdu, China. Her research interests include breast cancer diagnosis, ultrasound images analysis and artificial intelligence.

Yong Pi received the B.S. degree in Computer Science and Technology, Sichuan University in 2017. Now he is working on the process of Ph.D. degree in Computer Technology, and is also a research assistant at the Machine Intelligence Laboratory, College of Computer Science, Sichuan University. His research interests include neural networks and computer-assisted medical applications.

Yuanyuan Chen received the Ph.D. degree from the Department of Computer Science, Sichuan University, in 2015. She is currently an Assistant Professor with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. Her research interests include neural networks and computer-aided diagnosis.

Jixiang Guo received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, in 2011. She is currently an Assistant Professor with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. Her research interests include neural networks and computer-assisted medical applications.

Jianyong Wang received his Ph.D. degree in computer science from the Sichuan University in 2018. He is currently a Post-Doctoral Research Fellow in Machine Intelligence Lab at College of Computer Science, Sichuan University, Chengdu, China. His current research interests include Neural Networks and Big Data.

Quan Guo received a Ph.D. degree in machine intelligence from Sichuan University in 2017. He was a postdoc (Research Associate) at Tulane University and Michigan State University during 2019 and 2020. Currently, he is an associate professor at Sichuan University, Department of Artificial Intelligence, College of Computer Science. His current research interests are neural networks, deep learning, and structured learning.

Jilan Li received the Master's Degree from the Department of Computer Science, Sichuan University, in 2020. Her current research interests include deep neural networks and medical image analysis.

Yi Chen is currently a Research Assistant with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. Her current research interests include deep neural networks and medical image analysis.

Qing Lv received the Bachelor's Degree of Medicine from West China Hospital, Sichuan University, in 1985. She is currently a chief physician and the director of the Department of Galactophore Surgery, West China Hospital, Sichuan University, Chengdu, China. Her research interests include breast cancer diagnosis, operation and artificial intelligence.

Zhang Yi received a Ph.D. degree in mathematics from the Institute of Mathematics, The Chinese Academy of Science, Beijing, China, in 1994. Currently, he is a Professor at the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. He is the co-author of three books: Convergence Analysis of Recurrent Neural Networks (Kluwer Academic Publishers, 2004), Neural Networks: Computational Models and Applications (Springer, 2007), and Subspace Learning of Neural Networks (CRC Press, 2010). He was an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems (2009–2012), and he is an Associate Editor of IEEE Transactions on Cybernetics (2014-). His current research interests include Neural Networks and Big Data. He is a fellow of IEEE.