Tian *et al. BMC Cancer* (2025) 25:537
https://doi.org/10.1186/s12885-025-13960-0

BMC Cancer

RESEARCH Open Access

Two-stage augmentation for detecting malignancy of BI-RADS 3 lesions in early breast cancer

Huanhuan Tian[1], Li Cai[1], Yu Gui[2], Zhigang Cai[1], Xianfeng Han[1], Jianwei Liao[1], Li Chen[2*] and Yi Wang[1*]

Abstract

Objectives In view of inherent attributes of breast BI-RADS 3, benign and malignant lesions are with a subtle difference and the imbalanced ratio (with a very small part of malignancy). The objective of this study is to improve the detection rate of BI-RADS 3 malignant lesions on breast ultrasound (US) images using deep convolution networks.

Methods In the study, 1,275 lesions out of 1,096 patients were included from Southwest Hospital (SW) and Tangshan Hospital (TS). In which, 629 lesions, 218 lesions and 428 lesions were utilized for the development dataset, the internal and external testing set. All malignant lesions were biopsy-confirmed, while benign lesions were verified through biopsy or stable (no significant changes) over a three-year follow-up. And each lesion had both B-mode and color Doppler images. We proposed a two-step augmentation method, covering malignancy feature augmentation and data augmentation, and further verified its feasibility on a dual-branches ResNet50 classification model named Dual-ResNet50. We conducted a comparative analysis between our model and four radiologists in breast imaging diagnosis.

Results After malignancy feature and data augmentations, our model achieved a high area under the receiver operating characteristic curve (AUC) of 0.881 (95% CI: 0.830–0.921), the sensitivity of 77.8% (14/18), in the SW test set, and an AUC of 0.880 (95% CI: 0.847–0.910), a sensitivity of 71.4% (5/7) in the TS test set. Compared to four radiologists with over 10-years of diagnostic experience, our model outperformed their diagnoses.

Conclusions Our proposed augmentation method can help the deep learning (DL) classification model to improve the breast cancer detection rate in BI-RADS 3 lesions, demonstrating its potential to enhance diagnostic accuracy in early breast cancer detection. This improvement aids in a timely adjustment of subsequent treatment for these patients in clinical practice.

Keywords Breast cancer, Ultrasound screening image, BI-RADS 3, Deep learning, Artificial intelligence

*Correspondence:
Li Chen
chenli@tmmu.edu.cn
Yi Wang
echowang@swu.edu.cn
[1]College of Computer and Information Science, Southwest University, No. 2, Tiansheng Road, Beibei District, Chongqing 400715, China
[2]Department of Breast and Thyroid Surgery, Southwest Hospital of Third Military Medical University, No. 30, Gaotan Yanzheng Street, Shapingba District, Chongqing 400380, China

Introduction

Breast cancer, as the most common malignant tumor for women in the world, seriously threatens women's physical and mental health [1]. Early screening and treatment can significantly improve the prognosis of patients and reduce mortality rates [2–4]. Hence, early detection through screening and prompt treatment are essential for effective management of this disease. Compared with mammography [5], US screening imaging [6] is a major screening routine with detailed information inside the breast for early detection of breast cancer in many Asian countries, improving the detection rate especially in women with dense breasts [7].

The breast imaging reporting and data system (BI-RADS) is a standardized system to differentiate the stages of cancer as seen on US imaging, ranging from category 1 (not cancer) to category 6 (high likelihood of cancer) [8]. Notably, BI-RADS 3 indicates a malignancy probability less than 2%, and the corresponding treatment recommendation is short-term follow-up. BI-RADS 3 lesions indicate an extremely imbalanced data distribution (with a very small part of malignancy), and similar features between benign and malignant lesions. Such factors complicated the detection of BI-RADS 3 malignant lesions in breast cancer screening, posing a significant challenge for precise diagnosis.

Artificial Intelligence (AI) [9] has significantly revolutionized various clinical challenges. At early stages, machine learning (ML) [10–12] algorithms were applied in computer-aided diagnosis, including breast cancer, relying on manually extracted features and specific image properties defined by domain experts. With the further development of AI, deep learning (DL), particularly deep convolutional neural networks (DCNNs) [13], has emerged a unique advantage in medical imaging. These networks are capable of autonomously learning complex feature representations from a vast amount of medical images, eliminating the need for predefined features or expert intervention. The feasibility of applying DL on the classification of breast US images has been demonstrated [14–16], highlighting its potential to identify subtle disease variances that may be invisible to radiologists.

However, there are few studies specifically designing benign and malignant classification frameworks for BI-RADS 3 breast lesions, due to the inherent properties of BI-RADS 3 lesions. First, the morphology of benign and malignant lesions classified by BI-RADS 3 exhibit a significant degree of overlap [17], which makes DL models difficult to distinguish precisely. Second, it commonly has a high detection rate of BI-RADS 3 lesions but a very low probability of malignancy. So, it is difficult to collect sufficient malignant data to train the classification model, resulting in overfitting problems during the training stage, and causing limited generalization on new datasets. Thus, training a benign-malignant classification model for BI-RADS 3 breast lesions is a challenging task. This study aims to propose a two-stage augmentation method, including malignancy feature augmentation and data augmentation, by using the processed US images to train the Dual-ResNet50 model, to better support benign and malignant classification for BI-RADS 3 breast lesions.

Patients and methods

Patient data

This retrospective study had been approved by the Ethics Committee of the First Affiliated Hospital of Army Medical University ([No. (B) KY202264]), and the requirement for informed consent from all patients was waived before study inclusion. The workflow of our investigation followed the BI-RADS guidelines, as detailed in (Additional file 1: Table S1). The inclusion criteria included all these patients who had undergone US examination. Following this, malignant samples were confirmed by biopsy, and benign samples were verified either directly via biopsy or no significant changes over a three-year follow-up. The cases that lacked of B-mode or color Doppler images of the same lesion, and the cases of particular patients (including pregnant, lactation, and local treatment history) were excluded.

As a result, a total of 742 patients, 847 lesion images were obtained in Southwest Hospital (SW). Among them, 629 lesion images (400 BI-RADS 3 benign, 20 BI-RADS 3 cancers, and 209 BI-RADS 4 A cancers) from 2015 to 2020 were divided into the training and validation datasets, and 218 lesion images (18 BI-RADS 3 cancers, 200 BI-RADS 3 benign) in 2021 were split into the internal test set. Additionally, 354 patients, 428 lesion images (7 BI-RADS 3 cancers, 421 BI-RADS 3 benign) were gathered as the external test set from Tangshan Hospital (TS), in 2021. Figure 1 illustrates the screening flow chart for breast cancer patients, and the characteristics of the screened patients are described in Table 1. The detailed US data and preprocessing operations are provided in Additional file 1: ROI extraction and preprocessing.

A Two-stage augmentation architecture

In this study, we proposed a two-stage augmentation method for enhancing the extraction of malignancy features and balancing negative/positive samples. Specifically, it first utilized BI-RADS 4 A malignant lesions to achieve feature augmentation, and then applied the CycleGAN model for data augmentation, to autonomously predict benign and malignant of BI-RADS 3 lesions by using the proposed Dual-ResNet50 classification model. The detailed workflow is depicted in Fig. 2.

For the feature augmentation method, benign and malignant regions of BI-RADS 3 lesions were with very

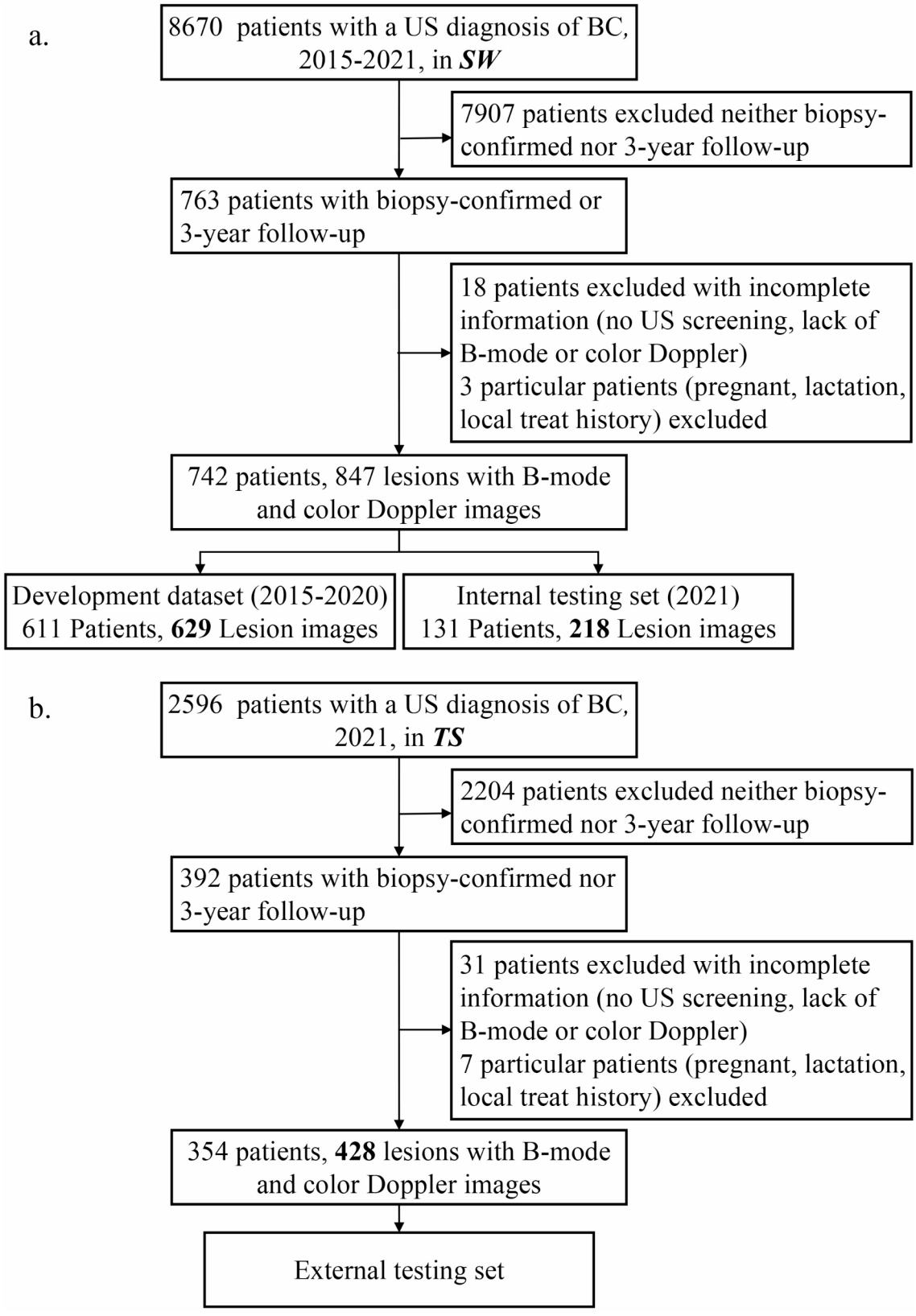

Fig. 1 The screening flow chart for breast cancer patients. *BC* breast cancer, *development dataset* containing training and validation sets by stratified sampling

Table 1 Breast cancer patient and lesion characteristics

Specifications	SW Dataset		TS Dataset
	Train & validate	Internal Test	External Test
Patients (742 patients in *SW*)			
Age			
< 30	273	36	95
30 ~ 49	306	78	211
50 ~ 69	31	17	47
≥ 70	1	0	1
Diagnostic methods			
Biopsy	431	131	354
Follow-up	180	0	0
Lesions (847 lesions from 742 patients in *SW*)			
Lesions size			
< 5	25	8	13
5 ~ 9.9	150	70	114
10 ~ 19.9	302	108	176
≥ 20	152	32	51
Lesions width			
< 5	177	60	98
5 ~ 9.9	320	111	180
10 ~ 19.9	125	46	74
≥ 20	7	1	2
Aspect ratio			
≥ 1	10	2	4
< 1	619	216	350
Boundary			
Clear	542	182	297
Others	87	36	57
Morphology			
Regular	564	193	313
Others	65	25	41
Blood Flow Spectrum			
Pulsating	13	3	349
Others	616	215	5

The training, validation, and internal dataset was collected by Southwest Hospital of China, while the external dataset was from Tangshan Hospital of China. In this table, lesion information was determined using existing screening and diagnostic reports. Note that the training and validation cohorts include 2015–2020 biopsy-confirmed lesions and 2015–2017 follow-up confirmed lesions

similar appearance, only subtle variations, leading to the missed cancer diagnosis. We observed BI-RADS 4 A is an adjacent category of BI-RADS 3 in the BI-RADS grading system, with a higher likelihood of malignancy between 2% and 10%, which requires a biopsy test for definitive diagnosis [18]. Therefore, we added all 209 BI-RADS 4 A cancers when training Dual-ResNet50, to extract more discriminative features by learning these similar features from BI-RADS 4 A, thereby assisting this model in the detection rate of BI-RADS 3 malignant lesions.

Furthermore, due to the extreme imbalance of benign and malignant samples in BI-RADS 3, the DL classification model demonstrated limited performance in identifying malignant lesions of BI-RADS 3. Thus, the data augmentation method was further applied to address the imbalanced nature of benign/malignant BI-RADS 3 lesions. The advantage of Generative adversarial network (GAN) [19] as the data augmentation method is to generate new data with diversity, instead of simply rotating, flipping, or cutting the original US images, thereby expanding the dataset, and improving the generalization ability of the model. We individually trained two CycleGAN models [20] named Bmode-GAN and Doppler-GAN using ten-fold cross-validation with respect to B-mode and color Doppler images, to realize the mutual transformation between BI-RADS 3 and BI-RADS 4 A. Therefore, from BI-RADS 4 A to BI-RADS 3, we could totally obtain 209 synthetic BI-RADS 3 malignant lesions (including B-mode and color Doppler images) by these two pre-trained CycleGAN models, achieving data augmentation of BI-RADS 3 malignant lesions.

Dual-ResNet50 classification model

In this study, we further developed a DL-based classification model based on dual-branches ResNet50, named Dual-ResNet50, to verify the impact of the proposed two-step augmentation method. Specifically, we evaluated the performance of the Dual-ResNet50 model under three distinct training conditions: DR-B, the base model trained Dual-ResNet50 only on original data without any augmentation; DR-F, the model trained Dual-ResNet50 by utilizing original data and the data with feature augmentation; DR-FD, the model obtained by training Dual-ResNet50 on original data and augmented data, which included both feature and data augmentations. We analyzed the effectiveness of these different augmentation methods, by the performance of these models under each of these three conditions.

Here, for Dual-ResNet50, both B-mode and color Doppler images of breast lesions were fed into the two ResNet50 [21] branches of the Dual-ResNet50 model to obtain their respective feature representations. Then, the dimensionality was reduced through two fully connected layers, and the probability between 0 and 1 could be output through a SoftMax function, as the final classification result. The detailed architecture of Dual-ResNet50 is provided in Fig. 2 and Additional file 1: Dual-ResNet50 model.

Comparison with radiologists

To further evaluate the clinical effect of the DR-FD model, the diagnostic performance of the model and four radiologists were compared in SW and TS test sets, which included patient data collected in 2021. We invited four radiologists, to independently diagnose whether the lesion is benign or malignant through analyzing B-mode

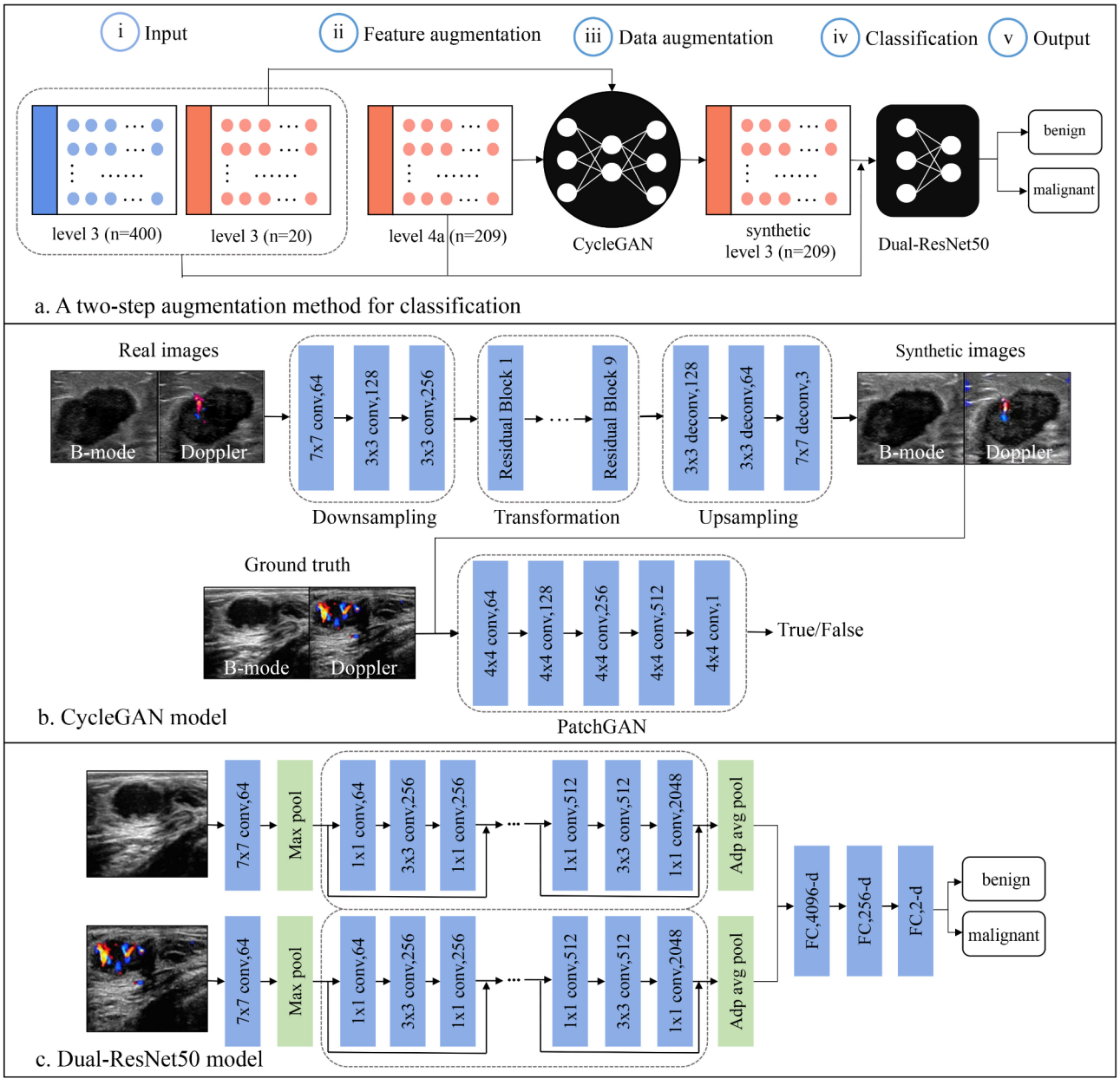

Fig. 2 Schematic diagram of the two-stage augmentation method for breast cancer classification. (**a**) A two-step augmentation method for classification. (**b**) CycleGAN model. (**c**) Dual-ResNet50 model. The two-stage augmentation method consists of feature augmentation and data augmentation. The original BI-RADS 3 lesions, along with BI-RADS 4 A malignant lesions after feature augmentation, and synthetic BI-RADS 3 malignant lesions after data augmentation using CycleGAN are input into the classification model named Dual-ResNet50, to predict the probability of BI-RADS 3

and color Doppler images of a specific lesion. The final diagnosis of all radiologists was determined if not less than two radiologists identified the lesion as malignant, we labeled it as malignant; otherwise, it was classified as benign.

Model interpretability

To alleviate the black-box nature of the DR-FD model, the visual tool named Grad-CAM [22, 23] was used to generate heatmaps for B-mode and color Doppler images. The saliency maps were attained by applying Grad-CAM on the last convolutional layers of two separate ResNet50 branches in DR-FD. The heatmap signals with higher brightness indicated areas that our DR-FD identified as being relevant to the target class or feature of interest. In clinical analysis, such heatmaps provided additional information for radiologists in identifying significant regions.

Statistical analysis

The statistical analysis was performed using MedCalc version 19.0.4, python packages. These indicators, including AUC, sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), F1-score, and false positive rate (FPR), were applied to evaluate the performance of the binary classification model, which were described in Additional file 1: Statistic metrics. We utilized DeLong's method to compare the difference of AUCs between our proposed model and other baselines, and provided the corresponding confidence intervals. A p-value of 0.05 or less meant that the null hypothesis was refused.

Results

Generation results of cyclegan model

A total of 209 synthetic BI-RADS 3 malignant lesion image pairs (including B-mode and color Doppler images of the same lesion) were generated by two pre-trained CycleGAN models. Several typical examples of real BI-RADS 4 A and corresponding synthetic BI-RADS 3 lesions are illustrated in Fig. 3.

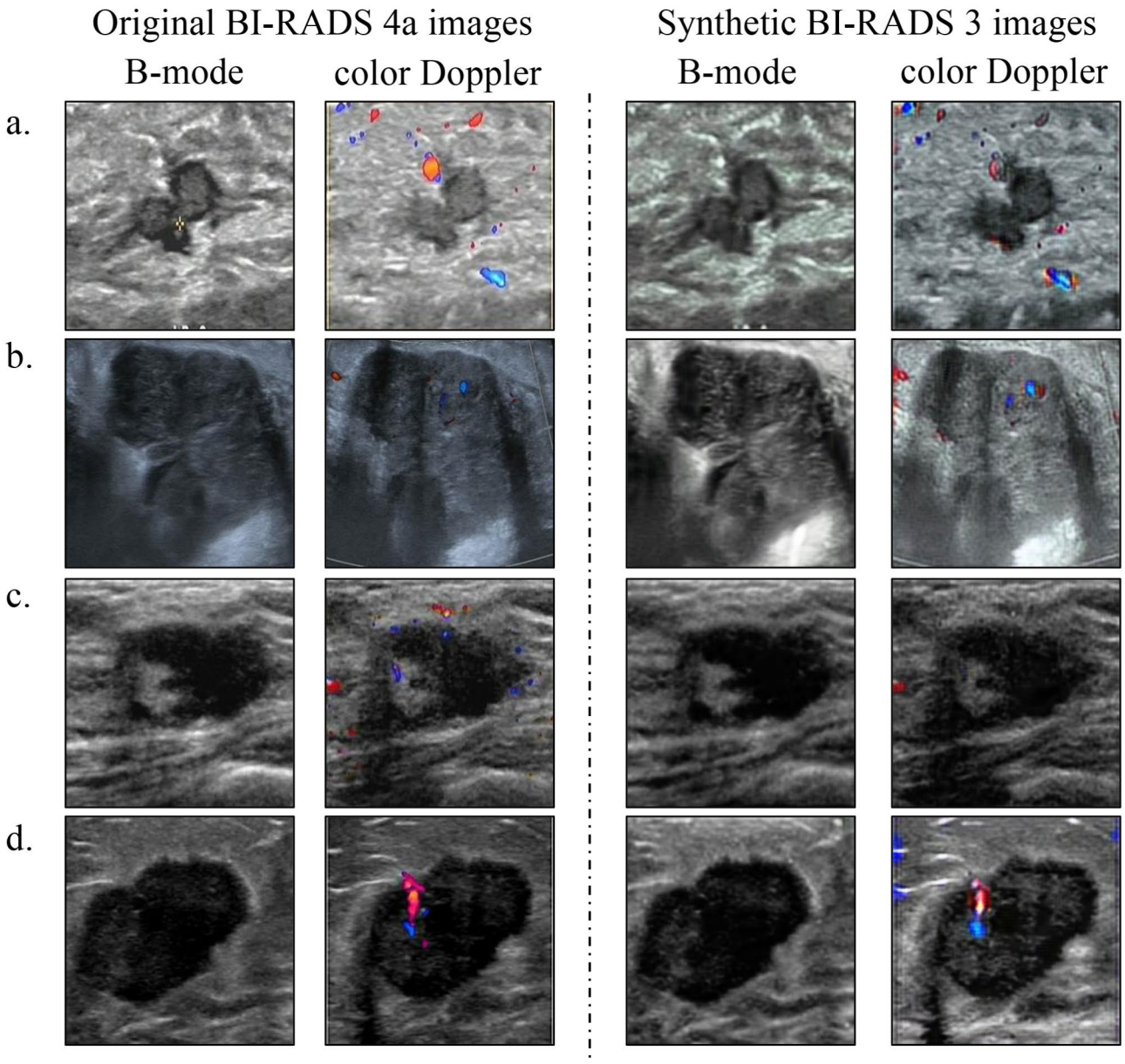

Fig. 3 Examples of real BI-RADS 4 A and corresponding synthetic BI-RADS 3 lesions. Real BI-RADS 4 A B-mode and color Doppler images are displayed in the first and second columns. Synthetic BI-RADS 3 B-mode and corresponding color Doppler images are presented in the third and fourth columns by Bmode-GAN and Doppler-GAN models, respectively

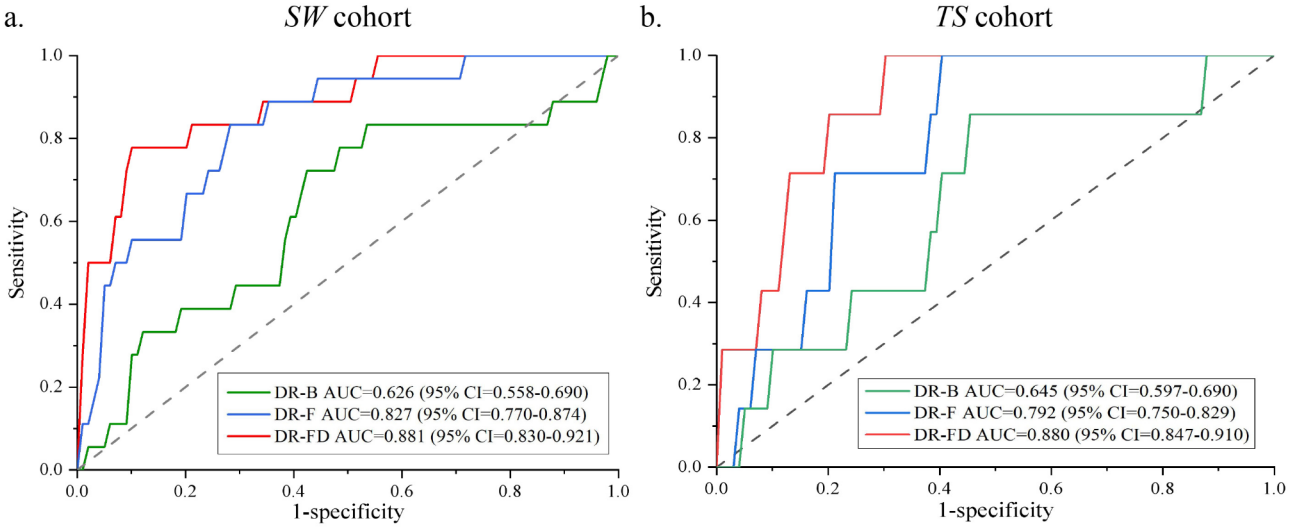

Fig. 4 Comparison of ROC curves among the DR-B, DR-F, and DR-FD models for predicting BI-RADS 3 benign and malignant lesions on SW and TS cohorts

Table 2 Performance of different augmentation methods for predicting BI-RADS 3 lesions in the internal and external test sets

	Method	AUC (95% CI)	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)	F1-score	FPR
SW cohort	DR-B	0.626 (0.558–0.690)	16.7	90.5	84.4	13.6	92.3	0.150	0.095
	DR-F	0.827 (0.770–0.874)	55.6	84.5	82.1	24.4	95.5	0.339	0.155
	DR-FD	0.881 (0.830–0.921)	77.8	87.5	86.7	35.9	97.8	0.491	0.125
	Radiologists	0.728 (0.663–0.786)	33.3	97.5	92.0	54.5	94.2	0.414	0.025
TS cohort	DR-B	0.645 (0.597–0.690)	14.3	90.5	89.3	2.4	98.4	0.041	0.095
	DR-F	0.792 (0.750–0.829)	28.6	84.1	83.2	2.8	98.6	0.052	0.159
	DR-FD	0.880 (0.847–0.910)	71.4	85.0	84.8	7.4	99.4	0.133	0.150
	Radiologists	0.696 (0.650–0.740)	14.3	98.5	97.4	16.7	98.5	0.154	0.012

AUC area under the receiver operating characteristic curve, *PPV* positive predictive value, *NPV* negative predictive value, *FPR* false positive rate, *CI* confidence interval, *Radiologists* all the four radiologists

The CycleGAN model accomplished a very complex image-to-image translation. From the perspective of visual perception, in the direction of BI-RADS 4 A to BI-RADS 3, the generated B-mode and color Doppler images effectively preserved the lesion information in BI-RADS 4 A, while changing in color, brightness, contrast, and other stylistic features of BI-RADS 3. Overall, the quality and variation of the synthetic images were consistent with our expectations. However, in actual clinical practice, it must never be assumed that the generated BI-RADS 3 lesion images had the same properties as the BI-RADS 3 lesion images of the specific actual patient, even though the synthetic images looked realistic. In our method, the synthetic BI-RADS 3 lesion images were only applied as data augmentation for training Dual-ResNet50 to alleviate the level of class imbalance, not for actual clinical diagnosis.

Performance of Dual-ResNet50 with augmentations

We compared the performance of the Dual-ResNet50 classification model before and after augmentations by examining various metrics (Fig. 4; Table 2). The DR-FD model exhibited outstanding performance in distinguishing malignant from BI-RADS 3, with AUCs of 0.881 (95% CI: 0.830–0.921) in the internal test set, and 0.880 (95% CI: 0.847–0.910) in the external test set. In comparison, both DR-B and DR-F models showed lower AUCs on SW cohort (DR-B: AUC = 0.626, 95% CI: 0.558–0.690, $P < 0.001$, DR-F: AUC = 0.827, 95% CI: 0.597–0.690, $P < 0.05$), and TS cohort (DR-B: AUC = 0.645, 95% CI: 0.597–0.690, $P < 0.005$; DR-F: AUC = 0.792, 95% CI: 0.750–0.829, $P < 0.05$).

The ability of the Dual-ResNet50 model in detecting malignant and benign cases before and after augmentations is further presented in Table 2; Fig. 5. On SW cohort containing 18 malignant and 200 benign lesions, DR-FD achieved a high sensitivity of 77.8% (14 out of 18 malignant cases). By contrast, DR-B had a low sensitivity of 16.9% (only 3 out of 18 malignant cases), while DR-F had a sensitivity of 55.6% (10 out of 18 malignant cases). Similarly, on TS cohort consisting of 421 benign and 7 malignant lesions, the DR-B model demonstrated a low sensitivity of 14.3% (only 1 true positive case). However, DR-F showed a sensitivity of 28.6% (2 true positive cases), and DR-FD achieved a sensitivity of 71.4% (5 true positive cases). These findings clearly demonstrated our

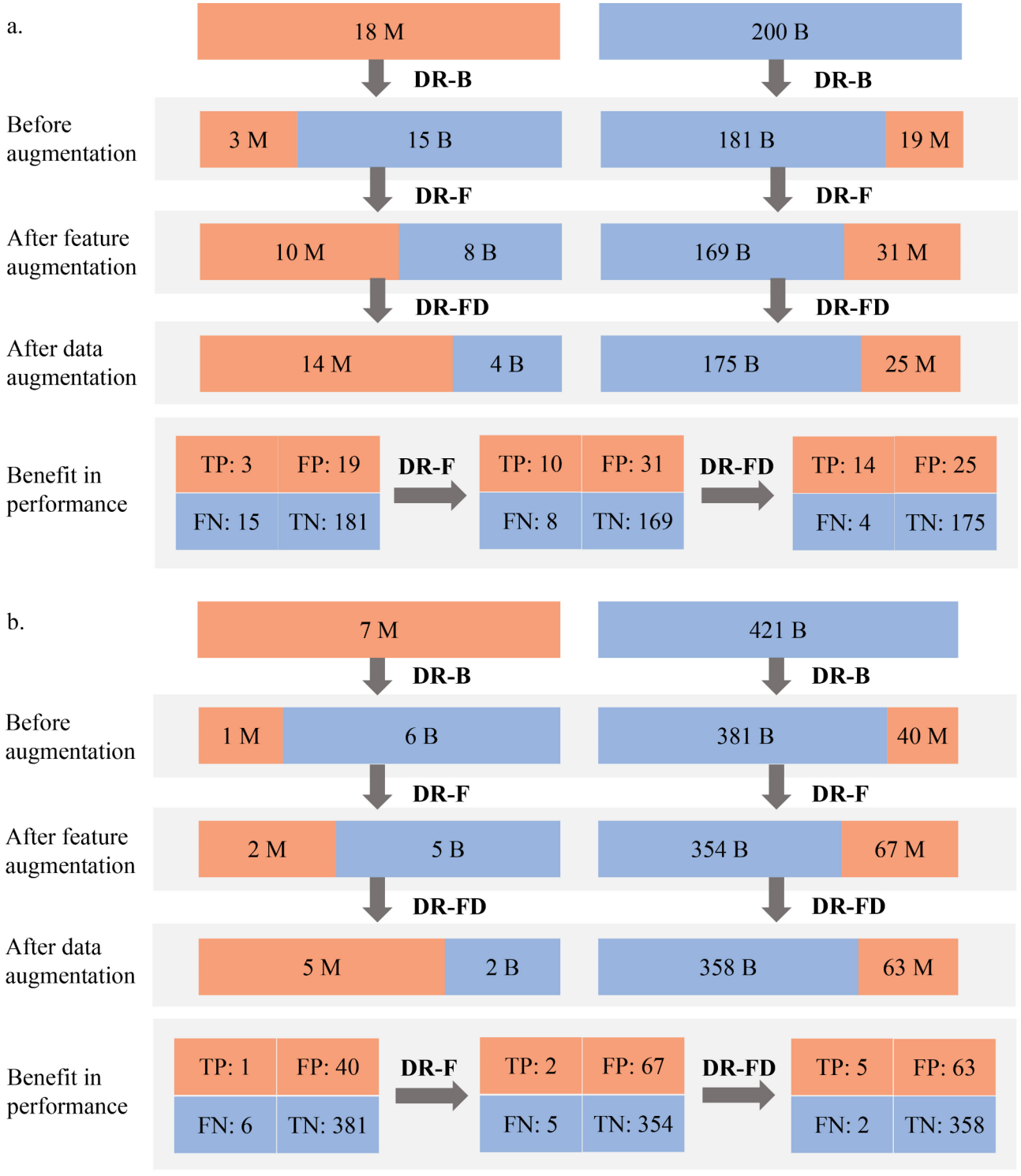

Fig. 5 Performance of Dual-ResNet50 model in detecting malignant and benign lesions with different augmentation methods on SW and TS cohorts. *M* malignancy, *B* benign, *TP* true positive, *FP* false positive, *FN* false negative, *TN* true negative. (**a**) the classification performance of Dual-ResNet50 in the SW test set using three different levels of augmentation: no augmentation, feature augmentation, and combined feature and data augmentation. (**b**) the classification performance of Dual-ResNet50 performance in the TS test set using three different levels of augmentation: no augmentation, feature augmentation, and combined feature and data augmentation

augmentation methods could improve the performance of DL classification models in identifying malignant lesions to various degrees.

In addition, we observed the performance of the DR-FD model trained separately by B-mode or color Doppler images on SW cohort (Additional file 1: Fig. S3). Single modality B-mode or color Doppler images effectively differentiated benign and malignant lesions (AUC = 0.795, 95% CI: 0.735–0.847, AUC = 0.789, 95% CI: 0.728–0.841). Notably, the bimodal information integrating both B-mode and color Doppler images further enhanced the diagnostic accuracy (AUC = 0.880, 95% CI: 0.830–0.921).

Comparison of DR-FD model and radiologists

According to Table 2, the DR-FD model exhibited excellent performance in terms of AUC, sensitivity, and particularly in identifying patients with malignant lesions. The final consensus diagnosis and the independent diagnoses of the four radiologists are provided (Table 2 and Additional file 1: Table S3), respectively. The average AUC of these radiologists was 0.728 (95% CI: 0.663–0.786, $p < 0.01$ for DR-FD vs. radiologists) and a corresponding sensitivity of only 33.3% on SW cohort. Meanwhile, on TS cohort, the average AUC was 0.696 (95% CI: 0.650–0.740, $p < 0.01$ for DR-FD vs. radiologists), and the sensitivity was merely 14.3%. However, radiologists exhibited a high specificity, indicating their tendency to give a benign diagnosis to BI-RADS 3 lesions, which may lead to missing some patients with malignant lesions. Therefore, our DR-FD model has the potential to serve as a valuable tool, alerting radiologists in paying attention to these predicted high-risk lesions.

Interpretability of the DR-FD model

To enhance the interpretability of our DR-FD model, and make radiologists understand the decision-making process, we incorporated heatmaps to highlight the important regions associated with DR-FD predictions. Each heatmap highlighted regions of US images that were vigorous in predicting malignant and benign microcalcification, with areas of strong emphasis marked in red and areas of weak emphasis in blue.

The heatmap signals presented by B-mode and color Doppler images of the same lesion did not overlap, when the DR-FD model was employed to predict breast US images (Fig. 6). For B-mode images, the heatmap mainly covered the area of the lesion, indicating a potential focus on the morphologic and acoustic features of breast masses. For color Doppler images, the heatmap primarily overlayed regions with relatively rich blood vessels. This revealed a specific attention on capturing blood flow information.

The interpretability features presented by the heatmap offered valuable guidance for radiologists. In cases where there was inconsistency between the initial assessment of radiologists and the predictions of the DR-FD model for the same lesion, the heatmap offered an opportunity for additional analysis. This facilitated a more comprehensive evaluation of suspicious lesions, ensuring that they received secondary attention.

Discussion

The increasing detection rate of early-stage breast cancer and suitable treatment effectively reduced the mortality rate. From the perspective of breast US imaging, BI-RADS 3 lesions have circumscribed margin, oval shape, and parallel orientation of the mass, showing benign characteristics [24]. The malignant masses diagnosed by follow-up patients with category 3 are typically early stage, small in size, and node-negative [25], and the imaging characteristics of malignant masses are very similar to those of benign category 3 lesions. Even experienced US radiologists could not easily detect breast cancer patients in BI-RADS 3 population in time, thus leading to potential misdiagnosis or delayed confirmation of cancer during follow-up. In this study, we aimed to design a two-stage augmentation method for BI-RADS 3 US images, and utilized the augmented images to train the Dual-ResNet50 classification network, achieving the classification of BI-RADS 3 benign and malignant lesions. Different from previous studies [26–29], we focused on the challenging task of breast lesions graded by BI-RADS 3, which was difficult to accurately identity the malignancy of breast ultrasound images with naked eyes or computer-aided diagnosis [30]. Our results demonstrated that the proposed two-stage augmentation method held promising potential as an effective tool to assist the DL model in enhancing the detection ability of BI-RADS 3 lesions. Thus, it could suggest follow-up observation for low-risk lesions, and promote early confirmation of breast cancer with double-checks or immediate biopsy for high-risk BI-RADS 3 lesions.

The performance of DR-B in classifying BI-RADS 3 lesions was unsatisfactory with low AUCs and sensitivities on both SW and TS cohorts (Table 2), which further highlights the challenges when training a DL model using BI-RADS 3 lesions in the clinical setting. BI-RADS categories 1–3 and 5, as visualized by mammography and US imaging, display obvious benign and malignant features, facilitating their accurate diagnosis as either benign or malignant by radiologists or DL models [31]. Moreover, the PPV of BI-RADS 4 A patients is less than 10% [32]. These findings indicate that BI-RADS 3 lesions typically exhibit benign features on US imaging, while the malignant features of BI-RADS 4 A are relatively weak. This relationship highlights they are closely related in terms of

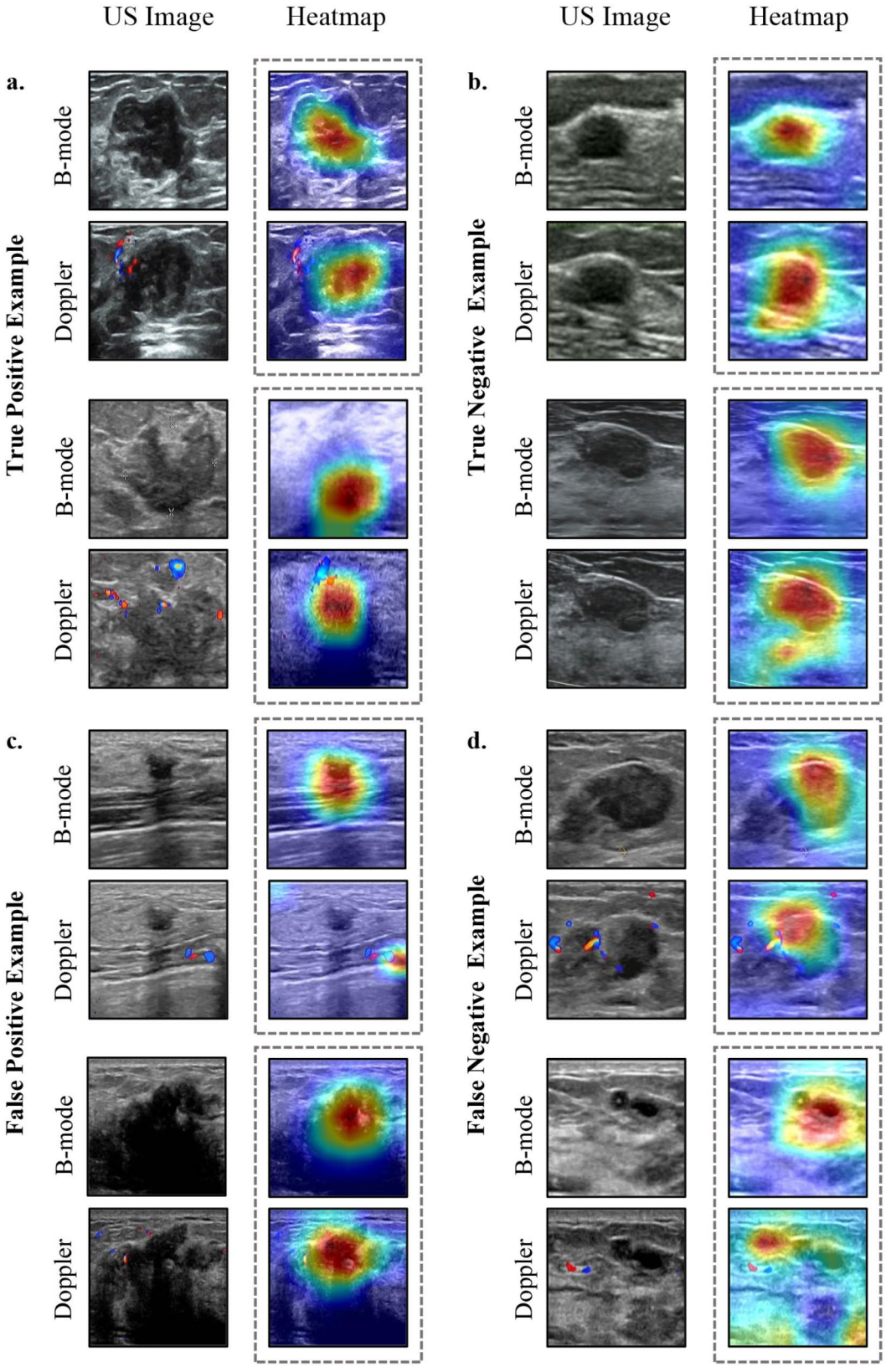

Fig. 6 Examples of biopsy-confirmed breast lesions and corresponding heatmaps on SW cohort. Each column of US Image contains B-mode and color Doppler images of two lesions, and each column of Heatmap represents the corresponding heatmap of the specified lesion predicted by DR-FD with Grad-CAM. (**a**) heatmaps of true positive lesions, (**b**) heatmaps of true negative lesions, (**c**) heatmaps of false positive lesions, (**d**) heatmaps of false negative lesions

malignant characteristics, particularly in the manifestation of early malignancy.

Based on this similarity, the incorporation of BI-RADS 4 A malignant lesions into the training process may potentially assist the DL classification model, in learning more obvious features associated with malignant lesions, thereby improving the identification of potential malignancies. Therefore, by utilizing our proposed feature augmentation method, the DR-F model achieved a significant improvement, especially in AUC and sensitivity. This finding reveals the potential of extracting similar features from existing malignant samples classified as BI-RADS 4 A, and employing them in the diagnosis of BI-RADS 3 lesions. As a result, the DL model more easily detects subtle features associated with BI-RADS 3 malignancy lesions. Additionally, data augmentation techniques based on GANs have emerged as effective approaches for mitigating data imbalance in computer vision tasks, thereby enhancing the performance of classification networks [33]. The generated images from GAN-based in the classification task of many clinical diseases, such as chest X-Rays images and colorectal histopathology images, provided a solution to address the issue of data imbalance [34–36] Through the combination of feature augmentation and data augmentation, DR-FD achieved a high AUC and sensitivity. This observation confirms the importance of appropriately addressing data imbalance for breast lesion classification, and demonstrates the potential of the CycleGAN generation method in enhancing the accuracy of breast lesion diagnosis at an early stage.

In breast US images, almost 99% of malignant lesions presented a color signal, whereas only 4% of benign lesions showed a detectable vascularization [37]. In case of benign lesions, vessel distribution was found to be equivalent for the core and periphery of the lesion, whereas malignant lesions had greater vascularization towards their center [38]. The heatmap provided valuable insights into the accuracy of our model in identifying and locating the lesion information. These accurate predictions by the DR-FD model in Fig. 6a and b imply a potential correlation between blood flow signals and malignant lesions, which is critical for the screening of breast cancer in BI-RADS 3 patients. Specifically, in the true positive (14 in the SW dataset and 5 in the TS dataset) and true negative (175 in the SW dataset and 358 in the TS dataset) cases, the heatmap accurately captured the morphology and blood flow signals of the lesion. Moreover, the signal enhancements of blood vessels and the velocity of blood flow in the heatmaps of color Doppler images were more substantial in malignant lesions compared to benign lesions. The color Doppler ultrasound represents the blood flow as a color map superimposed over matched B-mode image [39, 40]. Therefore, we speculate that the strong signal in the Doppler heatmap obtained by DR-FD is more likely to resemble the blood flow signal from the small blood vessels rather than the structural signal from the B-mode image. This may be attributed to the scattered or weak signals of microvessels, resulting in insufficient pixels to display red or blue colors. Additionally, it is possible that the blood flow velocity of microvessels is not fast enough to produce noticeable red or blue coloration. However, these disparities are observed between the predicted and actual lesion locations, as indicated by the cases of false positive (25 in the SW dataset and 63 in the TS dataset) and false negative (4 in the SW dataset and 2 in the TS dataset) (Fig. 6c and d). This reveals that the potential limitations of our models may lie in not effectively capturing crucial information, or overly emphasizing specific features during the prediction process.

Our study still has several limitations. First, we only collected datasets from two hospitals, which may not fully represent the characteristics of other regions and populations. Second, we integrated the bimodal information from B-mode and color Doppler images. However, it should be noted that US images alone may not fully reflect the disease. Clinicians typically considered multiple information sources, including the medical history, clinical manifestations, and other imaging examination results. Furthermore, we only compared the individual diagnostic performance of the model and radiologists in this retrospective study, but failed to determine whether the model had a positive impact on the clinical diagnosis.

In conclusion, our study demonstrated the potential of the proposed two-stage augmentation method in improving the ability of breast cancer detection in BI-RADS 3 lesions, and provided a promising tool for early diagnosis of breast cancer. In the future, we will transfer our augmentation method to other medical domains that have the feature of the scarcity of malignant samples, for identifying early cancers. This could improve cancer detection and help assessing the generalizability of our method. In addition, we plan to explore integrating imaging data with clinical information, such as patient records and imaging reports, through a multimodal deep learning framework to further enhance diagnostic performance.

Abbreviations

US	Ultrasound
BI-RADS	The breast imaging reporting and data system
AI	Artificial intelligence
DL	Deep learning
DCNNs	Deep convolutional neural networks
GAN	Generative adversarial network

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-13960-0.

Additional file 1: Electronic supplymentary material

Supplementary Figure 1: Training loss of two CycleGAN models named Bmode-GAN and Doppler-GAN generative model under ten-fold cross-validation

Supplementary Figure 2: Training loss of DR-FD model

Supplementary Figure 3: Comparison of ROC curves for Dual-ResNet50 model using B-mode images, color Doppler images, and bimodal images on SW cohort

Acknowledgements
Not applicable.

Author contributions
CL, WY, and LJW played a key role in the research design, including the formulation of the research questions and the development of the overall methodology. CL and GY provided the data. THH, CL, and CZG actively contributed to the data preprocessing, collaborating with other authors in the application of appropriate statistical methods. HXF and LJW designed the deep learning model. THH, LJW, GY, and CL wrote the manuscript with the assistance and feedback of all other co-authors.

Funding
This work was supported by the National Key R&D Program of China (no. 2018YFC1707503) and the Southwest University Research and Innovation Project (no. SWUB24052).

Data availability
The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request. The website of the retrospective datasets on a web-based rating platform is available at: http://birads3.ssdlab.cn/admin.

Declarations

Ethics approval and consent to participate
This retrospective study has been approved by the Ethics Committee of the First Affiliated Hospital of Army Medical University ([No. (B) KY202264]), and the requirement for informed consent from all patients was waived. All experimental procedures were carried out according to the Code of Ethics of the World Medical Association (Declaration of Helsinki).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 21 June 2023 / Accepted: 18 March 2025
Published online: 24 March 2025

References

1. Jemal A, Ward E, Thun MJ. Recent trends in breast cancer incidence rates by age and tumor characteristics among US women. Breast Cancer Res. 2007;9(3):1–6.
2. Ward ZJ, Atun R, Hricak H, Asante K, McGinty G, Sutton EJ, Norton L, Scott AM, Shulman LN. The impact of scaling up access to treatment and imaging modalities on global disparities in breast cancer survival: a simulation-based analysis. Lancet Oncol. 2021;22(9):1301–11.
3. Duggan C, Trapani D, Ilbawi AM, Fidarova E, Lavesanne M, Curigliano G, Bray F, Anderson BO. National health system characteristics, breast cancer stage at diagnosis, and breast cancer mortality: a population-based analysis. Lancet Oncol. 2021;22(11):1632–42.
4. Fitzgerald RC, Antoniou AC, Fruk L, Rosenfeld N. The future of early cancer detection. Nat Med. 2022;28(4):666–77.
5. Shen S, Zhou Y, Xu Y, Zhang B, Duan X, Huang R, Li B, Shi Y, Shao Z, Liao H, et al. A multi-centre randomised trial comparing ultrasound vs mammography for screening breast cancer in high-risk Chinese women. Br J Cancer. 2015;112(6):998–1004.
6. Van Sloun RJG, Cohen R, Eldar YC. Deep learning in ultrasound imaging. Proc. IEEE. 2019;108(1):11–29.
7. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, White E. Breast density as a predictor of mammographic detection: comparison of interval-and screen-detected cancers. J Natl Cancer Inst. 2000;92(13):1081–7.
8. Mendelson EB, Bohm-Velez M, Berg WA, Whitman GJ, Feldman MI, Madjar H, Rizzatto G, Baker JA, Zuley M, Stavros AT et al. ACR BI-RADS® Ultrasound. J Am Coll Radiol. 2013;149.
9. Szolovits P, Patil RS, Schwartz WB. Artificial intelligence in medical diagnosis. Ann Intern Med. 1988;108(1):80–7.
10. Saha A, Harowicz MR, Grimm LJ, Kim CE, Ghate SV, Walsh R, Mazurowski MA. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. Br J Cancer. 2018;119(4):508–16.
11. Yassin NIR, Omran S, Houby EMF, Allam H. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Comput Methods Programs Biomed. 2018;156:25–45.
12. Xu Y, Wang YX, Yuan J, Cheng Q, Wang XD, Carson PL. Medical breast ultrasound image segmentation by machine learning. Ultrasonics. 2019;91:1–9.
13. Rawat W, Wang ZJ. Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput. 2017;29(9):2352–449.
14. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2(10):719–31.
15. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9.
16. Chen ZH, Lin L, Wu CF, Li CF, Xu RH, Sun YJ. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. Cancer Commun. 2021;41(11):1100–15.
17. Raza S, Goldkamp AL, Chikarmane SA, Birdwell RL. US of breast masses categorized as BI-RADS 3, 4, and 5: pictorial review of factors influencing clinical management. Radiographics. 2010;30(5):1199–213.
18. Raza S, Chikarmane SA, Neilsen SS, Zorn LM, Birdwell RL. BI-RADS 3, 4, and 5 lesions: value of US in management—follow-up and outcome. Radiology. 2008;248(3):773–81.
19. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Commun ACM. 2020;63(11):139–44.
20. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 Proceedings of the IEEE international conference on computer vision(ICCV). 2017:2223–2232.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR). 2016:770–778.
22. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 Proceedings of the IEEE international conference on computer vision. 2017:618–626.
23. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: 2016 Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:2921–2929.
24. Song SE, Cho N, Chu A, Shin S, Yi A, Lee SH, Kim WH, Bae MS, Moon WK. Undiagnosed breast cancer: features at supplemental screening US. Radiology. 2015;277(2):372–80.
25. Sickles EA. Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases. Radiology. 1991;179(2):463–8.
26. Shen Y, Shamout FE, Oliver JR, Witowski J, Kannan K, Park J, Wu N, Huddleston C, Wolfson S, Millet A. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat Commun. 2021;12(1):5645.
27. Qian X, Pei J, Zheng H, Xie XX, Yan L, Zhang H, Han CG, Gao X, Zhang HQ, Zheng WW, et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat Biomed Eng. 2021;5(6):522–32.
28. Marinovich ML, Wylie E, Lotter W, Lund H, Waddell A, Madeley C, Pereira G, Houssami N. Artificial intelligence (AI) for breast cancer screening: breast screen population-based cohort study of cancer detection. EBioMedicine. https://doi.org/10.1016/j.ebiom.2023.104498

29. Gu Y, Xu W, Lin B, An X, Tian JW, Ran HT, Ren WD, Chang C, Yuan JJ, Kang CS, et al. Deep learning based on ultrasound images assists breast lesion diagnosis in China: a multicenter diagnostic study. Insights Imag. 2022;13(1):124.
30. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Bohm-velez M, Pisano ED, Jong RA, Evans WP, Morton MJ, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. JAMA. 2008;299(18):2151–63.
31. Lei C, Wei W, Liu Z, Xiong QQ, Yang CQ, Yang M, Zhang LL, Zhu T, Zhuang XS, Liu CL, et al. Mammography-based radiomic analysis for predicting benign BI-RADS category 4 calcifications. Eur J Radiol. 2019;121:108711.
32. Weaver DL, Rosenberg RD, Barlow WE, Lchikawa L, Carney PA, Kerlikowske K, Buist DSM, Geller BM, Key CR, Maygarden SJ, et al. Pathologic findings from the breast cancer surveillance consortium: Population-Based outcomes in women undergoing biopsy after screening mammography. Cancer. 2006;106(4):732–42.
33. Sampath V, Maurtua I, Aguilar Martin JJ, Gutierrez A. A survey on generative adversarial networks for imbalance problems in computer vision tasks. J Big Data. 2021;8:1–59.
34. Salehinejad H, Valaee S, Dowdell T, Colak E, Barfett J. Generalization of deep neural networks for chest pathology classification in X-Rays using generative adversarial networks. In: 2018 IEEE international conference on acoustics, speech and signal processing. 2018:990–994.
35. Wei J, Suriawinata A, Vaickus L, Ren B, Liu XY, Wei J, Hassanpour S. Generative image translation for data augmentation in colorectal histopathology images. In: Proceedings of machine learning research. 2019:10–24.
36. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. J Big Data. 2018;5(1):1–30.
37. Cosgrove DO, Kedar PP, Bamber JC, al-Murrani B, Davey JB, Fisher C, Mckinna JA, Svensson WE, Tohno E, Vagios E. Breast diseases: color doppler US in differential diagnosis. Radiology. 1993;189(1):99–104.
38. Athanasiou A, Tardivon A, Ollivier L, Thibault F, Khoury CE, Neuenschwander S. How to optimize breast ultrasound. Eur J Radiol. 2009;69(1):6–13.
39. Hooley RJ, Scoutt LM, Philpotts LE. Breast ultrasonography: state of the Art. Radiology. 2013;268(3):642–59.
40. Liao, J., Gui, Y., Li, Z., Deng, Z., Han, X., Tian, H.,... & Chen, L. (2023). Artificial intelligence-assisted ultrasound image analysis to discriminate early breast cancer in Chinese population: a retrospective, multicentre, cohort study. EClinicalMedicine, 60.

Publisher's note