Fundamental Research 5 (2025) 880–889

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

Artificial intelligence-based diagnosis of breast cancer by mammography microcalcification

Qing Lin[a,b,1], Wei-Min Tan[b,1], Jing-Yu Ge[a,c,1], Yan Huang[d,1], Qin Xiao[d], Ying-Ying Xu[e], Yi-Ting Jin[f], Zhi-Ming Shao[a,c], Ya-Jia Gu[d,*], Bo Yan[b,*], Ke-Da Yu[a,c,*]

[a] *Cancer Institute, Fudan University Shanghai Cancer Center, Shanghai 200032, China*
[b] *School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University, Shanghai 200438, China*
[c] *Department of Breast Surgery, Fudan University Shanghai Cancer Center, Shanghai 200032, China*
[d] *Department of Radiology, Fudan University Shanghai Cancer Center, Shanghai 200032, China*
[e] *Department of Breast Surgery, The First Affiliated Hospital of China Medical University, Shenyang 110001, China*
[f] *Department of General Surgery, Huashan Hospital, Fudan University, Shanghai 200040, China*

ARTICLE INFO

Article history:
Received 1 December 2022
Received in revised form 7 April 2023
Accepted 19 April 2023
Available online 18 June 2023

Keywords:
Microcalcification detection
Classification
Mammography
Deep learning
Breast cancer

ABSTRACT

Mammography is the mainstream imaging modality used for breast cancer screening. Identification of microcalcifications associated with malignancy may result in early diagnosis of breast cancer and aid in reducing the morbidity and mortality associated with the disease. Computer-aided diagnosis (CAD) is a promising technique due to its efficiency and accuracy. Here, we demonstrated that an automated deep-learning pipeline for microcalcification detection and classification on mammography can facilitate early diagnosis of breast cancer. This technique can not only provide the classification results of mammography, but also annotate specific calcification regions. A large mammography dataset was collected, including 4,810 mammograms with 6,663 microcalcification lesions based on biopsy results, of which 3,301 were malignant and 3,362 were benign. The system was developed and tested using images from multiple centers. The overall classification accuracy values for discriminating between benign and malignant breasts were 0.8124 for the training set and 0.7237 for the test set. The sensitivity values of malignant breast cancer prediction were 0.8891 for the training set and 0.7778 for the test set. In addition, we collected information regarding pathological sub-type (pathotype) and estrogen receptor (ER) status, and we subsequently explored the effectiveness of deep learning-based pathotype and ER classification. Automated artificial intelligence (AI) systems may assist clinicians in making judgments and improve their efficiency in breast cancer screening, diagnosis, and treatment.

1. Introduction

Breast cancer is the most commonly diagnosed cancer worldwide. According to the recent global cancer burden data, 2.3 million new cases were diagnosed in 2020 [1]. Early diagnosis and treatment can considerably improve the outcomes of this disease [2–5]. According to randomized clinical trials, the screening efficacy of mammography is effective in reducing breast cancer mortality [6–8]. Mammography is currently the mainstream imaging modality used to detect microcalcifications that may be associated with malignant changes [9,10]. However, interpretation of mammograms remains an arduous task owing to wide variations in the interpretative accuracy of radiologists [11,12]. Consequently, false positives can lead to unnecessary invasive procedures, increased expenditure, and heightened anxiety among patients, while false negatives may result in delayed diagnosis, disease progression, and impaired survival [13,14]. In addition, large number of mammography examinations being performed on an everyday basis also creates additional workload for the clinicians [15].

Computer-aided detection methods were introduced in the 1990s to improve the interpretative performance and work efficiency of the clinicians [16–18]. Nevertheless, it failed to produce a significant improvement in the performance of radiologists in clinical practice. In recent years, this field has evolved rapidly because of the development of artificial intelligence (AI). Approaches based on deep-learning convolutional neural networks have demonstrated their potential in other medical image analyses [19–21]. Several studies have developed deep learning-based systems for breast cancer detection using mammography, which have shown similar performance to that of human clinicians [22–24].

[*] Corresponding authors.
E-mail addresses: guyajia@126.com (Y.-J. Gu), byan@fudan.edu.cn (B. Yan), yukeda@fudan.edu.cn (K.-D. Yu).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.fmre.2023.04.018

However, most of the existing studies are based on relatively small datasets. Moreover, most previous studies focused only on the identification of lesions in the mammogram without distinguishing the target lesion as mass, microcalcifications, or abnormal structures, which may restrict further improvement of the relevant models. Breast cancer comprises a group of diseases with different intrinsic molecular subtypes [25,26]. Microcalcifications are small calcium deposits that appear as white specks on a mammogram. They may signify ductal carcinoma in situ (DCIS) or early breast cancer if they appear in specific patterns. The greatest advantage of mammography over ultrasound and magnetic resonance imaging (MRI) is its ability to detect microcalcifications.

In this study, we developed a deep-learning system for mammography that focused on interpreting microcalcifications. We also explored its ability to distinguish between invasive and non-invasive breast cancers and determine the molecular subtypes of breast carcinoma. Considering the advantages of mammography in detecting microcalcifications (compared to breast ultrasound and MRI), we hypothesized that a deep-learning model trained to recognize only microcalcifications (with or without masses) would have the ability to improve accuracy, facilitating early diagnosis of breast cancer. We anticipate that the prediction results of this model would assist clinicians in making judgments and improve their efficiency in population- or community-based breast cancer screening and treatment.

2. Methods

2.1. Ethical considerations

The independent research ethics committee (REC) of the participating centers (Fudan University Shanghai Cancer centre, Fudan University Huashan Hospital, and First Affiliated Hospital of China Medical University) approved this retrospective study protocol. RECs are there to protect the rights, safety, dignity and wellbeing of research participants. All the procedures performed in this study were conducted in accordance with good clinical practice and the Declaration of Helsinki, Finland. Written informed consent was obtained from all the patients prior to the commencement of the study.

2.2. Biopsy of microcalcification

Both stereotactic wire-localization-based open biopsy and stereotactic vacuum-assisted biopsy were used for breast microcalcification biopsy. Stereotactic vacuum-assisted biopsy was used in outpatient clinics for patients with single microcalcification, whereas stereotactic wire localization and open surgical biopsy was performed under general anesthesia for patients with multiple microcalcification lesions since repeated biopsies would cause intolerable pain and even bleeding.

Stereotactic wire localization of breast microcalcification was performed using a fenestrated compression plate and hook-wire needle under the guidance of digital mammography. All procedures were performed carefully to avoid tissue damage and bleeding in the breast. A localizing wire was placed on the target lesion of breast microcalcification. After the procedure, the patients were bandaged and immediately transferred to the operating room for subsequent surgical procedure. The excised specimen was viewed under magnification to confirm adequate removal of the lesion. Each microcalcification lesion with a localized wire was diagnosed based on histopathological examination.

2.3. Pathological examination

All microcalcification lesions in the collected mammograms were subjected to pathological examination at the Department of Pathology of each participating center. Two independent pathologists diagnosed the slides for each biopsied microcalcification. If there were disagreements regarding the pathological diagnosis, a consensus was reached after a re-review and discussion between the two pathologists. Ductal carcinoma in situ (DCIS) with foci of microinvasion (one or more foci of stromal invasion, none exceeding 1 mm in size) was classified as microinvasive carcinoma. Estrogen receptor (ER) status was identified based on immunohistochemical analysis of the tumor sections. The immunohistochemical cutoff for ER-negative status was less than 1% staining of the nuclei [23].

2.4. Data preprocessing

Image standardization was essential in this study to reduce the impact of different devices with different scanners or imaging protocols. Since the calcification regions are relatively small compared to the mammography image, we cropped the black background region in the image. First, we used an algorithm to sum up the image pixels in each column from right or left, and stopped when the sum of pixels in a column was zero. This helped us to automatically find the first valid pixel. We then cut out the images on the other side of the column to obtain an image without a black background. The basic properties exhibited by mammography images did not essentially differ for craniocaudal (CC) view or mediolateral oblique (MLO) view; therefore, we mixed all the images together for training and testing.

2.5. Model architecture

We used FasterRCNN [27] as our detection network, including Conv layers, Region Proposal Networks (RPN), Region of Interest (ROI) Pooling, and Classification Modules. In practice, we used Resnet50 [28] as the feature extractor to obtain the feature maps of the images. Since the calcification regions are generally small, we used the feature pyramid network (FPN) [29] to treat all the feature layers at different scales. The network summarized the results at all scales and provided precise predictions. The FPNNet was based on our detection network, which discriminated malignant calcifications from the detection box. If the detection model detected a malignant box in an image, the image was contemplated as malignant. Otherwise, the lesions were considered benign.

SPPNet is a basic deep-learning network that directly uses a single mammograph image as an input to provide predictions by observing the deep features of the image. We used Resnet18 [28] as the baseline model, which comprised 18 2D-convolution operations. The residual unit was added using a short-circuit mechanism that retained the image features while deepening the network for better classification. To better fuse the image features at different scales, we introduced a Spatial Pyramid Pooling (SPP) [30] structure at the end of this baseline network.

Both the FPNNet and SPPNet use a single mammograph image as input. The difference is that SPPNet only outputs the prediction of microcalcifications in the images, whereas FPNNet also flags the areas in which the network considers microcalcifications. During training, SPPNet used image-level binary tags as ground truth. In contrast, FPNNet used a microcalcification bounding box labeled by physicians as the ground truth.

2.6. Model training

We conducted the experiments on a machine equipped with Nvidia GTX 2080Ti GPU. Our framework was based on a PyTorch implementation. The parameters were optimized using ADAM. In a calcification-detection network, the RPN generates several possible bounding boxes, some of which may have low confidence or overlap with other bounding boxes. Therefore, we set both the score threshold and the Non-Maximum Suppression (NMS) threshold to 0.3 and achieved the best results in this setting.

2.7. Model testing

At the mammograph image level, we input the image directly into the trained model to obtain the prediction results for a single image. At the breast level, based on the prediction of two different positions of the breasts, we determined the status of the breasts according to the following rules and provided the prediction probabilities. If the predicted status of the two positions was consistent, the status of the breast was directly determined by the status of the mammograph images, and a higher prediction probability of the two position images was considered as the confidence of the breast. Otherwise, we focused on the predicted probability of the malignant image. If the probability of malignancy was greater than 0.5, we resolute the breast as malignant; otherwise, it was considered benign. The prediction confidence was given by the prediction probability of the selected position image based on the final result. At the microcalcification lesion level, we searched for regions with the highest intersection-over-union (IoU) values in the gold standard across all the predicted regions and assessed the results.

2.8. Metrics

For the classification task results, we briefly introduced four basic concepts. True Positive (TP) represented the sample predicted to be positive and was also actually positive. False Positive (FP) represented sample predicted as positive but was actually negative. False Negative (FN) represented sample predicted as negative but was actually positive. True Negative (TN) represented sample predicted to be negative and was also actually negative. Thus, precision was calculated as

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity was calculated as

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

F1 score is an indicator used in statistics to measure the accuracy of bicategorical models. It considers both precision and sensitivity of the classification model. The F1 score is the weighted average of the model precision and sensitivity, with a maximum of 1 and a minimum of 0. The F1 score was calculated as

$$F1 = 2 \cdot \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

We also used overall accuracy to assess the overall performance of the model, which was defined as

$$Overall\ ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

3. Results

3.1. Mammogram collection and labeling

A total of 4810 digital mammograms of microcalcifications were collected from 2448 consecutive female patients (1222 with non-malignant disease and 1226 with breast cancer) between June 2014 and August 2019 at the participating centers. Among them, 2362 patients had mammograms with two views (craniocaudal and mediolateral oblique), 28 had mammograms with a craniocaudal view, and 58 had mammograms with a mediolateral oblique view. Patients without apparent microcalcification lesions on mammograms were excluded from the study.

All patients randomly underwent mammography using Selenia Dimensions (Hologic Medical Systems, USA), MAMMOMAT Revelation (Siemens Healthcare, Germany), or Senographe DS (GE Healthcare, USA). All portable network graphics (PNG) data were exported from the Picture Archiving and Communication Systems.

Microcalcifications were categorized as Breast Imaging Reporting and Data System (BI-RADS) grade 3, 4, or 5 by two independent radiologists at each participating center. Patients underwent a biopsy based on the physician's decision. Some discontinuous microcalcifications or lesions with large-area microcalcifications decomposed during the biopsy examination. All collected mammograms showed at least one microcalcification, confirming the pathological diagnosis after biopsy examination. A total of 6663 microcalcification lesions were biopsied, of which 3301 were malignant and 3362 were benign. Correspondingly, 2420 and 2390 mammograms were identified to contain malignant and benign lesions, respectively.

Furthermore, to explore the predictive subtypes of breast cancer, we obtained immunohistochemical information of 1226 patients with confirmed malignant microcalcifications in the above cohort. Stereotactic wire localization and biopsy of breast microcalcifications were performed. Fig. 1 shows biopsy of microcalcification and some representative plots. Notably, atypical ductal hyperplasia (ADH) is neither a form of breast cancer nor a completely benign disease; thus, patients with ADH were excluded from this study.

When labeling microcalcifications, physicians drew ground-truth bounding boxes on the basis of the pathological results. The bounding boxes maximized the coverage of the entire area of the microcalcifications. The physicians then assigned the boxes an attribute, benign or malignant, against the pathological results. All labeled areas had corresponding pathological results to ensure labeling accuracy.

3.2. Detection of microcalcifications

Microcalcifications are often ignored during breast cancer screening due to their small granularity and indistinguishable characteristics. Malignant microcalcification lesions, particularly the ones with minimal calcification points where symptoms are not obvious, are easily misjudged by physicians. Computer-aided diagnosis has shown good prospects in the recent years due to the rapid development of AI. First, we constructed an AI system to detect and classify microcalcifications (Fig. 2a). The system included a pipeline comprising image standardization, microcalcification detection, and mammography classification. If a mammography image showed malignancy, the system also predicted the pathological type and ER status.

The structure of the calcification detection network is shown in Fig. 2b. We used FasterRCNN [27] as our detection network, including Conv layers, Region Proposal Networks (RPN), ROI (Region of Interest) Pooling, and Classification Modules. In practice, we used Resnet50 [28] as the feature extractor to obtain the feature maps of the images. Resnet50 contains 50 two-dimensional (2D)-convolution operations, which are often used to extract the deep features. Feature maps were shared between the subsequent RPN layers and the fully connected layers. The RPN generated region proposals, which were judged by Softmax to belong to either the positive or negative category, and it reused the bounding box regression to obtain exact proposals. ROI Pooling collected the input feature maps and proposals to extract the proposal feature maps, which were then sent to the subsequent fully connected layers to determine the target category. Finally, the classification module used the proposed feature maps to compute the proposed category. Since the calcification regions are generally small, we used the feature pyramid network (FPN) [29] to treat all the feature layers at different scales. The network summarized the results at all scales and provided precise predictions.

We divided the dataset into training and test sets in a 4:1 ratio. The samples from the training and test sets did not cross each other, and the cases of the test set were mainly from multiple centers. Ultimately, 1964 benign and 1970 malignant mammography images were included in the training set, and 426 benign and 450 malignant images in the test set, all of which were collected from multiple centers (Table 1). Fig. 3 shows some of our experimental results. Examples of benign and malignant calcifications are shown in Fig. 3a and Fig. 3c, respectively.

The benign calcification label given by the physicians only comprised a green box in the middle. However, our model predicted not only this annotation box, but also a small benign calcification in the upper right corner. For intermediate benign calcifications, the model

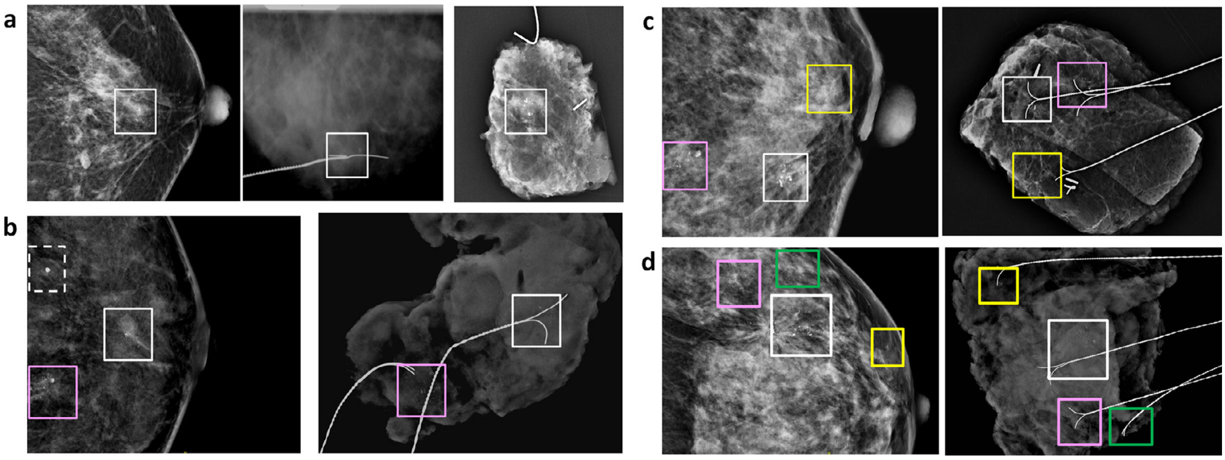

Fig. 1. Biopsy of microcalcification and representative plots. (a) Patient with one lesion of breast microcalcification. Left, microcalcification shown in the original mammograph picture. Middle, wire localization of breast microcalcification under an operator plate and a hook wire needle under the guidance of digital mammography. Right, microcalcification shown in the surgical biopsied specimen. A titanium clip was also placed beside the wire needle before performing mammography on the surgical biopsied specimen. The lesion in the white frame is pathologically confirmed benign. (b) Patient with two lesions of breast microcalcifications. Left, microcalcifications shown in the original mammograph picture. The calcification in the dash-line frame is categorized as Breast Imaging Reporting & Data System (BI RADS-2) (probably benign) and was not biopsied. Right, microcalcifications shown in the surgical biopsied specimen and each lesion of microcalcification was diagnosed pathologically. The lesion in the pink frame is Ductal carcinoma in situ (DCIS), and the lesion in the white frame is DCIS with microinvasion. (c) Patient with three lesions of breast microcalcifications. Left, microcalcifications shown in the original mammograph picture. Right, microcalcifications shown in the surgical biopsied specimen. The lesion in the pink frame is DCIS, the white frame is invasive cancer, and the yellow frame (with very shallow microcalcification and obvious under magnification) is benign. (d) Patient with four lesions of breast microcalcifications. Left, microcalcifications shown in the original mammograph picture. Right, microcalcifications shown in the surgical biopsied specimen. The lesion in the pink frame is invasive cancer, the white frame is invasive cancer, the yellow frame is DCIS, and the green frame is benign.

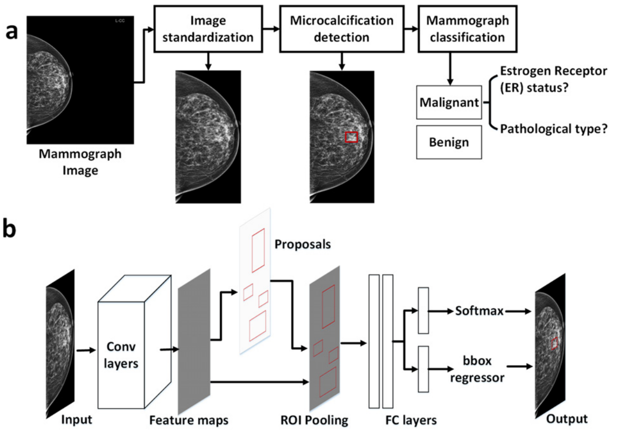

Fig. 2. The artificial intelligence (AI) system for detection and classification of microcalcifications. (a) Model development of AI system. The system includes a pipeline consisting of image standardization, microcalcification detection, and mammography classification. If the mammography image is malignant, the system will also predict the pathological type and estrogen receptor status. (b) Network structure of microcalcification detection method. We use FasterRCNN as our detection network, including the Conv layers, Region Proposal Networks (RPN), ROI (Region of Interest) Pooling, and Classification Modules with FC (fully connected) layers.

provided a confidence degree of 0.650. For small calcification points in the upper-right corner, the model yielded a confidence degree of 0.630. This demonstrates the advantages of the proposed model. Unlike the human eye, computer vision captures the deep features of an image through deep learning, has a full perspective of the image, and thus, outperforms prediction tasks. In addition, computers are not affected by fatigue, negligence, or other human factors, which can effectively improve the efficiency of medical diagnoses.

Our model detected the malignant calcification while labeling the possible benign calcification, as shown in Fig. 3c. Since physicians use histopathological results as the gold standard, they cannot accurately label benign calcifications in malignant calcification images. Although our model viewed a large number of benign samples during training, it could empirically infer possible benign calcifications in malignant calcification images. For the two predicted regions of malignant calcification, the model yielded confidence degrees of 0.831 and 0.426. For the predicted benign calcification regions, the model furnished confidence degrees of 0.708, 0.616, 0.476, and 0.382, respectively. After identification by experienced physicians, they were found to be typical of benign calcifications, which demonstrates the effectiveness of our model.

To validate the effectiveness of our deep-learning model, we visualized the feature extraction backbone of our detection network. We normalized and summed the output of the feature extraction module for each layer using the channel, thereby obtaining the feature image of the corresponding layer. Fig. 3e shows the feature visualization results.

For better detection and classification of small calcification regions, the most critical task of the feature extraction module of the detection network is to distinguish between calcification and background regions. As evident in Fig. 3e, most regions of the image were highlighted in the first layer of the network, except the region near the calcified lesion. As the network depth increased, the discrimination of the calcified regions became increasingly clear. Since deep networks lose some of their information with increasing depth, we added FPN to maximize the use of features at different scales. FPN uses the features of different scales as inputs, and can fuse more useful information. The final network therefore, yields accurate predictions.

To provide more intuitive results to help clinicians make judgments, we present the Class Activation Maps (CAM) of FPNNet in Fig. 4. The results showed that FPNNet can focus on microcalcifications and assist clinicians in making decisions.

Table 1
Datasets for training and testing of the deep-learning system.

Characteristics		Training set	Test set
Centers		Fudan University Shanghai Cancer Center	Fudan University Shanghai Cancer Center Fudan University Huashan Hospital The First Affiliated Hospital of China Medical University
Age, median (IQR), years		50 (43–58)	51 (44–59)
Number of patients		2010	438
Number of mammography images		3934	876
Number of breasts		2010	438
	Overall diagnosis: benign	1009	213
	Overall diagnosis: malignant	1001	225
Number of microcalcification lesions		5290	1373
	Benign disease	2674	688
	Malignant disease	2616	685
Malignant disease by pathology		2616	685
	Pure DCIS	201	30
	DCIS with microinvasion	275	74
	Invasive breast cancer*	1970	445
	Unknown pathology	170	136
Malignant disease by ER status		2616	685
	ER-positive	1451	416
	ER-negative	614	148
	Unknown ER	551	121

Abbreviations: DCIS, Ductal carcinoma in situ; ER, estrogen receptor; IQR, interquartile range.
* excluding DCIS with microinvasion.

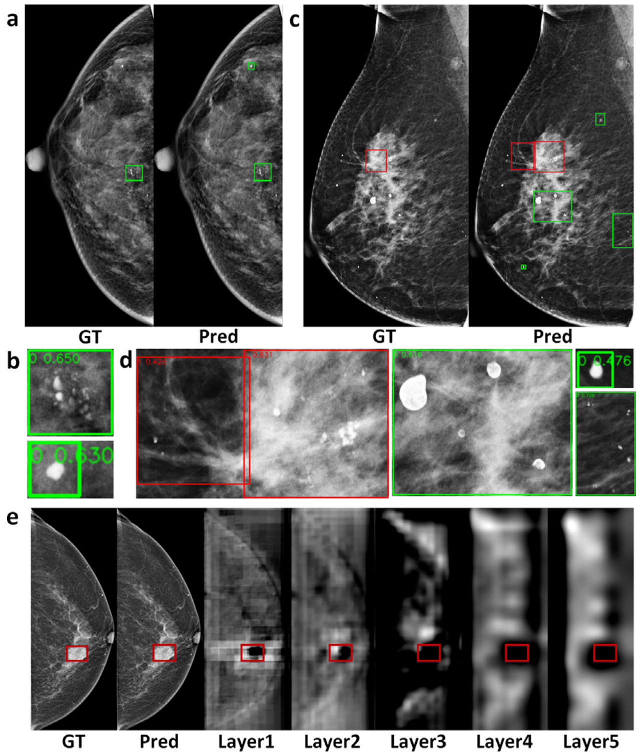

Fig. 3. Results and feature visualization of microcalcification detection experiments. (a) A benign example of calcification. The green boxes represent benign lesions. ‘GT’ refers to the gold standard label given by the doctor. ‘Pred’ refers to our model’s prediction. (b) Small patches obtained by enlarging the calcification regions in A. ‘0′ means benign, and the decimal number represents the confidence of the model for the given category of the prediction box. (c) An example of malignant calcification. The red boxes represent malignant calcification within the detected region. ‘GT’ refers to the gold standard label given by the doctor. ‘Pred’ means our model’s prediction. (d) Small patches obtained by enlarging the calcification regions in C. ‘1’ means malignant, and the decimal number represents the confidence of the model for the given category of the prediction box. (e) Visual analysis of the model intermediate feature layers. From left to right successively is the doctor-annotated gold standard label, our model predicted calcification regions and the first to fifth layers of our deep-learning model’s feature extraction backbone.

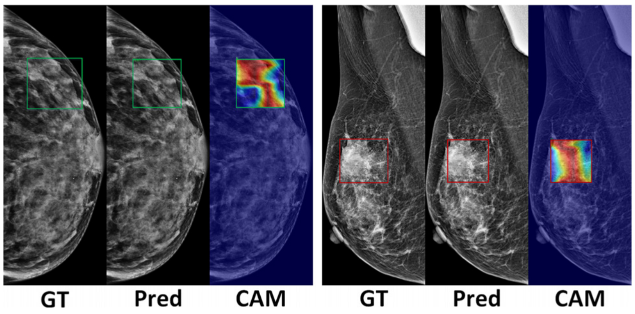

Fig. 4. The Class Activation Maps (CAM) of FPNNet. Left: benign. Right: malignant.

3.3. Mammography image classification results

Physicians cannot accurately determine whether a patient’s calcifications are benign or malignant using mammography alone. These patients require further pathological validation. Therefore, it is important to judge whether a mammography image is malignant using computer vision methods. In this study, we provided an annotation for each mammography image using the pathological outcome as the gold standard.

First, we constructed a basic deep-learning network that directly takes a mammograph image as the input to provide predictions by observing the deep features of the image. Empirically, the network depth is critical to the performance of the model. When the number of layers are increased, the network can extract more complex feature patterns. Thus, better results can theoretically be achieved when the model is deeper. However, the calcification point on a mammograph image may be very small. After several down-sampling of deep learning, such small calcification points may be ignored, leading to prediction errors. Therefore, we took Resnet18 [28] as the baseline model, which comprised 18 2D-convolution operations. The residual unit was added using a short-circuit mechanism that retains the image features while deepening the network for better classification. To better fuse the image features at different scales, we introduced a Spatial Pyramid Pooling (SPP) [30] structure at the end of this baseline network. Thus, this network was called SPPNet.

We used the same dataset partitioning method for the detection task to ensure fairness of the overall experiment. In other words, 1964 benign and 1970 malignant mammography images were included in the

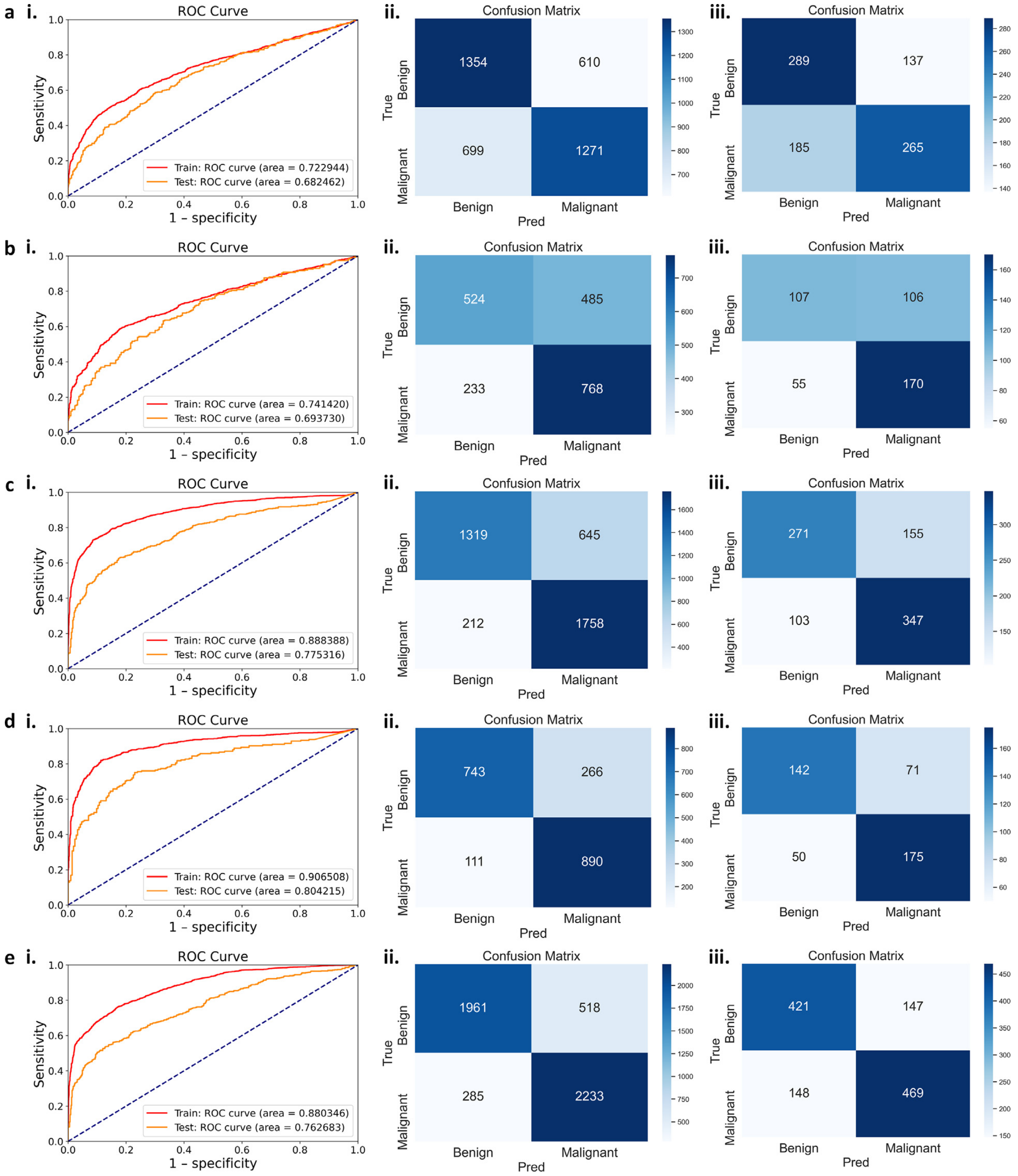

Fig. 5. Performance of different methods in discriminating between benign and malignant status on mammography image, one-side breast, and microcalcification lesion levels. The receiver operating characteristic (ROC) curve is a visualization tool to compare the quality of the two classification models. AUC (Area Under Curve) is defined as the area enclosed within the coordinate axis under the ROC curve. The larger the AUC, better is the result. (a–e) ROC curves (i) and normalized confusion matrices of classifications on the training (ii) and test set (iii). (a-b) The performance of the SPPNet. (c-e) The performance of FPNNet. (a, c) The performance on the mammograph image level. (b, d) The performance on the breast level. (e) The performance on the microcalcification lesion level.

training set. A total of 426 benign and 450 malignant mammography images were used in the test set. Accordingly, there were 1009 benign and 1001 malignant breasts in the training set, and 213 benign and 225 malignant breasts in the test set. Based on the pathological results, 2674 benign and 2616 malignant microcalcification lesions were included in the training set, and 688 benign and 685 malignant microcalcification lesions in the test set.

Since our study focused on the classification task, we used the classification report function in scikit-learn (a machine learning package in Python) to present a text report of the main classification metrics, including precision, sensitivity, and F1 score of each class. The SPPNet classification results showed an overall performance of Area Under the Curve (AUC) of 0.6825 (Fig. 5a) and achieved an overall accuracy of 0.6324 (Table 2) at the mammograph image level. Based on the pre-

Table 2
Classification results of different deep-learning models on mammography image, one-side breast, and microcalcification lesion level.

Set	Level	Model	Overall Accuracy	Precision		Sensitivity		F1-score	
				Benign	Malignant	Benign	Malignant	Benign	Malignant
Training	Mammogram	SPPNet	0.6673	0.6595	0.6757	0.6894	0.6452	0.6741	0.6601
		FPNNet	0.7822	0.8615	0.7316	0.6716	0.8924	0.7548	0.8040
	Breast	SPPNet	0.6428	0.6922	0.6129	0.5193	0.7672	0.5934	0.6815
		FPNNet	0.8124	0.8700	0.7699	0.7364	0.8891	0.7976	0.8252
	Microcalcification lesion	FPNNet	0.8393	0.8731	0.8117	0.7910	0.8868	0.8301	0.8476
Test	Mammogram	SPPNet	0.6324	0.6097	0.6592	0.6784	0.5889	0.6422	0.6221
		FPNNet	0.7055	0.7246	0.6912	0.6362	0.7711	0.6775	0.7290
	Breast	SPPNet	0.6324	0.6605	0.6159	0.5023	0.7556	0.5707	0.6786
		FPNNet	0.7237	0.7396	0.7114	0.6667	0.7778	0.7012	0.7431
	Microcalcification lesion	FPNNet	0.7511	0.7399	0.7614	0.7412	0.7601	0.7405	0.7607

‘Precision’ calculates the proportion of the number of correct images to the total number of positive class predictions. From the predictive perspective, it represents how many predictions are accurate. ‘Sensitivity’ determines the number of positive images predicted as positive to account for all annotated images. In medical analysis, it is equivalent to sensitivity. The ‘F1 score’ can be considered as a weighted average of the model precision and sensitivity.

diction of the two different positions of the breasts, we determined the status of the breasts according to the following rules and provided prediction probabilities: if the predicted status of the two positions was consistent, the status of the breast was directly determined by the status of the mammograph images, and a higher prediction probability of the two position images was considered as the confidence of the breast. Otherwise, we focused on the predicted probability of a malignant image. If the probability of malignancy was greater than 0.5, we determined the breast as malignant; otherwise, it was considered benign. The prediction confidence was determined by the prediction probability of the selected position image based on the final result. Thus, at the breast level, SPPNet exhibited an overall performance of an AUC of 0.6937 (Fig. 5b) and achieved an overall accuracy of 0.6324 (Table 2).

When physicians observe calcified lesions, they usually first determine the location of the lesion and then enlarge the image for careful evaluation. Therefore, we believe that detection-based classification methods can achieve better results. Using our detection network, we were able to discriminate between malignant calcifications and the detection box. If the detection model detected a malignant box in an image, we resolve that the image is malignant. Otherwise, the lesions were considered benign. This network is called the FPNNet. The results of the FPNNet were better than those of the SPPNet, which showed an overall performance of an AUC of 0.7753 (Fig. 5c) and achieved an overall accuracy of 0.7055 (Table 2) at the mammograph image level. The overall classification accuracy for discriminating between benign and malignant breasts was 0.7237 (Table 2) with an AUC of 0.8042 (Fig. 5d). The sensitivity of malignant breast prediction reached 0.7778 (Table 2), which was very helpful in reducing the missed detection rate of malignant calcification lesions. Multiple indicators showed significant improvement.

There were a total of 1373 calcified lesions in the test set according to pathology, the gold standard. Since the model-predicted calcification regions may be larger and more numerous than the annotated regions, we searched for regions with the highest intersection-over-union (IoU) values in the gold standard across all predicted regions. This yielded 1185 matching regions. We calculated the classification performance of the models for these calcified lesions. The overall accuracy of the microcalcification lesion level was 0.7511 (Table 2) with an AUC of 0.7627 (Fig. 5e). In addition, a five-fold cross-validation was conducted to ensure the effectiveness and non-overfitting of the network (Table 3).

3.4. Pathological type and ER status classification results

Under the premise that calcification could be judged as malignant or benign, we assumed that the pathological type (invasive vs. non-invasive) and ER status (ER-negative vs. ER-positive) could also be distinguished using deep-learning methods. This study has great medical prospects as a deeper task for exploring calcification classification. Based on our detection model, we cropped the malignant calcification regions into patches in accordance with the bounding boxes. Depending on the pathological type of the malignant calcified lesions, calcification images could be classified as DCIS or invasive cancer. We divided the malignant patches on the basis of the results of 2764 invasive cancer and 231 DCIS images. We divided the training and test sets in the manner similar to training the detection network. The images used for training in the detection task belonged to the training set in the subsequent classification task. Thus, we ensured that our test data did not cross the training set, ensuring the fairness of the classification task. Ultimately, the training set comprised 2245 invasive cancer images and 201 DCIS images, whereas the test set contained 519 invasive cancer images and 30 DCIS images. We used SPPNet for pathological type classification. Table 4 shows the classification results for the pathological types.

Since the microinvasion lesion (defined as an invasion lesion less than 1 mm) was very small and almost all microinvasions were accompanied by DCIS, it was very difficult to distinguish between DCIS with microinvasion and pure DCIS. Therefore, we classified DCIS with microinvasions into the DCIS category. Similarly, there were 1970 Class 1 and 476 Class 2 images in the training set, and 445 Class 1 and 104 Class 2 images in the test set. Class 1 refers to invasive cancer images excluding microinvasion and Class 2 refers to DCIS images with or without microinvasion (Table 4).

Based on ER expression, breast cancer can be divided into two subtypes: ER-positive and ER-negative. Clinically, approximately two-third of breast cancer patients are ER-positive. Therefore, the two ER categories were closer to the clinical value. Malignant patches were classified into two categories based on ER positivity or negativity. Ultimately, the training set contained 1451 ER-positive and 614 ER-negative images, while 416 ER-positive and 148 ER-negative images were included in the test set (Table 4).

The experimental results showed that the proposed model can achieve high overall accuracy. However, considering the imbalance in the sample size, a stochastic model could also achieve similar results. Therefore, we plotted the ROC curve for the proposed method, as shown in Fig. 6. The ROC curve is a visualization tool used to compare the quality of two classification models. AUC is defined as the area enclosed by the coordinate axis under the ROC curve, where a larger value is better. Theoretically, the AUC of the stochastic model was 0.5. The blue line represents the performance of the stochastic model and the yellow line represents the performance of our model. The AUC of our pathological-type classification (invasive cancer including microinvasion) (Fig. 6a) and pathological-type classification (invasive cancer excluding microinvasion) models (Fig. 6b) were 0.5444 and 0.5573, respectively, and the

Table 3
Five-fold cross validation results of the training set for different deep-learning models on mammography image, one-side breast, and microcalcification lesion level.

Level	Model	Five-Fold	Overall Accuracy	Precision		Sensitivity		F1-score	
				Benign	Malignant	Benign	Malignant	Benign	Malignant
Mammograph images	SPPNet	1	0.6031	0.8857	0.2787	0.5849	0.6800	0.7045	0.3953
		2	0.5751	0.3342	0.7904	0.5877	0.5704	0.4261	0.6626
		3	0.5598	0.4107	0.7408	0.6580	0.5087	0.5057	0.6032
		4	0.6247	0.4697	0.8270	0.7799	0.5444	0.5863	0.6566
		5	0.5304	0.8382	0.3326	0.4466	0.7619	0.5827	0.4631
		Average	0.5786	0.5877	0.5939	0.6114	0.6131	0.5611	0.5562
		STD	0.0368	0.2555	0.2656	0.1214	0.1050	0.1036	0.1206
	FPNNet	1	0.7010	0.9426	0.3724	0.6714	0.8267	0.7842	0.5135
		2	0.7430	0.5122	0.9472	0.8957	0.6870	0.6517	0.7964
		3	0.9466	0.9158	0.9630	0.9294	0.9555	0.9225	0.9592
		4	0.9720	0.9695	0.9733	0.9478	0.9846	0.9585	0.9789
		5	0.6506	0.8654	0.4118	0.6207	0.7333	0.7229	0.5274
		Average	0.8026	0.8411	0.7335	0.8130	0.8374	0.8080	0.7551
		STD	0.1470	0.1878	0.3121	0.1546	0.1315	0.1304	0.2256
One-sided breast	SPPNet	1	0.4877	0.9116	0.2490	0.4061	0.8333	0.5618	0.3835
		2	0.6516	0.3652	0.7676	0.3889	0.7491	0.3767	0.7583
		3	0.6225	0.4462	0.7074	0.4234	0.7262	0.4345	0.7167
		4	0.6967	0.5497	0.7863	0.6103	0.7414	0.5784	0.7632
		5	0.4049	0.8800	0.2970	0.2215	0.9159	0.3539	0.4485
		Average	0.5727	0.6305	0.5615	0.4100	0.7932	0.4611	0.6140
		STD	0.1219	0.2511	0.2655	0.1381	0.0803	0.1040	0.1831
	FPNNet	1	0.6936	0.9399	0.3657	0.6636	0.8205	0.7780	0.5059
		2	0.7920	0.5776	0.9370	0.8611	0.7663	0.6914	0.8431
		3	0.9700	0.9771	0.9665	0.9343	0.9886	0.9552	0.9774
		4	0.9850	0.9924	0.9813	0.9632	0.9962	0.9776	0.9887
		5	0.6272	0.8848	0.3972	0.5671	0.7944	0.6912	0.5296
		Average	0.8135	0.8744	0.7295	0.7979	0.8732	0.8187	0.7689
		STD	0.1608	0.1710	0.3184	0.1741	0.1105	0.1396	0.2365
Microcalcification lesions	FPNNet	1	0.7521	0.9542	0.4951	0.7062	0.8947	0.8117	0.6375
		2	0.8691	0.6553	0.9684	0.9059	0.8581	0.7605	0.9099
		3	0.9736	0.9625	0.9789	0.9559	0.9821	0.9592	0.9805
		4	0.9836	0.9870	0.9820	0.9620	0.9939	0.9744	0.9879
		5	0.6676	0.8900	0.4246	0.6282	0.7795	0.7365	0.5497
		Average	0.8492	0.8898	0.7698	0.8316	0.9017	0.8485	0.8131
		STD	0.1382	0.1359	0.2841	0.1542	0.0892	0.1115	0.2050

STD means Standard Deviation.

Table 4
Classification results according to pathological type and estrogen receptor (ER) status.

Set	Classification	Class 1	Class 2	Overall Accuracy	Precision		Sensitivity		F1-score	
					Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Train	Pathological type	Invasive cancer (including microinvasion)	Pure DCIS	0.8626	0.9196	0.1053	0.9318	0.0896	0.9257	0.0968
		Invasive cancer (excluding microinvasion)	DCIS with or without microinvasion	0.6648	0.8476	0.2828	0.7117	0.4706	0.7737	0.3533
	ER status	ER-positive	ER-negative	0.6993	0.7038	0.3793	0.9876	0.0179	0.8219	0.0342
Test	Pathological type	Invasive cancer (including microinvasion)	Pure DCIS	0.8889	0.9508	0.1220	0.9306	0.1667	0.9406	0.1408
		Invasive cancer (excluding microinvasion)	DCIS with or without microinvasion	0.6430	0.8152	0.2013	0.7236	0.2981	0.7667	0.2403
	ER status	ER-positive	ER-negative	0.7411	0.7402	1.0000	1.0000	0.0135	0.8507	0.0267

ER means estrogen receptor. DCIS means ductal carcinoma in situ.

AUC of our ER status classification model (Fig. 6c) was 0.5692. Since the pathological type and ER expression can be discerned at the molecular level and not at the visual level, they may be extremely difficult to distinguish using AI. Notably, even under the current sample imbalance conditions, our model achieved some results, proving that this direction deserves further exploration.

4. Discussion and conclusion

With the development of deep-learning algorithms, computer-aided diagnosis has played an important role in multiple medical domains. Yala et al. [24] demonstrated that deep-learning models could classify even a fraction of mammography scans as cancer-free, thereby improving the performance and workflow efficiency. Rodriguez-Ruiz et al. [23] used an AI system to assist in diagnosis, which yielded a significantly better performance compared to radiologists. Although AI systems have performed well in predicting breast cancer, most of the methods only consider image-level classification and cannot specifically locate regions where lesions may be present.

In this study, we developed a deep-learning system for mammography that focused on interpreting microcalcifications and was trained to recognize only microcalcification lesions (with or without masses). We built an AI system using deep-learning methods, including image standardization, microcalcification detection, and mammography classification. The microcalcification detection model was based on Faster RCNN [27], which automatically searches for calcification regions in mammog-

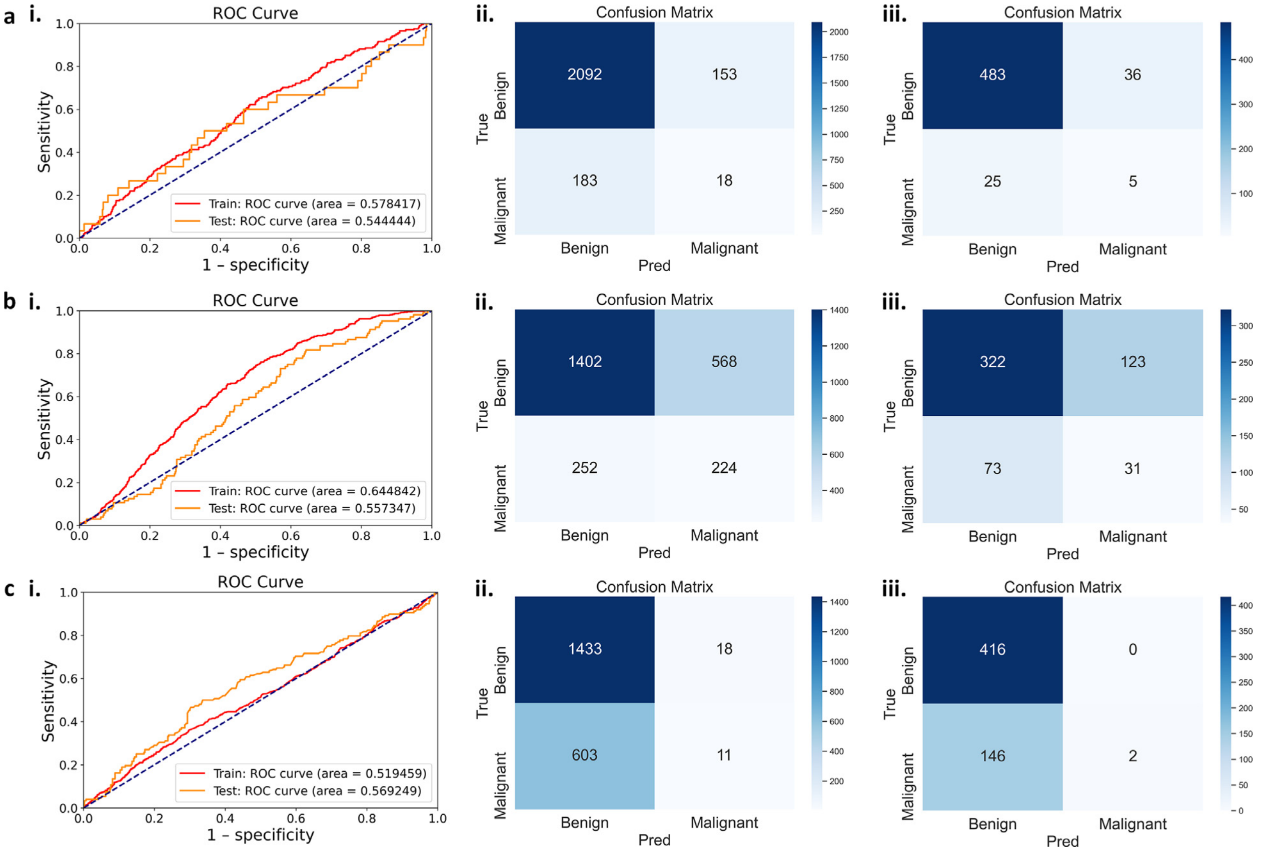

Fig. 6. Performance of SPPNet in different classification tasks. The receiver operating characteristic (ROC) curve is a visualization tool to compare the quality of the two classification models. AUC (Area Under Curve) is defined as the area enclosed with the coordinate axis under the ROC curve, where a larger value is better. (a–c) ROC curves (i) and normalized confusion matrices of image classifications on training set (ii) and test set (iii). (a) The performance of pathological type classification (invasive cancer including microinvasion). (b) The performance of pathological type classification (invasive cancer excluding microinvasion). (c) The performance of estrogen receptor (ER) status classification.

raphy and determines whether they are benign or malignant. As microcalcifications are very small, the FPN structure was used to fuse deep features at different scales. Based on the detection boxes and prediction results presented by the detection model, we determined the category of mammography images (benign or malignant), i.e., mammography with the presence of a malignant box was classified as malignant. The experiments demonstrated that our detection model accurately extracted calcification regions after learning large-scale mammography image features. The overall classification accuracy of the system for discriminating between benign and malignant breasts was 0.7237 with an AUC of 0.8042, and the sensitivity of malignant breast prediction reached 0.7778. The overall accuracy for the mammograph image level was 0.7055, with an AUC of 0.7753, and 0.7511 for the microcalcification lesion level with an AUC of 0.7627.

The ability of the AI system to predict the pathological type and ER status remains unexplored. We performed experiments using a basic classification model to classify pathological types and ER statuses. Experiments showed that while the model could achieve high overall accuracy in both classification tasks, its AUC was only close to that of a stochastic model. The high overall accuracy was attributed to an extreme imbalance in the data sample sizes of the two classes. As the lesion of microinvasion (defined as an invasion lesion less than 1 mm) was too small to distinguish from pure DCIS, we also tried another classification—i.e., categorizing DCIS with microinvasion into the category of DCIS. The AUC of our pathological type classification model was approximately 0.5–0.6. Compared to the mammography image classification results, AI achieved good performance from the image visual perspective, but it did not prove to be ideal for molecular-level classification.

In conclusion, we explored the ability of AI to distinguish between invasive and noninvasive breast cancers. Moreover, we also used AI to determine the molecular subtypes of breast carcinoma. We demonstrated that a deep-learning model trained to recognize only microcalcification lesions can achieve better accuracy in early diagnosis of breast cancer. We could not only provide the classification results of mammography images, but also annotate specific calcification regions. This practice has rarely been reported in the literature. However, our attempts at pathological type and ER status classification did not achieve good results and deserve further research.

Data availability

The main data supporting the results of this study, including the deidentified and anonymized data generated, are available at https://pan.baidu.com/s/1swu7-7zO2RooHDItk8cQ_A (psd: 18el).

Code availability

The codes for our AI system, including image standardization, microcalcification detection, and mammography classification, are available at https://github.com/AliceQLin/Microcalcification-detection-and-classification. The codes are available for download for noncommercial use.

CRediT authorship contribution statement

B.Y., K.Y., Y.G., Q.L., W.T., J.G. and Y.H. conceived the technique and the whole study. Q.L., W.T. and J.G. implemented the algorithm. Y.H., Q.X., Y.X., Y.J., and Z.S. collected and labeled the data. Q.L. and W.T. conceived the deep learning-based microcalcification detection and classification. B.Y., Q.L., and W.T. designed the validation experiments. Q.L. and W.T. trained the network and performed the validation experiments. B.Y., K.Y., and Y.G. supervised the project. All authors had access to the study, contributed to writing the manuscript, and have read the final version.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

We gratefully acknowledge the grants received from the National Natural Science Foundation of China (82072916, 82071878, U2001209, 61902076, and 61772137), the 2018 Shanghai Youth Excellent Academic Leader, the Fudan ZHUOSHI Project, Chinese Young Breast Experts Research project (CYBER-2021-A01), Shanghai Engineering Research Center of Artificial Intelligence Technology for Tumor Diseases, Xuhui District Artificial Intelligence Medical Hospital Cooperation Project (2020-008), Shanghai Science and Technology Foundation (19DZ1930502), Clinical Research Plan of SHDC (SHDC2020CR2008A), and Natural Science Foundation of Shanghai (21ZR1406600).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fmre.2023.04.018.

References

[1] H. Sung, J. Ferlay, R.L. Siegel, et al., Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 71 (3) (2021) 209–249.

[2] R. Rezk, R. Mar´ın-Garc´ıa, A.K. Gad, The fibrillar matrix: Novel avenues for breast cancer detection and treatment, Engineering 7 (10) (2021) 1375–1380.

[3] B. Dong, H. Xue, Y. Li, et al., Classification and diagnosis of cervical lesions based on colposcopy images using deep fully convolutional networks: A manmachine comparison cohort study, Fundamental Research 5 (1) (2025) 419–428.

[4] S. Tang, K. Yuan, L. Chen, Molecular biomarkers, network biomarkers, and dynamic network biomarkers for diagnosis and prediction of rare diseases, Fundamental Research 2 (2) (2022) 894–902.

[5] A.C. Dumitru, D. Mohammed, M. Maja, et al., Labelfree imaging of cholesterol assemblies reveals hidden nanomechanics of breast cancer cells, Adv. Sci. 7 (22) (2020) 2002643.

[6] S. Liu, Y. Wang, X. Yang, et al., Deep learning in medical ultrasound analysis: A review, Engineering 5 (2) (2019) 261–275.

[7] K. Dembrower, E. W˚ahlin, Y. Liu, et al., Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: A retrospective simulation study, Lancet Dig, Health 2 (9) (2020) e468–e474.

[8] J. Frisell, U. Glas, L. Hellstr¨om, et al., Randomized mammographic screening for breast cancer in stockholm, Breast Cancer Res. Treat. 8 (1) (1986) 45–54.

[9] L. Wilkinson, V. Thomas, N. Sharma, Microcalcification on mammography: Approaches to interpretation and biopsy, Br. J. Radiol. 90 (1069) (2017) 20160594.

[10] S. Zhang, H. Wang, X. Ding, et al., Bidirectional crosstalk between therapeutic cancer vaccines and the tumor microenvironment: Beyond tumor antigens, Fundamental Research 3 (6) (2023) 1005–1024.

[11] C.D. Lehman, R.F. Arao, B.L. Sprague, et al., National performance benchmarks for modern screening digital mammography: Update from the breast cancer surveillance consortium, Radiology 283 (1) (2017) 49.

[12] J.G. Elmore, S.L. Jackson, L. Abraham, et al., Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy, Radiology 253 (3) (2009) 641.

[13] A.N. Tosteson, D.G. Fryback, C.S. Hammond, Consequences of false-positive screening mammograms, JAMA Intern. Med. 174 (6) (2014) 954–961.

[14] N. Houssami, K. Hunter, The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening, NPJ Breast Cancer 3 (1) (2017) 1–13.

[15] A. Rimmer, Radiologist shortage leaves patient care at risk, warns royal college, BMJ: Br. Med. J. (2017) 359.

[16] C. Qiao, M. Lv, X. Li, et al., A novel human antibody, hf, against her2/erb-b2 obtained by a computer-aided antibody design method, Engineering 7 (11) (2021) 1566–1576.

[17] G.-D. Liu, Y.-C. Li, W. Zhang, et al., A brief review of artificial intelligence applications and algorithms for psychiatric disorders, Engineering 6 (4) (2020) 462–467.

[18] Y. Xu, M. Kong, W. Xie, et al., Deep sequential feature learning in clinical image classification of infectious keratitis, Engineering 7 (7) (2021) 1002–1010.

[19] J. Wang, Y. Wang, X. Tao, et al., Pca-u-net based breast cancer nest segmentation from microarray hyperspectral images, Fundamental Research 1 (5) (2021) 631–640.

[20] S. Zhao, C.-Y. Yan, H. Lv, et al., Deep learning framework for comprehensive molecular and prognostic stratifications of triple-negative breast cancer, Fundamental Research 4 (3) (2024) 678–689.

[21] X. Xu, X. Jiang, C. Ma, et al., A deep learning system to screen novel coronavirus disease 2019 pneumonia, Engineering 6 (10) (2020) 1122–1129.

[22] H.-E. Kim, H.H. Kim, B.-K. Han, et al., Changes in cancer detection and false-positive Sensitivity in mammography using artificial intelligence: A retrospective, multireader study, Lancet Dig. Health 2 (3) (2020) e138–e148.

[23] A. Rodriguez-Ruiz, K. L˚ang, A. Gubern-Merida, et al., Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists, J. Natl. Cancer Inst. 111 (9) (2019) 916–922.

[24] A. Yala, T. Schuster, R. Miles, et al., A deep learning model to triage screening mammograms: A simulation study, Radiology 293 (1) (2019) 38–46.

[25] A.C. Wolff, M.E.H. Hammond, D.G. Hicks, et al., Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of American pathologists clinical practice guideline update, Arch. Pathol. Lab. Med. 138 (2) (2014) 241–256.

[26] I. Fleming, J. Cooper, D. Jenson, et al., in: Ajcc (American Joint Committee on cancer) Cancer Staging Manual, Lippincott Williams & Wilkins, Philadelphia, 1997, pp. 53–55.

[27] S. Ren, K. He, R. Girshick, et al., Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015) 91–99.

[28] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[29] T.-Y. Lin, P. Doll´ar, R. Girshick, et al., Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[30] K. He, X. Zhang, S. Ren, et al., Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.

Author profile

Qing Lin (BRID: 02897.00.99583) received the B.E. degree from the School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2018. She is currently pursuing the Ph.D. degree with the School of Computer Science, Fudan University, Shanghai, China. Her research interests include digital image and video processing.

Ke-Da Yu, MD, PhD, is a professor now. He was supported by the National Natural Science Foundation of China (NSFC) with National Science Fund for Distinguished Young Scholars. In the past five years, he has published research papers as the corresponding author in *JAMA Oncol*, *Nat Commun*, *JNCI*, *Sci Adv*, and other international authoritative journals. According to Web of Science, the papers have been cited more than 5300 times, and the H index is 43. He also serves as a member of the Steering Committee of the Global Early Breast Cancer Trialists Cooperative Group (EBCTCG). His research field is clinical and translational research on breast cancer molecular subtype.