Medical Engineering and Physics 125 (2024) 104117

Contents lists available at ScienceDirect

Medical Engineering and Physics

journal homepage: www.elsevier.com/locate/medengphy

# Benign and malignant classification of breast tumor ultrasound images using conventional radiomics and transfer learning features: A multicenter retrospective study

Ronghui Tian [a,1], Guoxiu Lu [a,b,2], Shiting Tang [c,3], Liang Sang [d,4], He Ma [a,5], Wei Qian [a,5], Wei Yang [e,*]

[a] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
[b] Department of Nuclear Medicine, General Hospital of Northern Theatre Command, Shenyang, China
[c] Department of Orthopedics, Joint Surgery and Sports Medicine, The First Hospital of China Medical University, Shenyang, China
[d] Department of Ultrasound, The First Hospital of China Medical University, Shenyang, China
[e] Department of Radiology, Cancer Hospital of China Medical University, Liaoning Cancer, Hospital & Institute, Shenyang, China

## ARTICLE INFO

*Keywords:*
Radiomics
Transfer learning
Feature fusion
Classification
Breast cancer

## ABSTRACT

This study aims to establish an effective benign and malignant classification model for breast tumor ultrasound images by using conventional radiomics and transfer learning features. We collaborated with a local hospital and collected a base dataset (Dataset A) consisting of 1050 cases of single lesion 2D ultrasound images from patients, with a total of 593 benign and 357 malignant tumor cases. The experimental approach comprises three main parts: conventional radiomics, transfer learning, and feature fusion. Furthermore, we assessed the model's generalizability by utilizing multicenter data obtained from Datasets B and C. The results from conventional radiomics indicated that the SVM classifier achieved the highest balanced accuracy of 0.791, while XGBoost obtained the highest AUC of 0.854. For transfer learning, we extracted deep features from ResNet50, Inception-v3, DenseNet121, MNASNet, and MobileNet. Among these models, MNASNet, with 640-dimensional deep features, yielded the optimal performance, with a balanced accuracy of 0.866, AUC of 0.937, sensitivity of 0.819, and specificity of 0.913. In the feature fusion phase, we trained SVM, ExtraTrees, XGBoost, and LightGBM with early fusion features and evaluated them with weighted voting. This approach achieved the highest balanced accuracy of 0.964 and AUC of 0.981. Combining conventional radiomics and transfer learning features demonstrated clear advantages over using individual features for breast tumor ultrasound image classification. This automated diagnostic model can ease patient burden and provide additional diagnostic support to radiologists. The performance of this model encourages future prospective research in this domain.

---

* Corresponding author at: Department of Radiology, Cancer Hospital of China Medical University, Liaoning Cancer Hospital & Institute, No. 44 Xiaoheyan Road, Dadong District, Shenyang 110801, Liaoning Province, China.

*E-mail address:* lbywdf@163.com (W. Yang).

[1] College of Medicine and Biological Information Engineering, Northeastern University, No. 195 Chuangxin Road, Hunnan District, Shenyang 110819, Liaoning Province, China.

[2] Department of Nuclear Medicine, General Hospital of Northern Theater Command, No. 83 Wenhua Road, Shenhe District, Shenyang 110016, Liaoning Province, China.

[3] Department of Orthopedics, Joint Surgery and Sports Medicine, The First Hospital of China Medical University, No. 155 Nanjing North Street, Heping District, Shenyang 110001, Liaoning, China.

[4] Department of Ultrasound, The First Hospital of China Medical University, No. 155 Nanjing North Street, Heping District, Shenyang 110001, Liaoning Province, China.

[5] College of Medicine and Biological Information Engineering, Northeastern University, No. 195 Chuangxin Road, Hunnan District, Shenyang 110819, Liaoning, China.

https://doi.org/10.1016/j.medengphy.2024.104117
Received 17 August 2023; Received in revised form 25 January 2024; Accepted 13 February 2024
Available online 15 February 2024
1350-4533/© 2024 IPEM. Published by Elsevier Ltd. All rights reserved.

## 1. Introduction

According to the latest global cancer burden data released by the International Agency for Research on Cancer (IARC) of the World Health Organization in 2020, female breast cancer has surpassed lung cancer as the most common cancer worldwide, accounting for approximately 11.7 % of all new cancer cases [1]. Breast cancer early screening and diagnosis represent effective approaches to enhance survival rates. Commonly employed examination methods include mammography, screening ultrasonography, and screening with magnetic resonance imaging (MRI) [2]. Research has shown that mammography significantly reduces the mortality rate of breast cancer patients, especially those with fatty breasts [3]. For dense breast women, ultrasound is needed in combination with mammography for screening [4]. MRI is mainly used for breast monitoring in high-risk women [5]. In comparison, ultrasound examination has the advantages of high cost-effectiveness, no radiation, and minimal side effects. Thus, it is widely used for early screening and diagnosis.

Previous studies have confirmed the importance and effectiveness of conventional radiomics in medical image diagnosis [6]. To improve breast cancer diagnosis, researchers have developed many computer-aided diagnosis (CAD) systems in the past [7–13]. These CAD systems can be mainly divided into segmentation, detection, and classification systems based on their application [14–16]. However, many traditional CAD systems heavily rely on handcrafted features, significantly reducing the overall performance. Moreover, the robustness of these systems (i.e., their performance in the presence of new data) still needs improvement.

With the emergence of deep learning (DL), new CAD systems have addressed these challenges. Convolutional neural networks (CNNs) have achieved excellent performance in image recognition challenges [17–21]. DL methods are based on hierarchical representation learning, transforming low-level representations into high-level representations through the combination of nonlinear modules. For complex structures and high-dimensional feature extraction, DL outperforms traditional machine learning methods [22]. Additionally, compared to traditional machine learning methods, DL requires less human intervention, reducing the complexity of feature engineering. Currently, DL has also achieved significant advancements in breast ultrasound image diagnosis [23–27]. It is worth noting that transfer learning (TL) based on CNNs can demonstrate better performance with small datasets.

Furthermore, other studies have shown that deep features extracted using TL models can also improve the performance of traditional machine learning models [28]. Undeniably, due to the complex black-box nature of TL, clinicians find it challenging to comprehend the underlying logic of its decisions [29,30]. In contrast, traditional machine learning methods with interpretable results are more likely to be accepted by clinicians. Therefore, the combination of conventional radiomics and TL features to further enhance the performance of classification models is the focus of our research. Based on the effectiveness of conventional radiomics features and the high-dimensional abstraction of TL features, we integrate these two types of features and employ traditional machine learning algorithms for classification. The addition of conventional radiomics features to the fully connected layers of CNNs is also a part of our experimental exploration.

The main content of this paper includes (1) preprocessing ultrasound images and extracting tumor regions, (2) extracting, filtering, and selecting conventional radiomics features, (3) selecting and extracting features and reducing the data dimensionality with CNNs, (4) fusing features, (5) training, validating and evaluating the classification model, and (6) conducting multicenter validation.

## 2. Materials and methods

### 2.1. Ultrasound dataset

Dataset A consists of a total of 1050 cases of single-lesion breast ultrasound images, including 593 cases of benign tumor and 357 cases of malignant tumors. It is important to note that the ultrasound images were acquired from different ultrasound diagnostic devices, namely, a GE LOGIQ E9 and a PHILIPS EPIQ5, resulting in four resolutions: 649 × 850, 775 × 580, 720 × 960, and 768 × 1024. For privacy protection, all patient information in the images was anonymized. The boundaries of tumor regions were manually delineated by an experienced radiologist and served as lesion annotations, as shown in Fig. 1. For each case, a pathological examination was performed to classify tumors into benign or malignant. To better understand the data distribution, we conducted statistical analysis on benign and malignant tumors using the Kolmogorov−Smirnov test (KS test), including first-order statistics (FOS), shape-based (S-2D), gray level cooccurrence matrix (GLCM), gray level size zone matrix (GLSZM), and gray level dependence matrix (GLDM), as shown in Table 1. Based on the selected features, the majority of benign and malignant data had p values less than 0.05, indicating significant differences in most radiomics features between benign and malignant tumors.

Dataset B [31] was obtained from a public Kaggle project. It consists of breast ultrasound images from women aged between 25 and 75 years that were collected in 2018. The dataset includes 780 images with an average size of 500 × 500 pixels in PNG format from 600 female patients. Ground truth images are provided along with the original images, and the images are categorized into three classes: normal, benign, and malignant. In the experiment, we randomly selected 138 benign images and 72 malignant images from this dataset to form Dataset B.

Dataset C was obtained from Liaoning Cancer Hospital in China. This dataset consists of 69 benign images and 36 malignant images in JPG format and provides detailed information about the types of tumors present, including fibroadenoma and lipoma in benign images and invasive ductal carcinoma, mucinous carcinoma, medullary carcinoma, and intraductal papillary carcinoma in malignant images. The tumor annotations were performed by radiologists, each with over 5 years of experience, from two tertiary hospitals. The benign and malignant labels were confirmed based on pathological results.

### 2.2. Data preprocessing

Since Dataset B and Dataset C contain complete mask images, image preprocessing was applied to Dataset A. The original data include tumor ultrasound images and lesion masks in closed curves. Using image processing algorithms, we generated mask images from the lesion annotations. By locating the masks in the entire image, we cropped the rectangular tumor regions from the ultrasound images, as shown in Fig. 2. These regions encompassed the complete tumor boundaries and internal texture.

### 2.3. Feature extraction

Conventional radiomics features from regions of interest (ROIs) have significant value in classifying benign and malignant tumors [6]. Clinically, quantified tumor features are essential for accurate diagnosis by radiologists. In this section, we utilized the PyRadiomics library [32] in Python to extract ROI features, including FOS (8), S-2D (9), GLCM (24), GLDM (14), GLRLM (16), GLSZM (16), and NGTDM (5).

The deep features of the ROIs were automatically extracted as quantitative and high-throughput features by employing a transfer CNN model as a feature encoder. These features differ from conventional radiomics features in their abstract and discrete nature. We selected ResNet50 [17], Inception-v3 [18], DenseNet121 [19], MNASNet [20], and MobileNet [21] as the pretrained models for the experiments.

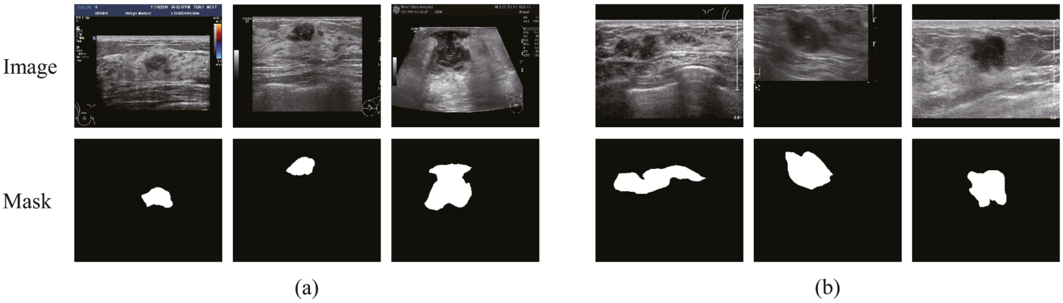


**Fig. 1.** Breast tumor ultrasound images: (a) benign images; (b) malignant images.

**Table 1**
KS test is conduct using radiomic features of benign and malignant data, and p value and statistic (D value) are calculated.

| Radiomic Features | | Statistic | *P* value |
|---|---|---|---|
| FOS | Energy | 0.1341 | 0.0003 |
| | Kurtosis | 0.0491 | 0.5979 |
| S-2D | Perimeter | 0.3922 | 6.48e-33 |
| | Maximum 2D diameter | 0.3278 | 6.10e-23 |
| GLCM | Cluster Prominence | 0.1174 | 0.0026 |
| | Difference Average | 0.2901 | 5.35e-18 |
| GLSZM | Size-Zone Non-Uniformity | 0.3320 | 1.54e-23 |
| | Zone Entropy | 0.0566 | 0.4178 |
| GLDM | Dependence Entropy | 0.0770 | 0.1142 |
| | High Gray Level Emphasis | 0.1472 | 6.27e-05 |

FOS: First-Order Statistics.
S-2D: Shape-based.
GLCM: Gray Level Co-occurrence Matrix.
GLSZM: Gray Level Size Zone Matrix.
GLDM: Gray Level Dependence Matrix.

Dataset A was randomly divided into training and testing sets (8:2). After setting appropriate hyperparameters and fine-tuning the models, we observed the dynamic testing results of each model. Upon the completion of the training, the data from Dataset A was fed into the different CNN models, and high-dimensional deep features were extracted. For ResNet50 and Inception-v3, we selected the values from the final average pooling layer (2048 features). For DenseNet121, we selected the values from the 16th convolutional layer of the fourth dense block (1568 features). For MNASNet, we selected the values from the last convolutional layer (1280 features). For MobileNet, we selected the values from the second classifier (1024 features). Thus, we obtained high-dimensional deep features from five CNN models.

### 2.4. Feature dimensionality reduction

Regarding conventional radiomics features, we concatenated the label information (0 and 1) with feature values to form the experimental dataset. The dataset was inspected for statistical information to detect any outliers. If outliers were found, all corresponding feature values for those data points were removed. All feature values were normalized to follow N ∼ (0, 1). Then, a clustering analysis was conducted based on the Pearson correlation coefficients, and features with a correlation coefficient threshold greater than 0.9 were preliminarily selected.

For TL tasks, we obtained high-dimensional tensors from CNN classification models as deep features. These deep features are more abstract and discrete compared to the features extracted conventionally. Before inputting the features into the classifier, principal component analysis (PCA) was employed for feature dimensionality reduction. After the first reduction, the dimension was reduced to half the number of original samples, and each subsequent reduction halved the dimension until an order of 100 was reached (corresponding to the size of the conventional radiomics feature dimension).

### 2.5. Feature fusion

We extracted 102 groups of conventional radiomics features using conventional imaging methods and selected 37 groups of feature values after feature weight ranking for Fusion-1. For TL, MNASNet provided 1280 groups of deep features, which were further reduced to 640 groups using PCA. Similarly, after feature screening, we retained 87 groups of

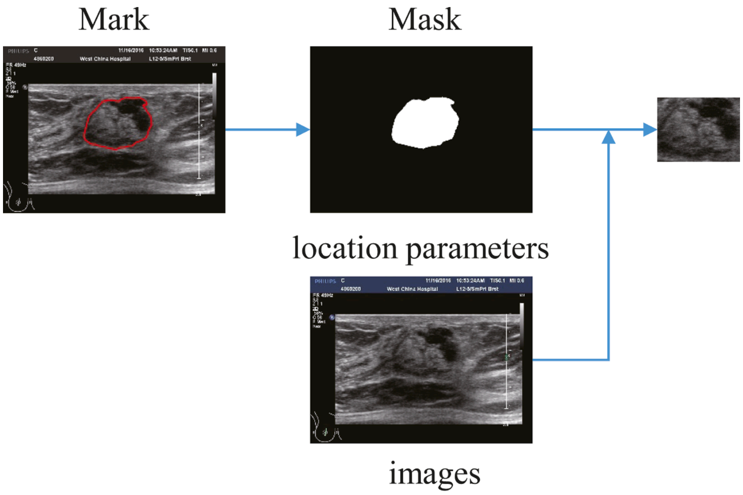


**Fig. 2.** Based on annotated ultrasound images, we employed advanced image processing algorithms to generate masked images and accurately segment the ROI encompassing the tumor. This segmentation process enabled us to precisely isolate the tumor area and extract relevant features for subsequent analysis and classification tasks.

deep features for Fusion-2. In the early fusion approach, the retained deep features and selected conventional radiomics features were concatenated into a new feature set with the same dimension. In terms of late fusion, we used two fusion strategies: ensemble and stacking. Ensemble is based on accuracy weighted voting, which uses Softmax normalization of accuracy for weighting. The higher the accuracy, the greater the weight. Stacking involves adding the results of the training and testing sets as features to the data, and then using another machine learning algorithm for classification. In addition, to compare the performance of softmax classifiers with TL, we added 37 neurons in the final fully connected layer of MNASNet, allowing for the fusion of conventional radiomics and deep features at the CNN model classification level.

### 2.6. Classification model

The complete experimental process includes conventional radiomics feature extraction, transfer learning feature extraction, feature dimensionality reduction, feature fusion, and classification model construction, as shown in Fig. 3.

MNASNet, a lightweight network proposed by Google, was used for transfer learning. It uses a new decomposed hierarchical search space, which encourages diversity in layers throughout the network while maintaining a balance between layer flexibility and search space size. Under typical mobile latency constraints, MNASNet exhibited the best accuracy for ImageNet classification and COCO object detection.

After feature extraction and selection, 37 groups of conventional radiomics features and 87 groups of deep features were retained. In the early fusion approach, we fused deep features and selected conventional radiomics features at the data level and then used an SVM classifier for classification. The fused features enriched the feature dimension of the tumor, allowing the benign and malignant classification of tumors.

### 2.7. Evaluation metrics

In the experiment, the selected evaluation metrics for the classification model included balanced accuracy, sensitivity (true positive rate, TPR), specificity (true negative rate, TNR), positive predictive value (PPV), negative predictive value (NPV), and F1-score.

$$Sensitivity \Big/ TPR = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity \Big/ TNR = \frac{FP}{FP + TN} \tag{2}$$

$$PPV = \frac{TP}{TP + FP} \tag{3}$$

$$NPV = \frac{TN}{FN + TN} \tag{4}$$

$$F1_{score} = \frac{2Precision * Recall}{Precision + Recall} \tag{5}$$

$$Balanced\_accuracy = \frac{Sensitivity + Specificity}{2} \tag{6}$$

Here, TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative. PPV represents positive predictive value, NPV represents negative predictive value. Furthermore, to address the anomaly detection problem caused by data imbalance, we calculated Balanced_ Accuracy, which helps account for class imbalance.

## 3. Results

### 3.1. Conventional radiomics classification

Previous studies have demonstrated the importance of conventional radiomics features in medical imaging diagnosis. We extracted 102 features for the benign and malignant classification of breast tumor ultrasound images. After feature selection and ranking based on feature importance, 37 features were retained. Using these features, we tested different machine learning classification algorithms [33–36], and the experimental results are shown in Table 2. The results indicate that based on the selected 37 features, the SVM classifier achieved the highest balanced accuracy of 0.791, and the XGBoost classifier achieved the highest AUC value of 0.854.

### 3.2. Transfer learning classification

We extracted deep features from five CNN models, including ResNet50, Inception-v3, DenseNet121, MNASNet, and MobileNet, for the benign and malignant classification of breast tumor ultrasound images, as shown in Table 3. The results indicate that the MNASNet model with deep features achieved the optimal performance when compressed to 640 dimensions, with a balanced accuracy of 0.866, AUC value of 0.937, sensitivity of 0.819, and specificity of 0.913. Based on these results, we selected MNASNet as the best TL model and used its extracted deep features as part of the fusion features.

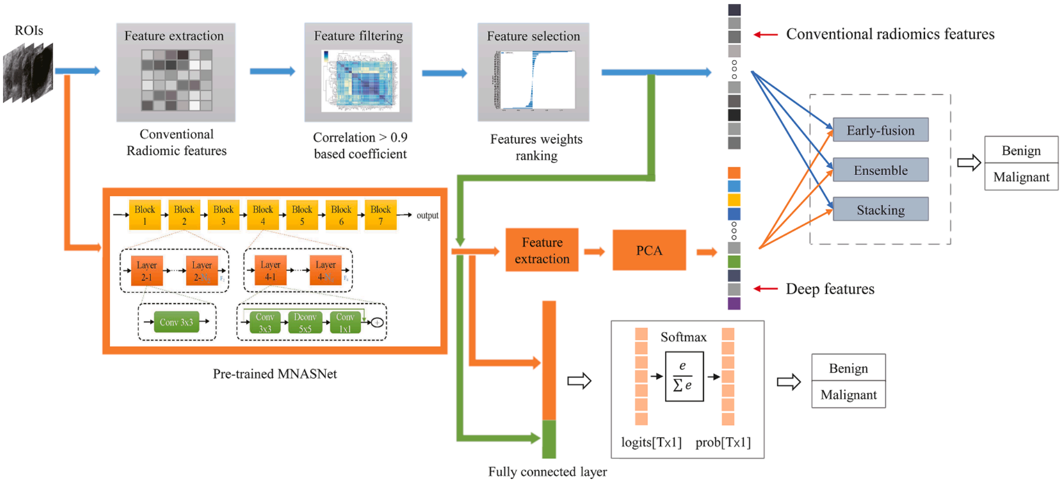


**Fig. 3.** In the classification pipeline, conventional radiomics and deep features are combined through the following tasks: extraction of conventional radiomics features, extraction of deep features through transfer learning, and feature fusion classification (early fusion, ensemble and stacking).

**Table 2**
Classification of benign and malignant tumors from breast ultrasound using conventional radiomics features under different machine learning classifiers. T means training cohort, and V means validation cohort. 95 %CI: Confidence Interval. Bal_acc means Balanced accuracy.

| Classifier | | Bal_acc | TPR | TNR | PPV | NPV | F1 | AUC | 95 %CI |
|---|---|---|---|---|---|---|---|---|---|
| KNN | T | 0.788 | 0.891 | 0.685 | 0.592 | 0.925 | 0.711 | 0.876 | 0.853–0.898 |
| | V | 0.744 | 0.805 | 0.683 | 0.563 | 0.869 | 0.662 | 0.805 | 0.744–0.866 |
| DT | T | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | Nan-Nan |
| | V | 0.778 | 0.555 | 1.000 | 0.588 | 0.774 | 0.571 | 0.676 | 0.609–0.743 |
| RF | T | 0.990 | 0.989 | 0.991 | 0.983 | 0.995 | 0.986 | 0.999 | 0.999–1.000 |
| | V | 0.761 | 0.861 | 0.661 | 0.563 | 0.900 | 0.681 | 0.830 | 0.773–0.888 |
| ET | T | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | Nan-Nan |
| | V | 0.774 | 0.722 | 0.826 | 0.684 | 0.850 | 0.702 | 0.817 | 0.755–0.879 |
| XGB | T | 0.949 | 0.979 | 0.919 | 0.861 | 0.988 | 0.916 | 0.990 | 0.985–0.994 |
| | V | 0.796 | 0.861 | 0.731 | 0.626 | 0.909 | 0.725 | 0.854 | 0.801–0.907 |
| LGB | T | 0.871 | 0.874 | 0.868 | 0.773 | 0.931 | 0.820 | 0.951 | 0.938–0.964 |
| | V | 0.775 | 0.819 | 0.731 | 0.614 | 0.885 | 0.702 | 0.832 | 0.774–0.889 |
| SVM | T | 0.813 | 0.796 | 0.829 | 0.705 | 0.888 | 0.748 | 0.876 | 0.851–0.901 |
| | V | 0.791 | 0.763 | 0.818 | 0.687 | 0.869 | 0.723 | 0.843 | 0.788–0.898 |

**Table 3**
Classification of benign and malignant tumors from breast ultrasound using transfer learning features under the optimal machine learning classification algorithm and features dimension.

| Model/Dim | | Bal_acc | TPR | TNR | PPV | NPV | F1 | AUC | 95 %CI |
|---|---|---|---|---|---|---|---|---|---|
| ResNet50 / 256 | T | 0.971 | 0.980 | 0.962 | 0.930 | 0.989 | 0.954 | 0.992 | 0.988–0.997 |
| | V | 0.850 | 0.873 | 0.827 | 0.720 | 0.927 | 0.790 | 0.912 | 0.872–0.953 |
| InceptionV3 / 512 | T | 0.980 | 0.982 | 0.978 | 0.959 | 0.991 | 0.971 | 0.997 | 0.994–0.999 |
| | V | 0.867 | 0.944 | 0.789 | 0.701 | 0.964 | 0.805 | 0.927 | 0.894–0.960 |
| Densenet121 / 392 | T | 0.981 | 0.989 | 0.980 | 0.962 | 0.995 | 0.976 | 0.998 | 0.997–0.999 |
| | V | 0.840 | 0.847 | 0.833 | 0.726 | 0.912 | 0.782 | 0.867 | 0.813–0.922 |
| MNASNet / 640 | T | 0.975 | 0.996 | 0.953 | 0.916 | 0.998 | 0.955 | 0.996 | 0.993–0.999 |
| | V | 0.866 | 0.819 | 0.913 | 0.830 | 0.906 | 0.825 | 0.937 | 0.906–0.968 |
| MobileNet / 256 | T | 0.982 | 0.982 | 0.981 | 0.965 | 0.991 | 0.973 | 0.996 | 0.994–0.999 |
| | V | 0.862 | 0.847 | 0.876 | 0.782 | 0.916 | 0.813 | 0.905 | 0.862–0.948 |

### *3.3. Feature fusion classification*

For the early fusion features, we randomly selected 840 images as the training set and 210 images as the test set based on the benign and malignant ratio of the data. The experiments used the SVM algorithm for classification, and this model was named MNASNet + Rad + SVM.

When selecting MNASNet as the TL model, we added conventional radiomics features to the classification model. The merged model was named MNASNet + Rad + Softmax. The conventional radiomics features were added to the last fully connected layer of MNASNet by increasing the number of neurons to match the number of conventional radiomics features.

In terms of late fusion, we adopted the ensemble and stacking fusion strategies. We compared the performance of the four fusion strategies with the test set, as shown in Table 4. The experiment demonstrated that the MNASNet + Rad + Ensemble model achieved the highest balanced accuracy of 0.964 and an AUC value of 0.981. The ROC curves of the four feature fusion classification models are shown in Fig. 4(a). We present the confusion matrices and decision curve analysis (DCA) for the four classification models with the test cohort in Fig. 5.

### *3.4. Interpretability of the transfer learning model*

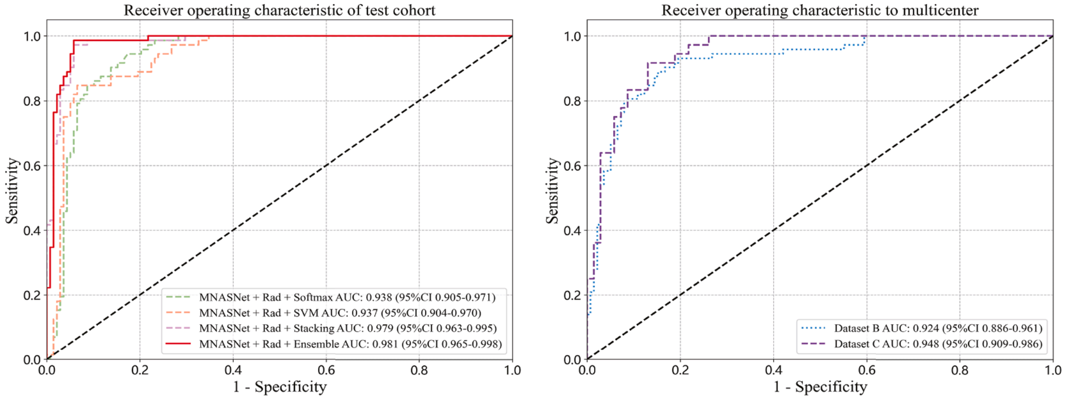
The deep features extracted by the TL model play an essential role in the benign and malignant classification of tumors. We attempted to visualize specific convolutional layers to interpret the effectiveness of the CNN model. Gradient-weighted class activation mapping (Grad-CAM) [37] is capable of explaining the decision-making process of TL models by generating heatmaps to locate the tumor region that the model focuses on. The last convolutional layer of the MNASNet model, with a depth coefficient of 1.0, is chosen to generate the Grad-CAM heatmap. The 1280-dimensional deep features of this layer were considered important fusion features and had significant reference value. As shown in Fig. 6, we found that the MNASNet model focuses on the interior and boundaries of the tumor in the ROI. Moreover, it is also a significant reason why the fusion of conventional radiomics features with MNASNet deep features can enhance the classification performance.

### *3.5. Multicenter validation*

To validate the model's performance across different data centers, we collected data from two groups of different institutions. With the Kaggle public dataset of 210 cases (Dataset B: 138 benign and 72 malignant cases) and local hospital dataset of 105 cases (Dataset C: 69 benign and 36 malignant cases) as the test set, we evaluated the generalizability of the best model. The test results showed that with Dataset B, the model achieved a balanced accuracy of 0.859 and an AUC value of 0.924 (95 % CI: 0.893–0.961), while with Dataset C, the model achieved a balanced accuracy and AUC value of 0.875 and 0.948 (95 %

**Table 4**
Evaluation results of four classification models.

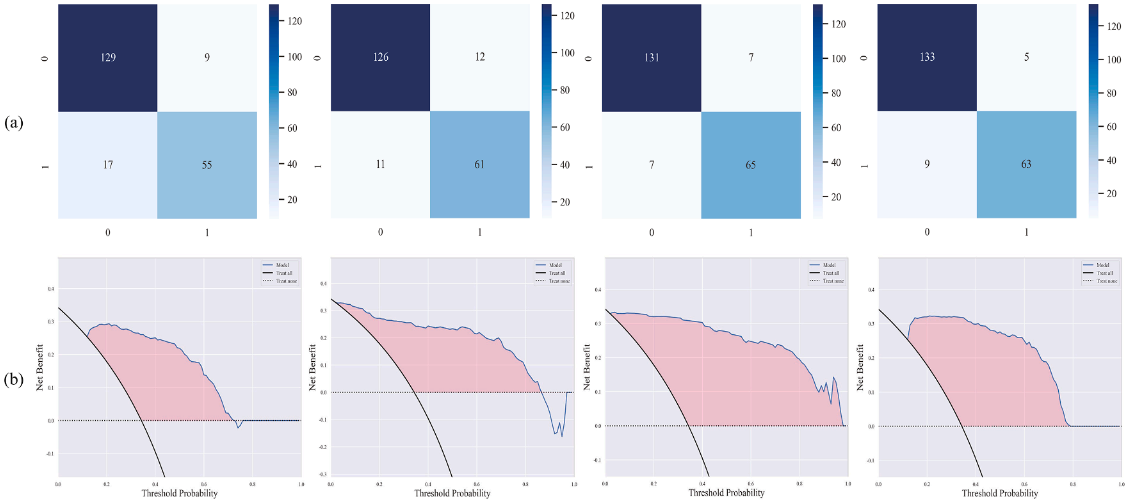
| Model_name | Bal_acc | AUC (95 %CI) | Sensitivity (95 %CI) | Specificity (95 %CI) | PPV (95 %CI) | NPV (95 %CI) | F1score |
|---|---|---|---|---|---|---|---|
| MNASNet+Rad+Softmax | 0.888 | 0.938[0.905–0.971] | 0.944[0.891–0.997] | 0.832[0.762–0.889] | 0.739[0.649–0.828] | 0.966[0.933–0.998] | 0.829 |
| MNASNet+Rad+SVM | 0.891 | 0.937[0.904–0.970] | 0.847[0.764–0.930] | 0.934[0.893–0.975] | 0.871[0.793–0.949] | 0.921[0.876–0.966] | 0.859 |
| MNASNet+Rad+ Stacking | 0.957 | 0.979[0.963–0.995] | 0.972[0.934–1.000] | 0.942[0.903–0.981] | 0.897[0.830–0.964] | 0.985[0.964–1.000] | 0.933 |
| MNASNet+Rad+Ensemble | 0.964 | 0.981[0.965–0.998] | 0.986[0.959–1.000] | 0.942[0.903–0.981] | 0.899[0.832–0.965] | 0.992[0.977–1.000] | 0.940 |



(a)

(b)

**Fig. 4.** (a). ROC curves of four classification models. Under the ensemble fusion strategy, the model with the combination of deep features from the MNASNet model and conventional radiomic features achieved the highest AUC value. The experiments demonstrated that the MNASNet+Rad+Ensemble classifier had the best performance. (b) ROC curve analysis across two independent data centers. Our model exhibited decreased discriminative ability with Datasets B and C compared to that of Dataset A, as evidenced by lower AUC values.



MNASNet+Rad+Softmax MNASNet+Rad+SVM MNASNet+Rad+ Stacking MNASNet+Rad+Ensemble

**Fig. 5.** Evaluation of four radiomics-based classification models with the independent test set. The MNASNet+Radiomics+Ensemble fusion model exhibited superior discriminative performance. (a) The confusion matrix, summarizing correct and incorrect predictions for each class, and (b) DCA, quantifying the clinical utility over a range of threshold probabilities. Together, these metrics demonstrate that the MNASNet+Radiomics+Ensemble model achieves the optimal trade-off between discrimination ability, generalizability, and clinical usefulness for our application.

CI: 0.875–0.988), respectively. Table 5 shows the results of the test sets. The ROC curves for different data centers are shown in Fig. 4(b). In this experiment, the predicted values of each sample and their corresponding true results were plotted, as shown in Fig. 7, and the legend indicates the true labels of the categories.

## 4. Discussion

Based on conventional radiomics features, we modeled Dataset A using seven machine learning algorithms. Regarding data allocation, the proportion of training and testing set was 8:2. In terms of balanced accuracy, the SVM classifier performed the best overall. For sensitivity and specificity, the XGBoost classifier achieved relatively good scores. Regarding PPV and NPV, XGBoost and SVM obtained the optimal scores. For the AUC values, RF, XGBoost, LGB, and SVM achieved relatively high scores. Additionally, SVM performed the best in terms of the F1-score. Overall, SVM demonstrated better overall performance in classifying conventional radiomics features.

Extracting deep features from the TL model is different from extracting conventional radiomics features. First, we conducted TL tasks using Dataset A. After model training, we saved the model locally. Next, we visualized each layer of the model, selecting the last pooling layer for deep feature extraction and saving it. Since the number of deep features far exceeds that of conventional radiomics features, we used the PCA-reduced deep features for classification. The experimental results showed that different TL models had varying evaluation results with different dimensions. When using the SVM classifier, MNASNet with 640-dimensional deep features yielded the best result, achieving a balanced accuracy of 0.866 and an AUC value of 0.937 (95 % CI: 0.906–0.968). Inception-v3 performed best in terms of sensitivity and specificity. Furthermore, MNASNet also achieved the best performance in terms of PPV and NPV. In conclusion, MNASNet with extracted deep features showed more robust performance in the classification task.

Fusing conventional radiomics features into the TL model was another experimental strategy. We added a fully connected layer with the number of neurons equal to the number of selected conventional radiomics features after the MNASNet pooling layer. During the model training, conventional radiomics features and deep features were fused through softmax for prediction. The test results indicated that compared to using only conventional radiomics or deep features of TL, the TL

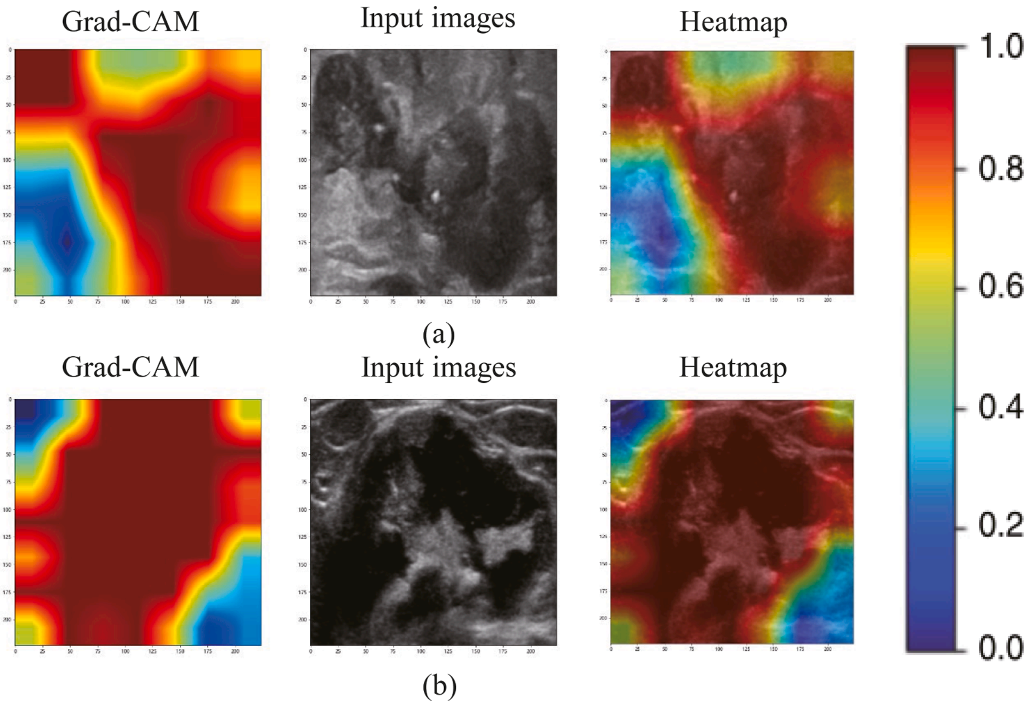


**Fig. 6.** Heatmaps visualizing model predictions for benign and malignant tumors using ultrasound images via gradient-weighted class activation mapping (Grad-CAM). Representative examples are provided for (a) a benign tumor, correctly classified by the model through emphasis on the oval, lobulated shape and circumscribed margins indicative of a benign morphology; and (b) a malignant tumor, accurately classified by the model through the identification of an irregular shape, spiculated margins, and an echogenic halo, which are suggestive of a malignant ultrasound phenotype.

**Table 5**
Performance of models in different data centers.

| Dataset | Bal_acc | AUC (95 %CI) |
|---|---|---|
| Dataset B (Public data from the Kaggle) | 0.859 | 0.924 (0.893–0.961) |
| Dataset C (Local data from cooperative hospital) | 0.875 | 0.948 (0.909–0.986) |

model with fused conventional radiomics features improved the prediction results, with a balanced accuracy of 0.888 and an AUC value of 0.938.

The optimal classification model in this study was obtained with the late fusion strategy of conventional radiomics features and MNASNet deep features. We used SVM, ExtraTrees, XGBoost, and LightGBM as four machine learning algorithms for task learning and performed ensemble fusion for all algorithms. From the results, the fusion of conventional radiomics features and MNASNet deep features in the classifier weighted voting strategy surpassed the former in both balanced accuracy and AUC values. The experimental results showed the excellent performance of the model in benign and malignant classification.

To test and evaluate the model's performance and generalizability across data from different data centers, we used Dataset B and Dataset C. Dataset B was from the Kaggle competition data, representing public datasets, while Dataset C was from a local cooperative hospital, representing private clinical data. It should be noted that Dataset A was collected according to our review criteria and differed from datasets B and C. From the validation results, the evaluation indicators of the model with Datasets B and C were lower than those with the Dataset A test set. This could be due to differences in data collection devices and parameters. This highlights a center dependence of the model, with performance declining with external test sets from different institutions relative to the initial institution. Further optimization of the model generalizability across diverse centers representing real-world heterogeneity is warranted. Nevertheless, for the initial diagnosis and screening of breast tumors, our classification model still showed promising prospects. It should be noted that although different from models using other databases, we are using proprietary databases [38–40]. Relatively speaking, our optimal model achieved higher values in AUC and sensitivity. From the perspective of evaluation indicators, our model has higher balanced accuracy in detecting positive cases.

This study still has some limitations. The training of the model using Dataset A, which was obtained from a single center, somewhat weakened the robustness of the model. Our next goal is to adopt multicenter data to establish a benign and malignant classification model. Additionally, features were only extracted from each tumor's ROI region, i.e., the intratumor area, while the influence of features at different distances from the tumor margin on the classification model was ignored. In clinical diagnosis, changes in the boundary region between lesions and surrounding tissues, as well as the echogenic features of tissues behind the tumor, are also important considerations. Next, we will study the combined effects of intratumor and peri-tumor features on the classification model. Moreover, based on the experimental data, we will further collect patient diagnostic reports, design key information points, extract and quantify text features, and enrich the sources of extracted features. Additionally, we will upgrade the experimental plan, particularly in the area of TL. Currently, our team is conducting experiments on transformer-based classification tasks, providing a direction for our future research.

## 5. Conclusion

The establishment of a breast ultrasound tumor benign and malignant classification model based on both conventional radiomics and deep features outperforms single-feature models. Moreover, our research presents an automated diagnostic model that can alleviate the burden on patients and provide additional diagnostic means for radiologists. Given the performance in this study, we expect to achieve significant results in prospective studies.

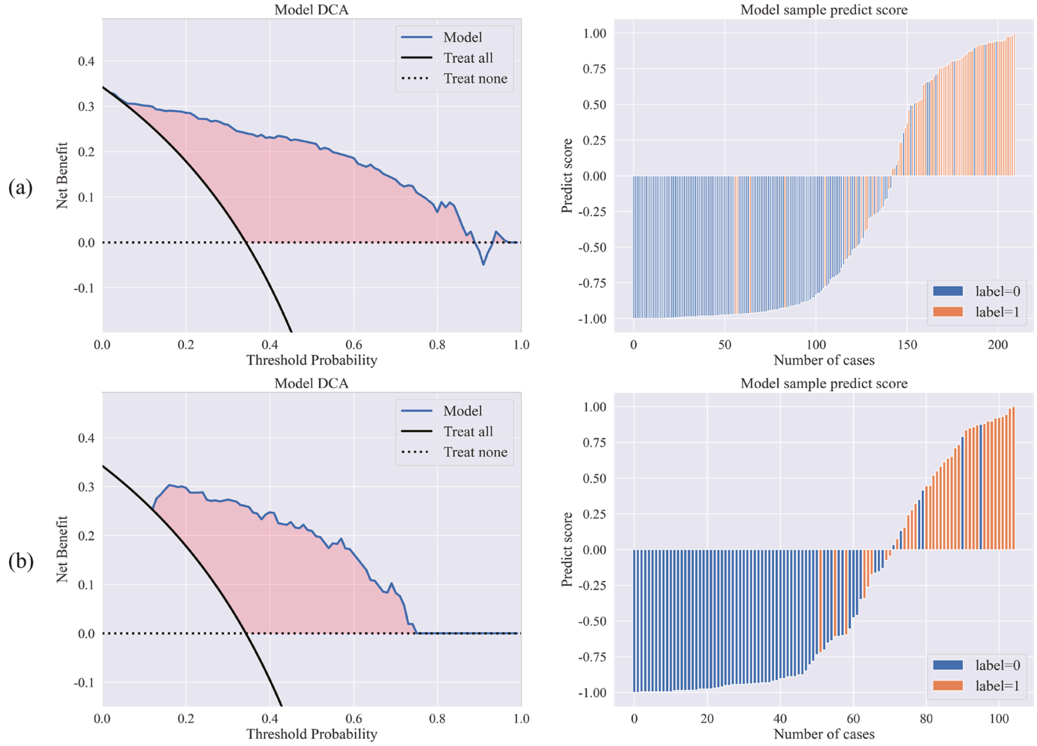


**Fig. 7.** Model predictions and ground truth labels for individual samples across multiple independent data centers, with DCA to assess clinical utility. Shown are predictions with (a) Dataset B, exhibiting a downward shift in predictions for the malignant class, reducing discriminative ability and (b) Dataset C, with improved calibration but decreased sensitivity compared to those with Dataset B, indicating center-dependent performance. Ongoing model optimization and validation across diverse real-world data are required to improve the generalizability of the model.

## Funding

This work was supported by the Liaoning Natural Science Foundation 2022-YGJC-52.

## Ethical approval

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Northeastern University, Shenyang, Liaoning, China (Date: September 23, 2021/No. NEU-EC-2021B019S).

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We acknowledge the artificial intelligence model Claude 2 from Anthropic for assistance with English grammar and spelling in academic writing.

## References

[1] The International Agency for Research on Cancer, World Health Organization. Latest global cancer data: cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020. 2020.

[2] Warner E. Breast-cancer screening. N Engl J Med 2011;365:1025–57.

[3] Løberg M, Lousdal ML, Bretthauer M, Kalager M. Benefits and harms of mammography screening. Breast Cancer Res 2015;17. https://doi.org/10.1186/s13058-015-0525-z.

[4] Brem RF, Lenihan MJ, Lieberman J, Torrente J. Screening breast ultrasound: past, present, and future. Am J Roentgenol 2015;204:234–40. https://doi.org/10.2214/AJR.13.12072.

[5] Mann RM, Kuhl CK, Moy L. Contrast-enhanced MRI for breast cancer screening. J Magn Resonance Imaging 2019;50:377–90. https://doi.org/10.1002/jmri.26654.

[6] Conti A, Duggento A, Indovina I, Guerrisi M, Toschi N. Radiomics in breast cancer classification and prediction. Semin Cancer Biol 2021;72:238–50. https://doi.org/10.1016/j.semcancer.2020.04.002.

[7] Jalalian A, Mashohor SBT, Mahmud HR, Saripan MIB, Ramli ARB, Karasfi B. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. Clin Imaging 2013;37:420–6. https://doi.org/10.1016/j.clinimag.2012.09.024.

[8] Kaisar Alam S, Feleppa EJ, Rondeau M, Kalisz A, Garra BS. Ultrasonic multi-feature analysis procedure for computer-aided diagnosis of solid breast lesions. Ultrason Imaging 2011;33:17–38. https://doi.org/10.1177/016173461103300102.

[9] Muhtadi S, Chowdhury A, Razzaque RR, Shafiullah A. Analyzing the texture of nakagami parametric images for classification of breast cancer. In: 2021 IEEE National Biomedical Engineering Conference (NBEC). Institute of Electrical and Electronics Engineers Inc.; 2021. p. 100–5. https://doi.org/10.1109/NBEC53282.2021.9618762.

[10] Chowdhury A, Razzaque RR, Muhtadi S, Shafiullah A, Ul Islam Abir E, Garra BS, et al. Ultrasound classification of breast masses using a comprehensive Nakagami imaging and machine learning framework. Ultrasonics 2022;124. https://doi.org/10.1016/j.ultras.2022.106744.

[11] Hu J, Heidari AA, Zhang L, Xue X, Gui W, Chen H, et al. Chaotic diffusion-limited aggregation enhanced grey wolf optimizer: insights, analysis, binarization, and feature selection. Int J Intell Syst 2022;37:4864–927. https://doi.org/10.1002/int.22744.

[12] Ru J, Lu B, Chen B, Shi J, Chen G, Wang M, et al. Attention guided neural ODE network for breast tumor segmentation in medical images. Comput Biol Med 2023;159. https://doi.org/10.1016/j.compbiomed.2023.106884.

[13] Liu L, Zhao D, Yu F, Heidari AA, Ru J, Chen H, et al. Performance optimization of differential evolution with slime mould algorithm for multilevel breast cancer image segmentation. Comput Biol Med 2021;138. https://doi.org/10.1016/j.compbiomed.2021.104910.

[14] Xu Y, Wang Y, Yuan J, Cheng Q, Wang X, Carson PL. Medical breast ultrasound image segmentation by machine learning. Ultrasonics 2019;91:1–9. https://doi.org/10.1016/j.ultras.2018.07.006.

[15] Pavithra S, Vanithamani R, Justin J. Computer aided breast cancer detection using ultrasound images. Mater Today Proc 2020;33:4802–7. https://doi.org/10.1016/j.matpr.2020.08.381. Elsevier Ltd.

[16] Moon WK, Chen IL, Chang JM, Shin SU, Lo CM, Chang RF. The adaptive computer-aided diagnosis system based on tumor sizes for the classification of breast tumors detected at screening ultrasound. Ultrasonics 2017;76:70–7. https://doi.org/10.1016/j.ultras.2016.12.017.

[17] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 770–8. https://doi.org/10.1109/CVPR.2016.90.

[18] Szegedy C, Vanhoucke V, Ioffe S, Shlens J. Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition; 2016.

[19] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017.

[20] Tan M, Chen B, Pang R, Vasudevan V, Sandler M, Howard A, et al. MnasNet: platform-aware neural architecture search for mobile. In: IEEE Conference on Computer Vision and Pattern Recognition; 2019. p. 2820–8.

[21] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: efficient convolutional neural networks for mobile Vision Applications. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017.

[22] Vial A, Stirling D, Field M, Ros M, Ritz C, Carolan M, et al. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. Transl Cancer Res 2018;7:803–16. https://doi.org/10.21037/tcr.2018.05.02.

[23] Wang Y, Choi EJ, Choi Y, Zhang H, Jin GY, Ko SB. Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning. Ultrasound Med Biol 2020;46:1119–32. https://doi.org/10.1016/j.ultrasmedbio.2020.01.001.

[24] Ma H, Tian R, Li H, Sun H, Lu G, Liu R, et al. Fus2Net: a novel Convolutional Neural Network for classification of benign and malignant breast tumor in ultrasound images. Biomed Eng Online 2021;20. https://doi.org/10.1186/s12938-021-00950-z.

[25] Han S, Kang HK, Jeong JY, Park MH, Kim W, Bang WC, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol 2017;62:7714–28. https://doi.org/10.1088/1361-6560/aa82ec.

[26] Byra M, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, et al. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Med Phys 2019;46:746–55. https://doi.org/10.1002/mp.13361.

[27] Muhtadi S, Haque ST, Gallippi CM. Combined B-mode and Nakagami images for improved discrimination of breast masses using deep learning. In: IEEE International Ultrasonics Symposium. IEEE Computer Society; 2022. https://doi.org/10.1109/IUS54386.2022.9957624. vol. 2022- October.

[28] Zheng X, Yao Z, Huang Y, Yu Y, Wang Y, Liu Y, et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nat Commun 2020;11. https://doi.org/10.1038/s41467-020-15027-z.

[29] Kim M, Yun J, Cho Y, Shin K, Jang R, Bae HJ, et al. Deep learning in medical imaging. Neurospine 2019;16:657–68. https://doi.org/10.14245/ns.1938396.198.

[30] Tanaka H, Chiu SW, Watanabe T, Kaoku S, Yamaguchi T. Computer-aided diagnosis system for breast ultrasound images using deep learning. Phys Med Biol 2019;64. https://doi.org/10.1088/1361-6560/ab5093.

[31] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief 2020;28. https://doi.org/10.1016/j.dib.2019.104863.

[32] Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77:e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.

[33] Pisner DA, Schnyer DM. Support vector machine. Machine learning: methods and applications to brain disorders. Elsevier; 2019. p. 101–21. https://doi.org/10.1016/B978-0-12-815739-8.00006-7.

[34] Alfian G, Syafrudin M, Fahrurrozi I, Fitriyani NL, Atmaji FTD, Widodo T, et al. Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. Computers 2022;11. https://doi.org/10.3390/computers11090136.

[35] Li M, Fu X, Li D. Diabetes prediction based on XGBoost algorithm. In: IOP Conf Ser Mater Sci Eng. 768. Institute of Physics Publishing; 2020. https://doi.org/10.1088/1757-899X/768/7/072093.

[36] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. In: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31st Conference on Neural Information Processing Systems; 2017.

[37] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vision 2020;128:336–59. https://doi.org/10.1007/s11263-019-01228-7.

[38] Fang Z, Zhang W, Ma H. Breast cancer classification with ultrasound images based on SLIC. Frontier computing: theory, technologies and applications (FC 2019) 8. 2020. p. 235–48.

[39] Mishra AK, Roy P, Bandyopadhyay S, Das SK. Breast ultrasound tumour classification: a machine learning—radiomics based approach. Expert Syst 2021;38. https://doi.org/10.1111/exsy.12713.

[40] Zeimarani B, Costa MGF, Nurani NZ, Bianco SR, De Albuquerque Pereira WC, Filho CFFC. Breast lesion classification in ultrasound images using deep convolutional neural network. IEEE Access 2020;8:133349–59. https://doi.org/10.1109/ACCESS.2020.3010863.