BMC Medical Imaging

RESEARCH

Open Access

Predicting malignancy in breast lesions: enhancing accuracy with fine-tuned convolutional neural network models

Li Li[1], Changjie Pan[1], Ming Zhang[1], Dong Shen[1], Guangyuan He[1] and Mingzhu Meng[1*]

Abstract

Background This study aims to explore the accuracy of Convolutional Neural Network (CNN) models in predicting malignancy in Dynamic Contrast-Enhanced Breast Magnetic Resonance Imaging (DCE-BMRI).

Methods A total of 273 benign lesions (benign group) and 274 malignant lesions (malignant group) were collected and randomly divided into a training set (246 benign and 245 malignant lesions) and a testing set (28 benign and 28 malignant lesions) in a 9:1 ratio. An additional 53 lesions from 53 patients were designated as the validation set. Five models—VGG16, VGG19, DenseNet201, ResNet50, and MobileNetV2—were evaluated. Model performance was assessed using accuracy (Ac) in the training and testing sets, and precision (Pr), recall (Rc), F1 score (F1), and area under the receiver operating characteristic curve (AUC) in the validation set.

Results The accuracy of VGG19 on the test set (0.96) is higher than that of VGG16 (0.91), DenseNet201 (0.91), ResNet50 (0.67), and MobileNetV2 (0.88). For the validation set, VGG19 achieved higher performance metrics (Pr 0.75, Rc 0.76, F1 0.73, AUC 0.76) compared to the other models, specifically VGG16 (Pr 0.73, Rc 0.75, F1 0.70, AUC 0.73), DenseNet201 (Pr 0.71, Rc 0.74, F1 0.69, AUC 0.71), ResNet50 (Pr 0.65, Rc 0.68, F1 0.60, AUC 0.65), and MobileNetV2 (Pr 0.73, Rc 0.75, F1 0.71, AUC 0.73). S4 model achieved higher performance metrics (Pr 0.89, Rc 0.88, F1 0.87, AUC 0.89) compared to the other four fine-tuned models, specifically S1 (Pr 0.75, Rc 0.76, F1 0.74, AUC 0.75), S2 (Pr 0.77, Rc 0.79, F1 0.75, AUC 0.77), S3 (Pr 0.76, Rc 0.76, F1 0.73, AUC 0.75), and S5 (Pr 0.77, Rc 0.79, F1 0.75, AUC 0.77). Additionally, S4 model showed the lowest loss value in the testing set. Notably, the AUC of S4 for BI-RADS 3 was 0.90 and for BI-RADS 4 was 0.86, both significantly higher than the 0.65 AUC for BI-RADS 5.

Conclusions The S4 model we propose has demonstrated superior performance in predicting the likelihood of malignancy in DCE-BMRI, making it a promising candidate for clinical application in patients with breast diseases. However, further validation is essential, highlighting the need for additional data to confirm its efficacy.

Keywords BI-RADS, Convolutional Neural Networks, Deep transfer learning, Breast lesions, Magnetic resonance imaging

Background

In 2022, breast cancer was one of the most frequently diagnosed cancers worldwide, accounting for 11.6% of all cancer cases globally [1]. Although there has been a significant decline in breast cancer mortality in the United States, with a 40% reduction from 1989 to 2017, recent years have seen a slight annual increase in incidence rates, largely due to rising rates of local stage and

*Correspondence:
Mingzhu Meng
mengmz0202@njmu.edu.cn
[1] Department of Radiology, The Affiliated Changzhou No.2 People's Hospital of Nanjing Medical University, Changzhou 213164, China

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

hormone receptor-positive diseases [2]. This trend is also emerging in China, where an aging population and increasingly westernized lifestyles are expected to lead to a surge in breast cancer cases [3, 4]. These shifts underscore the growing complexity of breast cancer diagnosis, necessitating early detection and accurate diagnosis to improve treatment outcomes and reduce mortality [5, 6].

The Breast Imaging Reporting and Data System (BI-RADS), developed by the American College of Radiology (ACR), has been instrumental in standardizing breast cancer diagnosis across mammography, ultrasound, and magnetic resonance imaging (MRI) [7–9]. The system's fifth edition, released in 2013, provides clear guidelines for categorizing breast lesions, particularly in BI-RADS categories 3 (probably benign), 4, and 5, which guide clinical decision-making regarding follow-up and biopsy. However, many lesions categorized as BI-RADS 3, 4, or 5 are ultimately benign, leading to unnecessary medical procedures and psychological stress for patients. Dynamic contrast-enhanced breast magnetic resonance imaging (DCE-BMRI) offers the potential to better differentiate between malignant and benign lesions, which could reduce the need for unnecessary follow-ups and biopsies. However, the current literature on DCE-BMRI's effectiveness in this regard is inconsistent, highlighting the need for further research in this area.

In recent years, artificial intelligence has made notable strides in aiding the differentiation between benign and malignant lesions in DCE-BMRI. Zhang et al. [10] developed a radiomics model based on DCE-BMRI that achieved an AUC of 0.836 for distinguishing between benign and malignant breast lesions, outperforming models based on T2WI (0.791) and ADC map (0.770). However, a more recent study found that a deep learning model (ResNet34) based on DCE-BMRI had a slightly lower AUC (0.865) compared to a DTL model based on the ADC map (0.770) [11]. Despite these advancements, further research is needed to refine the application of DTL in distinguishing between benign and malignant lesions in DCE-BMRI images. This study aims to fill this gap by assessing the effectiveness of pre-trained convolutional neural networks (CNNs) in predicting malignancy in DCE-BMRI images. By focusing on DCE-BMRI, this research seeks to enhance the current understanding and improve the diagnostic accuracy in breast cancer imaging.

Methods

Dataset 1: training and testing set

We collected data from 530 patients with complete DCE-BMRI and pathological information, spanning January 2017 to December 2020. This included 17 patients with bilateral lesions (both benign and malignant lesions on one side). All lesions were confirmed using permanent specimens and categorized into benign or malignant groups. These were then randomly assigned to a training set (benign: 246 lesions, malignant: 245 lesions) and a testing set (benign: 28 lesions, malignant: 28 lesions) in a 9:1 ratio (refer to Fig. 1). Variables such as age, pathological type, and tumor diameter were compared between groups. Table 1 details the pathological distribution of breast lesions. Inclusion criteria were: ① Patients not subjected to preoperative chemotherapy or chemoradiotherapy before MRI, ② Absence of puncture or surgical procedures prior to MRI. Due to space constraints, clinical presentation details are omitted. To minimize bias

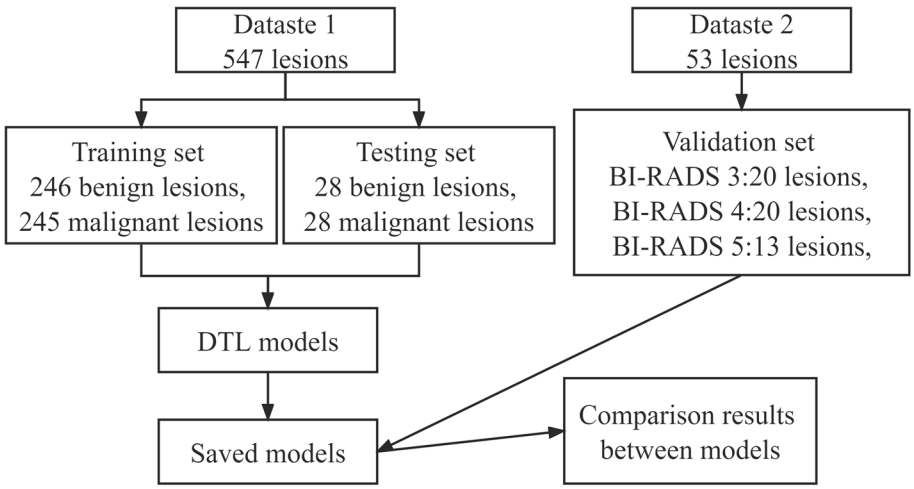

Fig. 1 Dataset structure diagram. This figure presents a schematic representation of the dataset arrangement, illustrating how data is categorized and structured for analysis

Table 1 The pathological distribution of breast lesions

Pathological diagnosis	Lesions	Percent (%)
Malignant lesions		
Invasive ductal carcinoma	220	80.29
Intraductal carcinoma	33	12.04
Invasive lobular carcinoma	7	2.55
Mucinous carcinoma	10	3.65
Lymphoma	1	0.36
Papillary carcinoma	3	1.09
Total	274	100.00
Benign lesions		
Cyst	26	9.52
Adenosis	42	15.38
Fibroadenoma	176	64.47
Chronic inflammation	6	2.20
Intraductal papilloma	20	7.33
Lobular tumor	3	1.10
Total	273	100.00

from bilateral lesions, only unilateral DCE-BMRI images were used.

Dataset 2: validation set

Simultaneously, 53 lesions from 53 patients were included as Dataset 2, using the same MRI scanner as Dataset 1, but unseen during training. Dataset 2 comprised three subsections: BI-RADS 3, 4, and 5 (see Fig. 1). Lesions with pathological results were all confirmed using permanent specimens. Absence of surgery with imaging stability was deemed indicative of no associated cancer. Follow-up adhered to referenced criteria [9, 12]. Correct classification of a lesion required accurate classification in six out of ten images. Table 2 lists the specific details of Dataset 2.

MRI techniques

We employed two 3T MRI scanners with dedicated breast coils in a prone position. Gd-DTPA (0.1 mmol/kg, 2.50 mL/s) was injected through the elbow vein. The process involved six dynamic enhancement phases (one pre-contrast, five post-contrast). MRIs were conducted preoperatively and before initiating therapy. Detailed scanning parameters are outlined in Table 3.

Table 2 Partial clinical information of patients in Dataset 2

Category	N		Confirmed	
	B	M	pathologically	follow up
BI-RADS 3	10	10	4	16
BI-RADS 4	10	10	14	6
BI-RADS 5	3	10	13	0

Readers

Five experienced radiologists from our department, each with over five years of breast MRI interpretation experience and specialized training in breast imaging, were enlisted. The BI-RADS score for a mass is primarily based on the lesion's shape, margin, and internal enhancement characteristics. For detailed criteria, see reference [12, 13]. MRI image analyses were conducted using the GOLDPACS viewer (www.jinpacs.com).

Proposed model

The study utilized a computer equipped with an Intel (R) Core (TM) i7-10700F, NVIDIA RTX 2060 GPU, running on Windows 10 Enterprise 64-bit with 6 GB RAM. All extraneous programs were closed during model operation. Each network underwent identical data testing and training for consistent comparison. Malignant images were identified based on a threshold of ≥ 0.5, while images below this threshold were considered benign.

We selected five commonly used pretrained models (VGG16, VGG19, DenseNet201, ResNet50, and MobileNetV2) and employed five-fold cross-validation to assess model performance, selecting the best-performing model. This cross-validation process was then applied to Dataset 2. Additionally, we enhanced model performance using various fine-tuning strategies. The architecture of the proposed DTL with the five models for breast lesion classification is depicted in Fig. 2.

Initially, the images underwent random shuffling. Data augmentation techniques (rotation, shear range, zoom range, and horizontal flip) were applied prior to training.

Table 3 Scan parameters for the two magnetic resonance scanners

Parameter	Philips Achieva	GE Healthcare
Field strength	3.0 T	3.0 T
No. of coil channels	8	8
Acquisition plane	Axial	Axial
Pulse sequence	3D gradient echo (Thrive)	Enhanced fast gradient echo 3D
Repetition time (ms)	5.5	9.6
Echo time (ms)	2.7	2.1
Flip angle	10°	10°
No. of postcontrast sequence	5	5
Fat suppression	Yes	Yes
Scan time	570 s	500 s

3D three dimensional, *ms* millisecond, *s* second

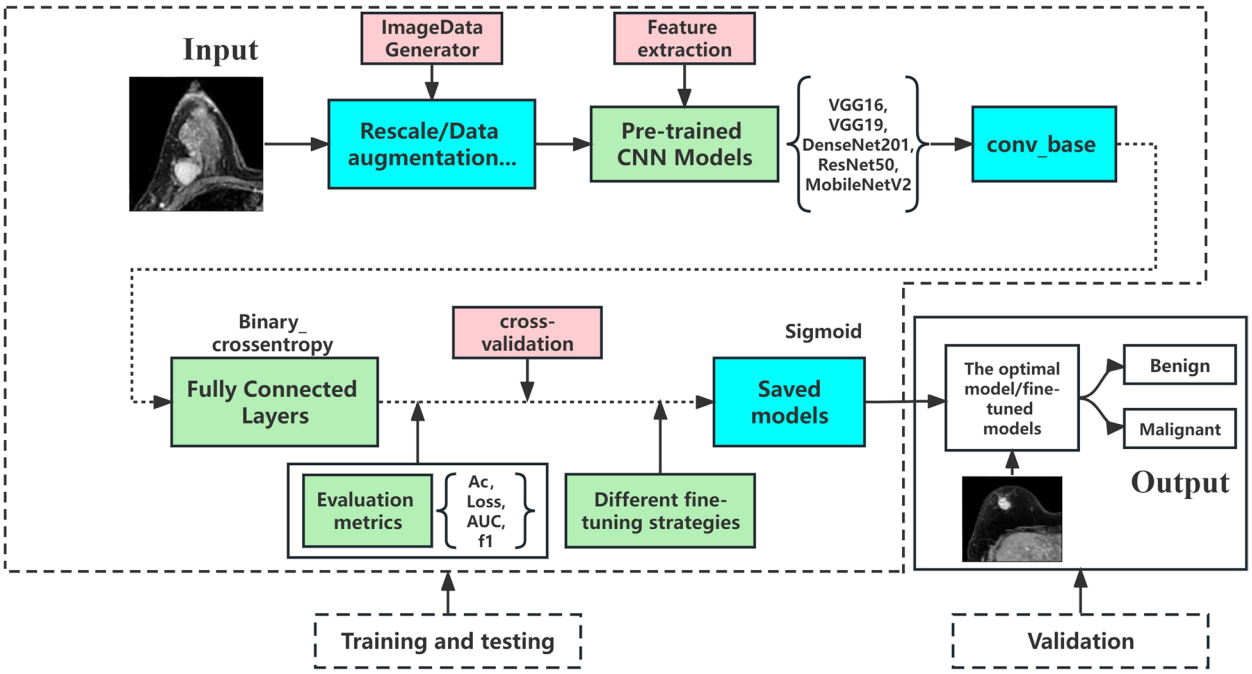

Fig. 2 Deep transfer learning network architecture. This figure depicts the architecture of the DTL network, highlighting its role in determining the likelihood of tumor malignancy. It emphasizes that validation sets do not have to mirror training sets and outlines the three-step data analysis process: feature extraction from the image network, training and testing of data, and data validation

The binary cross-entropy loss function was used, and the training process was optimized using the Adam optimizer with a learning rate of 0.001. Our model required 200 epochs for training on DCE-BMRI images, with a batch size of 64 images. Activation functions included ReLU and sigmoid, as detailed in Eqs. 1 and 2

$$\mathrm{Relu}(x) = f(x) = \begin{cases} \max(0, x), & |x \geq 0 \\ 0, & |x < 0 \end{cases} \tag{1}$$

$$\mathrm{Sigmoid}(x) = f(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

Evaluation metrics

We assessed the effectiveness of Deep Transfer Learning (DTL) models using five performance metrics: accuracy (Ac), precision (Pr), recall rate (Rc), F1 score (F1), and the area under the receiver operating characteristic curve (AUROC) [14]. For this analysis, cases were classified as either malignant or benign, representing positive and negative cases, respectively. True positives (TP) and true negatives (TN) denote the proportion of correctly diagnosed malignant and benign cases. False positives (FP) and false negatives (FN) indicate lesions misdiagnosed as benign and malignant, respectively. The formulas for these metrics are as follows:

$$\mathrm{Ac} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\mathrm{Pr} = \frac{TP}{TP + FP} \tag{4}$$

$$\mathrm{Rc} = \frac{TP}{TP + FN} \tag{5}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Pr} \times \mathrm{Rc}}{\mathrm{Pr} + \mathrm{Rc}} \tag{6}$$

Notably, the accuracy metric (Ac) does not account for data distribution. The F1 score is a balanced measure that considers both precision and recall, making it particularly useful in datasets with imbalanced classes.

Statistical analysis

Statistical analyses were conducted using SPSS 23.0 software (IBM). For data adhering to a normal distribution, counting data were presented as mean ± standard deviation ($\bar{x} \pm s$). One-way analysis of variance (ANOVA) was employed for variance analysis between groups. The Mann–Whitney U test was applied for data not meeting the normal distribution criteria. The chi-square test was utilized for comparing frequency counts between

malignant and benign groups in the datasets (training and testing sets). A *P*-value of < 0.05 (two-tailed) was considered statistically significant.

Results

Age and lesion diameter

Age and lesion diameter did not conform to normal distribution. The age difference between the malignant group (46.40 ± 10.90 years) and the benign group (44.84 ± 10.20 years) was not statistically significant ($P = 0.136$). However, lesion diameters were significantly smaller in the malignant group (25.06 ± 11.54 mm) compared to the benign group (33.44 ± 16.69 mm) ($P < 0.001$). No significant variance was observed in lesion distribution between the training and testing sets across both groups ($P = 0.988$).

Cross validation

We evaluated five models (VGG16, VGG19, DenseNet201, ResNet50, and MobileNetV2) through five-fold cross-validation in Dataset 1 (see Table 4 for results). The DenseNet201 and MobileNetV2 models achieved perfect accuracy (1.00) in the training set, but their testing set accuracies were lower at 0.91 and 0.88, respectively, both below VGG19's 0.96. Despite similar architectures, VGG19 outperformed VGG16 (0.91). However, both VGG16 and VGG19 exhibited premature loss increases with epoch advancement, indicating non-convergence on Dataset 1 and potential overfitting. Similar trends were observed for MobileNetV2 and DenseNet201. ResNet50 showed the lowest accuracy among the models (0.92 training, 0.67 testing). Figures 3 and 4 illustrate the learning curves and heat maps.

Fine-tuning strategy

Given these findings, we focused on enhancing the VGG19 model through five distinct fine-tuning strategies (Fig. 5). The fine-tuning involved activating neural network parameters for training, while keeping certain layers frozen. We noted that the accuracy achieved was 1.0, for all five fine-tuning models(S1-5) on the training set, but S4 obtained the highest test accuracy of 0.97 on the testing set.

In addition, the loss value was the lowest in the testing set for S4. These results reveal that the S4 model has a better generalization ability than the other fine-tuned models.

ROC analysis on validation set

As mentioned earlier, among the five models (VGG16, VGG19, DenseNet201, ResNet50, and MobileNetV2), VGG19 achieved the highest Ac (0.96) on the test set. However, its AUC on the validation set was only 0.76, indicating that the robustness of the VGG19 model may be limited. Among the fine-tuned models, S4 attained the highest AUC (0.89) on the validation set, marking a 13% improvement over the original VGG19 (Fig. 6). Further analysis of S4 across BI-RADS categories 3, 4, and 5 showed notably higher AUCs for BI-RADS 3 (0.90) and 4 (0.86) compared to 5 (0.65) (Fig. 7).

Classification reports on validation set

Classification reports for the five models and S1-5 strategies are provided in Table 5. For the validation set, VGG19 achieved higher performance metrics (Pr 0.75, Rc 0.76, F1 0.73, AUC 0.76) compared to the other models. Strategy S4 outperformed all others on the validation set with Pr 0.89, Rc 0.88, F1 0.87, and AUC 0.89.

Discussion

In this study, we evaluated five pre-trained convolutional neural network models using a fivefold cross-validation approach on our DCE-BMRI dataset. Our goal was to identify the best-performing model, defined as the one that excels across all predefined evaluation metrics. After selecting the top model, we fine-tuned it to further enhance its performance and assessed its generalization capability on a validation set.

Our findings revealed that the VGG19 model demonstrated superior performance, achieving accuracies of

Table 4 The results of the five-fold cross-validation in dataset 1

Folds	Accuracies of the training set					Accuracies of the testing set				
	model1	model2	model3	model4	model5	model1	model2	model3	model4	model5
Fold1	1.00	1.00	1.00	0.92	1.00	0.91	0.96	0.91	0.67	0.88
Fold2	1.00	1.00	1.00	0.93	0.99	0.90	0.96	0.91	0.67	0.87
Fold3	1.00	1.00	1.00	0.92	1.00	0.91	0.96	0.91	0.67	0.88
Fold4	1.00	1.00	1.00	0.92	0.99	0.90	0.96	0.91	0.67	0.87
Fold5	1.00	1.00	1.00	0.93	1.00	0.91	0.96	0.91	0.67	0.88

Model1, VGG16; Model2, VGG19; Model3, DenseNet201; Model4, ResNet50; Model5, MobileNetV2

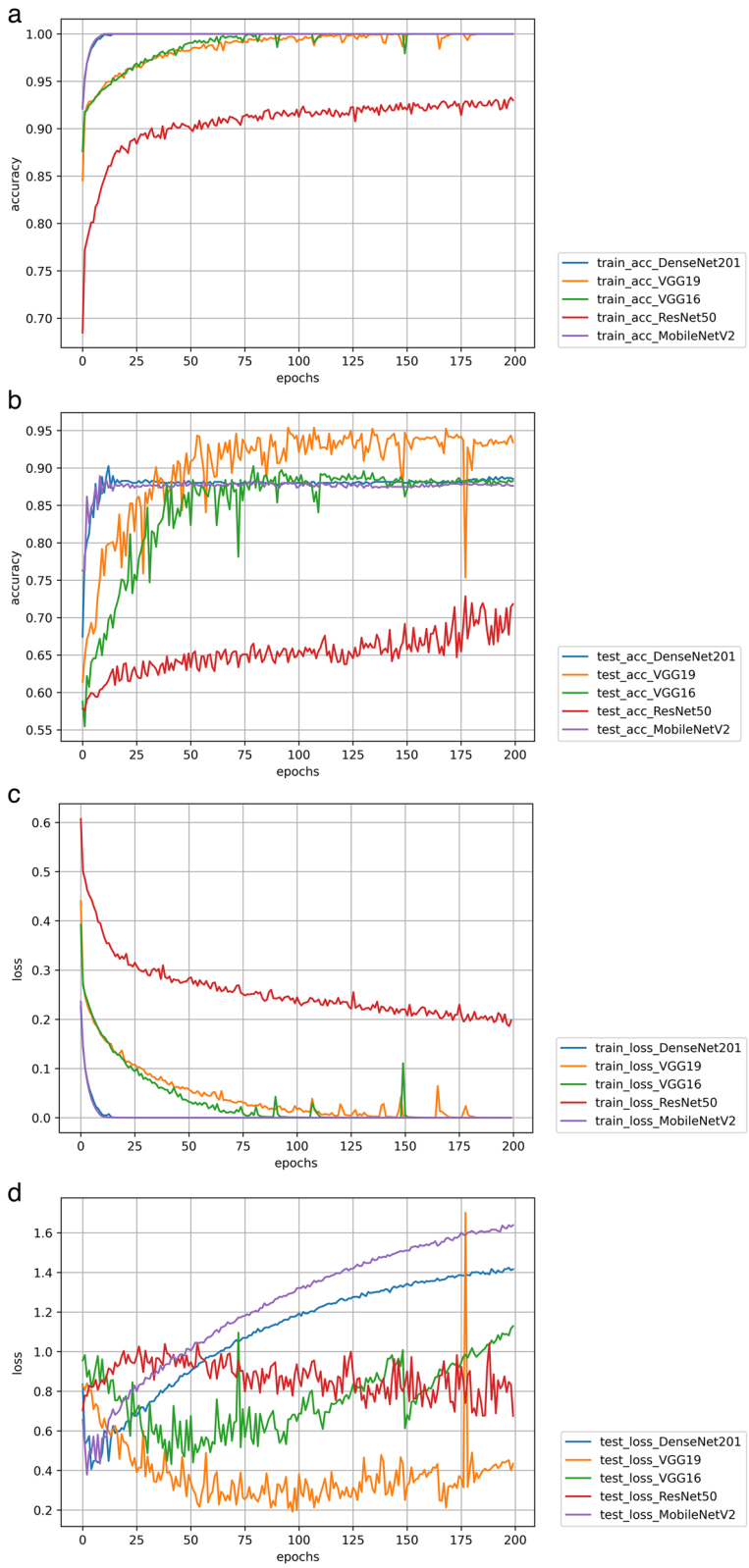

Fig. 3 Learning curves for the five pre-trained models. This figure displays learning curves for each of the five pre-trained models over various epochs, showing: **a**) training accuracy, **b**) testing accuracy, **c**) training loss, and **d**) testing loss. It notably illustrates that the VGG19 model achieved the highest accuracy in the testing set, while ResNet50 had the lowest

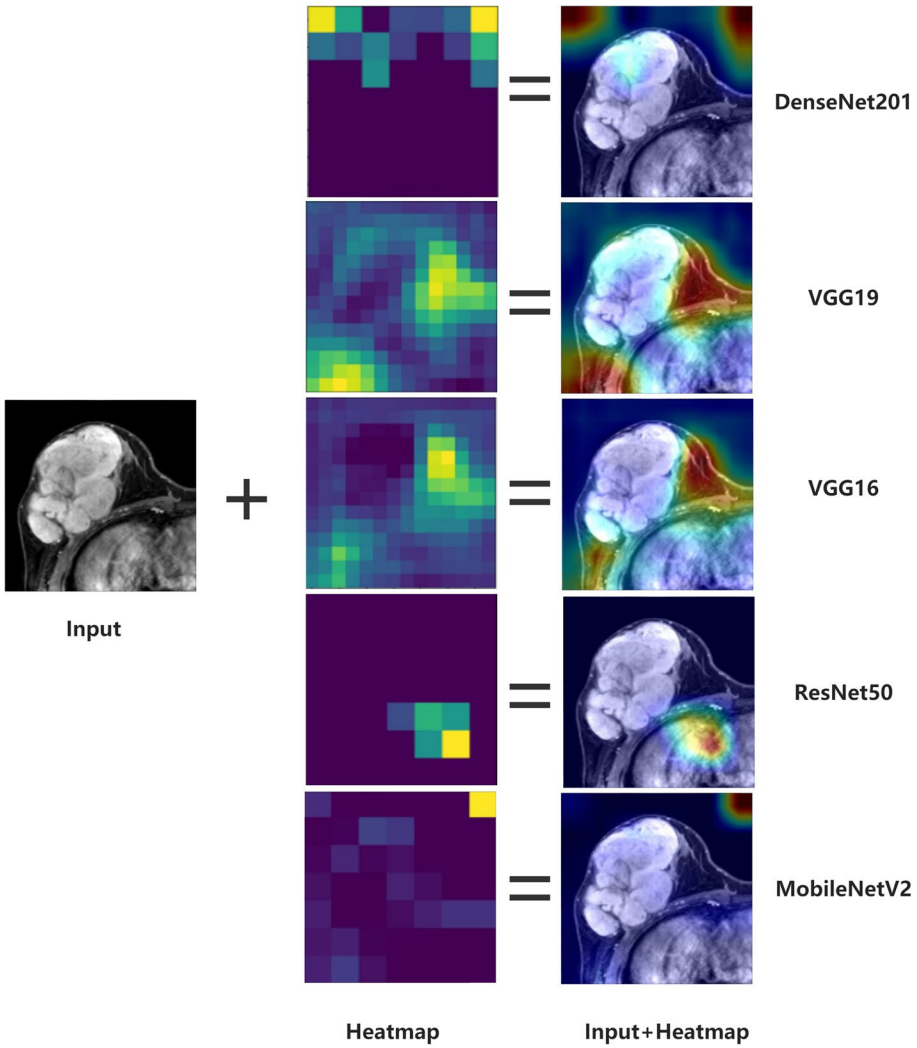

Fig. 4 Heatmaps of the five models. The figure provides heatmaps illustrating the activated-zone boundaries for each model. It shows that the activated zones for DenseNet201 and MobileNetV2 are located outside the input image, while ResNet50's activated zone is relatively small. The heatmaps for VGG19 and VGG16 display similar locations of activation zones, with VGG19 showing greater activation

1.00 on the training set and 0.96 on the test set. However, despite these high accuracies, the model's AUC on the validation set was only 0.76, indicating significant limitations in its generalization ability. This observation aligns with previous research, which suggests that fine-tuning can enhance the accuracy and precision of models but may not always translate into improved generalization across diverse datasets [15–18]. To address this, we implemented five distinct fine-tuning strategies for VGG19, aiming to identify a more robust approach. Among these strategies, the S4 model emerged as the most successful, achieving the highest test accuracy (0.97) and the lowest test loss, while also avoiding overfitting. These results suggest that the S4 model has superior generalization capability compared to the other strategies. Furthermore, the S4 model achieved the highest AUC (0.89) on the validation set, indicating improved performance in distinguishing between classes. These findings highlight the potential of the S4 fine-tuning strategy for enhancing the accuracy of medical image classification diagnostics, especially in complex datasets like DCE-BMRI.

Model fine-tuning is an effective method to overcome overfitting. Overfitting is a common issue encountered when training deep learning models on small datasets [19, 20]. In our study, all models employed data augmentation, regularization, and dropout to prevent overfitting. Despite these measures, some models, such as MobileNetV2 and DenseNet201, still experienced overfitting. Our results show that fine-tuning strategy 4(S4) was

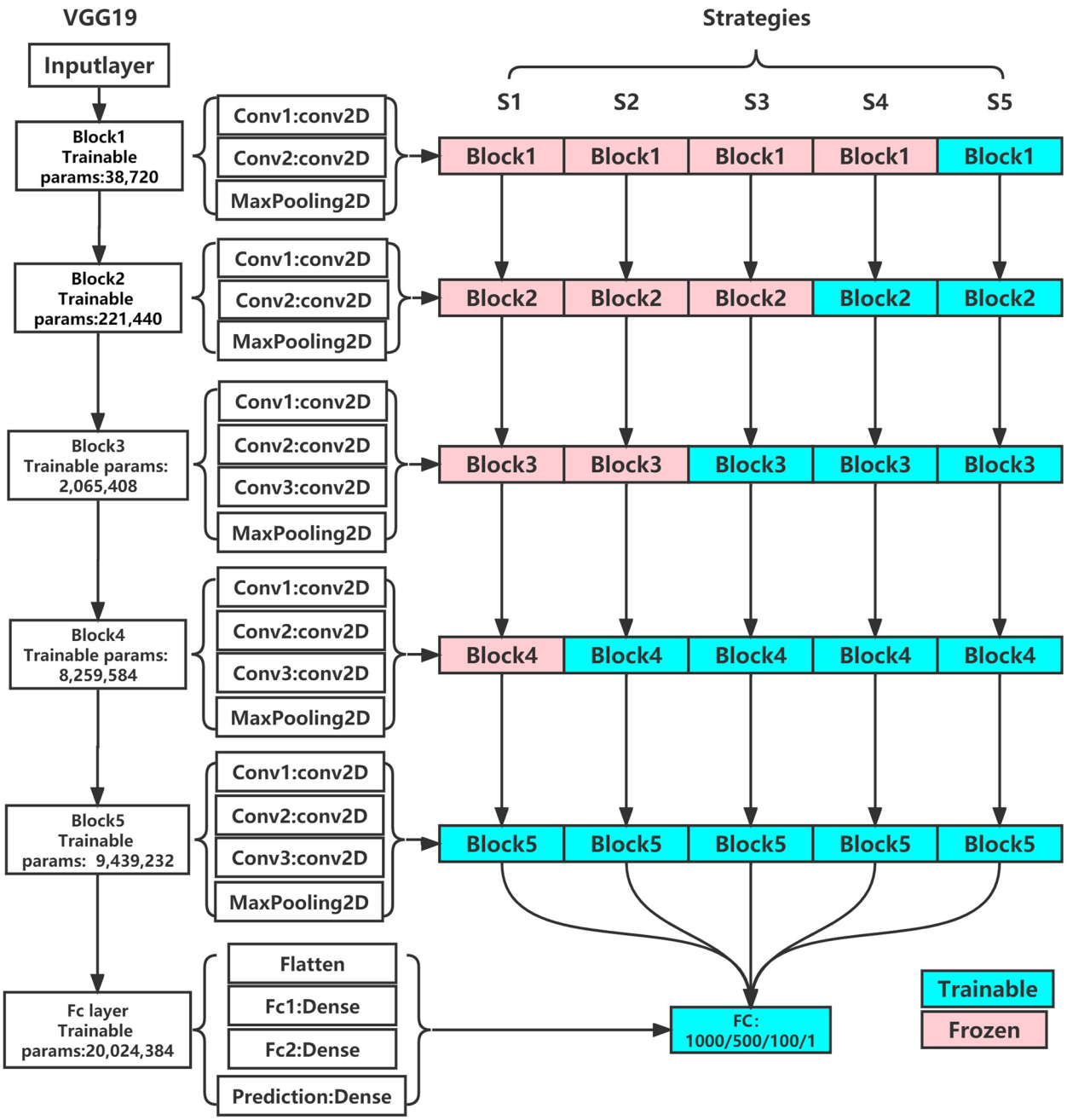

Fig. 5 Schematic of fine-tuning strategies for VGG19. This figure outlines the five different fine-tuning strategies applied to the VGG19 model, detailing the number of trainable parameters, the activated layers (trainable), and the non-trainable (frozen) layers of the neural network. It also highlights the full connection (Fc) layer

effective in preventing overfitting, consistent with our previous study where fine-tuning the Inception V3 model reduced the biopsy rate for BI-RADS 4 lesions [18].

We also explored whether the S4 model exhibited varying AUC scores across BI-RADS categories 3, 4, and 5. The results showed that S4 performed best in BI-RADS 3 (AUC 0.90), followed by BI-RADS 4 (AUC 0.86), and had the lowest performance in BI-RADS 5 (AUC 0.65). The differences in performance among these categories may stem from the model's learning capacity, feature extraction ability, or inherent characteristics of the data. For instance, BI-RADS 5 cases typically have more distinct malignant features, which might require more sophisticated feature extraction

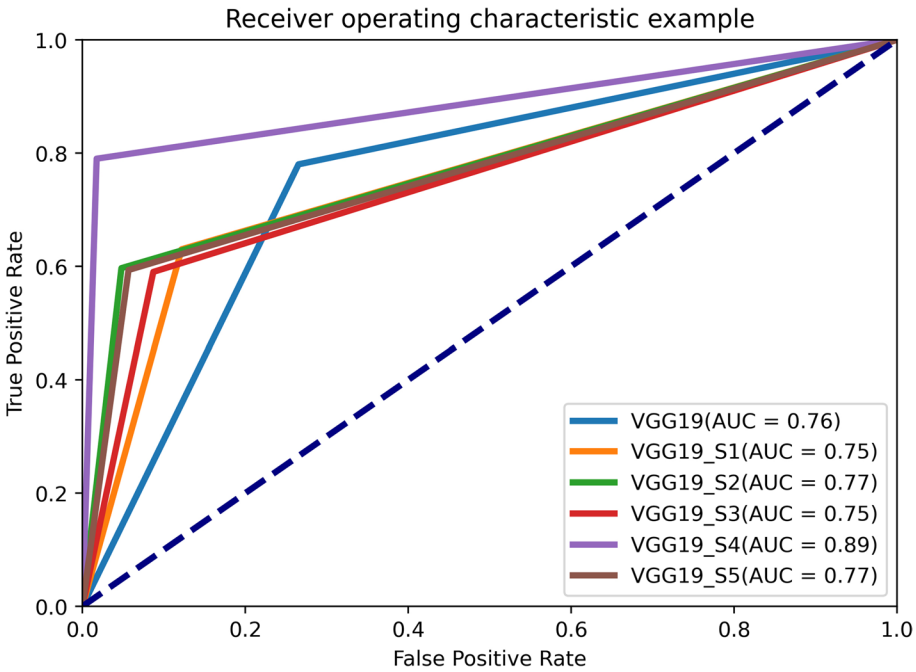

Fig. 6 AUC analysis of the proposed S1-5 and VGG19 models. This figure showcases the Area Under the Curve (AUC) analyses for the proposed S1-5 strategies and the VGG19 model, allowing for a comparative assessment of their performance

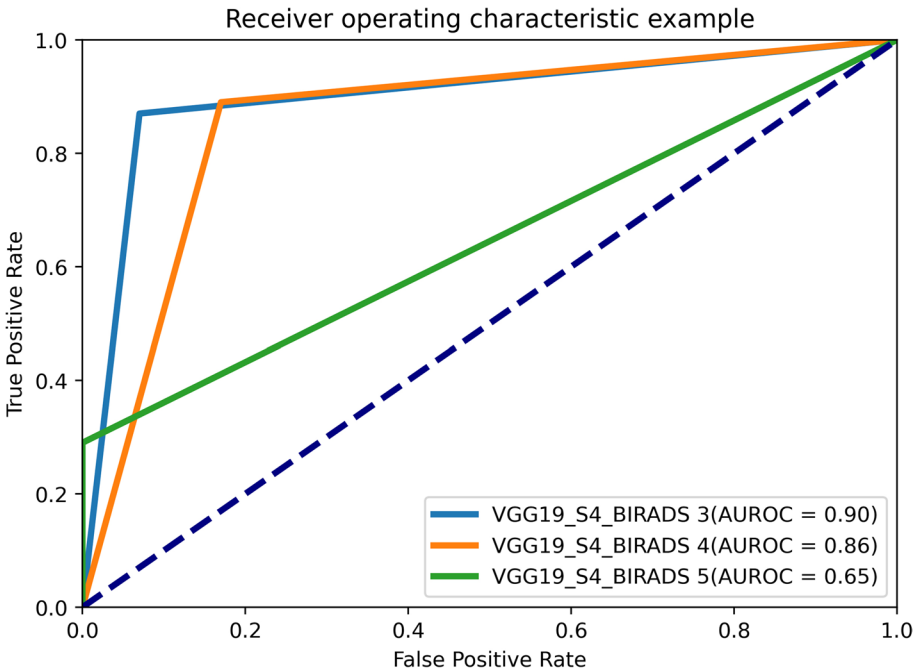

Fig. 7 AUC comparison in BI-RADS Subcategories for the S4 model. The figure compares the AUC scores of the S4 model across different BI-RADS categories (3, 4, and 5), offering insights into the model's performance in each category

techniques, whereas BI-RADS 3 cases involve subtler, often ambiguous features [12, 21]. This variability suggests that further improvements in data balancing, feature extraction, and model optimization are necessary to enhance model performance across all BI-RADS categories.

Table 5 Classification report of deep transfer learning models in validation set

DTL models	Pr			Rc			F1			AUC
	group1	group22	avg	group1	group2	avg	group1	group2	avg	
model1	0.93	0.53	0.73	0.60	0.91	0.75	0.73	0.67	0.70	0.73
model2	0.91	0.59	0.75	0.63	0.90	0.76	0.75	0.71	0.73	0.76
model3	0.91	0.52	0.71	0.59	0.88	0.74	0.72	0.65	0.69	0.71
model4	0.90	0.39	0.65	0.53	0.84	0.68	0.67	0.53	0.60	0.65
model5	0.91	0.55	0.73	0.61	0.89	0.75	0.73	0.68	0.71	0.73
S1	0.88	0.63	0.75	0.65	0.87	0.76	0.74	0.73	0.74	0.75
S2	0.95	0.60	0.77	0.64	0.94	0.79	0.77	0.73	0.75	0.77
S3	0.91	0.60	0.76	0.63	0.91	0.76	0.74	0.71	0.73	0.75
S4	0.98	0.79	0.89	0.78	0.98	0.89	0.87	0.88	0.87	0.89
S5	0.94	0.59	0.77	0.64	0.93	0.79	0.76	0.73	0.75	0.77

Group1, benign group; Group2, malignant group; Avg, average; Model1, VGG16; Model2, VGG19; Model3, DenseNet201; Model4, ResNet50; Model5, MobileNetV2

Our study found that the S4 model achieved the highest recall rate (0.89) among all the DTL models, which is particularly noteworthy given the relatively limited class diversity in our dataset. Recall, also known as sensitivity, measures the completeness of a classifier. A lower recall value suggests that the classifier has limited capability in managing a high number of false positives (FP). Recent publications have introduced new and updated performance benchmarks, replacing outdated metrics in the latest edition. Consequently, the recall rate benchmark has been revised. Initially, about half of all radiologists were unable to meet the 10% benchmark for recall rate, leading to a revision to a more achievable target of 12%, a standard now met by over 75% of radiologists [7].

This study, however, is not without its limitations. Firstly, the training set included a relatively small number of images, particularly with a scarcity of rare lesion types. As a result, our dataset may not fully represent the broader spectrum of breast disease cases, potentially affecting the accuracy of the DTL model. To address this, further analysis with larger and more diverse datasets is necessary to thoroughly evaluate the model's robustness. Secondly, our study focused exclusively on static DCE-BMRI images, without incorporating other routine diagnostic procedures such as clinical evaluations, breast ultrasounds, and mammography. Thirdly, we limited our investigation to only five pre-trained models; future research should explore a broader range of models to assess their robustness on larger datasets. Lastly, while this paper does not explore the various methods of fine-tuning CNN models, these aspects will be the focus of our future studies.

Conclusions

In this study, the S4 model demonstrated superior accuracy in classifying BI-RADS categories 3 and 4, outperforming its performance in category 5. This outcome is particularly significant as it suggests the potential to reduce the frequency of follow-up sessions for BI-RADS 3 cases and decrease unnecessary biopsies for benign lesions in BI-RADS 4. These findings underscore the promise of fine-tuned deep learning models in improving diagnostic accuracy in breast imaging. However, the results require further validation with larger and more diverse datasets. Future research will focus on exploring more robust models and expanding the dataset to enhance the generalizability and reliability of these findings.

Abbreviations

MRI	Magnetic resonance imaging
DL	Deep learning
DTL	Deep transfer learning
ROC	Receiver operating characteristic
AUC	Area under the ROC curve
BI-RADS	Breast Imaging Reporting and Data System
DCE-BMRI	Dynamic contrast enhanced breast MRI

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12880-024-01484-1.

Supplementary Material 1.

Acknowledgements

The authors wish to thank Shiquan Ge for his technical assistance in operating the Python programming code.

Authors' contributions

Li Li and Mingzhu Meng carried out the literature search, and designed and wrote the manuscript. Mingzhu Meng and Changjie Pan conceived of the project, and participated in its design and coordination, and helped to draft the manuscript. Ming Zhang, Dong Shen and Guangyuan He are responsible for figures processing. Both authors read and approved the final manuscript.

Funding

This study was supported by the Program of Bureau of Science and Technology Foundation of Changzhou (No. CJ20220260). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
This study was approved by the Second Hospital of Changzhou Affiliated to Nanjing Medical University of Chinese Medicine Ethics Review Committee (Ethics Number: [2023]KY313-01).

Consent for publication
This study was a retrospective analysis and informed consent was waived.

Competing interests
The authors declare no competing interests.

Received: 7 February 2024 Accepted: 29 October 2024
Published online: 11 November 2024

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63.
2. DeSantis CE, Ma J, Gaudet MM, et al. Breast cancer statistics, 2019. CA Cancer J Clin. 2019;69(6):438–51.
3. Cao W, Chen HD, Yu YW, et al. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. Chin Med J (Engl). 2021;134(7):783–91.
4. Rm F, Yn Z, Sm C, et al. Current cancer situation in China: good or bad news from the 2018 Global Cancer Statistics? Cancer Commun (Lond). 2019;39(1):22.
5. Gao Y, Heller SL. Abbreviated and Ultrafast Breast MRI in Clinical Practice. Radiographics. 2020;40(6):1507–27.
6. Berdzuli N. Breast cancer: from awareness to access. BMJ. 2023;380:290.
7. Mercado CL. BI-RADS Update. Radiol Clin North Am. 2014;52(3):481–7.
8. Sedgwick EL, Ebuoma L, Hamame A, et al. BI-RADS update for breast cancer caregivers. Breast Cancer Res Treat. 2015;150(2):243–54.
9. Pesce K, Orruma MB, Hadad C, et al. BI-RADS terminology for mammography reports: what residents need to know. Radiographics. 2019;39(2):319–20.
10. Zhang Q, Peng Y, Liu W, et al. Radiomics based on multimodal mri for the differential diagnosis of benign and malignant breast lesions. J Magn Reson Imaging. 2020;52(2):596–607.
11. Du Y, Wang D, Liu M, et al. Study on the differential diagnosis of benign and malignant breast lesions using a deep learning model based on multimodal images. J Cancer Res Ther. 2024;20(2):625–32.
12. Lee SE, Lee JH, Han K, et al. BI-RADS category 3, 4, and 5 lesions identified at preoperative breast MRI in patients with breast cancer: implications for management. Eur Radiol Exp. 2020;30(5):2773–81.
13. Eghtedari M, Chong A, Rakow-Penner R, et al. Current status and future of BI-RADS in multimodality breast imaging, from the ajr special series on radiology reporting and data systems. Am J Roentgenol. 2020;216(4):860–73.
14. Wang Z, Li X, Yao M, et al. A new detection model of microaneurysms based on improved FC-DenseNet. Sci Rep. 2022;12(1):950.
15. Tan T, Li Z, Liu H, et al. Optimize transfer learning for lung diseases in bronchoscopy using a new concept: sequential fine-tuning. IEEE J Transl Eng Health Med. 2018;6:1800808.
16. Ahamed KU, Islam M, Uddin A, et al. A deep learning approach using effective preprocessing techniques to detect COVID-19 from chest CT-scan and X-ray images. Comput Biol Med. 2021;139:105014.
17. Montaha S, Azam S, Rafid A, et al. BreastNet18: a high accuracy fine-tuned vgg16 model evaluated using ablation study for diagnosing breast cancer from enhanced mammography images. Biology (Basel). 2021;10(12):1347.
18. Meng M, Li H, Zhang M, et al. Reducing the number of unnecessary biopsies for mammographic BI-RADS 4 lesions through a deep transfer learning method. BMC Med Imaging. 2023;23(1):82.
19. Balasubramanian PK, Lai WC, Seng GH, et al. APESTNet with Mask R-CNN for Liver Tumor Segmentation and Classification. Cancers (Basel). 2023;15(2):330.
20. Zhou Q, Zhu W, Li F, et al. Transfer learning of the ResNet-18 and DenseNet-121 model used to diagnose intracranial hemorrhage in CT canning. Curr Pharm Des. 2022;28(4):287–95.
21. L M. BI-RADS category 3 is a safe and effective alternative to biopsy or surgical excision. Radiology. 2020;296(1):42–3.

Publisher's Note

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