CSUM 2022
中华医学会第二十二次全国超声医学学术会议
The 22nd Annual Meeting of Chinese Society of Ultrasound in Medicine
2022年11月3-6日
四川 · 成都
主办单位：中华医学会 中华医学会超声医学分会
承办单位：四川省医学会
协办单位：四川大学华西医院

Deep learning based classification of breast lesions with ultrasound images: a multicenter study

Yang Gu [1] Ting Liu [2] Xing An [2] Hongyan Wang [1] Yuxin Jiang [1]

1. Department of Ultrasound, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College

2. Department of Medical Imaging Advanced Research, Beijing Research Institute, Shenzhen Mindray Bio-Medical Electronics Co., Ltd.

Purpose

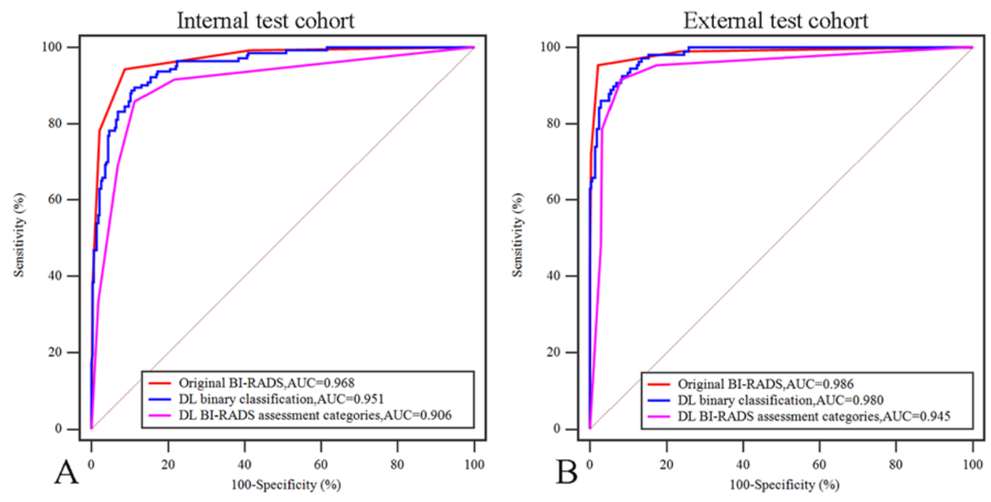
To establish a breast lesion risk stratification system using ultrasound images to predict breast malignancy and assess Breast Imaging Reporting and Data System (BI-RADS) categories simultaneously.

Methods

This multicenter study prospectively collected a dataset of ultrasound images in 5012 patients in thirty-two hospitals from December 2018 to December 2020. A large dataset with pathology information and BI-RADS categories interpreted by board-certified radiologists from multiple hospitals was utilized to train our deep learning (DL) model. A DL model was developed to conduct binary categorization (benign and malignant) and BI-RADS categories (2, 3, 4a, 4b, 4c and 5) simultaneously. The training set of 4212 patients and the internal test set (ITS) of 416 patients were from thirty hospitals. The remaining two hospitals of 384 patients were used as an external test set (ETS). We measured agreement between the DL binary classification and the pathological finding and between the DL BI-RADS categories and the radiologist's assessment across the six BI-RADS categories. The receiver operating characteristic analyses were performed to calculate the area under the receiver operating characteristic curve (AUC) to assess the diagnostic performance of the DL binary categorization, DL six-way categorizations and the radiologists based on the probabilities of malignancy, DL BI-RADS categories and radiologist's BI-RADS categories, respectively. The following evaluation metrics were used to estimate the classification performance of the DL model based on the confusion matrices: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy.

Results

For DL binary categorization of benign and malignant lesions, the DL model showed substantial agreement (κ=0.759) with the pathology results in the ITS and almost perfect agreement (κ=0.823) in the ETS. For the DL BI-RADS categories, the DL model showed moderate agreement with the radiologists in the ITS and ETS (κ=0.626 and κ=0.669, respectively). In the ETS, for the DL binary categorization, the DL model achieved an AUC of 0.980, which was comparable to that of the radiologists (0.986, P=0.4203). The DL model achieved superior specificity (93.84% vs. 76.45%, P<0.0001), PPV (85.09% vs. 62.21%, P<0.0001) and accuracy (92.71% vs. 82.81%, P<0.0001) for diagnosing breast lesions compared with the radiologists, and inferior sensitivity (89.81% vs. 99.07%, P=0.0020) and NPV (95.93% vs. 99.53%, P=0.0118) compared with the radiologists. For the DL six-way categorizations, the AUC was inferior to that of the radiologists (0.945 vs. 0.986, P=0.0009). The DL model was superior in specificity (82.61% vs. 76.45%, P=0.0464) to that of the radiologists and was similar in sensitivity (95.37% vs 99.07%, P=0.1250), PPV (68.21% vs. 62.21%, P=0.2598), NPV (97.85% vs. 99.53%, P=0.1266) and accuracy (86.20% vs. 82.81%, P=0.1480) to that of the radiologists.



AUCs for the diagnostic performance of the DL binary classification and BI-RADS assessment categories and the original radiologists in the ITC (A) and ETC (B).
AUCs, areas under the receiver operating characteristic curve; DL, deep learning; BI-RADS, Breast Imaging Reporting and Data System; ITC, internal test cohort; ETC, external test cohort.

Conclusions

The DL model performed well in distinguishing benign from malignant breast lesions and yielded similar outcomes to the radiologists. This indicates the potential applicability of the DL model in clinical diagnosis.