www.nature.com/scientificreports

scientific reports

OPEN

A deep learning-based multimodal medical imaging model for breast cancer screening

Junwei Chen[1,3,13], Teng Pan[2,13], Zhengjie Zhu[2,13], Lijue Liu[1,3]✉, Ning Zhao[4], Xue Feng[5], Weilong Zhang[6], Yuesong Wu[7], Cuidan Cai[2], Xiaojin Luo[2], Bihai Lin[2], Xuewei Wang[2], Qiaoru Ye[8], Rui Gao[2], Zizhen Zhou[2], Richard Beatson[9], Jin Tang[1,3], Ruijie Ming[10], Dan Wang[11], Jinhai Deng[7,9,12]✉ & Guanglin Zhou[2]✉

In existing breast cancer prediction research, most models rely solely on a single type of imaging data, which limits their performance. To overcome this limitation, the present study explores breast cancer prediction models based on multimodal medical images (mammography and ultrasound images) and compares them with single-modal models. We collected medical imaging data from 790 patients, including 2,235 mammography images and 1,348 ultrasound images, and conducted a comparison using six deep learning classification models to identify the best model for constructing the multimodal classification model. Performance was evaluated using metrics such as area under the receiver operating characteristic curve (AUC), sensitivity, specificity, precision, and accuracy to compare the multimodal and single-modal classification models. Experimental results demonstrate that the multimodal classification model outperforms single-modal models in terms of specificity (96.41% (95% CI:93.10%-99.72%)), accuracy (93.78% (95% CI:87.67%-99.89%)), precision (83.66% (95% CI:76.27%-91.05%)), and AUC (0.968 (95% CI:0.947-0.989)), while single-modal models excel in sensitivity. Additionally, heatmap visualization was used to further validate the classification performance of the multimodal model. In conclusion, our multimodal classification model shows strong potential in breast cancer screening tasks, effectively assisting physicians in improving screening accuracy.

Keywords Digital mammography, Ultrasound, Deep learning, Multimodal

Breast cancer is one of the most common malignant tumors among women, predominantly affecting those over the age of 50. According to the Global Cancer Statistics 2022[1], breast cancer ranks among the top five most diagnosed cancers in women worldwide, with 2.3 million new cases and 665,700 deaths reported. In many high-income countries, mortality rates have been declining since the early 1990s, thanks largely to therapeutic advancements and increased screening for early detection[1].With few identified modifiable risk factors for breast cancer, the primary focus of disease control is to enhance access to early diagnosis, screening, and timely,

[1]School of Automation, Central South University, Changsha, Hunan 410083, China. [2]Longgang District Maternity & Child Healthcare Hospital of Shenzhen City, Longgang Maternity and Child Institute of Shantou University Medical College, Shenzhen 518172, China. [3]Xiangjiang Laboratory, Changsha 410205, China. [4]National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China. [5]Department of Respiratory and Critical Care Medicine, Tianjin Chest Hospital, Tianjin 300222, China. [6]Department of Hematology, Lymphoma Research Center, Peking University Third Hospital, Beijing 100191, China. [7]Clinical Research Center (CRC), Medical Pathology Center (MPC), Cancer Early Diagnosis and Treatment Center (CEDTC) and Translational Medicine Research Center (TMRC), Chongqing University Three Gorges Hospital, School of Medicine, Chongqing 404100, China. [8]The Third People's Hospital of Longgang District Shenzhen, Shenzhen 518112, China. [9]Richard Dimbleby Department of Cancer Research, Comprehensive Cancer Centre, Kings College London, London SE1 1UL, United Kingdom. [10]Department of Oncology, Chongqing University Three Gorges Hospital, Chongqing 404010, China. [11]Richard Dimbleby Laboratory of Cancer Research, Randall Division and Division of Cancer and Pharmaceutical Sciences, King's College London, London, UK. [12]Guangzhou Baiyunshan Pharmaceutical Holding Co., Ltd. Baiyunshan Pharmaceutical General Factory/Guangdong Province Key Laboratory for Core Technology of Chemical Raw Materials and Pharmaceutical Formulations, Guangzhou 510515, China. [13]Junwei Chen, Teng Pan and Zhengjie Zhu contributed equally to this work. ✉email: ljliu@csu.edu.cn; jinhaideng_kcl@163.com; zgl3822@lg.gov.cn

comprehensive treatment[2].Early detection and diagnosis of breast cancer typically involves screening through various imaging modalities, such as ultrasound (US), digital mammography (DM), magnetic resonance imaging (MRI), and digital breast tomosynthesis (DBT), with mammography and breast US being the most prevalent[3]. Radiologists play a crucial role in examining breast tissue images to identify potential masses or abnormalities based on their density, structure, and morphology, determining their benign or malignant nature, in order to determine the benign or malignant nature of these masses. This task requires extensive professional knowledge and experience, with factors like image overlap, artifacts, and visual perception impacting diagnostic accuracy[4]. Furthermore, breast cancer screening covers a vast population, involving the examination of thousands of medical images, a process that is both time-consuming and labor-intensive.

Recent advances in artificial intelligence (AI), particularly deep learning, have significantly revolutionized medical image analysis and diagnosis[5], significantly enhancing early detection and screening efforts for breast cancer[6]. These models autonomously learn features from large medical image datasets, enhancing early detection capabilities. This automation improves efficiency in breast cancer screening and alleviates the workload on healthcare providers[7,8]. Unlike human radiologists, deep learning models excel at extracting complex image features, including subtle lesions that may be difficult to detect, thus increasing sensitivity in early detection and reducing unnecessary biopsies. Convolutional neural networks (CNNs), a cornerstone of deep learning models, are widely employed in medical image processing. They are adept at managing high-dimensional data and have been prominently used in breast cancer research, especially using DM and US imaging modalities[9–11]. DM is a pivotal tool for early breast cancer screening, renowned for its effectiveness and standardization in detecting and diagnosing breast abnormalities in women[12,13]. Previous studies like those by Arefan et al.[14], who developed a CNN-based breast cancer risk prediction model using DM data, and Mendel et al.[15], who extracted features from regions of interest in DM for breast cancer detection, showed accepted results with high area under the curve (AUC) values. US imaging, often used complementarily with DM, has also demonstrated high accuracy in detecting benign and malignant masses[13,16]. Recent studies revealed the efficacy of the application of CNN framework[17] and GoogleNet architecture with US images for localizing and classifying masses, and classifying breast cancer US images as benign or malignant, respectively[18].

Despite the promising results, several limitations are evident in the aforementioned studies utilizing DM and US images. These studies typically analyze one type of medical images, which inherently restricts the amount of structural information and local detail captured from breast tissue, particularly in cases involving occult or deep-seated lesions. In contrast, the integration of multimodal medical imaging, combining different imaging types, offers a more comprehensive view by providing unique perspectives, essential for detecting complex cases such as occult or deep-seated lesions[19,20]. For instance, while US excels at identifying larger benign calcifications ($\geq 0.5mm$), it may miss smaller microcalcifications ($< 0.5mm$) that are detectable by DM. In scenarios where DM struggles, such as with gland margins, breast color US can offer complementary insights, especially beneficial in dense breast scenarios where mammographic differentiation of lumps or nodules is challenging due to overlapping glands. Thus, combining DM with US significantly enhances breast cancer detection rates, which has been proved by a study establishing a multimodal classification model[21], which integrated US and DM images to predict malignancy degree in BI-RADS US 4A lesions. Experimental results demonstrated superior accuracy in malignancy prediction compared to single-modality approaches[21]. However, the model's generalizability is constrained by its focus on BI-RADS US 4A lesion patients. Therefore, this study proposes developing a deep learning-based multimodal breast cancer screening model that combines DM and US images. By leveraging these techniques, the goal is to reduce physician workload and enhance screening accuracy for all breast cancer patients.

Method

Data pre-processing

Our study has been performed in accordance with the Declaration of Helsinki. The US and DM images were evaluated by two radiologists; one with less than five years of experience and no specialized breast imaging training, and the other with over 20 years of experience and specialized training in breast imaging from Longgang District Maternity & Child Health care Hospital of Shenzhen City. The criteria for all pathological examinations were based on the 2019 World Health Organization (WHO) classification of breast tumors. The screened data of 790 breast cancer patients contained 2,235 DM images (1,832 benign and 403 malignant) and 1,348 US images (1,139 benign and 209 malignant).

Localization and cropping of tumor regions in US images

In breast cancer US images, the critical information for assessing tumor benignity and malignancy primarily resides within the tumor region. Outside this region, there exists irrelevant interference such as textual data, colored spots, and other artifacts. Therefore, we conducted tumor region cropping on the US images before inputting them into the deep learning model for training as shown in Fig. 1. This cropping process allows us to eliminate significant interfering information and extract effective features from the tumor region. Additionally, it helps to reduce the computational burden on the deep learning model.

Manually delineating tumor regions in US images is a labor-intensive process requiring precise identification of tumor boundaries and a comprehensive understanding of image structures and abnormal features. Given the intricate nature of breast cancer US imaging, this task demands operators with specialized medical expertise and extensive practical experience. To alleviate the manual workload, we employed the YOLOv8 target detection model to autonomously localize tumor regions in breast cancer US images. YOLOv8, the latest iteration in the YOLO series of target detection models, is an end-to-end real-time model renowned for its efficacy in detecting targets[22].

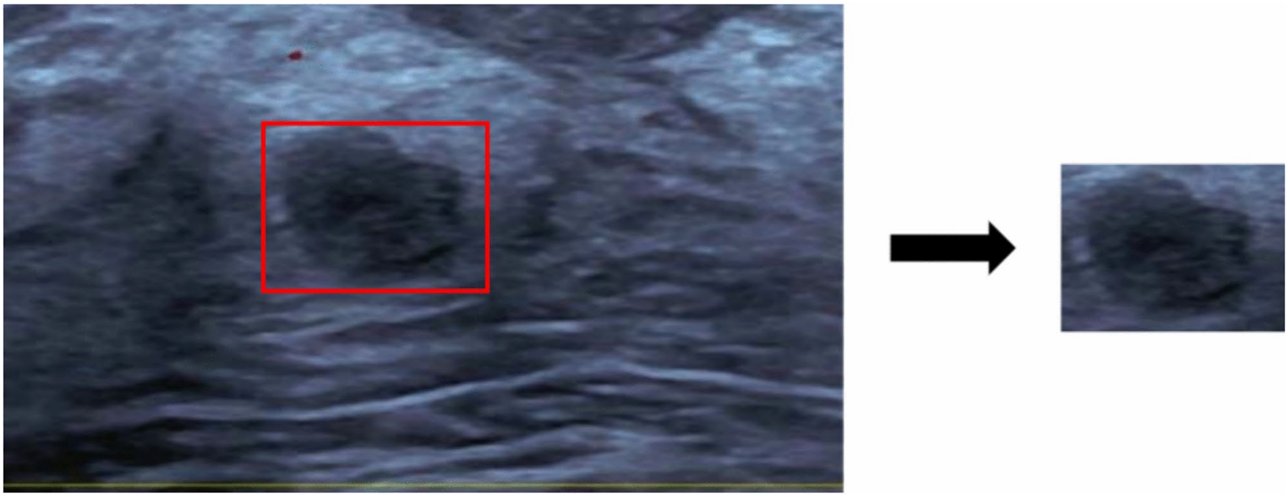

Fig. 1. ROI region cropping of US image.

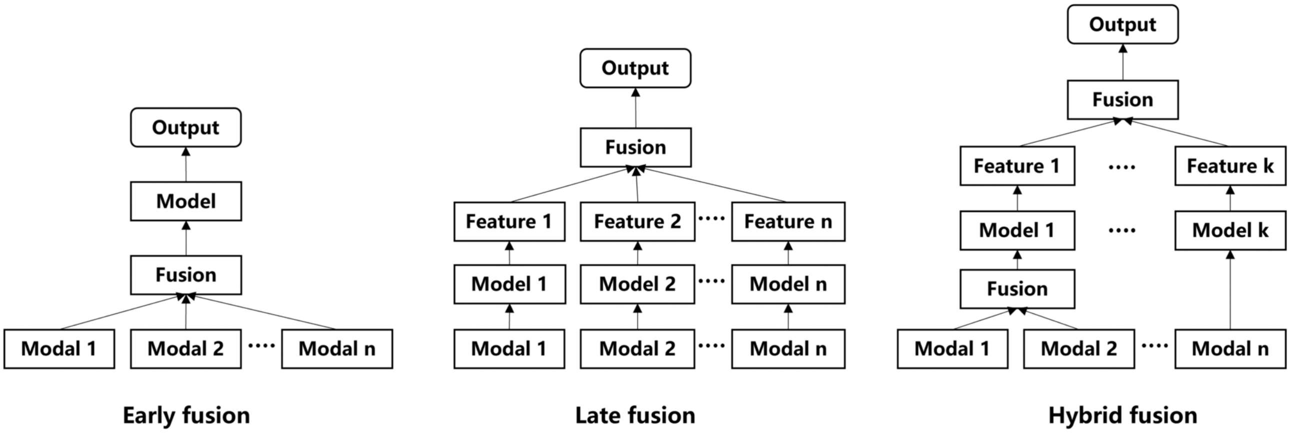

Fig. 2. Different multimodal fusion strategies.

Image grayscale normalization

Breast cancer medical images obtained from hospitals often exhibit inconsistent grayscale distributions due to variations in equipment acquisition and imaging factors. These discrepancies can adversely affect the training efficacy of deep learning models. Therefore, grayscale normalization is necessary for DM and US images to standardize their grayscale levels and mitigate variations among them. In this study, we employed the min-max normalization method to achieve this standardization. Equation (1) illustrates the formula used, where x_{min} represents the minimum grayscale value in the image and x_{max} represents the maximum grayscale value in the image.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Data augmentation

To address the issues of data imbalance and insufficient data in DM and US images, this study employs data augmentation techniques to increase the number of malignant samples[23]. The data augmentation operations include random horizontal flipping, random vertical flipping, elastic deformation, and combined random horizontal and vertical flipping. These operations aim to enhance and expand the existing samples, thereby increasing the diversity and quantity of the data. This, in turn, improves the robustness and performance of the model.

Model construction

The multimodal feature fusion strategies mainly include early fusion, late fusion, and hybrid fusion[24] as shown in Fig. 2. Early fusion concatenates features from multiple modalities at the shallow layers (or input layers) of the model, followed by a cascaded deep network structure, and ultimately connects to the classifier or other models[25–27]. Early fusion is an early exploration of multimodal fusion that learns the correlations between the low-level features of each modality. As it only requires training a single unified model, its complexity is manageable. However, early fusion faces challenges in feature concatenation due to the different sources of data from multiple modalities. Directly concatenating raw data may result in high feature dimensions, necessitating extensive data preprocessing. Late fusion involves independently training multiple models for each modality, where each modality undergoes feature extraction through separate models. The extracted features are then fused and connected to a classifier for final classification[28]. Hybrid fusion combines the principles of both early and late fusion. Since early fusion integrates multiple modalities at the shallow layers or input layers, it is suitable for cases with minimal differences between the modalities[29]. Given that this study involves only two modalities, DM and US images, which exhibit significant differences, the late fusion strategy is adopted.

The overall workflow of the multimodal model is shown in Fig. 3. Initially, the DM and cropped US image undergo resizing and grayscale normalization. Subsequently, the images are augmented using three methods: vertical flipping, horizontal flipping, and both horizontal and vertical flipping. The augmented DM and US images are used to train the multimodal model. Iteratively, the parameters of the baseline models and classifiers in both channels of the multimodal model are optimized to achieve the final multimodal model.

Selection of baseline model

In order to identify an appropriate baseline model for multimodal model channels, this study conducted comparative experiments involving six widely adopted deep learning classification models: ResNet-18, ResNet-50, ResNext-50, Inception v3, VGG16, and GoogleNet[30–34].

We trained six models on datasets of DM and US images. The US dataset consisted of tumor regions cropped using the YOLOv8 model. Fine-tuning deep learning models after transfer learning yields superior performance compared to training from scratch[35]. Therefore, prior to training our deep learning classification model on DM and US images, we pre-trained it on the ImageNet dataset of natural images. The model weights were initialized based on this pre-training for the natural image classification task. However, due to the discrepancy in image channel numbers between natural color images (3 channels) and grayscale DM and US images (1 channel), adjustments were necessary for the mammography and US datasets. The grayscale images were replicated across three channels to simulate three-channel images. During training, we employed the Adam optimizer with an initial learning rate of 0.0001, a batch size of 8, and a maximum of 200 iterations. Following comparative experiments, we selected ResNet-18 as our baseline model for the multimodal approach. The number of parameters of the model is about 11.7M.

Construction of multimodal model

The overall framework of our proposed multimodal model is shown in Fig. 4. This model consists of two channels, each utilizing a ResNet-18 architecture. Initially, two different medical images are fed into their respective ResNet-18 networks for feature extraction. After the average pooling operation, the output feature maps from the two channels are summed and fused to combine the feature information from both medical images. The fused feature maps are then passed through a fully connected layer and subsequently classified using a softmax layer to generate the final output. The input to the model is a pair of cropped US images and a whole DM image from the same patient, resized to 1024x1024 and 512x512 pixels, respectively. Prior to feeding

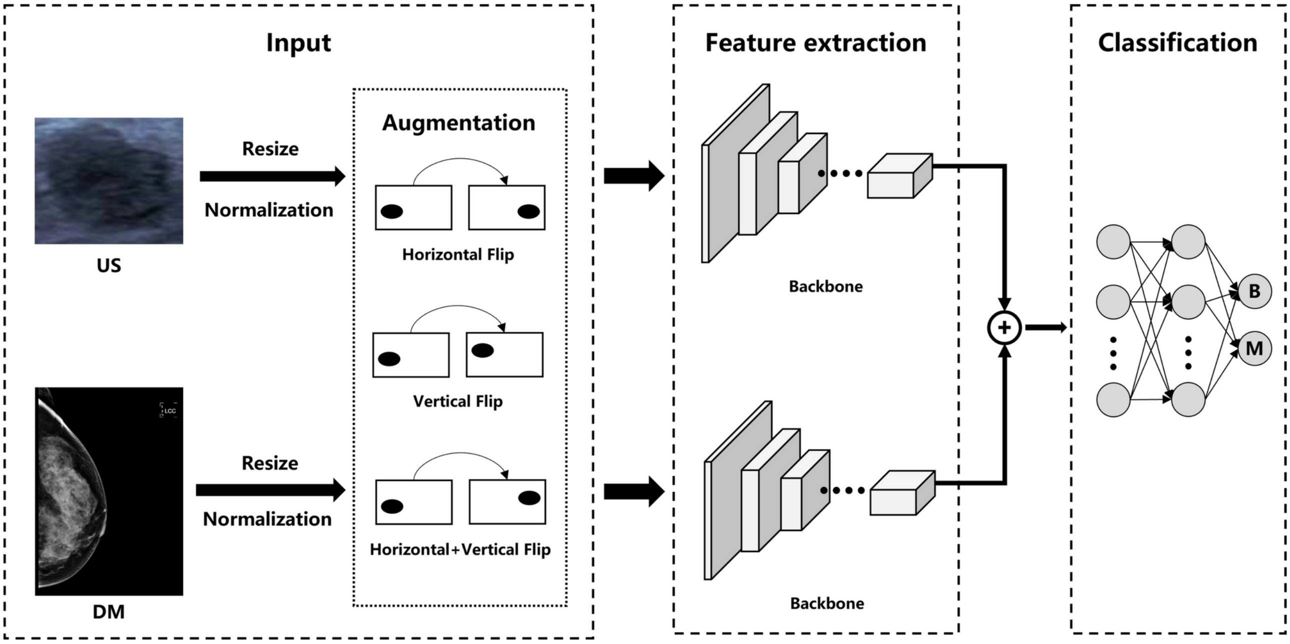

Fig. 3. The overall pipeline of multimodal model.

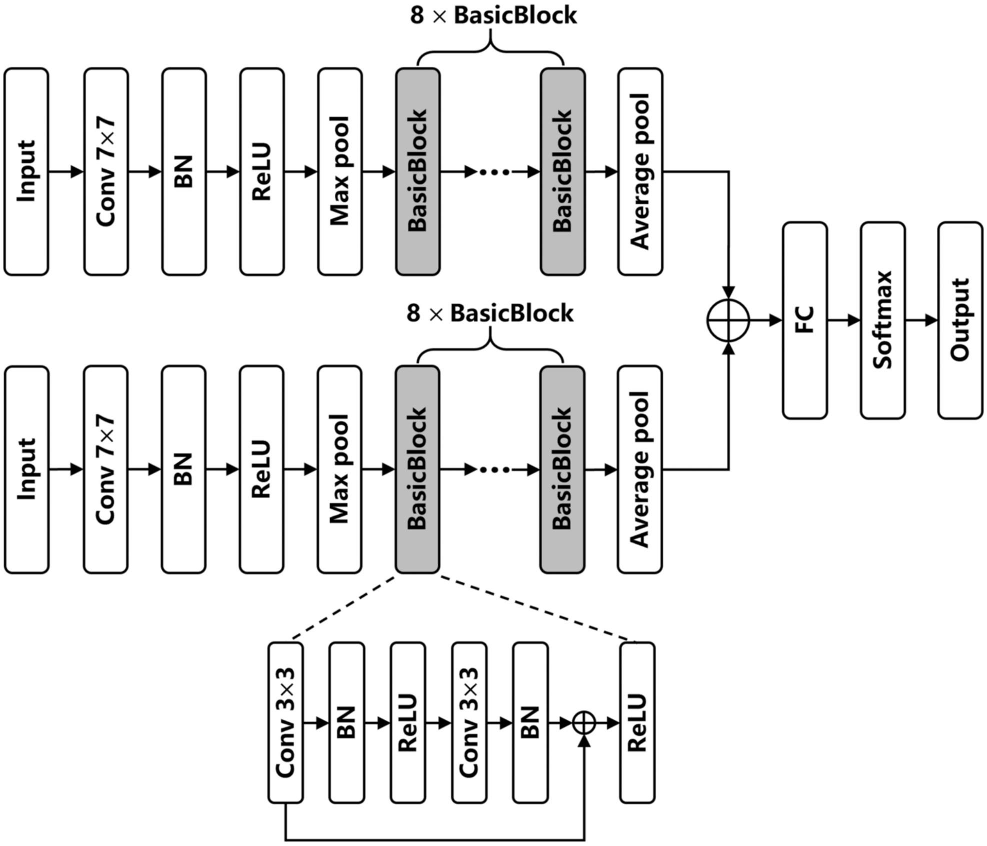

Fig. 4. The framework of multimodal model.

these images into the model, data augmentation techniques such as horizontal flip, vertical flip, and combined horizontal-vertical flip are applied to enhance the training data volume and improve the model's generalization performance. In the provided dataset, we observed that some patients had multiple US and DM images. To address this, we employed a cross-combination approach, matching each US image with a corresponding DM image to construct the inputs for the multimodal model. This resulted in a multimodal dataset comprising 6,105 samples, with 5,369 samples in the training set (2,759 benign, 2,880 malignant) and 466 samples in the test set (388 benign, 78 malignant).For training, the number of parameters of multimodal model is about 22.92M, we utilized the Adam optimizer to update the model parameters, with an initial learning rate of 0.0001, a batch size of 16, and a maximum of 200 iterations.

Model evaluation metric

In this study, we used sensitivity, specificity, accuracy, and precision as evaluation metrics to assess model performance, aligning with both the requirements of medical research and the standard evaluation metrics used in deep learning. Additionally, the Area Under the Receiver Operating Characteristic (AUC-ROC) curve was used to compare the classification performance of different models. The specific formulas for these evaluation metrics are shown in Equations (2–5).

$$sensitivity = \frac{TP}{TP + FN} \tag{2}$$

The ratio of the number of correctly predicted positive samples to the number of all positive samples.

$$specificity = \frac{TN}{TN + FP} \tag{3}$$

The ratio of the number of correctly predicted negative samples to the number of all negative samples.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

The ratio of the number of correctly predicted samples to the number of all samples.

$$precision = \frac{TP}{TP + FP} \tag{5}$$

The ratio of the number of correctly predicted positive samples to the number of all samples predicted as positive.

Where TP is true positive; TN is true negative; FP is false positive; FN is false positive.

Results

Images collection

All US and DM images were collected from 802 breast cancer patients. After screening out images with excessive artifacts or low resolution, we retained imaging data from 790 patients, comprising 585 benign and 205 malignant cases. These patients were randomly divided into training and validation cohorts, and a testing cohort at an 8:2 ratio. The patient inclusion workflow is detailed in Fig. 5.

Localization cropping of tumor regions in US images based on YOLOv8

To streamline the analysis of US images by focusing solely on tumor regions, we employed the YOLOv8 object detection model. This approach helped significantly reduce the manual effort involved by automatically cropping these regions. The model was trained using 1,348 US images, divided into 1,098 for training and 250 for testing. The training set was expanded to 4,392 images through data augmentation techniques, including horizontal flip, vertical flip, and combined horizontal-vertical flip. The training parameters were configured as follows: an initial learning rate of 0.01, a pre-trained model yolov8n.pt, a maximum of 300 iterations, image resizing to 640, a batch size of 8, a confidence threshold of 0.25, and an IoU threshold of 0.7.

The YOLOv8 model exhibited robust performance on the test set, achieving an accuracy of 0.913, a sensitivity of 0.887, and a mean Average Precision (mAP50) of 0.938, and the predictions of the model are shown in Fig. 6. Although the boundaries between tumors and normal tissues in US images are often indistinct, YOLOv8 was proved to effectively detect tumor regions in breast cancer US images for several reasons. Firstly, the multi-scale feature detection capability of YOLOv8 enables it to extract features at different levels, identifying tumors of various sizes and capturing subtle differences in the images[36]. Furthermore, YOLOv8 effectively separates the foreground (tumor region) from the background through its refined feature extraction and classification mechanism, which counters background noise and artifacts in US images.

Construction of deep learning model based on DM images

Next, we tried to classify benign and malignant digital mammography (DM) images. To select the optimal network for constructing a classification model, we experimentally compared six common classification networks: ResNet-18, ResNet-50, ResNeXt-50, Inception v3, VGG16, and GoogleNet. We utilized a total of 2,235 DM image datasets in our study. To equalize the distribution of benign and malignant cases in the training set, we implemented data augmentation, resulting in an expanded training set of 2,738 images (1,446 benign and

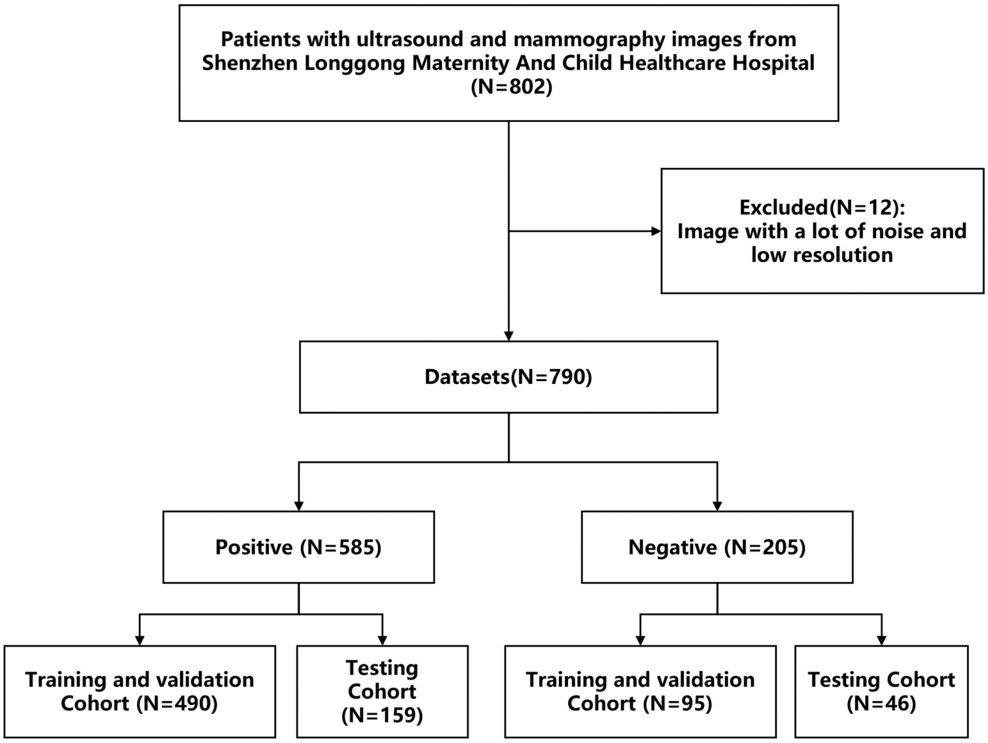
Patients with ultrasound and mammography images from Shenzhen Longgong Maternity And Child Healthcare Hospital (N=802)
Excluded(N=12): Image with a lot of noise and low resolution
Datasets(N=790)
Positive (N=585)
Negative (N=205)
Training and validation Cohort (N=490)
Testing Cohort (N=159)
Training and validation Cohort (N=95)
Testing Cohort (N=46)

Fig. 5. The patient inclusion workflow.

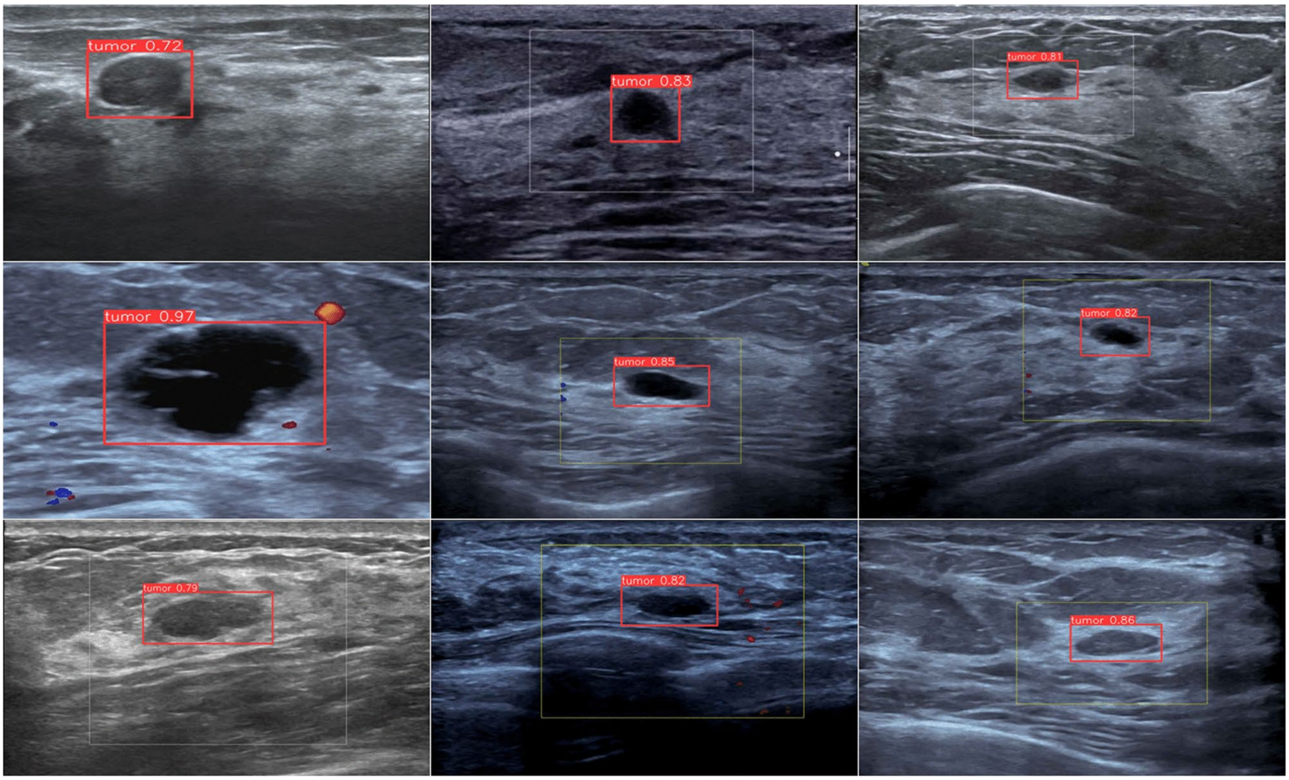

Fig. 6. YOLOv8 predictions for the test set.

Models	AUC	Sensitivity(%)	Specificity(%)	Precision(%)	Accuracy(%)
ResNet-18	0.950 ± 0.024	92.51 ± 5.41	90.75 ± 6.01	68.57 ± 6.16	91.09 ± 5.72
ResNet-50	0.896 ± 0.031	81.37 ± 6.36	84.32 ± 5.88	51.62 ± 6.41	83.98 ± 6.24
ResNext-50	0.953 ± 0.029	76.25 ± 6.19	90.42 ± 6.02	66.11 ± 5.77	87.71 ± 5.74
Inception v3	0.941 ± 0.026	80.04 ± 5.73	91.48 ± 5.52	67.22 ± 5.82	89.38 ± 5.92
VGG16	0.934 ± 0.022	76.58 ± 5.82	89.48 ± 6.04	69.30 ± 5.68	89.78 ± 5.54
GoogleNet	0.938 ± 0.019	65.06 ± 6.26	93.62 ± 5.22	68.24 ± 6.12	82.28 ± 5.46

Table 1. The performance of different models on DM images. CI = 95%, $P < 0.01$.

1,292 malignant) and a test set of 446 images (336 benign and 80 malignant). Each network was initially pre-trained on the ImageNet dataset prior to training on the DM dataset.

The results showed that ResNet-18 surpassed other models, achieving the highest sensitivity of 92.51% (95% CI:87.10%-97.92%), accuracy of 91.09% (95% CI:85.37%-96.81%), and AUC of 0.950 (95% CI:0.926-0.974) (Table 1). In contrast, GoogleNet recorded the highest specificity at 93.62% (95% CI:88.40%-98.84%) but a notably lower accuracy of 82.28% (95% CI:76.82%-87.74%). These findings indicate that ResNet-18 is superior in processing the DM image dataset. This result underscores that the complexity and size of the DM image dataset do not necessarily demand a deep network architecture for effective feature learning.. The relatively shallow network structure of ResNet-18 appears better adapted to DM images, potentially reducing the risk of overfitting. Additionally, the resolution and detailed features of DM images influence feature distribution, making ResNet-18's simpler network particularly effective at capturing essential features.

Construction of deep learning model based on US images

Also, we aimed to classify benign and malignant tissues from US images. Six different networks aforementioned have also been tested to identify the optimal network for constructing the classification model. A total of 1,348 US datasets were used. Data augmentation was also performed on the malignant images within the training set to balance the distribution of benign and malignant cases. This adjustment resulted in a training set expanded to 1,584 images (912 benign and 672 malignant) and a test set comprising 268 images (227 benign and 41 malignant). Similar to what we have done on DM images, each model underwent pre-training on the ImageNet dataset prior to training on the US datasets.

The findings demonstrated that ResNet-18 achieved the highest sensitivity of 92.68% (95% CI:86.45%-98.91%), specificity of 92.07% (95% CI:86.53%-97.61%), accuracy of 92.16% (95% CI:86.84%-97.48%), and

Models	AUC	Sensitivity(%)	Specificity(%)	Precision(%)	Accuracy(%)
ResNet-18	0.955 ± 0.017	92.68 ± 6.23	92.07 ± 5.54	67.86 ± 6.63	92.16 ± 5.32
ResNet-50	0.936 ± 0.023	85.37 ± 6.74	88.55 ± 6.33	57.38 ± 8.01	88.06 ± 7.72
ResNext-50	0.881 ± 0.027	78.05 ± 6.34	90.31 ± 6.21	59.26 ± 7.53	88.43 ± 6.35
Inception v3	0.941 ± 0.021	86.33 ± 5.52	91.32 ± 5.32	67.74 ± 7.63	89.24 ± 6.22
VGG16	0.934 ± 0.018	79.45 ± 5.24	89.75 ± 5.73	64.71 ± 7.74	87.78 ± 6.44
GoogleNet	0.939 ± 0.025	75.55 ± 5.44	90.54 ± 6.11	68.47 ± 6.71	88.45 ± 5.43

Table 2. The performance of different models on US images. CI = 95%, P < 0.01.

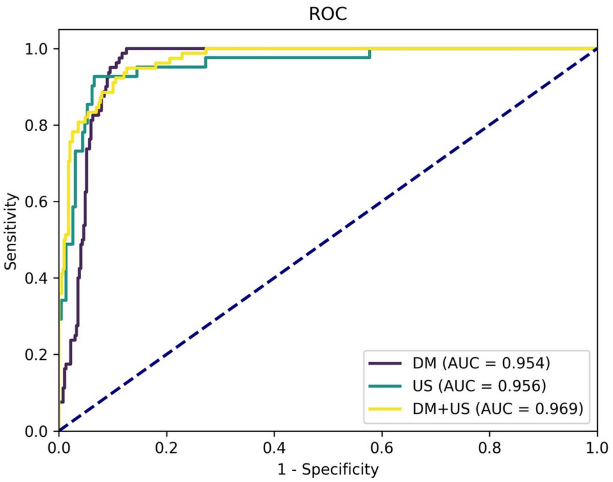

Fig. 7. The ROC curves of different models.

AUC of 0.955 (95% CI:0.938-0.972). Conversely, GoogleNet achieved the highest precision of 68.47% (95% CI:61.76%-75.18%) (Table 2). These results suggest that ResNet-18 excels in processing the US image dataset. US images are typically characterized by greater noise and lower image quality, factors that can negatively impact deeper, more complex models and increase the likelihood of overfitting. In contrast, the simpler architecture of ResNet-18 is less susceptible to overfitting, particularly with smaller datasets, making it a more effective choice for US image classification.

Construction of multimodal deep learning model based on both DM and US images

Compared to single type of medical images, multi-modal medical data can provide integrate characteristic information of different modes, better improving clinical applicability[37]. Thus, we developed a multimodal classification model combining digital mammography (DM) or ultrasound (US) images, and assessed this model's performance by comparing it with single-modal models that utilize either US or DM images exclusively. The experiment used data from 205 patients in the test cohort as a common test set. The Receiver Operating Characteristic (ROC) curves and confusion matrices for the three models are depicted in Figs. 7 and 8, with detailed test results displayed in Table 3. The results revealed that although the sensitivity of the multimodal model is lower than that of the single-modal models, it demonstrated superior performance in terms of AUC, specificity, precision, and accuracy. Specifically, the multimodal model achieved an AUC of 0.968 (95% CI:0.947-0.989), an accuracy of 93.78% (95% CI:87.67%-99.89%), a specificity of 96.41% (95% CI:93.10%-99.72%), and a precision of 83.66% (95% CI:76.27%-91.05%). From the confusion matrix, it can be seen that the number of false negatives in the results predicted by the multimodal model is higher than that of the single-modal model, and the number of false positives is lower than that of the single-modal model. The complexity of multimodal models makes their determination of malignant tumors more rigorous. When information conflicts between different models, multimodal models may tend to classify them as benign to avoid erroneous malignant diagnoses. The high AUC and accuracy indicate that the multimodal model is more effective than the single-modal models in classifying benign and malignant tumors. Its increased specificity suggests that the model excels at identifying benign cases during screenings, thereby reducing false positives and minimizing unnecessary biopsies, which enhances the overall screening process's efficiency and effectiveness.

Validation of multimodal model with heat map

To further validate whether our constrcuted multimodal model was able to accurately focus on the tumor regions in DM and US images, we visualized its predictions using Grad-CAM[38] and generated corresponding heat maps. These heat maps, as illustrated in Fig. 9, displayed three distinct colored regions indicating different levels of contribution to the model's predictive accuracy: the red region signifies the highest contribution and

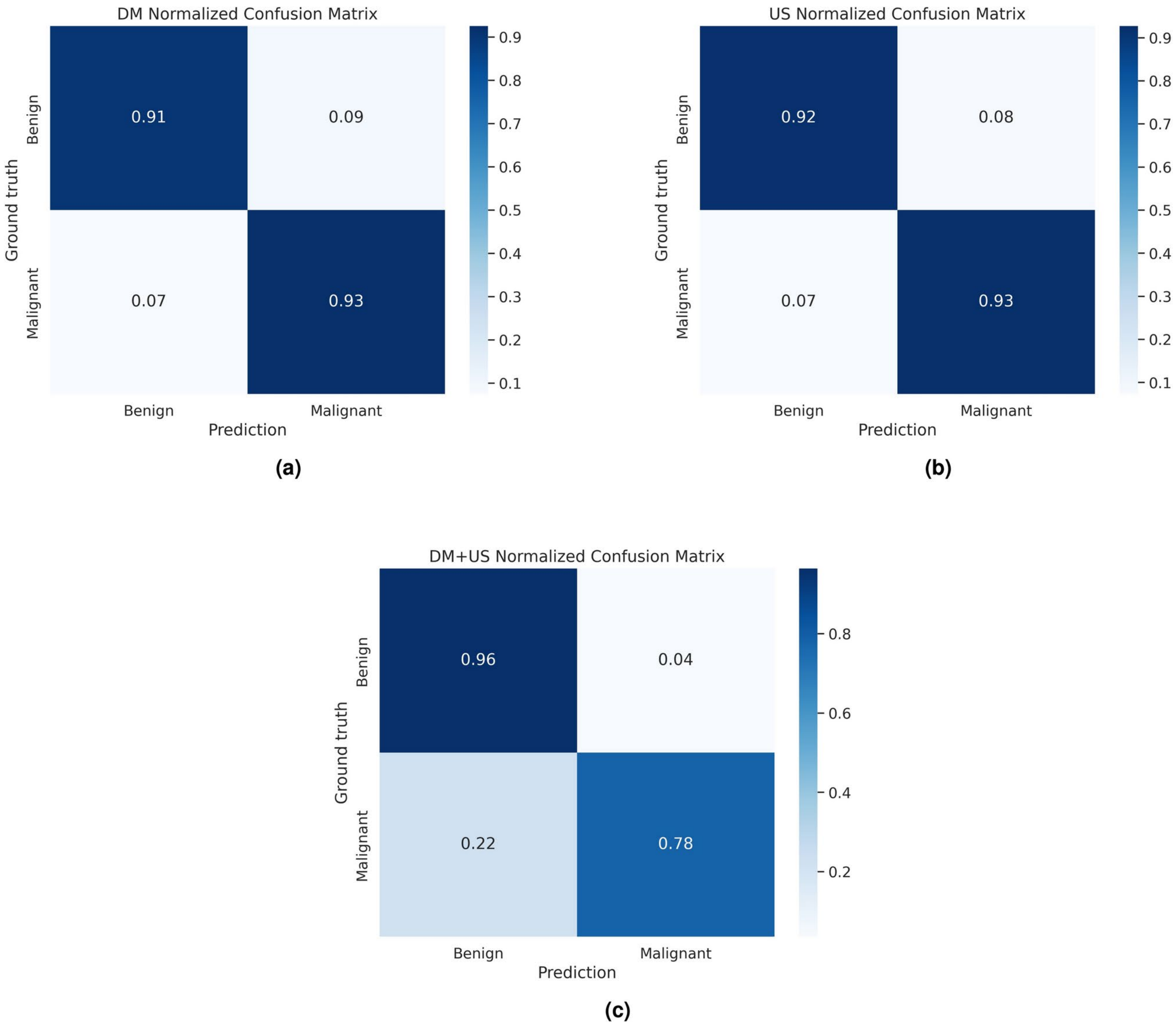

Fig. 8. The normalized confusion matrix of different models.

Models	AUC	Sensitivity(%)	Specificity(%)	Precision(%)	Accuracy(%)
DM Only	0.950 ± 0.024	92.48 ± 5.06	90.51 ± 6.71	68.42 ± 7.83	91.03 ± 6.15
US Only	0.955 ± 0.017	92.63 ± 4.13	92.01 ± 6.53	67.76 ± 7.11	92.16 ± 6.56
DM+US	0.968 ± 0.021	78.21 ± 6.24	96.41 ± 3.31	83.66 ± 7.39	93.78 ± 6.11

Table 3. The performance of models from different modalities in the test cohort. CI = 95%, P < 0.01.

captures the model?s primary focus, the yellow region indicates moderate contribution, and the blue region, which contributes the least, is typically deemed as redundant information. The results confirmed that the multimodal model could effectively focus on the tumor regions in both US and DM images, demonstrating its capacity to accurately localize and classify tumor regions. Notably, these heat maps can assist clinicians in making better diagnoses by highlighting the areas of the image that significantly influence the model's final classification decisions.

Discussion

Early screening for breast cancer is essential for reducing mortality rates. The screening process, however, involves a broad population base and relies significantly on the manual efforts of radiologists, consuming substantial human and material resources[37]. Moreover, effective breast cancer screening demands considerable professional knowledge and clinical experience from imaging doctors, with an inherent rate of misdetection. In this study,

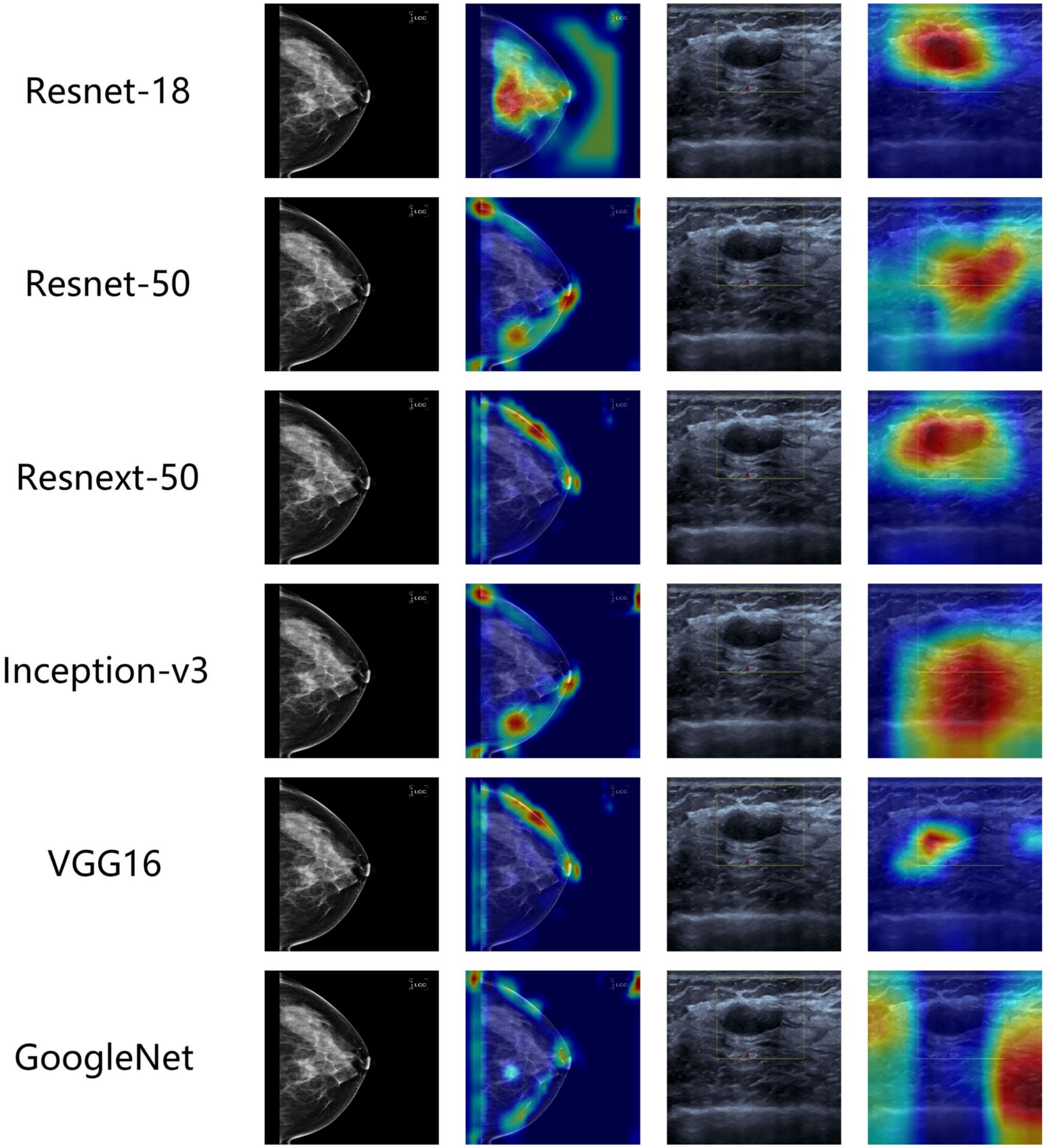

Fig. 9. The heat map of different models.

we developed a breast cancer screening model that integrates digital mammography (DM) and ultrasound (US) images using advanced deep learning techniques. This model aimed to alleviate the workload on clinicians and improve the efficiency and accuracy of screenings.

Here, we employed the YOLOv8 object detection model to localize and crop tumor regions in US images, thereby eliminating extraneous interference. This approach reduced the necessity for manual cropping and minimized redundant and interfering information, allowing the model to focus more effectively on the tumor characteristics. Additionally, we also analyzed the selection of multimodal feature fusion strategies. DM and US images differ significantly, possessing unique features and noise characteristics. Early fusion involves merging data from multiple modalities at the initial network level, requiring handling differences and noise across modalities. This can increase network complexity and potentially lead to the loss of crucial modal feature information, negatively impacting classification performance[29]. Consequently, we opted for a late fusion strategy.

Author	AUC	Sensitivity(%)	Specificity(%)	Precision(%)	Accuracy(%)
This paper	0.968	78.21	96.41	83.66	93.78
Yang et al.[21]	0.940	92.00	88.20	52.27	88.67
Arefan et al.[14]	0.730	-	-	-	-
Mendel et al.[15]	0.881	-	-	-	-

Table 4. The performance comparison with other related studies.

This approach allows independent networks for each modality to first extract specific features individually, without interference. This enables each network to concentrate on its unique features, maximizing their respective strengths. By integrating these features at a later stage, the model can leverage the complementary information from both modalities, improving overall classification performance.

In selecting a baseline model for our multimodal approach, we compared the classification performance of six models, finding that ResNet-18 performed optimally for both DM and US images. Despite the visual differences and noise characteristics between DM and US images, ResNet-18's deep structure and residual connections are well-suited to handle such variations. Residual connections help in deeper information propagation and fusion within the network, efficiently processing various medical images.

Next, we evaluated the performance of the multimodal model against single-modal models (DM only and US only). The results demonstrated that the multimodal model excelled in classifying benign and malignant breast cancer cases, achieving an AUC of 0.968 (95% CI:0.947-0.989), a sensitivity of 78.21% (95% CI:71.97%-84.45%), an accuracy of 93.78% (95% CI:87.67%-99.89%), a specificity of 96.41% (95% CI:93.10%-99.72%), and a precision of 83.66% (95% CI:76.27%-91.05%) in the test cohort. Except for sensitivity, all performance metrics of the multimodal model surpassed those of the single-modal models. The lower sensitivity of the multimodal model compared to the single-modal model may be attributed to its higher complexity. This increased complexity makes the multimodal model more sensitive to the size and balance of the training set. However, in terms of overall performance, the multimodal model outperforms the single-modal model. The enhanced precision of the multimodal model indicates a significantly reduced rate of false positives for benign lesions, substantially lowering the number of unnecessary biopsies. In addition, we compared the model performance with other related studies mentioned above, and all other indexes of our model except sensitivity exceeded other models, showing its excellent performance, as shown in Table 4. Compared to other studies, our model effectively utilizes the informative features of both DM and US images, improving its classification capabilities. Moreover, the model's versatility makes it applicable to all 35-65 year old breast cancer patients with both US and DM images, rendering it a reliable tool for breast cancer screening.

However, this study has several limitations. First, it is a single-center study, and our findings need validation with data from multiple centers. Second, the sample size may be insufficient, necessitating the recruitment of more breast cancer cases to further validate our results, thus enhancing the study's generalizability and clinical applicability. Third, there is uncertainty in image interpretation, influenced by factors such as the healthcare organization's level, examination equipment clarity, and the imaging physician's training. Lastly, our deep learning model does not include magnetic resonance imaging (MRI), which, due to its high cost, is not feasible for widespread implementation. In future studies, we aim to assist doctors in diagnosis by combining AI techniques, minimizing errors caused by manual examinations. This will help to address these limitations and improve the robustness and applicability of our model in clinical settings.

Conclusion

In this study, we constructed a multimodal breast cancer screening model by combining DM and US images using deep learning techniques. Our model demonstrated superior performance in classifying benign and malignant breast lesions, Compared with the single-modal models (DM only and US only), although the sensitivity is lower, the AUC, specificity, accuracy, and precision are higher. Unlike previously reported studies, Our model makes full use of the advantages of multimodal medical images, and is applicable to all 35-65 year old breast cancer patients with both US and DM images. In summary, our proposed model shows great potential to improve screening accuracy and reduce the burden on physicians in breast cancer screening tasks. Future studies should involve recruiting more multicenter breast data to further validate our findings, thereby enhancing the model's generalizability and clinical applicability.

Data availability

The medical imaging data used in this study are not publicly available due to patient privacy considerations but are available from the corresponding author on reasonable request.

Received: 11 November 2024; Accepted: 21 April 2025
Published online: 26 April 2025

References

1. Bray, F. et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024).
2. Jokhadze, N., Das, A. & Dizon, D. S. Global cancer statistics: A healthy population relies on population health. *CA Cancer J. Clin.***74** (2024).

3. Tarique, M., ElZahra, F., Hateem, A. & Mohammad, M. Fourier transform based early detection of breast cancer by mammogram image processing. *J. Biomed. Eng. Med. Imaging* **2**, 17 (2015).
4. Sadoughi, F. *et al.* Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. *Breast Cancer: Targets and Therapy* 219–230 (2018).
5. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
6. Luo, L. *et al.* Deep learning in breast cancer imaging: A decade of progress and future directions. *IEEE Reviews in Biomedical Engineering* (2024).
7. Yu, X., Zhou, Q., Wang, S. & Zhang, Y.-D. A systematic survey of deep learning in breast cancer. *Int. J. Intell. Syst.* **37**, 152–216 (2022).
8. Chougrad, H., Zouaki, H. & Alheyane, O. Deep convolutional neural networks for breast cancer screening. *Comput. Methods Programs Biomed.* **157**, 19–30 (2018).
9. Nasser, M. & Yusof, U. K. Deep learning based methods for breast cancer diagnosis: a systematic review and future direction. *Diagnostics* **13**, 161 (2023).
10. Abhisheka, B., Biswas, S. K. & Purkayastha, B. A comprehensive review on breast cancer detection, classification and segmentation using deep learning. *Arch. Computat. Methods Eng.* **30**, 5023–5052 (2023).
11. Siddique, M., Liu, M., Duong, P., Jambawalikar, S. & Ha, R. Deep learning approaches with digital mammography for evaluating breast cancer risk, a narrative review. *Tomography* **9**, 1110–1119 (2023).
12. Liu, J. et al. Radiation dose reduction in digital breast tomosynthesis (dbt) by means of deep-learning-based supervised image processing. In *Medical imaging 2018: Image processing* Vol. 10574 (ed. Liu, J.) 89–97 (SPIE, 2018).
13. Jalalian, A. et al. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clin. Imaging* **37**, 420–426 (2013).
14. Arefan, D. et al. Deep learning modeling using normal mammograms for predicting breast cancer risk. *Med. Phys.* **47**, 110–118 (2020).
15. Mendel, K. R., Li, H., Sheth, D. & Giger, M. L. Transfer learning with convolutional neural networks for lesion classification on clinical breast tomosynthesis. In *Medical Imaging 2018: Computer-Aided Diagnosis* Vol. 10575 (ed. Mendel, K. R.) 186–191 (SPIE, 2018).
16. Afrin, H., Larson, N. B., Fatemi, M. & Alizad, A. Deep learning in different ultrasound methods for breast cancer, from diagnosis to prognosis: current trends, challenges, and an analysis. *Cancers* **15**, 3139 (2023).
17. Shin, S. Y., Lee, S., Yun, I. D., Kim, S. M. & Lee, K. M. Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. *IEEE Trans. Med. Imaging* **38**, 762–774 (2018).
18. Han, S. et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Phys. Med. Biol.* **62**, 7714 (2017).
19. Haribabu, M., Guruviah, V. & Yogarajah, P. Recent advancements in multimodal medical image fusion techniques for better diagnosis: an overview. *Curr. Med. Imaging* **19**, 673–694 (2023).
20. Azam, M. A. et al. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput. Biol. Med.* **144**, 105253 (2022).
21. Yang, Y. et al. Deep learning combining mammography and ultrasound images to predict the malignancy of bi-rads us 4a lesions in women with dense breasts: a diagnostic study. *Int. J. Surg.* **110**, 2604–2613 (2024).
22. Terven, J., Córdova-Esparza, D.-M. & Romero-González, J.-A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Mach. Learn. Knowl. Extract.* **5**, 1680–1716 (2023).
23. Maharana, K., Mondal, S. & Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Trans. Proc.* **3**, 91–99 (2022).
24. Zhou, T., Ruan, S. & Canu, S. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array* **3**, 100004 (2019).
25. Zhang, Y., Sidibé, D., Morel, O. & Mériaudeau, F. Deep multimodal fusion for semantic image segmentation: A survey. *Image Vis. Comput.* **105**, 104042 (2021).
26. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit. Med.* **3**, 136 (2020).
27. Gadzicki, K., Khamsehashari, R. & Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)* (ed. Gadzicki, K.) 1–6 (IEEE, 2020).
28. Trong, V. H., Gwang-hyun, Y., Vu, D. T. & Jin-young, K. Late fusion of multimodal deep neural networks for weeds classification. *Comput. Electron. Agric.* **175**, 105506 (2020).
29. Pereira, L. M., Salazar, A. & Vergara, L. On comparing early and late fusion methods. In *International Work-Conference on Artificial Neural Networks* (ed. Pereira, L. M.) 365–378 (Springer, 2023).
30. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
31. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proc. IEEE conference on computer vision and pattern recognition*, 1492–1500 (2017).
32. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proc. IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).
33. Simonyan, K. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556 (2014).
34. Szegedy, C. *et al.* Going deeper with convolutions. In *Proc. IEEE conference on computer vision and pattern recognition*, 1–9 (2015).
35. Iman, M., Arabnia, H. R. & Rasheed, K. A review of deep transfer learning and recent advancements. *Technologies* **11**, 40 (2023).
36. Sohan, M., Sai Ram, T., Reddy, R. & Venkata, C. A review on yolov8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics* (ed. Sohan, M.) 529–545 (Springer, 2024).
37. Pei, X., Zuo, K., Li, Y. & Pang, Z. A review of the application of multi-modal deep learning in medicine: bibliometrics and future directions. *Int. J. Computat. Intell. Syst.* **16**, 44 (2023).
38. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE international conference on computer vision*, 618–626 (2017).

Acknowledgements

The authors thank the Longgang District Maternity & Child Healthcare Hospital of Shenzhen City for data support and guidance.

Author contributions

JC, LL and JD contributed to the conception and design. JC, TP, JD, GZ and LL contributed to the development of the methodology. JC, LL, TP, ZZ, NZ, XF, WZ, YW, CC, XL, BL, XW, QY, RG, ZZ and RB contributed to material preparation and data collection. JC, TP, JD, RM, DW contributed to the writing and revision of the manuscript. JC, JT, LL, GZ, RM, DW contributed to the study supervision. All authors read and approved the final manuscript.

Funding

This work was supported by Chongqing Natural Science Foundation (Grant No. CSTB2024NSCQ-MSX0761), Research Initiation Funding Project of Longgang District Maternity & Child Healthcare Hospital of Shenzhen City (Grant No. Y2024013), Scientific Research Project of Xiang Jiang Lab (Grant No. 23XJ02002), and Hunan Natural Science Foundation (Grant No. 2025JJ50396).

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study was reviewed and approved by the Ethics Committee Longgang District Maternity and Child Healthcare Hospital of Shenzhen City (LGFYKYXMLL-2024-42). The study used data from breast cancer patients at Shenzhen Longgang Maternity and Child Healthcare Hospital, and it received approval from the Ethics Committee. All data were authorized for use by Shenzhen Longgang Maternity and Child Health Hospital, and informed consent was obtained from all patients involved. Shenzhen Longgang Maternity and Child Health Hospital is a Grade 3A general hospital under the Health and Wellness Bureau of Shenzhen Longgang District, ensuring the reliability and applicability of the data.

Additional information

Correspondence and requests for materials should be addressed to L.L., J.D. or G.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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