ARTICLE

Clinical Studies

Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: a multicentre study

Ning Mao[1,2,7], Haicheng Zhang[2,7], Yi Dai[3], Qin Li[4], Fan Lin[1], Jing Gao[1], Tiantian Zheng[1], Feng Zhao[5], Haizhu Xie[1], Cong Xu[6]✉ and Heng Ma[1]✉

BACKGROUND: This study aims to develop an attention-based deep learning model for distinguishing benign from malignant breast lesions on CESM.
METHODS: Preoperative CESM images of 1239 patients, which were definitely diagnosed on pathology in a multicentre cohort, were divided into training and validation sets, internal and external test sets. The regions of interest of the breast lesions were outlined manually by a senior radiologist. We adopted three conventional convolutional neural networks (CNNs), namely, DenseNet 121, Xception, and ResNet 50, as the backbone architectures and incorporated the convolutional block attention module (CBAM) into them for classification. The performance of the models was analysed in terms of the receiver operating characteristic (ROC) curve, accuracy, the positive predictive value (PPV), the negative predictive value (NPV), the F1 score, the precision recall curve (PRC), and heat maps. The final models were compared with the diagnostic performance of conventional CNNs, radiomics models, and two radiologists with specialised breast imaging experience.
RESULTS: The best-performing deep learning model, that is, the CBAM-based Xception, achieved an area under the ROC curve (AUC) of 0.970, a sensitivity of 0.848, a specificity of 1.000, and an accuracy of 0.891 on the external test set, which was higher than those of other CNNs, radiomics models, and radiologists. The PRC and the heat maps also indicated the favourable predictive performance of the attention-based CNN model. The diagnostic performance of two radiologists improved with deep learning assistance.
CONCLUSIONS: Using an attention-based deep learning model based on CESM images can help to distinguishing benign from malignant breast lesions, and the diagnostic performance of radiologists improved with deep learning assistance.

British Journal of Cancer (2023) 128:793–804; https://doi.org/10.1038/s41416-022-02092-y

INTRODUCTION

Breast cancer is one of the most common malignant tumours among women [1]. Early detection and classification of breast lesions is closely related to improving the survival rate of women. According to the American College of Radiology, a short-term follow-up observation is usually recommended for Breast Imaging Reporting and Data System (BI-RADS) category 3 lesions, while biopsy is usually recommended for BI-RADS category 4 and 5 lesions. Accurate classification of breast lesions could help change the BI-RADS category so that patients could benefit from avoiding biopsy.

Mammography is associated with the reduction of breast cancer mortality and recommended for the early diagnosis of breast cancer [2]. However, mammography still has potential issues, such as the overdiagnosis of inert tumours, the requirement of additional testing, patient anxiety, and radiation exposure [3]. Contrast enhanced spectral mammography (CESM) is a new breakthrough in mammography technology that can effectively detect breast lesions, but its specificity in the diagnosis of breast cancer is limited [4].

Radiologists usually distinguish between benign and malignant breast lesions through morphological features, such as the diameter, volume, and edge of the lesions. However, these morphological features often overlap, preventing their accurate classification based on expertise and experience. Therefore, the results vary greatly among different radiologists. Considering that, an automatic and noninvasive tool needs to be developed for identifying benign and malignant breast lesions.

[1]Department of Radiology, Yantai Yuhuangding Hospital, Qingdao University, 264000 Yantai, Shandong, P. R. China. [2]Big Data and Artificial Intelligence Laboratory, Yantai Yuhuangding Hospital, Qingdao University, 264000 Yantai, Shandong, P. R. China. [3]Department of Radiology, Peking University Shenzhen Hospital, 518000 Shenzhen, P. R. China. [4]Department of Radiology, Fudan University Cancer Center, 200433 Shanghai, P. R. China. [5]School of Computer Science and Technology, Shandong Technology and Business University, 264005 Yantai, Shandong, P. R. China. [6]Physical Examination Center, Yantai Yuhuangding Hospital, Qingdao University, 264000 Yantai, Shandong, P. R. China. [7]These authors jointly supervised this work: Ning Mao, Haicheng Zhang. ✉email: 616574369@qq.com; 827341627@qq.com

Received: 1 May 2022 Revised: 21 November 2022 Accepted: 24 November 2022
Published online: 15 December 2022

Radiomics have been used for breast cancer diagnosis and evaluation for nearly a decade [5–9]. Our research and literature reports showed that CESM-based radiomics can be used to distinguish benign from malignant breast lesions [10, 11]. However, in these previous studies, "hand-crafted" features, such as shape and texture, may not capture the full range of information contained within the images and are limited by low reproducibility. Deep learning extracts deeper and more comprehensive information directly from raw images, and has important clinical potential in breast cancer diagnosis, preoperative prediction, and therapeutic effect prediction [12–14]. The exploration of deep learning on CESM is still at the initial phase, and few recently published studies showed its extraordinary ability in this area [15, 16]. However, the sample size of these studies was relatively small, and they were all single-centre studies. In addition, many works have demonstrated that the attention-based convolutional neural networks (CNNs) could focus the attention of the networks on the objects of interest and improve the diagnostic performance of deep learning models [17, 18]. Therefore, we reason that the discriminatory ability of CNNs could be strengthened by introducing the attention mechanism. However, the literature shows no studies with focus on CESM.

In this study, we developed an attention-based deep learning model to preoperatively discriminate benign from malignant breast lesions on CESM images and validated its performance on a multicentre data set. We adopted DenseNet 121, Xception, and ResNet 50 as the backbone architectures and incorporated the convolutional block attention module (CBAM) [19] into them for classification.

MATERIALS AND METHODS

Patients and data sets

This retrospective multicentre study was approved by the institutional review board of Yantai Yuhuangding Hospital, and the patient informed consent was waived. For the primary cohort, we assessed Yantai Yuhuangding Hospital's database of medical records from July 2017 to May 2020 to identify patients with histologically confirmed breast lesions. The inclusion criteria for our study were as follows: (1) patients with suspected breast lesions after physical examination, mammography screening, or ultrasound; (2) patients referred for CESM as part of diagnostic imaging; (3) pathology confirmed breast lesions after surgery; (4) patients with multiple lesions all benign or all malignant; and (5) CESM images in low-energy, recombined image within 2 weeks before surgery. The exclusion criteria for our study were as follows: (1) tumour diameter <5 mm, which hinders lesion observation; (2) biopsy or surgery was performed before CESM; (3) chemotherapy, radiotherapy, or hormone treatment were performed before CESM; and (4) non-mass lesions without delineate boundaries. Additionally, the largest lesion was selected for evaluation for patients with more than one lesion [20]. Finally, 1093 patients were included in the training and validation sets, and 100 were included in the internal test set according to the CESM examination time. From June 2018 to May 2019, an independent external test on 46 patients from Fudan University Cancer Center was enrolled with the same criteria used for the primary cohort. The CESM images were obtained from the Picture Archiving and Communication System at Yantai Yuhuangding Hospital and Fudan University Cancer Center. The CESM image acquisition is shown in Supplementary Materials 1. The patient inclusion workflow is shown in Fig. 1a.

Image segmentation

The regions of interest (ROIs) were manually drawn by a senior radiologist (Segmenter 1) with 10 years of experience in breast imaging diagnosis on low-energy and recombined images of cranio-caudal (CC) view, respectively. Therefore, each lesion had two ROIs. Image segmentation was performed using ITK-SNAP (version 3.8.0). The dice similarity coefficient (DSC) was utilised for evaluating the agreement of image segmentation. Two other radiologists (Segmenters 2 and 3), with 10 and 12 years of experience in breast imaging diagnosis, respectively, randomly segmented 150 images to evaluate the agreement. An illustration of image segmentation is shown in Fig. S1.

Data preprocessing

We cropped bounding boxes 5 mm from the edge of the ROIs and then resized them to different size according to different CNNs. The cropped image patches grey value were normalised to 0–255 and subjected to histogram equalisation.

To increase the training data and prevent network overfitting, random geometric image transformations were applied for data augmentation, including flipping, rotation, scaling, and shifting. This process was performed in Python (version 3.6.6; Python Software Foundation, Wilmington, Del) using the Keras ImageDataGenerator (https://keras.io/preprocessing/image/).

Models building and testing

The CBAM-based CNN architecture is presented in Fig. 1b. We selected and evaluated the models on the training and validation sets using fivefold cross-validation. Training in the training and validation sets resulted in five models; the averaged predictions from all five models were used for evaluation in the test sets. Given their advantages, three representative deep CNNs, namely, DenseNet 121 [21], Xception [22], and ResNet 50 [23], were used as the backbone architectures. These models were all pre-trained with ImageNet data set [24]. Since the pre-trained model expected a three-channel input, the low-energy image, recombined image and lesion mask of each patient were input to the red, green, and blue channel of an image, respectively. CNNs consist of input, convolution, pooling, full connected, and output layers. Detailed information on the CNNs can be found in Supplementary Materials 2. Then, we added a CBAM before the global average pooling layer on the backbone architectures. CBAM, a lightweight and general attention module, can be integrated into any feed-forward CNN to improve the performance of classification tasks. CBAM consists of two independent submodules, namely, the channel and spatial attention modules, which carry out channel and spatial attention tasks, respectively [19]. This module can ensure integration with any CNN architecture as a plug and play module. The detailed information of CBAM can be found in Supplementary Materials 3 and Fig. 1c. The detailed network architecture of each CBAM-based model are shown in Supplementary Material (Fig. S2)

All models were trained using Keras 2.2.0 with TensorFlow 1.9.0 as the backend. The weight and biases of the model was initialised according to the weights from the pre-trained model with ImageNet. The RMSProp optimiser was used to train the network with a batch size of 12. The initial learning rate was set to 0.0001 and decayed by a factor of 10 each time when the accuracy of the validation set showed no further improvement for 10 continuous epochs. Finally, the models with the lowest validation loss were selected. The loss function of the CNNs was cross-entropy. All the programs were performed in Python version 3.6.6.

Moreover, we also explored the predictive performance of the deep learning models with and without tumour segmentation, as detailed in Supplementary Materials 4.

Radiologists' performance and comparison with deep learning

Two radiologists (Radiologist 1 and 2), with 5 and 12 years of breast imaging experience, respectively, reviewed the data of the test sets and judged whether they were malignant or benign. Only the age, medical history, family history, and CESM images were open to the radiologists. Then the radiologists' predictive performance was compared with that of the deep learning models. To evaluate the inter-rater variability, the Cohen's kappa values were calculated between two radiologists.

Radiomics model building and evaluation

Radiomics features were extracted from the ROIs of each patient's CESM images (low-energy and recombined images) using the RadiomicsFeatureExtractor toolbox in PyRadiomics, which is a flexible open-source platform implemented in Python. In the training and validation sets, radiomics features were selected from two ROIs separately using Spearman correlation analysis, analysis of variance, and least absolute shrinkage and selection operator (LASSO) logistic regression. Four machine learning classifiers, namely, *K*-nearest neighbour (KNN), logistics regression (LR), decision tree (DT), and support vector machine (SVM) were used to establish radiomics models for comparison with deep learning models. Detailed information about radiomics features extraction and features selection can be found in Supplementary Materials 5.

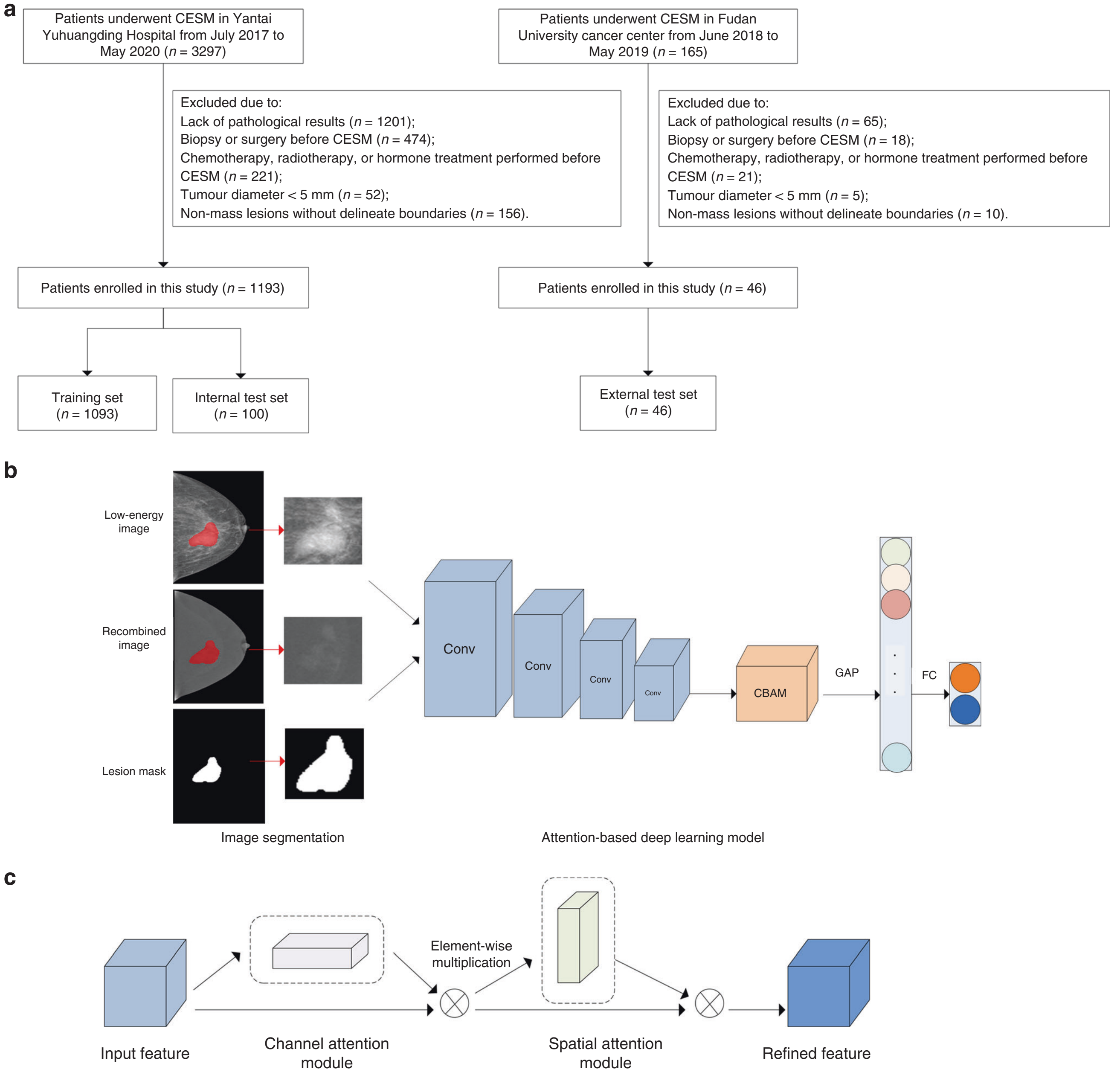
a
Patients underwent CESM in Yantai Yuhuangding Hospital from July 2017 to May 2020 (n = 3297)
Excluded due to:
Lack of pathological results (n = 1201);
Biopsy or surgery before CESM (n = 474);
Chemotherapy, radiotherapy, or hormone treatment performed before CESM (n = 221);
Tumour diameter < 5 mm (n = 52);
Non-mass lesions without delineate boundaries (n = 156).
Patients enrolled in this study (n = 1193)
Training set (n = 1093)
Internal test set (n = 100)
Patients underwent CESM in Fudan University cancer center from June 2018 to May 2019 (n = 165)
Excluded due to:
Lack of pathological results (n = 65);
Biopsy or surgery before CESM (n = 18);
Chemotherapy, radiotherapy, or hormone treatment performed before CESM (n = 21);
Tumour diameter < 5 mm (n = 5);
Non-mass lesions without delineate boundaries (n = 10).
Patients enrolled in this study (n = 46)
External test set (n = 46)
b
Low-energy image
Recombined image
Lesion mask
Conv
Conv
Conv
Conv
CBAM
GAP
FC
Image segmentation
Attention-based deep learning model
c
Element-wise multiplication
Input feature
Channel attention module
Spatial attention module
Refined feature

Fig. 1 The study flowchart. The patient inclusion workflow (**a**), an illustration of the CBAM-based convolutional neural network architecture (**b**), and the architectures of CBAM (**c**). GAP global average pooling layer, FC fully connected layer, CBAM convolutional block attention module.

Radiologists' performance with deep learning assistance

To evaluate the incremental benefit of deep learning to radiologists, their review was repeated independently with access to the best-performing CNN model with a 2-month washout period, recording new predictions. The model provided radiologists with predicted probability of tumour malignancy [25], and the age, medical history, family history, and CESM images were also open to the radiologists. After taking consideration of the CBAM-based Xception model-predicted probability of malignancy, the two radiologists evaluated the malignancy status of each patient once again.The deep learning-assisted performance was compared with that of the radiologists alone with the same metrics as above. To evaluate the inter-rater variability with the assistance of the model, the Cohen's kappa values were calculated between two radiologists.

Statistical analysis

All statistical tests were performed using R (version 3.6.2; www.r-project.org). The chi-square or Fisher's exact test was used as appropriate to compare the differences in categorical variables, whereas the two-sample *t* test was utilised to compare the differences in continuous variables. The performance of all the models was assessed in terms of the receiver operating characteristic (ROC) curve, precision recall curve (PRC) and confusion matrix. Accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score were calculated from the ROC curve according to the cut off value that maximises the Youden index (sensitivity + specificity − 1), and their 95% confidence intervals (CIs) were reported. In addition, a high sensitivity (95% sensitivity) and a high specificity (95% specificity) thresholds were defined. We also performed stratified analysis on the Breast Imaging Reporting and Data System (BI-RADS) 4 subgroup and different lesion diameter subgroups. The specificity and sensitivity points of the radiologists on the test sets were plotted in the same ROC curves. Comparisons between AUCs were made by using the method devised by DeLong test. The radiologists' performance with or without deep learning assistance were compared using McNemar's χ^2 test. To interpret the CNN predictions, gradient-weighted class activation mapping (Grad-CAM)

Table 1. Demographic data for 1239 patients.

Patient cohort	**Training and validation sets ($n = 1093$)**		***P* value**	**Internal test set ($n = 100$)**		***P* value**	**External test set ($n = 46$)**		***P* value**
	Benign ($n = 288$)	**Malignant ($n = 805$)**		**Benign ($n = 50$)**	**Malignant ($n = 50$)**		**Benign ($n = 13$)**	**Malignant ($n = 33$)**	
Age, years (mean ± SD)	43.26 ± 11.05	54.61 ± 10.25	<0.001*	42.82 ± 11.96	54.74 ± 9.01	<0.001*	46.54 ± 6.15	53.30 ± 8.27	0.005*
Diameter, cm (%)	2.11 ± 1.64	2.49 ± 1.22	<0.001*	2.10 ± 2.26	2.39 ± 0.99	0.41	1.63 ± 0.67	2.63 ± 0.93	<0.001*
≤1	89 (30.9)	35 (4.3)		20 (40)	3 (6)		2 (15.4)	0	
1–2	95 (33)	301 (37.4)		13 (26)	18 (36)		9 (69.2)	8 (24.2)	
>2	104 (36.1)	469 (58.3)		17 (34)	29 (58)		2 (15.4)	25 (75.8)	
BI-RADS (%)			<0.001*			0.22			0.87
3	17 (5.9)	1 (0.1)		0	0		0	0	
4A	105 (36.5)	21 (2.6)		24 (48)	0		6 (46.2)	2 (6.1)	
4B	108 (37.5)	122 (15.2)		13 (26)	6 (12)		6 (46.2)	10 (30.3)	
4C	52 (18)	296 (36.8)		13 (26)	21 (42)		1 (7.6)	12 (36.4)	
5	6 (2.1)	365 (45.3)		0	23 (46)		0	9 (27.2)	
Lesion types (%)			–			–			–
Benign lesions									
Fibroadenoma	124 (43.1)	–		23 (46)	–		6 (46.2)	–	
Adenosis	68 (23.6)	–		12 (24)	–		7 (53.8)	–	
Intraductal papilloma	42 (14.6)	–		4 (8)	–		0	–	
Inflammation	9 (3.1)	–		2 (4)	–		0	–	
Phyllodes tumour	9 (3.1)	–		2 (4)	–		0	–	
Tubular adenoma	0	–		1 (2)	–		0	–	
Fibrocystic disease	14 (4.9)	–		1 (2)	–		0	–	
Unknown/other	22 (7.6)	–		5 (10)	–		0	–	
Malignant lesions									
IDC	–	689 (85.59)		–	47 (94)		–	33 (100)	
DCIS	–	30 (3.73)		–	1 (2)		–	0	
ILC	–	17 (2.11)		–	2 (4)		–	0	
Papillary carcinoma	–	17 (2.11)		–	0		–	0	
MAC	–	17 (2.11)		–	0		–	0	
Unknown/other	–	35 (4.35)		–	0		–	0	

DCIS ductal carcinoma in situ, *IDC* invasive ductal carcinoma, *ILC* invasive lobular carcinoma, *MAC* mucinous adenocarcinoma.
*$P < 0.05$.

was used to produce heat maps to visualise the most indicative regions [26]. $P < 0.05$ was statistically significant.

Ethics statement

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the institutional review board of Yantai Yuhuangding Hospital and Fudan University Cancer Center.

RESULTS

Clinical characteristics

A total of 1239 patients were enrolled in this study. Of the 1239 patients, 805 malignant breast lesions and 288 benign lesions formed the training and validation sets, 50 malignant breast lesions and 50 benign lesions formed the internal test set, and 13 malignant breast lesions and 33 benign lesions formed the external test set. The age was 51.62 ± 11.62 years (range, 17–85 years) for the training and validation sets, 48.0 ± 12.38 years (range, 17–80 years) for internal test set, and 51.39 ± 8.26 years (range, 36–70 years) for external test set. The patient characteristics of the three sets are shown in Table 1.

Segmentation similarity

The DSC between Segmenters 1 and 2 was 0.87. The DSC between Segmenters 1 and 3 was 0.88. The DSC between Segmenters 2 and 3 was 0.89. The average DSC across all segmenters was 0.90 in benign lesions and 0.87 in malignant lesions.

Performance of deep learning models

All CNN models exhibited good performance based on CESM on the training and test sets. After incorporating the attention mechanism into the CNNs models, these models exhibited improved performance, with an area under the ROC curves (AUCs) of 0.912 for the CBAM-based Xception model, 0.934 for the CBAM-based DenseNet model, and 0.872 for the CBAM-based ResNet model on the internal test set; with AUCs of 0.970 for the CBAM-based Xception model, 0.918 for the CBAM-based DenseNet model, and 0.876 for the CBAM-based ResNet model on external test set (Fig. 2). The performance comparison of different CNN models is shown in Fig. 2 and Table 2. The confusion matrices and the PRCs also indicated the favourable predictive performance of the CNN models (Fig. 3 and S3).

The best-performing deep learning model, that is, CBAM-based Xception, achieved an AUC of 0.912, an area under the PRC (AUPRC) of 0.923, a F1 score of 0.835, an accuracy of 0.850, a sensitivity of 0.760, a specificity of 0.940, a PPV of 0.970, and a NPV of 0.797 on the internal test set; and an AUC of 0.970, an AUPRC of 0.988, a F1 score of 0.918, an accuracy of 0.891, a sensitivity of 0.848, a specificity of 1.000, a PPV of 1.000, and a NPV of 0.722 on the external test set. The high-sensitivity threshold, high-specificity threshold, and cutoff point maximising the Youden index were calculated as 0.60, 0.97, and 0.95, respectively. The performance of CBAM-based Xception model for high-sensitivity and high-specificity thresholds is shown in Table 3. The high-sensitivity threshold of the model could prevent patients with benign breast lesions from undergoing unnecessary biopsy on the premise of avoiding missed diagnosis, with specificity of 0.620 and NPV of 0.939 in the internal test set, and specificity of 0.769 and NPV of 0.909 in the external test set, respectively. Moreover, the high-specificity threshold of the model could reduce missed diagnosis rate of malignant breast lesions on the premise of avoiding misdiagnose, with sensitivity of 0.700 and PPV of 0.923 in the internal test set, and sensitivity of 0.848 and PPV of 1.000 in the external test set, respectively.

As shown in Fig. 4, the heat maps obtained by the CBAM-based Xception model from the CESM images visualised the most indicative regions. The red and yellow regions have higher predictive significance than the green and blue regions, indicating that CNN focuses on the most predictive information for distinguishing benign from malignant breast lesions.

Moreover, we analysed the failure cases misdiagnosed by the CBAM-based Xception model, and ideally in comparison with radiologists. In some cases, glandular structures overlapped on low-energy images resulting in edge occlusion. In some cases, the edges of enhancing lesions were blurred. Two representative cases were shown in Fig. 4d, e.

Comparison with radiologists' performance

The predictive performance of the deep learning models was compared with that of two radiologists. Table 4 shows the accuracy, the sensitivity, and the specificity for Radiologist 1 and 2 on the internal and external tests. The inter-rater agreement between two radiologists was moderately high, with the Cohen's kappa values of 0.793 and 0.748 in the internal and external test sets, respectively (Table 4). The ROC curve of the best-performing model, that is, CBAM-based Xception, is shown in Fig. 5. The specificity and sensitivity points of the radiologists on the test sets were plotted in the ROC curve. The performance of the radiologists lies below that of our proposed model in the same ROC curve. On CESM images, the attention-based CNN performed best and can effectively predict benign from malignant breast lesions.

Comparison with radiomics models

Although the LR exhibited the best performance in radiomics models, with respective AUCs of 0.795 and 0.825 on the internal and external test sets, it performed worse than the CNN models ($P = 0.029$, 0.027, for CBAM-based Xception vs LR-based radiomics

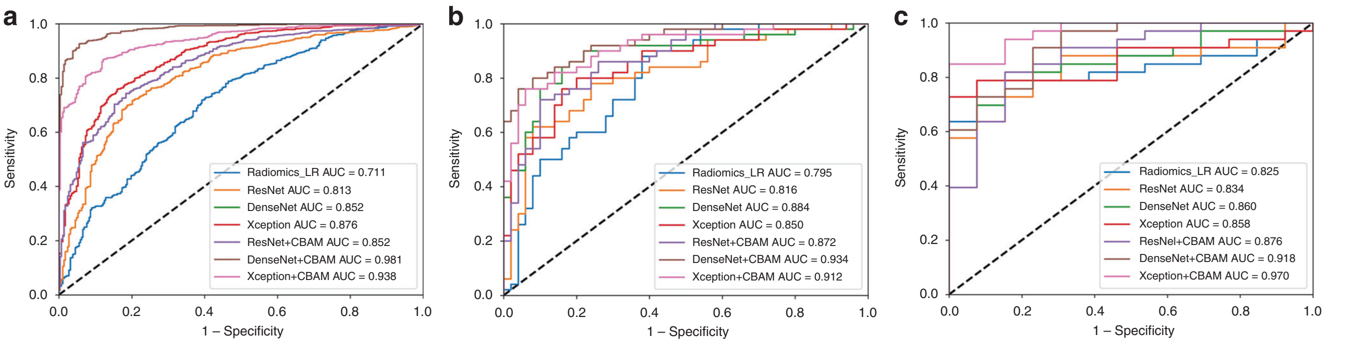

Fig. 2 Receiver operating characteristic curves of different models. Receiver operating characteristic curves of different convolutional neural network models and best-performing radiomics model in training and validation sets (**a**), internal (**b**), and external (**c**) test sets.

Table 2. The performance of different CNN models and best-performing radiomics model.

Performance metric	LR	ResNet	DenseNet	Xception	ResNet + CBAM	DenseNet + CBAM	Xception + CBAM
Training and validation sets							
AUC (95% CI)	0.711 (0.681–0.743)	0.813 (0.784–0.842)	0.852 (0.828–0.877)	0.876 (0.852–0.899)	0.852 (0.828–0.877)	0.981 (0.974–0.987)	0.938 (0.925–0.952)
ACC (95% CI)	0.689 (0.661–0.716)	0.738 (0.711–0.764)	0.762 (0.736–0.787)	0.780 (0.754–0.804)	0.762 (0.736–0.787)	0.931 (0.914–0.945)	0.871 (0.850–0.890)
SENS (95% CI)	0.718 (0.685–0.749)	0.717 (0.684–0.747)	0.743 (0.711–0.772)	0.758 (0.726–0.787)	0.743 (0.711–0.858)	0.925 (0.905–0.942)	0.868 (0.843–0.891)
SPEC (95% CI)	0.608 (0.548–0.664)	0.799 (0.746–0.842)	0.816 (0.765–0.858)	0.840 (0.792–0.880)	0.816 (0.765–0.858)	0.944 (0.910–0.967)	0.878 (0.834–0.912)
PPV (95% CI)	0.836 (0.806–0.863)	0.909 (0.883–0.929)	0.919 (0.894–0.938)	0.930 (0.907–0.948)	0.919 (0.894–0.938)	0.979 (0.965–0.988)	0.952 (0.934–0.966)
NPV (95% CI)	0.435 (0.386–0.485)	0.502 (0.455–0.549)	0.532 (0.484–0.579)	0.554 (0.506–0.601)	0.532 (0.484–0.579)	0.819 (0.773–0.858)	0.705 (0.654–0.751)
F1 score	0.773	0.801	0.821	0.835	0.821	0.951	0.908
P value	–	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Internal test set							
AUC (95% CI)	0.795 (0.719–0.868)	0.816 (0.731–0.901)	0.884 (0.815–0.953)	0.850 (0.774–0.926)	0.872 (0.805–0.940)	0.934 (0.890–0.979)	0.912 (0.857–0.968)
ACC (95% CI)	0.750 (0.653–0.831)	0.770 (0.675–0.848)	0.786 (0.775–0.797)	0.800 (0.708–0.873)	0.810 (0.719–0.882)	0.860 (0.776–0.921)	0.850 (0.765–0.914)
SENS (95% CI)	0.880 (0.750–0.950)	0.620 (0.472–0.750)	0.900 (0.774–0.963)	0.800 (0.659–0.895)	0.720 (0.573–0.833)	0.800 (0.659–0.895)	0.760 (0.615–0.865)
SPEC (95% CI)	0.620 (0.472–0.750)	0.920 (0.799–0.974)	0.780 (0.637–0.880)	0.800 (0.659–0.895)	0.900 (0.774–0.963)	0.920 (0.799–0.895)	0.940 (0.825–0.984)
PPV (95% CI)	0.698 (0.568–0.804)	0.886 (0.723–0.963)	0.804 (0.659–0.895)	0.800 (0.659–0.895)	0.878 (0.730–0.954)	0.909 (0.774–0.970)	0.927 (0.790–0.981)
NPV (95% CI)	0.838 (0.673–0.932)	0.708 (0.580–0.811)	0.886 (0.659–0.895)	0.800 (0.659–0.895)	0.763 (0.631–0.860)	0.821 (0.692–0.907)	0.797 (0.668–0.886)
F1 score	0.779	0.729	0.849	0.800	0.791	0.851	0.835
P value	–	0.746	0.115	0.365	0.188	0.006	0.029
External test set							
AUC (95% CI)	0.825 (0.708–0.909)	0.834 (0.717–0.952)	0.860 (0.754–0.967)	0.858 (0.751–0.965)	0.876 (0.764–0.989)	0.918 (0.835–1.000)	0.970 (0.929–1.000)
ACC (95% CI)	0.717 (0.565–0.840)	0.696 (0.543–0.823)	0.826 (0.686–0.922)	0.804 (0.661–0.906)	0.826 (0.686–0.922)	0.870 (0.737–0.951)	0.891 (0.764–0.964)
SENS (95% CI)	0.606 (0.422–0.766)	0.576 (0.394–0.740)	0.818 (0.639–0.924)	0.727 (0.542–0.861)	0.818 (0.818–0.846)	0.909 (0.745–0.976)	0.848 (0.673–0.943)
SPEC (95% CI)	1.000 (0.717–1.000)	1.000 (0.717–1.000)	0.846 (0.537–0.973)	1.000 (0.717–1.000)	0.846 (0.537–0.973)	0.769 (0.460–0.938)	1.000 (0.717–1.000)
PPV (95% CI)	1.000 (0.800–1.000)	1.000 (0.791–1.000)	0.931 (0.758–0.988)	1.000 (0.828–1.000)	0.931 (0.758–0.988)	0.901 (0.745–0.976)	1.000 (0.850–1.000)
NPV (95% CI)	0.500 (0.304–0.696)	0.481 (0.292–0.676)	0.647 (0.386–0.847)	0.591 (0.367–0.785)	0.647 (0.386–0.847)	0.769 (0.460–0.938)	0.722 (0.464–0.893)
F1 score	0.755	0.731	0.871	0.842	0.871	0.909	0.918
P value	–	0.913	0.668	0.690	0.540	0.211	0.027

CNN convolutional neural network, *DT* decision tree, *CBAM* convolutional block attention module, *CI* confidence interval, *AUC* area under the curve, *ACC* accuracy, *SENS* sensitivity, *SPEC* specificity, *PPV* positive predictive value, *NPV* negative predictive value.

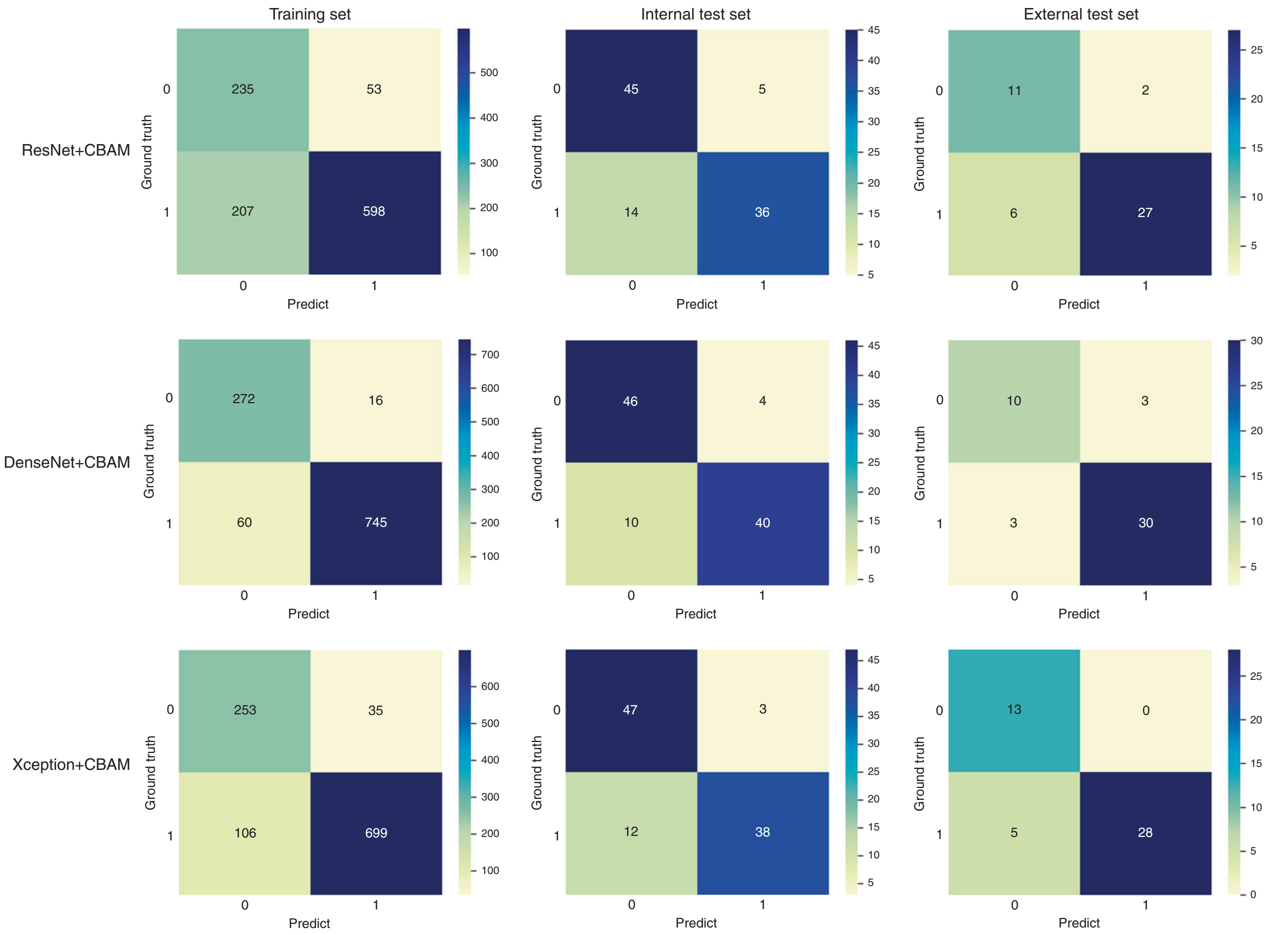

Fig. 3 The confusion matrices for CBAM-based CNN models across the three sets. CBAM convolutional block attention module, CNN convolutional neural network.

Table 3. The performance of CBAM-based Xception model for high-sensitivity and high-specificity thresholds.

Performance metric	AUC (95% CI)	ACC (95%CI)	SENS (95% CI)	SPEC (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 score
High sensitivity							
Internal test set	0.912 (0.857–0.968)	0.790 (0.697–0.865)	0.960 (0.851–0.993)	0.620 (0.472–0.750)	0.716 (0.591–0.817)	0.939 (0.784–0.989)	0.820
External test set	0.970 (0.929–1.000)	0.913 (0.795–0.962)	0.970 (0.825–0.998)	0.769 (0.460–0.938)	0.914 (0.758–0.978)	0.909 (0.571–0.995)	0.939
High specificity							
Internal test set	0.912 (0.857–0.968)	0.830 (0.742–0.898)	0.700 (0.552–0.817)	0.940 (0.825–0.984)	0.923 (0.780–0.980)	0.770 (0.642–0.865)	0.795
External test set	0.970 (0.929–1.000)	0.891 (0.764–0.964)	0.848 (0.673–0.943)	1.000 (0.717–1.000)	1.000 (0.850–1.000)	0.722 (0.464–0.893)	0.918

AUC area under the curve, *ACC* accuracy, *SENS* sensitivity, *SPEC* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *CI* confidence interval.

model on the internal and external test sets, respectively). The comparison is shown in Fig. 2 and Table 2. The selected radiomics features and the performance of different radiomics models are shown in Supplementary Materials 6, Tables S1 and S2, and Fig. S4.

Radiologists' performance with deep learning assistance

The performance of the radiologists with CBAM-based Xception model assistance is presented in Fig. 5 and Table 5. For Radiologist 1, CBAM-based Xception model assistance led to increase in accuracy from 0.780 to 0.890 ($P = 0.005$) on the internal test set and from 0.804 to 0.891 ($P = 0.134$) on the external test set; and in specificity from 0.680 to 0.920 ($P = 0.005$) and from 0.692 to 0.923 ($P = 0.083$), respectively. For Radiologist 2, CBAM-based Xception model assistance led to increase in accuracy from 0.800 to 0.900 ($P = 0.004$) on the internal test set and from 0.847 to 0.891 ($P = 0.480$) on the external test set; and in specificity from 0.720 to 0.900 ($P = 0.003$) and from 0.769 to 0.923 ($P = 0.157$), respectively. For Radiologist 1 and 2, deep learning assistance resulted no significant changes in sensitivity. The inter-rater agreement between two radiologists was significantly improved after the CBAM-based Xception model assistance, with the Cohen's kappa values of 0.900 and 0.808 in the internal and external test sets, respectively (Table 5). The results showed that the CBAM-based Xception model could effectively reduce inter-rater variability.

BI-RADS 4 subgroup analysis

According to the American College of Radiology, BI-RADS 4 category is used for findings that have ≥2% but <95% chance of malignancy and for which biopsy is recommended [27]. Using CESM images, attention-based CNN model can effectively predict benign from malignant breast lesions in the BI-RADS 4 subgroup. The CBAM-based Xception model achieved AUCs of 0.906 on the

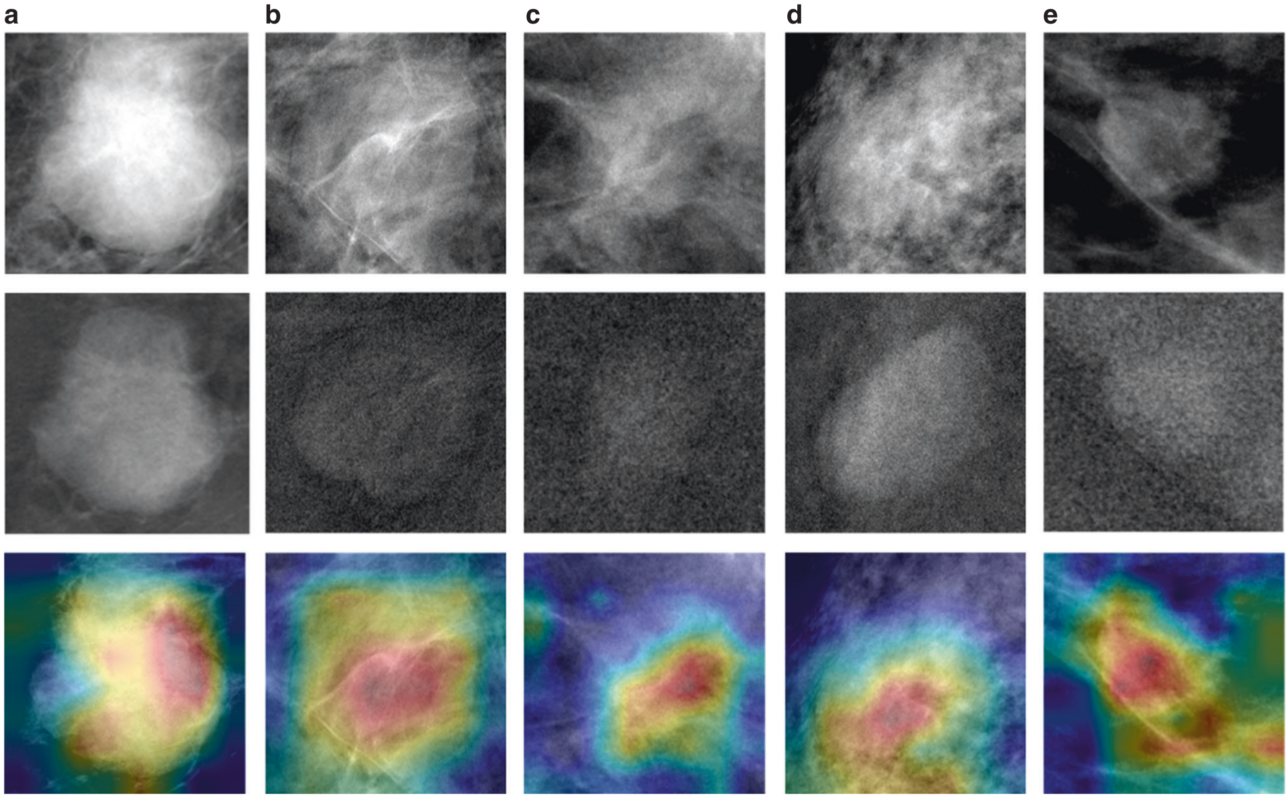

Fig. 4 CESM images and heat maps of three breast lesions. The heat maps obtained by the deep learning from the CESM images visualised the most indicative regions. The red and yellow regions have higher predictive significance than the green and blue regions. **a** Images in a 41-year-old woman with invasive ductal carcinoma. CESM images shows a 4.2-cm mass, BI-RADS 4C. The deep learning model and two radiologists correctly predicted the malignant lesion. **b** Images in a 45-year-old woman with fibroadenoma. CESM image shows a 2.7-cm mass, BI-RADS 4A. The deep learning model and two radiologists correctly predicted the benign lesion. **c** Images in a 48-year-old woman with intraductal papillomas. CESM images shows a 1.2-cm mass, BI-RADS 4B.The deep learning model correctly predicted the benign lesion, whereas two radiologists made the wrong prediction. **d** Images in a 47-year-old woman with breast tubular adenoma. CESM images shows a 2.7-cm mass, BI-RADS 4B.The deep learning model erroneously predicted the malignancy lesion, whereas two radiologists made the correct prediction. **e** Images in a 45-year-old woman with breast adenosis tumour. CESM images shows a 0.8-cm mass, BI-RADS 4A.The deep learning model erroneously predicted the malignancy lesion, whereas two radiologists made the correct prediction. CESM contrast enhanced spectral mammography, BI-RADS Breast Imaging Reporting and Data System.

internal test set and 0.971 on the external test set. The performance is shown in Fig. 6, Fig. S5, and Table S3.

Different lesion diameter subgroups analysis

Attention-based CNN models can effectively predict benign from malignant breast lesions in the different lesion diameter subgroups. Within the lesion diameter ≤2 cm subgroup, the CBAM-based Xception model achieved AUCs of 0.887 on the internal test set and 0.977 on the external test set. Within the lesion diameter >2 cm subgroup, the CBAM-based Xception model achieved AUCs of 0.927 on the internal test set and 0.940 on the external test set. The performance is shown in Fig. 6, Fig. S5, and Table S3.

DISCUSSION

In this paper, we present an attention-based deep learning model for distinguishing benign from malignant breast lesions based on CESM with a satisfactory performance. The best-performing model, that is, the CBAM-based Xception model, yielded an AUC of 0.970 on the external test set. In addition, this model was tested on multicentre data, suggesting the strong generalisation ability of the model. Moreover, the model performed well on the BI-RADS 4 subgroup and different lesion diameter subgroups. Finally, we demonstrated the superior performance of deep learning over the radiologists with specialised breast imaging experience, and deep learning assistance may improve the performance of radiologists. To the best of our knowledge, this is the first study to apply attention-based deep learning in distinguishing benign from malignant breast lesions and the largest sample size study on CESM.

In recent years, artificial intelligence has made substantial progress in medical decision support. Several studies developed some radiomics models based on traditional machine learning, such as LR, KNN, and SVM, to preoperatively discriminate benign and malignant breast lesions [28–36]. However, these radiomic features were hand-craft and defined based on experience. In addition, traditional machine learning method may not improve the model performance further with the increase of the training samples. Deep learning, especially CNN, can automatically extract high-throughput features and has been widely used in the field of breast cancer detection and diagnosis [36–41]. Recently, Truhn et al. [12] used deep learning and radiomics to differentiate benign from malignant breast lesions on multiparametric MRIs. Their results show that CNN's performance (AUC, 0.88) is superior to that of radiomics (AUC, 0.81). Dalmiş et al. [42] used deep learning and clinical information to differentiate benign from malignant breast lesions on multiparametric MRIs. Their results show that the combination of deep learning and clinical information has a significant contribution to the differentiation of benign from malignant breast lesions on breast MRIs with an AUC of 0.88. These studies are limited by the single centre and the small sample size. In the present study, our deep learning model

Table 4. The performance of best-performing CNN model was compared with two radiologists.

Performance metric	Xception + CBAM	R 1	*P* value	R 2	*P* value
Internal test set					
ACC (95% CI)	0.850 (0.765–0.914)	0.780 (0.686–0.857)	0.324	0.800 (0.708–0.873)	0.499
SENS (95% CI)	0.760 (0.615–0.865)	0.880 (0.760–0.950)	0.157	0.880 (0.760–0.950)	0.157
SPEC (95% CI)	0.940 (0.825–0.984)	0.680 (0.532–0.801)	0.003	0.720 (0.573–0.833)	0.008
PPV (95% CI)	0.927 (0.790–0.981)	0.733 (0.601–0.835)	0.001	0.759 (0.625–0.857)	0.015
NPV (95% CI)	0.797 (0.668–0.886)	0.850 (0.695–0.938)	0.455	0.857 (0.708–0.941)	0.387
F1 score	0.835	0.767	–	0.792	–
Cohen's kappa	0.793				
External test set					
ACC (95% CI)	0.891 (0.764–0.964)	0.804 (0.661–0.906)	0.423	0.847 (0.702–0.939)	0.773
SENS (95% CI)	0.848 (0.673–0.943)	0.848 (0.673–0.943)	1	0.879 (0.709–0.960)	0.739
SPEC (95% CI)	1.000 (0.717–1.000)	0.692 (0.389–0.896)	0.046	0.769 (0.460–0.938)	0.083
PPV (95% CI)	1.000 (0.850–1.000)	0.875 (0.701–0.959)	0.046	0.906 (0.738–0.975)	0.083
NPV (95% CI)	0.722 (0.464–0.893)	0.643 (0.356–0.860)	0.589	0.714 (0.420–0.904)	0.955
F1 score	0.918	0.861	–	0.892	–
Cohen's kappa	0.748				

CNN convolutional neural network, *CBAM* convolutional block attention module, *R* radiologist, *AUC* area under the curve, *ACC* accuracy, *SENS* sensitivity, *SPEC* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *CI* confidence interval.

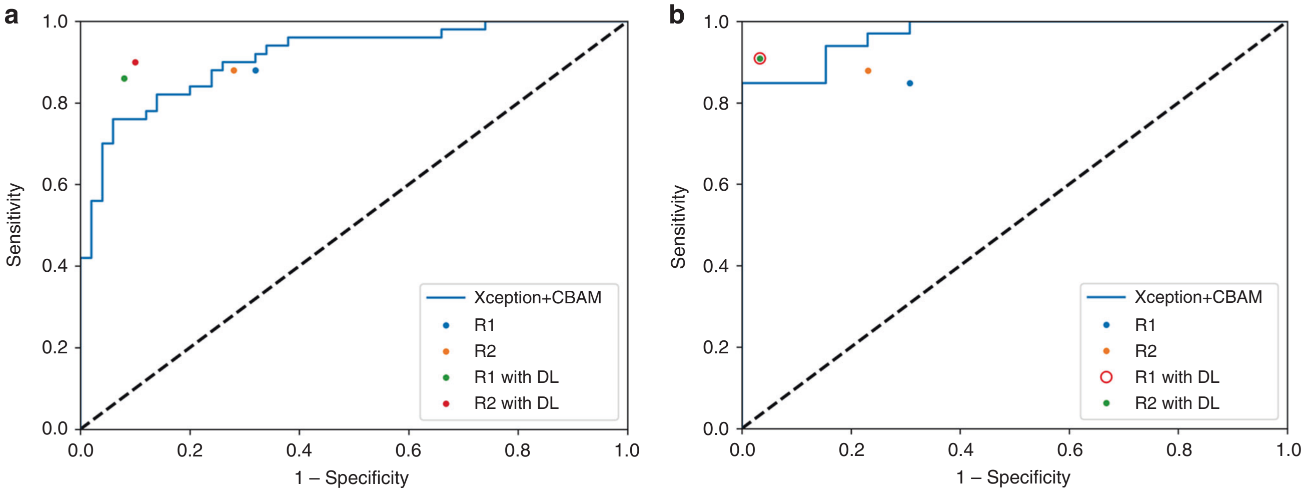

Fig. 5 Receiver operating characteristic curves of CBAM-based Xception model and radiologists' performance. Receiver operating characteristic curves of CBAM-based Xception model and radiologists' performance in internal (**a**) and external (**b**) test sets. CBAM convolutional block attention module.

performed well and was tested on multicentre data. For comparing the performance of the models in the present study, we listed all previous studies on the classification of benign and malignant breast lesions in Table S4.

Our study has several differences from previous studies. First, we selected CESM instead of MRI or ultrasound. Patel et al. [10] were the first to apply radiomics on CESM and develop a computer-aided diagnostic tool to test whether false positives in breast cancer screening can be reduced. The SVM-based radiomics model has higher specificity and accuracy than the prediction performance of the radiologists, with an overall accuracy of 90%. Perek et al. [15] and Song et al. [16] applied deep learning on CESM, and developed a deep learning-based decision support system to test whether the performance in breast cancer diagnosis can be improved. However, the sample size of these studies was small, and these studies lacked external test set. Although CESM is a relatively new technology, we still collected the largest data set compared with the currently published studies and tested the generalisation ability of our model on multicentre data. Second, we incorporated the attention module into different deep learning networks to increase the effective weight features, thereby reducing the invalid weight features and improving the performance of the networks. Previous studies showed that squeeze-and-excitation networks (SENet) with a channel attention module can improve the predictive performance of breast imaging [43]. In the present study, the CBAM not only includes a channel attention module but also a spatial attention module. The three attention-based CNNs achieved a better performance than the CNNs without an attention mechanism. Third, previous studies showed that integrating clinical information can improve the performance of models in other diseases [13, 39, 42, 44]. However, to address the instability of the model, we did not integrate clinical information. Fourth, we compared the performance of the deep learning models with that of conventional

Table 5. Radiologists' performance with deep learning assistance.

Performance metric	R 1 with DL	*P* value	R 2 with DL	*P* value
Internal test set				
ACC (95% CI)	0.890 (0.812–0.944)	0.005	0.900 (0.824–0.951)	0.004
SENS (95% CI)	0.860 (0.726–0.937)	0.317	0.900 (0.774–0.962)	0.317
SPEC (95% CI)	0.920 (0.799–0.974)	0.0005	0.900 (0.774–0.962)	0.003
PPV (95% CI)	0.915 (0.787–0.972)	0.0006	0.900 (0.774–0.962)	0.002
NPV (95% CI)	0.868 (0.740–0.941)	0.449	0.900 (0.774–0.962)	0.083
F1 score	0.887	–	0.900	–
Cohen's kappa	0.900			
External test set				
ACC (95% CI)	0.891 (0.764–0.964)	0.134	0.891 (0.764–0.964)	0.480
SENS (95% CI)	0.879 (0.709–0.960)	0.317	0.879 (0.709–0.960)	1
SPEC (95% CI)	0.923 (0.621–0.996)	0.083	0.923 (0.621–0.996)	0.157
PPV (95% CI)	0.967 (0.809–0.998)	0.080	0.967 (0.809–0.998)	0.157
NPV (95% CI)	0.750 (0.474–0.917)	0.122	0.750 (0.474–0.917)	0.248
F1 score	0.921	–	0.921	–
Cohen's kappa	0.808			

R Radiologist, *DL* deep learning, *AUC* areas under the curve, *ACC* accuracy, *SENS* sensitivity, *SPEC* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *CI* confidence interval.

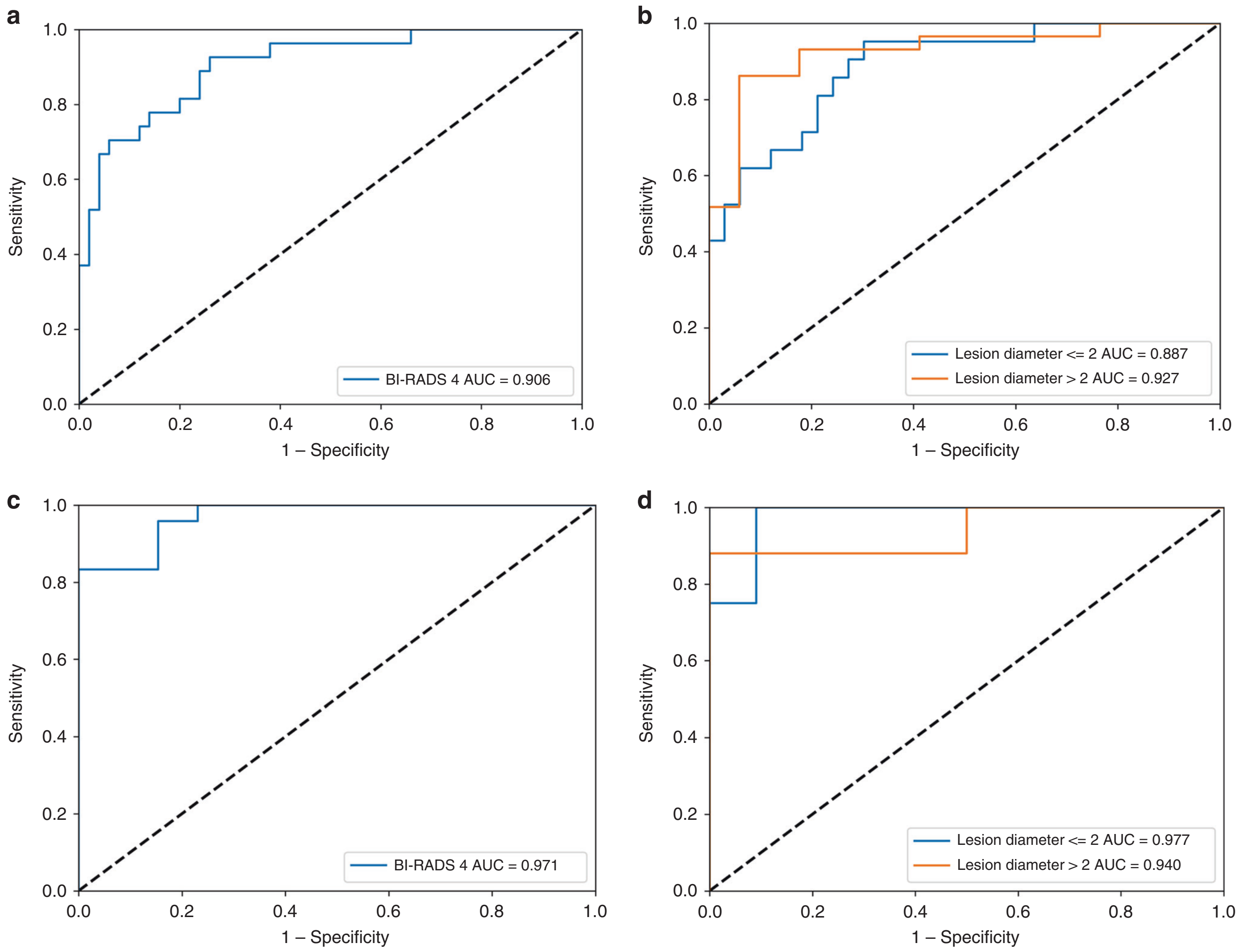

Fig. 6 BI-RADS 4 and different lesion diameter subgroups analysis. a, b Receiver operating characteristic curves of CBAM-based Xception model in internal test set. **c, d** Receiver operating characteristic curves of CBAM-based Xception model in external test set. BI-RADS Breast Imaging Reporting and Data System, CBAM convolutional block attention module.

radiomics model and radiologists and investigated the radiologist's performance with deep learning assistance, which few previous studies have tried. The deep learning model performed better than the conventional radiomics model and the radiologists, displaying clinical application prospects. In addition, the radiologists' performance improved with deep learning assistance. Lastly, to avoid missed diagnosis and biopsy, we defined the high-sensitivity and high-specificity thresholds.

Our study has several limitations. First, the patients enrolled in this study were collected from only two centres. Further study will explore the diagnostic performance of the model in more centres and larger sample sizes of external test sets. Second, the manual method was applied to image segmentation, although satisfactory DSC was achieved. However, automatic method is without human intervention from imagine segmentation to prediction model construction, which can improve the repeatability of research and practicability of the model. We did not find any literature reporting on the automatic segmentation of breast lesion on CESM. Third, this study only enrolled the largest lesion of patients with multiple masses, and the model was not applied to breast lesions smaller than 5 mm. Future studies will try to incorporate all lesions as possible. Lastly, only the CC view image was selected for analysis. Compared with single view, multiple views can provide more information and further improve model performance. Automatic segmentation and multiple views analysis with a larger sample size will be incorporated in future study.

Therefore, in this multicentre study, we developed an attention-based deep learning model using CESM to noninvasively distinguish between benign and malignant breast lesions. The diagnostic performance of radiologists improved with deep learning assistance. More centres in different countries must be used and a fully end-to-end deep learning model should be performed to improve the performance and for high-level evidence of clinical application in future work.

DATA AVAILABILITY

Some or all data, models, or code generated or used during the study are available from the corresponding author by request.

REFERENCES

1. DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast cancer statistics, 2019. CA Cancer J Clin. 2019;69:438–51.
2. Pace LE, Keating NL. A systematic assessment of benefits and risks to guide breast cancer screening decisions. JAMA. 2014;311:1327–35.
3. Puliti D, Duffy SW, Miccinesi G, de Koning H, Lynge E, Zappa M, et al. Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. J Med Screen. 2012;19:42–56.
4. Lee-Felker SA, Tekchandani L, Thomas M, Gupta E, Andrews-Tang D, Roth A, et al. Newly diagnosed breast cancer: comparison of contrast-enhanced spectral mammography and breast MR imaging in the evaluation of extent of disease. Radiology. 2017;285:389–400.
5. Mao N, Yin P, Li Q, Wang Q, Liu M, Ma H, et al. Radiomics nomogram of contrast-enhanced spectral mammography for prediction of axillary lymph node metastasis in breast cancer: a multicenter study. Eur Radiol. 2020;30:6732–9.
6. Mao N, Yin P, Wang Q, Liu M, Dong J, Zhang X, et al. Added value of radiomics on mammography for breast cancer diagnosis: a feasibility study. J Am Coll Radiol. 2019;16:485–91.
7. Mao N, Wang Q, Liu M, Dong J, Xiao C, Sun N, et al. Computerized image analysis to differentiate benign and malignant breast tumors on magnetic resonance diffusion weighted image: a preliminary study. J Comput Assist Tomogr. 2019;43:93–97.
8. Liu Z, Li Z, Qu J, Zhang R, Zhou X, Li L, et al. Radiomics of multiparametric MRI for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. Clin Cancer Res. 2019;25:3538–47.
9. Li H, Zhu Y, Burnside ES, Drukker K, Hoadley KA, Fan C, et al. MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, Oncotype DX, and PAM50 gene assays. Radiology. 2016;281:382–91.
10. Patel BK, Ranjbar S, Wu T, Pockaj BA, Li J, Zhang N, et al. Computer-aided diagnosis of contrast-enhanced spectral mammography: a feasibility study. Eur J Radiol. 2018;98:207–13.
11. Lin F, Wang Z, Zhang K, Yang P, Ma H, Shi Y, et al. Contrast-enhanced spectral mammography-based radiomics nomogram for identifying benign and malignant breast lesions of sub-1 cm. Front Oncol. 2020;10:573630.
12. Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI. Radiology. 2019;290:290–7.
13. Zheng X, Yao Z, Huang Y, Yu Y, Wang Y, Liu Y, et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nat Commun. 2020;11:1236.
14. Jiang M, Li CL, Luo XM, Chuan ZR, Lv WZ, Li X, et al. Ultrasound-based deep learning radiomics in the assessment of pathological complete response to neoadjuvant chemotherapy in locally advanced breast cancer. Eur J Cancer. 2021;147:95–105.
15. Perek S, Kiryati N, Zimmerman-Moreno G, Sklair-Levy M, Konen E, Mayer A. Classification of contrast-enhanced spectral mammography (CESM) images. Int J Comput Assist Radiol Surg. 2019;14:249–57.
16. Song J, Zheng Y, Zakir Ullah M, Wang J, Jiang Y, Xu C, et al. Multiview multimodal network for breast cancer diagnosis in contrast-enhanced spectral mammography images. Int J Comput Assist Radiol Surg. 2021;16:979–88.
17. Wang H, Wang S, Qin Z, Zhang Y, Li R, Xia Y. Triple attention learning for classification of 14 thoracic diseases using chest radiography. Med Image Anal. 2021;67:101846.
18. Zhang R, Duan H, Cheng J, Zheng Y. A study on tuberculosis classification in chest X-ray using deep residual attention networks. Annu Int Conf IEEE Eng Med Biol Soc. 2020;2020:1552–5.
19. Woo S, Park J, Lee J-Y, Kweon IS. CBAM: convolutional block attention module. arXiv:1807.06521v2 [Preprint]. 2018 [cited 2018 Jul 18]: [17 p]. Available from: https://arxiv.org/abs/1807.06521
20. Chen H, Yang BW, Qian L, Meng YS, Bai XH, Hong XW, et al. Deep learning prediction of ovarian malignancy at US compared with O-RADS and expert assessment. Radiology 2022;304:106–13.
21. Huang G, Liu Z, Maaten Lvd, Weinberger KQ. Densely connected convolutional networks. arXiv:1608.06993v5 [Preprint]. 2018 [cited 2018 Jan 28]: [9 p]. Available from: https://arxiv.org/abs/1608.06993
22. Chollet F. Xception: deep learning with depthwise separable convolutions. arXiv:1610.02357v3 [Preprint]. 2018 [cited 2018 Apr 4]: [8 p]. Available from: https://arxiv.org/abs/1610.02357
23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv:1512.03385v1 [Preprint]. 2015 [cited 2015 Dec 10]: [12 p]. Available from: https://arxiv.org/abs/1512.03385
24. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115:211–52.
25. Jiang Y, Zhang Z, Yuan Q, Wang W, Wang H, Li T, et al. Predicting peritoneal recurrence and disease-free survival from CT images in gastric cancer with multitask deep learning: a retrospective study. Lancet Digital Health. 2022;4:e340–50.
26. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128:336–59.
27. Lee CH, Phillips J, Sung JS, Lewin JM, Newell MS. Contrast enhanced mammography (CEM) (A supplement to ACR BI-RADS® Mammography 2013). American College of Radiology. 2022. https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/BIRADS_CEM_2022.pdf
28. Whitney HM, Taylor NS, Drukker K, Edwards AV, Papaioannou J, Schacht D, et al. Additive benefit of radiomics over size alone in the distinction between benign lesions and luminal a cancers on a large clinical breast MRI dataset. Acad Radiol. 2019;26:202–9.
29. Parekh VS, Jacobs MA. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. NPJ Breast Cancer. 2017;3:43.
30. Bickelhaupt S, Jaeger PF, Laun FB, Lederer W, Daniel H, Kuder TA, et al. Radiomics based on adapted diffusion kurtosis imaging helps to clarify most mammographic findings suspicious for cancer. Radiology. 2018;287:761–70.
31. Zhang Q, Xiao Y, Suo J, Shi J, Yu J, Guo Y, et al. Sonoelastomics for breast tumor classification: a radiomics approach with clustering-based feature selection on sonoelastography. Ultrasound Med Biol. 2017;43:1058–69.
32. Qiao M, Li C, Suo S, Cheng F, Hua J, Xue D, et al. Breast DCE-MRI radiomics: a robust computer-aided system based on reproducible BI-RADS features across the influence of datasets bias and segmentation methods. Int J Comput Assist Radiol Surg. 2020;15:921–30.
33. Zhao S, Zhang X, Zhong H, Qin Y, Li Y, Song B, et al. Background parenchymal enhancement on contrast-enhanced spectral mammography: influence of age,

breast density, menstruation status, and menstrual cycle timing. Sci Rep. 2020;10:8608.

34. Lei C, Wei W, Liu Z, Xiong Q, Yang C, Yang M, et al. Mammography-based radiomic analysis for predicting benign BI-RADS category 4 calcifications. Eur J Radiol. 2019;121:108711.
35. Wang S, Sun Y, Li R, Mao N, Li Q, Jiang T, et al. Diagnostic performance of perilesional radiomics analysis of contrast-enhanced mammography for the differentiation of benign and malignant breast lesions. Eur Radiol. 2021;32:639–49.
36. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18:500–10.
37. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol. 2017;52:434–40.
38. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 2019;292:60–6.
39. Zhang Q, Peng Y, Liu W, Bai J, Zheng J, Yang X, et al. Radiomics based on multimodal MRI for the differential diagnosis of benign and malignant breast lesions. J Magn Reson Imaging. 2020;52:596–607.
40. Guo X, Liu Z, Sun C, Zhang L, Wang Y, Li Z, et al. Deep learning radiomics of ultrasonography: Identifying the risk of axillary non-sentinel lymph node involvement in primary breast cancer. EBioMedicine. 2020;60:103018.
41. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577:89–94.
42. Dalmis MU, Gubern-Merida A, Vreemann S, Bult P, Karssemeijer N, Mann R, et al. Artificial intelligence-based classification of breast lesions imaged with a multiparametric breast MRI protocol with Ultrafast DCE-MRI, T2, and DWI. Invest Radiol. 2019;54:325–32.
43. Qian X, Pei J, Zheng H, Xie X, Yan L, Zhang H, et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat Biomed Eng. 2021;5:522–32.
44. Xi IL, Zhao Y, Wang R, Chang M, Purkayastha S, Chang K, et al. Deep learning to distinguish benign from malignant renal lesions based on routine MR imaging. Clin Cancer Res. 2020;26:1944–52.

AUTHOR CONTRIBUTIONS

(1) Guarantor of integrity of the entire study: HM and CX. (2) Study concepts and design: all authors. (3) Literature research: all authors. (4) Clinical studies: all authors. (5) Experimental studies/data analysis: all authors. (6) Statistical analysis: HZ. (7) Manuscript preparation: all authors. (8) Manuscript editing: all authors.

FUNDING

The study was supported by the National Natural Science Foundation of China (82001775 and 62176140), the Natural Science Foundation of Shandong Province of China (ZR2021MH120 and ZR2022MH274), the Special Fund for Breast Disease Research of Shandong Medical Association (YXH2021ZX055), and Taishan Scholars (tsqn202103197).

COMPETING INTERESTS

The authors declare no competing interests.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This retrospective multicentre study was approved by the institutional review board of Yantai Yuhuangding Hospital, and the patient informed consent was waived.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41416-022-02092-y.

Correspondence and requests for materials should be addressed to Cong Xu or Heng Ma.

Reprints and permission information is available at http://www.nature.com/reprints

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