The Establishment of Transformer-Based Computer-Aided Diagnosis Model to Improve the Classification Consistency of BI-RADS-US3–5 Nodules Among Radiologists: A Multi-Center Study

Hongtao Ji  
Beijing Tongren Hospital, Capital Medical University

Qiang Zhu (✉ qzhu@263.net)  
Beijing Tongren Hospital, Capital Medical University

Teng Ma  
Beijing Tongren Hospital, Capital Medical University

Yun Cheng  
Beijing Tongren Hospital, Capital Medical University

Shuai Zhou  
Beijing Tongren Hospital, Capital Medical University

Wei Ren  
Beijing Tongren Hospital, Capital Medical University

Huilian Huang  
Beijing Tongren Hospital, Capital Medical University

Wen He  
Beijing Tiantan Hospital, Capital Medical University

Haitao Ran  
Second Affiliated Hospital of Chongqing Medical University

Litao Ruan  
First Affiliated Hospital of Xi’an Jiaotong University

Yanli Guo  
Southwest Hospital

Jiawei Tian  
Second Affiliated Hospital of Harbin Medical University

Wu Chen  
First Hospital of Shanxi Medical University

Luzeng Chen  
Peking University First Hospital

Zhiyuan Wang  
Hunan Cancer Hospital

Qi Zhou  
Second Affiliated Hospital of Xi’an Jiaotong University
Research Article

Keywords: Breast Imaging Reporting and Data Systems (BI-RADS), Computer-Aided Diagnosis (CAD), Ultrasound, Transformers

Posted Date: December 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1179922/v1

License: ©  This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Background: Significant differences exist in classification outcomes for radiologists using ultrasonography-based breast imaging-reporting and data systems for diagnosing category 3–5 (BI-RADS-US 3–5) breast nodules, due to a lack of clear and distinguishing image features. As such, this study investigates the use of a transformer-based computer-aided diagnosis (CAD) model for improved BI-RADS-US 3–5 classification consistency.

Methods: Five radiologists independently performed BI-RADS-US annotations on a breast ultrasonography image set collected from 20 hospitals in China. The data were divided into training, validation, testing, and sampling sets. The trained transformer-based CAD model was then used to classify test images, for which sensitivity, specificity, and accuracy were calculated. Variations in these metrics among the 5 radiologists were analyzed by referencing BI-RADS-US classification results for the sampling test set, provided by CAD, to determine whether classification consistency (the kappa value) sensitivity, specificity, and accuracy had improved.

Results: Classification accuracy for the CAD model applied to the test set was 95.7% for category 3 nodules, 97.6% for category 4A nodules, 95.60% for category 4B nodules, 94.2% for category 4C nodules, and 97.5% for category 5 nodules. Adjustments were made to 1,583 nodules, as 905 were classified to a higher category and 678 to a lower category in the sampling test set. As a result, the accuracy, sensitivity, and specificity of classification by each radiologist improved, with the consistency (kappa values) for all radiologists increasing to >0.60.

Conclusions: The proposed transformer-based CAD model improved BI-RADS-US 3–5 nodule classification by individual radiologists and increased diagnostic consistency.

Background

Breast cancer exhibits the highest incidence among malignancies in women and its early diagnosis and treatment significantly reduces mortality rates.[1, 2] Among several commonly used breast examination techniques, ultrasonography is the most convenient and most economical modality with no radiation and relatively low cost. However, the quality of ultrasonography directly depends on operator expertise and experience, especially on her/his scanning techniques, ability of lesion detection, and description and interpretation of images.[3]. The emergence of breast imaging reporting and data systems on ultrasonography (BI-RADS-US) is an attempt to normalize and standardize the terminology used to describe a series of appearances in ultrasound images and then classify breast nodules from category 1 through category 6 depending on the probability of malignancy[4]. However, this probability for category 4 nodules varies widely (2–95%) and specific classification criteria for subcategories 4a, 4b, and 4c lack clear definitions. In addition, there is a lack of clear criteria of the classification in distinguishing categories 4a and 3 or 4c and 5 breast nodules.[5, 6] Hence, BI-RADS-US classifications of category 3–5 nodules differ significantly between hospitals and individual radiologists. As a result, a given breast nodule may be over- or under-treated in response to a diagnosis. For example, a misclassification of benign nodules into category 4 or above increases psychological burden and medical expenses for a patient, while a misclassification of malignant nodules into category 3 can cause life-threatening delays in treatment.
Computer-aided diagnosis (CAD) models can bypass conventional subjective diagnoses by humans. In recent years, the expanded availability of breast imaging datasets has facilitated end-to-end deep learning, thereby achieving objective diagnosis of breast nodules. While CAD models can be a highly effective aid for assisting radiologists in diagnosing disease, their performance is closely related to the size of the training set, as larger sets comprised of higher quality images produce better diagnostic outcomes. While acquiring annotated images marked by experienced radiologists can be difficult, open databases have allowed the application of machine learning in a variety of fields. For example, convolutional neural networks (CNNs) have been applied to the segmentation of ultrasound images,[7] the diagnosis of benign and malignant breast nodules,[6, 8, 9] and BI-RADS-US classification,[10, 11] achieving satisfactory results. However, the self-attention-based mechanism of a transformer has outperformed conventional CNNs for visual tasks.[12] Transformers were originally developed for machine learning in natural language processing,[13] but have since been applied to medical imaging research.[14, 15] It has also been suggested that transformers focus more on shape recognition and exhibit higher computational efficacy and scalability than CNNs for texture recognition. As such, this study utilized a transformer to establish a CAD model for the classification of BI-RADS-US 3–5 nodules, thereby providing classification references for radiologists in an attempt to improve the diagnostic level and consistency.

Methods

A total of 21,332 images of 5,057 breast nodules were collected from 3,978 female patients from December 2006 to December 2019. The age of the patients ranged from 12 to 95 years old, with a mean age of 47.81 ± 14.55 years. All nodules were surgically removed from the patients and pathological results were made available from 20 hospitals in China, including Beijing Tongren Hospital affiliated with Capital Medical University, the Second Affiliated Hospital of Chongqing Medical University, the First Affiliated Hospital of Xi’an Jiaotong University, Southwest Hospital of Army Medical University, the Second Affiliated Hospital of Harbin Medical University, the First Hospital of Shanxi Medical University, Beijing Tiantan Hospital of Capital Medical University, Peking University First Hospital, Hunan Cancer Hospital, the Second Affiliated Hospital of Xi’an Jiaotong University, the Cancer Hospital at the Chinese Academy of Medical Sciences, the Third Affiliated Hospital of Guangxi Medical University, the First Affiliated Hospital of Hebei North University, Sichuan Provincial People’s Hospital, the First Affiliated Hospital of Zhengzhou University, China-Japan Union Hospital of Jilin University, Qilu Hospital of Shandong University, the Second Xiangya Hospital of Central South University, the Second Hospital of Lanzhou University, and the First Affiliated Hospital of Nanchang University. This retrospective clinical study only involved the collection of age data, breast nodule images, imaging system models, and pathological results for patients. It did not interfere with individual treatment plans and an exemption from informed patient consent was approved by the hospital ethics committee.

Inclusion criteria were satisfied by images (1) produced using a high-frequency probe (≥12 MHz), (2) containing only one nodule, and (3) exhibiting nodules with identifiable boundaries. Exclusion criteria were applied to images exhibiting (1) no nodules, (2) clear cysts, (3) more than one nodule, (4) nodules too large to display a complete outline, and (5) poor quality or unclear nodule boundaries. The open-source “cornerstonejs” and “cornerstoneTools” JavaScript frameworks were used to establish a breast nodule image annotation platform. Five ultrasound radiologists independently performed BI-RADS-US classification and labeling of all sample images, with 8 (DR1), 11 (DR2), 12 (DR3), 15 (DR4), and 19 (DR5) years of experience in breast
ultrasonography. A senior radiologist with 21 years of experience served as the “referee” for final classification. Consistency was achieved to the degree possible as the 6 radiologists jointly discussed and formulated specific criteria for the classification of BI-RADS-US 3–5 nodules, based on their experience and recent literature.[4, 5, 16, 17] See Table 1 for further details.

The five radiologists who were annotating on images performed BI-RADS-US classification based on image features, without knowing patient ages, clinical symptoms, or pathological results. A rectangular box on the labeling platform was used to mark the nodule margin prior to classification and label selection (see Figure 1). The following two strategies were adopted when the BI-RADS-US results were inconsistent. (1) Images with a 4:1 consensus were classified according to the majority opinion. (2) All other disagreements were finalized by the referee, who was allowed to access pathological results in making the final diagnosis to provide a more accurate database for the model. Sample images were randomly divided into development and test sets using a 7:3 ratio. The development set was subsequently and randomly subdivided into training and validation sets using an 8:2 ratio.

A CAD model was constructed by modifying the hierarchical vision transformer architecture for the localization of breast nodules and BI-RADS-US classification. We introduced the multi-resolution feature extraction process to extract the lesion features, and classify them through the attention mechanism. The model included four sets of block 1 and one set of block 2 networks used for feature extraction from input images at different resolutions. In the first four transformer structures, each set of blocks included a window-based multi-head self-attention mechanism for feature extraction. The last group was composed of two transformer structures, the first of which was used to encode the input feature vector, while the second was used to obtain BI-RADS-US categories (e.g., 3, 4a, 4b, 4c, and 5) by decoding encoded feature vectors generated in the previous step. Input images of size 224×224×3 pixels were equally divided into 56×56 image blocks of size 4×4. The images were input to the first block 1 set for feature extraction, producing 56×56×96 feature maps. The first output was then divided into 28×28 image blocks of size 8×8, which were input to the second block 1 set for feature extraction and generation of 28×28×192 feature maps. Similarly, the third and fourth block 1 sets produced 14×14×384 and 7×7×768 feature maps, respectively. The 7×7×768 feature maps were subsequently input to block 2 for BI-RADS-US category classification (e.g., 3, 4a, 4b, 4c, and 5). The input feature maps for block 1 sets 1–4 were first divided into image blocks of a specified size. The image blocks were then shifted, which modified feature information distributions for blocks of varying sizes, allowing the attention to be focused on a wider area. In block 2, position coding information was first added to the coded image block, which was then converted into a one-dimensional vector and input to another transformer. BI-RADS-US categories were used as inputs to query detected nodular areas in each category. Final output results included the detected nodular areas and BI-RADS-US categories to which each node belonged. Figure 2 provides a flowchart of the data processing model.

After CAD model training, 500 breast nodule images with consistent diagnostic results from both the model and the radiologists were randomly selected from the test set to form a sampling test set, which consisted of 100 images for each BI-RADS-US 3–5 category. The five radiologists then re-classified the sampling test set by referencing BI-RADS-US classification results provided by the CAD model. Changes in diagnostic sensitivity, specificity, and accuracy for various nodule categories, including their categorical adjustments, were observed to determine whether classification consistency had been improved.
The SPSS28.0 software package (IBM, Armonk, NY) was used for statistical analysis in this study. The sensitivity, specificity, and accuracy of test set for category 3–5 BI-RADS-US breast nodule classification were determined for the proposed CAD model. A chi-square test was used to analyze the statistical significance of each metric for the sampling set, before and after the CAD model was referenced by the five radiologists. Corresponding Kappa coefficients were calculated and compared to identify any improvements, with values of 0–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1.00 representing poor consistency, normal consistency, moderate consistency, moderately strong consistency, and strong consistency, respectively. All statistical methods were considered significant when the P value was below 0.05.

**Results**

This study involved 3,317 patients exhibiting 1 breast nodule and 661 patients exhibiting 2 or more breast nodules. Each nodule was represented in 1–90 images, with an average of 5.36 ± 5.91 images. The set of 5,057 total breast nodules included 2,390 benign nodules (10,041 images) and 2,667 malignant nodules (11,291 images), as shown in Table 2. The maximum diameter of the nodules ranged from 0.30 to 10.98 cm, with an average diameter of 2.00 ± 1.13 cm. A total of 34 ultrasonography machines were included in the study (see Figure 3).

**Table 1 BI-RADS-US-3–5 classification standards for breast nodules**
Category 3
Solid nodules with clear and sharp margins in an oval shape, parallel to the skin.
Isolated complex cysts or clustered microcysts.

Category 4
*Morphology:*
Round in shape
Irregular in shape*

*Nodular orientation:*
Nodules not parallel to the skin. The anteroposterior diameter is greater than or equal to the transverse diameter*.

*Non-smooth margins:*
Blurred – no clear margin between the mass and the surrounding tissues.
Angled – part of or all the margin is at an acute angle.
Tiny lobes – with small bumps on the margin, crenated.
Burr-like – protruding sharp needles on the nodular margins*.

*Internal echo:*
Uneven echo, with anechoic zones or microcalcifications.

*Posterior echo:*
Attenuation*

*Changes in the surrounding tissue:*
Disordered structure of the surrounding tissue, with normal anatomical layers being destroyed, thickening, or a rigid Cooper ligament.
Catheter changes (unusual pipe diameter or branch-like changes).
>2-mm skin thickening or skin receding, sunken surface, or unclear margins showing tautly.
Edema, enhanced echo of the surrounding tissues, or tissue thickening.

Category 4a
Satisfying one of the above requirements.

Category 4b
Satisfying two of the above requirements or any one item marked with “*”.

Category 4c
Satisfying three of the above requirements or any two items marked with “*”.

Category 5
Satisfying four or more requirements.
Table 2. The distribution of pathological results for 21,332 images.

<table>
<thead>
<tr>
<th>Pathological result</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Malignant</strong></td>
<td></td>
</tr>
<tr>
<td>Non-specific invasive carcinoma</td>
<td>8,690</td>
</tr>
<tr>
<td>Ductal carcinoma <em>in situ</em></td>
<td>1,072</td>
</tr>
<tr>
<td>Mucinous carcinoma</td>
<td>390</td>
</tr>
<tr>
<td>Invasive lobular carcinoma</td>
<td>387</td>
</tr>
<tr>
<td>Intraductal papilloma</td>
<td>179</td>
</tr>
<tr>
<td>Medullary carcinoma</td>
<td>146</td>
</tr>
<tr>
<td>Other malignancy</td>
<td>427</td>
</tr>
<tr>
<td><strong>Benign</strong></td>
<td></td>
</tr>
<tr>
<td>Fibroadenomas</td>
<td>5,434</td>
</tr>
<tr>
<td>Adenopathy</td>
<td>2,491</td>
</tr>
<tr>
<td>Intraductal papillary carcinoma</td>
<td>999</td>
</tr>
<tr>
<td>Breast abscess</td>
<td>265</td>
</tr>
<tr>
<td>Granulomatous inflammation</td>
<td>326</td>
</tr>
<tr>
<td>Other benign nodules</td>
<td>524</td>
</tr>
</tbody>
</table>

Table 3 shows the category 3–5 breast nodule distributions for the training set, validation set, and test set. The sensitivity, specificity, and accuracy of CAD for various categories in the test set, determined using BI-RADS-US classification standards finalized by the referee radiologists, are provided in Table 4. As seen in Table 5, referencing the CAD model significantly improved the diagnostic sensitivity, specificity, and accuracy of the five radiologists (P <0.05). As a result, the radiologists changed a total of 1,583 classifications, including 905 changes to a lower category, with an average of 181 modifications per radiologist. There were 272 classifications adjusted from category 4a to category 3, with an average of 54.5 modifications per radiologist. Among these 272 modifications, 223 (81.99%) exhibited benign pathological results. An additional 678 classifications were adjusted to a higher category, with an average of 135.6 modifications per radiologist. There were 147 classifications adjusted from category 3 to 4a or above, with an average of 29.4 modifications per radiologist, 68 (46.26%) of which were malignant. Of the remaining 917 classifications without adjustments, 657 were classified as category 4a or above. Among these results, 411 (62.56%) were diagnosed as malignant by a pathological examination. In addition, 260 results were category 3, and 244 (93.80%) of which were benign tumors by a pathological examination.

Table 3. Imaging composition for category 3–5 nodules in the BI-RADS-US dataset.
Table 4. Sensitivity, specificity, and accuracy for the CAD model.

<table>
<thead>
<tr>
<th>BIRADS-US</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>95.86%</td>
<td>95.64%</td>
<td>95.70%</td>
</tr>
<tr>
<td>4a</td>
<td>86.88%</td>
<td>98.85%</td>
<td>97.62%</td>
</tr>
<tr>
<td>4b</td>
<td>87.01%</td>
<td>97.36%</td>
<td>95.56%</td>
</tr>
<tr>
<td>4c</td>
<td>91.21%</td>
<td>95.88%</td>
<td>94.25%</td>
</tr>
<tr>
<td>5</td>
<td>85.35%</td>
<td>99.46%</td>
<td>97.49%</td>
</tr>
</tbody>
</table>

Among the 21,332 classified images, 1,416 (6.64%) exhibited the same annotations by all five radiologists and 5,620 (26.3%) included a 4:1 inconsistency. A weighted Kappa test indicated the $k$ value of any two radiologists for all images was less than 0.6 ($0.33–0.57$), representing normal or moderate consistency. The $k$ values for the sampling test set also suggested normal or moderate consistency ($0.30–0.62$). Only the $k$ values of radiologists DR1 vs DR2 and DR1 vs DR3 were slightly higher than 0.6. After referencing the CAD model, the sampling test set images were readjusted and the $k$ values were increased to above 0.60 ($0.67–0.85$), as shown in Figure 4.

Table 5. Variations in sensitivity(SEN), specificity(SPE), and accuracy(ACC) among five radiologists before and after referring to the CAD model.
<table>
<thead>
<tr>
<th>Radiologist</th>
<th>BIRADS-US</th>
<th>SEN before</th>
<th>SEN after</th>
<th>P</th>
<th>SPE before</th>
<th>SPE after</th>
<th>P</th>
<th>ACC before</th>
<th>ACC after</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR1</td>
<td>3</td>
<td>0.51</td>
<td>0.62</td>
<td>0.002</td>
<td>0.91</td>
<td>0.91</td>
<td>0.441</td>
<td>0.83</td>
<td>0.85</td>
<td>0.341</td>
</tr>
<tr>
<td></td>
<td>4a</td>
<td>0.28</td>
<td>0.59</td>
<td>0.000</td>
<td>0.81</td>
<td>0.90</td>
<td>0.000</td>
<td>0.71</td>
<td>0.84</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>4b</td>
<td>0.24</td>
<td>0.53</td>
<td>0.000</td>
<td>0.86</td>
<td>0.88</td>
<td>0.000</td>
<td>0.73</td>
<td>0.81</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>4c</td>
<td>0.43</td>
<td>0.67</td>
<td>0.000</td>
<td>0.70</td>
<td>0.89</td>
<td>0.000</td>
<td>0.65</td>
<td>0.84</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.13</td>
<td>0.57</td>
<td>0.000</td>
<td>0.86</td>
<td>0.91</td>
<td>0.000</td>
<td>0.72</td>
<td>0.84</td>
<td>0.000</td>
</tr>
<tr>
<td>DR2</td>
<td>3</td>
<td>0.58</td>
<td>0.74</td>
<td>0.000</td>
<td>0.68</td>
<td>0.86</td>
<td>1.000</td>
<td>0.67</td>
<td>0.84</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>4a</td>
<td>0.32</td>
<td>0.58</td>
<td>0.000</td>
<td>0.86</td>
<td>0.91</td>
<td>0.009</td>
<td>0.75</td>
<td>0.84</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>4b</td>
<td>0.23</td>
<td>0.47</td>
<td>0.000</td>
<td>0.92</td>
<td>0.91</td>
<td>0.520</td>
<td>0.78</td>
<td>0.82</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>4c</td>
<td>0.62</td>
<td>0.75</td>
<td>0.481</td>
<td>0.60</td>
<td>0.95</td>
<td>0.393</td>
<td>0.60</td>
<td>0.91</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.06</td>
<td>0.72</td>
<td>0.000</td>
<td>0.93</td>
<td>0.93</td>
<td>0.000</td>
<td>0.76</td>
<td>0.89</td>
<td>0.000</td>
</tr>
<tr>
<td>DR3</td>
<td>3</td>
<td>0.71</td>
<td>0.67</td>
<td>0.000</td>
<td>0.82</td>
<td>0.80</td>
<td>1.000</td>
<td>0.80</td>
<td>0.78</td>
<td>0.441</td>
</tr>
<tr>
<td></td>
<td>4a</td>
<td>0.19</td>
<td>0.47</td>
<td>0.000</td>
<td>0.94</td>
<td>0.86</td>
<td>0.000</td>
<td>0.79</td>
<td>0.78</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>4b</td>
<td>0.13</td>
<td>0.36</td>
<td>0.000</td>
<td>0.90</td>
<td>0.90</td>
<td>0.262</td>
<td>0.75</td>
<td>0.79</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td>4c</td>
<td>0.32</td>
<td>0.21</td>
<td>0.000</td>
<td>0.75</td>
<td>0.96</td>
<td>0.008</td>
<td>0.67</td>
<td>0.82</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.36</td>
<td>0.73</td>
<td>0.000</td>
<td>0.77</td>
<td>0.84</td>
<td>0.000</td>
<td>0.69</td>
<td>0.82</td>
<td>0.000</td>
</tr>
<tr>
<td>DR4</td>
<td>3</td>
<td>0.01</td>
<td>0.74</td>
<td>1.000</td>
<td>1.00</td>
<td>0.93</td>
<td>0.000</td>
<td>0.80</td>
<td>0.89</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>4a</td>
<td>0.17</td>
<td>0.79</td>
<td>0.682</td>
<td>0.88</td>
<td>0.92</td>
<td>0.002</td>
<td>0.74</td>
<td>0.90</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>4b</td>
<td>0.39</td>
<td>0.58</td>
<td>0.000</td>
<td>0.74</td>
<td>0.91</td>
<td>0.301</td>
<td>0.67</td>
<td>0.84</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>4c</td>
<td>0.67</td>
<td>0.75</td>
<td>0.725</td>
<td>0.55</td>
<td>0.91</td>
<td>0.000</td>
<td>0.57</td>
<td>0.88</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.11</td>
<td>0.58</td>
<td>0.064</td>
<td>0.92</td>
<td>0.95</td>
<td>1.000</td>
<td>0.76</td>
<td>0.87</td>
<td>0.000</td>
</tr>
<tr>
<td>DR5</td>
<td>3</td>
<td>0.43</td>
<td>0.29</td>
<td>0.761</td>
<td>0.93</td>
<td>0.94</td>
<td>1.000</td>
<td>0.83</td>
<td>0.81</td>
<td>0.552</td>
</tr>
<tr>
<td></td>
<td>4a</td>
<td>0.24</td>
<td>0.34</td>
<td>0.000</td>
<td>0.85</td>
<td>0.87</td>
<td>0.022</td>
<td>0.73</td>
<td>0.76</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td>4b</td>
<td>0.32</td>
<td>0.35</td>
<td>0.002</td>
<td>0.82</td>
<td>0.78</td>
<td>1.000</td>
<td>0.72</td>
<td>0.69</td>
<td>0.330</td>
</tr>
<tr>
<td></td>
<td>4c</td>
<td>0.39</td>
<td>0.60</td>
<td>0.000</td>
<td>0.78</td>
<td>0.81</td>
<td>0.006</td>
<td>0.71</td>
<td>0.77</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.35</td>
<td>0.55</td>
<td>0.132</td>
<td>0.79</td>
<td>0.89</td>
<td>0.690</td>
<td>0.70</td>
<td>0.82</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Discussion**

This study established a multi-center breast nodule image dataset, used to train radiologists for classification with a transformer-based CAD model. This approach increased the accuracy, sensitivity, and specificity of BI-RADS-US classification for category 3–5 breast nodules. The diagnostic consistency among five radiologists
was also improved. In this process, BI-RADS-US was used to interpret image features and breast nodule
descriptions, striving for a more uniform definition among ultrasound radiologists. However, a clear clinical
standard for BI-RADS-US classification has yet to be established. Such discrepancies are common in practice,
especially for category 3–5 breast nodules, affecting diagnosis and treatment of the disease.

In this study, classification criteria were agreed upon in advance, to reach a common consensus. However,
Kappa values across 21,332 images were as high as 0.574, which was consistent with the findings of Berg et
al and Jales et al [18, 19]. This demonstrates that individual radiologists produced varying BI-RADS-US
classification results, due to differences in the recognition of specific image features, despite preemptively
establishing diagnostic standards. To improve classification consistency, radiologists in the study of Berg et al.
provided feedback to correct their results and increased Kappa values from 0.53 to 0.59. The study also
suggested that immediate feedback was helpful for correcting classifications, thereby improving consistency
regardless of radiologist experience.

The emergence of Artificial intelligence (AI) has provided a mechanism for avoiding subjective effects on the
classification of breast nodules. By training with large data quantities, AI forms a unique algorithm to provide
radiologists objective reference information for disease diagnosis. This study involved 14,296 (67.01%) images
in which the final result could not be determined by simply following the majority opinion. The referee
radiologist finalized classification by referencing pathological results, to provide a more accurate dataset for
the CAD model. This approach produced an accuracy in the identification of category 3–5 breast nodules
comparable to that of the referee radiologist, indicating improved classification capabilities. Lee et al also
demonstrated that classification accuracy for benign and malignant breast nodules was significantly higher
with the aid of a CAD model. In the study, the area under the curve (AUC) for inexperienced radiologists
increased from 0.65 to 0.71, while the AUC for the experienced group increased from 0.83 to 0.84.[20] Similar
studies have shown that young radiologists and radiologists lacking experience in breast ultrasound diagnosis
have benefited significantly from the use of a CAD model, especially for category 4a breast nodules, thereby
minimizing unnecessary biopsies.[21, 22]

The radiologists participating in this study each had several years of experience and classification consistency
among them reached values greater than 0.6 after referencing the CAD model. In addition, the sensitivity,
specificity, and accuracy of classification improved with the use of CAD, informing adjustments to the initial
diagnoses. A reduction to category 3 occurred in 54.4 instances on average, exhibiting a consistency rate of
81.99% with pathological diagnoses. These categorical decreases avoid unnecessary biopsies, medical costs,
and psychological burden for patients.[16] An increase to category 4a or above occurred in 29.4 instances on
average, exhibiting a consistency rate of 46.26% with pathological diagnoses, the rate is less than 50%, which
may be related to the large range of positive predictive values for category 4a–5 breast nodules (2–95%).
These increased categorizations provide patients and surgeons with a more credible basis for biopsy or
surgery, thereby avoiding delays in the diagnosis and treatment of breast cancer. Results also indicated CAD to
be a useful diagnostic tool for senior radiologists.

This multi-center study included images collected from 20 hospitals in China and involved a variety of cases
and pathological results. Conventional ultrasonography systems were used, providing a foundation for the
establishment of a model with good robustness and generalizability. The number of images for a single breast

nodule ranged from 1 to 90, which may be a result of varying routines used by radiologists at the same or different hospitals in collecting images. Unlike the previous other studies reported, [6, 8, 23-25] image features, such as nodule shape, orientation, edges, margin, internal echo, posterior echo, and variability in surrounding tissue, were not extracted as part of this study, in order to minimize the influence of human-selected rules in the machine learning algorithm. Rather, this study was based on task-specific feature extraction and the long-distance feature capture capabilities of self-attention and multi-head mechanisms used in transformer technology, to achieve end-to-end autonomous learning.[26]

A primary limitation of this study involved basing breast nodule classification on individual images rather than individual cases, since multi-section scanning is typically adopted in clinical practice. As a result, different sections of the same nodule may produce different classification results due to variations in image features. A correlation analysis was not conducted between the classification and diagnostic results for the same nodule in different sections. In addition, previous studies have shown that AI combined with 2-dimensional ultrasonography and blood flow data from color Doppler ultrasonography may achieve better diagnostic results in benign and malignant breast nodules.[8] This study did not involve blood flow data, spectral Doppler, elastography, or contrast-enhanced ultrasonography, which should be included in future research to further improve the CAD model.

Conclusions

This multi-center study introduced a CAD model, based on transformer technology, offering high accuracy for the classification of BI-RADS-US 3–5 nodules. Application of this model significantly improved diagnostic consistency among radiologists, confirming the value of AI for breast ultrasonography.

Abbreviations

CAD: Computer aided diagnosis, CNN: Convolutional neural network, BI-RADS: Breast imaging reporting and data systems, AI: Artificial intelligence

Declarations

Ethics approval and consent to participate

This retrospective clinical study only involved the collection of age data, breast nodule images, imaging system models, and pathological results for patients. It did not interfere with individual treatment plans and an exemption from informed patient consent was approved by the hospital ethics committee.

Consent for publication

Not applicable

Availability of data and materials

Due to the privacy of patient data, the data set generated or analyzed in this study is not available to the public. The authors will provide relevant data upon reasonable request.
Competing interests

All authors declare no conflicts of interest.

Funding

This project was supported by the National Key Research and Development Plan of China (2016YFC0104803).

Authors' contributions

HTJ participated in the design of the study, analyzed the data of patients, interpreted the results and drafted the manuscript. QZ and WH initiated, organized and designed the study, and reviewed the manuscript. TM, YC, SZ, WR and HLH annotated on the sonographic images of breast nodules according to BI-RADS-US. JHT, QZ, WH, HTR, LTR, YLG, JWT, WC, LJC, ZYW, LZN, WZ, RMY, QC, RFZ, HW, LL, MHL, FN and AYZ were responsible for selecting the sonographic images of breast lesions having pathological results from their respective hospitals according to the inclusion and exclusion criteria.

Acknowledgements:

Not applicable

References


Figures

Figure 1

The annotation platform. The yellow rectangular outline in the middle of the right panel indicates marked nodular areas. BI-RADS-US classification was performed in the upper right corner (indicated by the yellow arrow)
Figure 2

A flowchart for data processing in the CAD model. The name and primary structure of each data processing block are included in the flowchart.

Figure 3

Bar chart showing data processing in the CAD model.
Various ultrasonography machines and corresponding image quantities used in the study.

**Figure 4**

$k$ values for the full dataset and the sampling test set, before and after referring to the CAD model. The green, blue, and orange histograms represent $k$ values for the whole dataset, the sampling test set before referring to the CAD model, and the sampling test set after referring to the CAD model, respectively. The p values were all below 0.05.