Diagnostic Performance of an Artificial Intelligence System in Breast Ultrasound

Avice M. O'Connell, MD, Tommaso V. Bartolotta, MD, PhD, Alessia Orlando, MD, PhD, Sin-Ho Jung, PhD, Jihye Baek, MS, Kevin J. Parker, PhD

Objectives—We study the performance of an artificial intelligence (AI) program designed to assist radiologists in the diagnosis of breast cancer, relative to measures obtained from conventional readings by radiologists.

Methods—A total of 10 radiologists read a curated, anonymized group of 299 breast ultrasound images that contained at least one suspicious lesion and for which a final diagnosis was independently determined. Separately, the AI program was initialized by a lead radiologist and the computed results compared against those of the radiologists.

Results—The AI program's diagnoses of breast lesions had concordance with the 10 radiologists' readings across a number of BI-RADS descriptors. The sensitivity, specificity, and accuracy of the AI program's diagnosis of benign versus malignant was above 0.8, in agreement with the highest performing radiologists and commensurate with recent studies.

Conclusion—The trained AI program can contribute to accuracy of breast cancer diagnoses with ultrasound.

Key Words—artificial intelligence (AI); breast cancer; computer-aided detection (CADe); computer-assisted diagnosis (CADx); machine learning; ultrasound

The incidence and mortality of breast cancer across the globe creates an imperative to improve diagnosis and treatments. Ultrasound imaging presents a relatively affordable, accessible, and non-ionizing method for detection of lesions with reasonable sensitivity and specificity. Breast ultrasound has some advantages over X-ray mammography in certain cases, especially the dense breast. Given these factors, the drive to improve breast ultrasound using computer-assisted analyses has produced a number of approaches over the past few decades. The earlier approaches concentrated on the extraction of features of lesions such as size, shape, texture, and boundaries within a clustering or classification or rule-based decision making algorithms.[1–4] More recent developments in artificial intelligence (AI), machine learning, and deep learning systems have utilized a variety of approaches and extensive training sets to produce differentiated output classifications.[5–8]

Careful assessment of the introduction of these technologies into radiology practice is very important. Hence, the purpose of this study is to assess the diagnostic performance of an AI-based program and to compare the results against radiologists. We used

Received November 3, 2020, from the Department of Imaging Sciences, University of Rochester Medical Center, Rochester, New York, USA (A.M.O.); Department of Radiology, University Hospital, Palermo, Italy (T.V.B., A.O.); Fondazione Istituto G. Giglio Hospital, Cefalù, Italy (T.V.B.); Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, USA (S.-H.J.); and Department of Electrical and Computer Engineering, University of Rochester, Rochester, New York, USA (J.B., K.J.P.). Manuscript accepted for publication February 18, 2021.

This study could not have been completed without the dedicated work of Ms. Tammy Russell, Ms. Cheryl Vance, Dr. Hongmei Yang, and Dr. Juvenal Ormachea. Ten dedicated radiologists in the upstate New York area are thanked for their skilled assessments and dedication to the research study. Samsung Medison is thanked for the loan of the RS80A and RS85 ultrasound systems and S-Detect™ for Breast software, and for support of the research project (grant numbers GR504321 and GR504322).

Address correspondence to Kevin J. Parker, PhD, University of Rochester, Computer Studies Building 724, Box 270231, Rochester, NY 14627-0231.

E-mail: kevin.parker@rochester.edu

Abbreviations

AI, artificial intelligence; ANOVA, a one-way analysis of variance; ROC, receiver operating characteristic; IDC, invasive ductal carcinoma

doi:10.1002/jum.15684

© 2021 American Institute of Ultrasound in Medicine.

S-Detect™ for Breast, a software based on a convolutional neural network (Samsung Medison Co., Ltd., South Korea) that has been trained to classify lesions using over 10,000 breast scans against "gold standard" biopsy assessments. S-Detect™ for Breast can be used interactively with the reading radiologist by presenting a choice of boundaries and a likely classification as to benign or malignant. In addition, descriptors are generated to indicate features related to shape, orientation, margin, posterior features, and echo patterns of the lesion[9-18] on B-scan images.

In the current study, breast ultrasound images from Italy and the United States were studied in two settings. In one, all images were reviewed and assigned Breast Imaging Reporting and Data Systems (BI-RADS) lexicon descriptors and scores manually by the radiologists in conventional reading sessions (manual session). Radiologists were selected and grouped according to two levels of experience: more than 10 years in mammography or less than 5 years, all board-certified or board-eligible. Separately, each image was analyzed by the S-Detect™ for Breast AI program after initiation by the principal investigator (PI, Dr. O'Connell) to automatically perform a classification (automatic session). The study enabled us to quantify some measures of performance of S-Detect™ for Breast in comparison to a group of radiologists. Specifically, the concordance rate, sensitivity, specificity, and accuracy were examined. The details are provided in the following sections.

Materials and Methods

Enrollment of the Research Population

Subjects were drawn from those patients whose standard-of-care breast ultrasound revealed at least one suspicious lesion, and who were recommended to have either a biopsy or biannual ultrasound imaging follow-up. Hundred and fifty subjects were prospectively enrolled at the University of Rochester, and 149 subjects were prospectively enrolled at the University Hospital Palermo, Italy during the timeframe 2018–2019. The research protocol was approved by and conducted under the requirements of informed consent of the Research Subjects Review Boards at the University of Rochester and Policlinico P. Giaccone, University of Palermo.

Anonymized and de-identified images were shared consistent with HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) regulations. If a subject had more than one suspicious lesion, each could be chosen by the radiologist attending as suitable for "second review." The study's inclusion criteria were:

- Adult females or males recommended for ultrasound-guided breast lesion biopsy or ultrasound follow-up with at least one suspicious lesion.
- Age $\geq$ 18 years.
- Able to provide informed consent.

The study's exclusion criteria were:

- Unable to read and understand English (at the University of Rochester).
- Unable or unwilling to provide informed consent.
- A patient with current or previous diagnosis of breast cancer in the same quadrant.
- Unable or unwilling to undergo study procedures.

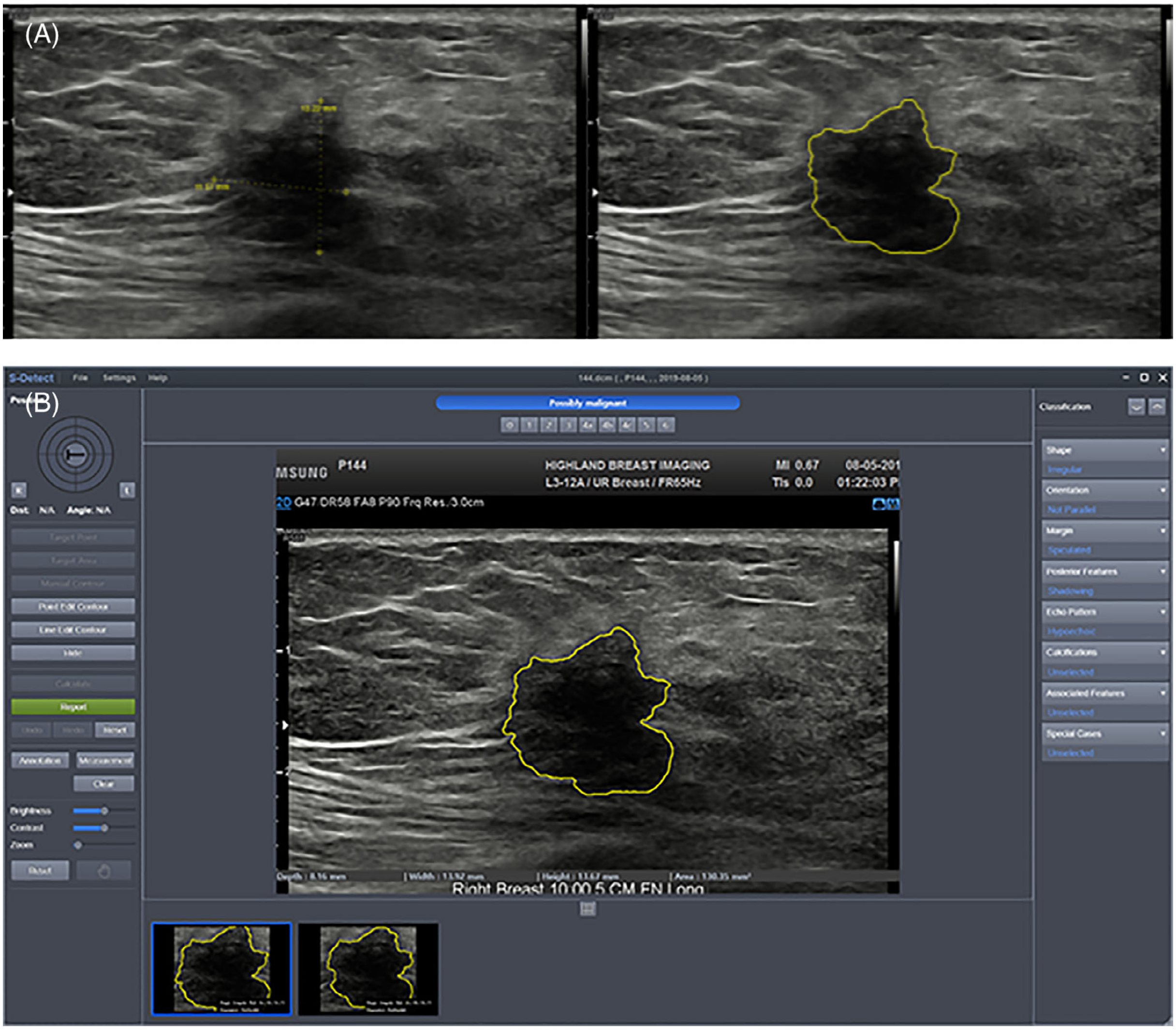
Image Acquisition

Two hundred and ninety nine patients (mean age 52.3 years) underwent ultrasonographic examinations using either the RS80A with Prestige (Samsung Medison Co., Ltd., South Korea) in Palermo, Italy or the RS85 (Samsung Medison Co., Ltd., South Korea) in Rochester, NY, USA with a 3–12 MHz linear array transducer (L3-12A) to acquire a static image of the suspicious lesion/mass, after first examining the lesion in 2 planes (transverse and longitudinal) and obtaining a cine loop of the region.

Image Analysis

The 299 patients and the breast lesion images obtained during the study received final reviews by board-certified or board-eligible radiologists (five with over 10 years of experience, five with less than 5 years of experience) and the S-Detect™ for Breast system (Figure 1) per the following schedule: The de-identified images were first reviewed by the radiologists in a reading room during two scheduled sessions. The radiologists were tasked with assigning BI-RADS lexicons and scores manually without any assistance from AI. The same images were separately processed by the AI program after initialization by the PI (Dr. O'Connell). This initialization consisted of identifying the interior of the lesion with a graphics

Figure 1. Example of S-Detect™ for Breast. **A**, Sample breast lesion of a patient with lesion area identified by a radiologist (left) and contoured boundary by S-Detect™ for Breast program (right), **B**, Full screen of S-Detect™ for Breast including BI-RADS Feature Classification and Assessment Category.



Shape: Irregular
Orientation: Not Parallel
Margin: Spiculated
Posterior Features: Shadowing
Echo Pattern: Hypoechoic
Calcifications: Unselected
Associated Features: Unselected
Special Cases: Unselected

interface, then selecting the most appropriate version of several lesion boundary outlines suggested by S-Detect™ (Figure 1A). After this initial step, there was no further intervention by the PI or the radiologists. S-Detect™ then automatically proposed feature classifications such as shape, orientation, margin, echo pattern, and posterior feature and suggested the final assessment in a dichotomized form, "possibly benign" or "possibly malignant" (Figure 1B). The details of the descriptors are given in Table A1. The two sets of BI-RADS lexicon labels, one from the radiologists and the other from the AI program, were then compared to statistically quantify concordance. Ultimately, all decisions were compared with the ground truths generated from the biopsy results or a 24-month follow-up. These comparisons are detailed in the following section.

Statistical Tests

The study was conducted and approved to examine the following hypotheses related to concordance: In order to assess the general agreement between the output of S-Detect™ for Breast and that of the radiologists, the sample size and related tests were defined and calculated as follows. Let p_r and p_s denote the concordance rate among radiologists as readers and that between readers and S-Detect™ for Breast, respectively. Let δ_0 = 10% denote the non-inferiority margin, that is, we initially consider a difference in concordance rate of 10% or smaller to be acceptable. We want to test

$$\begin{aligned} &H_0: \; p_s \leq p_r - \delta_0 \\ &\qquad \text{versus} \\ &H_1: \; p_s > p_r - \delta_0 \end{aligned} \tag{1}$$

Let m = 10 denote the number of readers. For subject i (=1, …, n) and reader j (=1, …, m), let $r_{ijj'} = 1$ if readers j and j' concord for subject i, and $r_{ijj'} = 0$ otherwise, and let s_{ij} = 1 if reader j and S-Detect™ concord for subject i, and s_{ij} = 0 otherwise. We have $p_r = E\left(r_{ijj'}\right)$ and $p_s = E(s_{ij})$. The concordance rates for subject i was estimated by

$$r_i = \frac{\sum_{j=1}^{m-1}\sum_{j'=j+1}^{m} r_{ijj'}}{m(m-1)/2} \tag{2}$$

among m readers, and

$$s_i = \frac{\sum_{j=1}^{m} s_{ij}}{m} \tag{3}$$

between readers and S-Detect™ for Breast. The concordance between S-Detect™ for Breast and readers is estimated by

$$\hat{p}_s = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij} = \frac{i}{n}\sum_{i=1}^{n} s_i \tag{4}$$

A 100(1 − α)% confidence interval (CI) for the concordance rate between S-Detect™ for Breast and m readers was obtained by

$$\hat{p}_s \pm z_{1-\alpha/2}\hat{\sigma}\left(\hat{p}_s\right) \tag{5}$$

where

$$\hat{\sigma}^2\left(\hat{p}_s\right) = \frac{\sum_{i=1}^{n}\left\{\sum_{j=1}^{m}\left(s_{ij} - \hat{p}_s\right)\right\}^2}{(mn)^2} = \frac{\sum_{i=1}^{m}\left(s_i - \hat{p}_s\right)^2}{n^2} \tag{6}$$

On the other hand, let $r_{ijj'}$ denote the concordance score between reader j and j' ($1 \leq j < j' \leq \mathrm{m}$) for image i. Then, the concordance rate among m readers was estimated by

$$\hat{p}_r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m-1}\sum_{j'=j+1}^{m} r_{ijj'}}{nm(m-1)/2} \tag{7}$$

A 100(1 − α)% CI for the concordance rate among m readers was obtained by

$$\hat{p}_r \pm z_{1-\alpha/2}\hat{\sigma}\left(\hat{p}_r\right) \tag{8}$$

where

$$\hat{\sigma}^2\left(\hat{p}_r\right) = \frac{\sum_{i=1}^{n}\left\{\sum_{j=1}^{m-1}\sum_{j'=j+1}^{m}\left(r_{ijj'} - \hat{p}_r\right)\right\}^2}{\{nm(m-1)/2\}^2} = \frac{\sum_{i=1}^{n}\left(r_i - \hat{p}_r\right)^2}{n^2} \tag{9}$$

A 100(1 − α)% CI for $p_s - p_r$ was estimated by

$$\hat{p}_s - \hat{p}_r \pm z_{1-\alpha/2}\hat{\sigma}\left(\hat{p}_s - \hat{p}_r\right) \tag{10}$$

where

$$\hat{\sigma}^2\left(\hat{p}_s - \hat{p}_r\right) = \frac{1}{n^2}\sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{m}\left(s_{ij} - \hat{p}_s\right)}{m} - \frac{\sum_{j=1}^{m-1}\sum_{j'=j+1}^{m}\left(r_{ijj'} - \hat{p}_r\right)}{m(m-1)/2}\right\}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\left(s_i - \hat{p}_s - r_i - \hat{p}_r\right)^2 \tag{11}$$

We reject H_0 if the Z-score is

$$Z=\frac{n^{-1/2}\sum_{i=1}^{n}\left(s_i-r_i+\delta_0\right)}{\sqrt{n^{-1}\sum_{i=1}^{n}\left(s_i-r_i+\delta_0\right)^2}}>z_{1-\alpha} \tag{12}$$

Note that we are using one-sided alpha level for a non-inferiority test, as usual. To convert our analysis results to those for a two-sided non-inferiority test, we simply divide the one-sided p-values by two.

Separate calculations confirmed that this protocol had an appropriate power for the wide range of p_r and correlation ρ values with an estimated sample size of 299 images. Once the data from the reading sessions were available, it was also possible to examine the sensitivity, specificity, and accuracy of the S-Detect™ for Breast and the radiologists using the receiver operating characteristic (ROC) analyses found in MATLAB (MathWorks Inc., Natick, MA, USA).

ROC Comparisons

From the total 299 cases, major categories of benign and malignant cases (n = 226) were selected by an experienced radiologist for further study as an empirical ROC analysis. This subset of cases incorporated the most commonly occurring categories, for example, invasive ductal carcinoma (IDC), and excluded rare and single occurring pathologies. This subset is listed in Table A2. In comparing readings within these major categories, S-Detect™ for Breast classified breast lesions as benign or malignant, whereas the radiologists provided BI-RADS scores. Therefore, to compare the sensitivity and specificity, the BI-RADS scores were used as a basis for discriminating benign versus malignant; initially the scores from 4a to 5 were considered as malignant, and the scores from 1 to 3 were benign. This threshold (the BI-RADS score below which lesions are considered benign) was then varied within the ROC analysis.[19] Furthermore, the 10 radiologists were divided into two reader groups based on their years of experience in breast ultrasound or mammography. The more experienced group included 5 radiologists with over 10 years (mean 24.6 ± standard deviation (SD) 8.6 years), and the less experienced group included the other 5 radiologists with less than 5 years (mean 2.6 ± SD 2.3 years) experience. The performance of the two reader groups was compared using a one-way analysis of variance (ANOVA).

Results

Population Characteristics

The 150 patients scanned at the University of Rochester were self-reported as 70.7% Caucasian, 22.0% African-American, 2.0% Asian, and 5.3% others. Their average age was 53 and, of the lesions studied, 95 were benign and 55 were malignant. The 149 patients scanned in Palermo, Italy were 100% Caucasian, their average age was 52, and of the lesions studied 54 were benign and 95 were malignant. Overall, 50.2% of this group of 299 lesions were verified as malignant based on biopsy results or a 24-month follow-up assessment.

Table 1. The Concordance Rate Between S-Detect™ for Breast and the 10 Readers ($\hat{p}_s$), Concordance Rate Among the 10 Readers ($\hat{p}_r$), $N(0, 1)$ Non-inferiority Test Statistic (Z), and Two-sided P-value

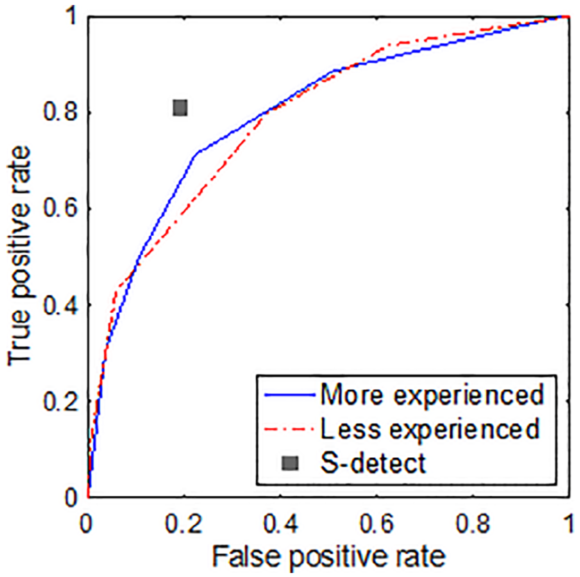
Lexicon Classifications	$\hat{p}_s$ (95% CI)	$\hat{p}_r$ (95% CI)	$\hat{p}_s-\hat{p}_r$ (95% CI)	Z	P-value
Shape	0.6291 (0.5917, 0.6665)	0.6757 (0.6496, 0.7018)	−0.0466 (−0.0779, −0.0152)	3.2780	.0010
Orientation	0.7769 (0.7441, 0.8098)	0.8015 (0.7762, 0.8267)	−0.0246 (−0.0509, −0.0018)	5.3346	.0000
Margin	0.3756 (0.3388, 0.4124)	0.4155 (0.3863, 04448)	−0.0399 (−0.0717, −0.0082)	3.6243	.0002
Posterior Feature	0.6314 (0.5946, 0.6683)	0.6687 (0.6409, 0.6965)	−0.0372 (−0.0698, −0.0047)	3.6943	.0002
Echo Pattern	0.5234 (0.4851, 0.5617)	0.6017 (0.5748,0.6286)	−0.0783 (−0.1164,-0.0402)	1.1146	.2650
Total (5 lexicons combined)	**0.5873** **(0.5703, 0.6043)**	**0.6326** **(0.6185, 0.6467)**	**−0.0453** **(−0.0598, −0.0309)**	**6.8129**	**.000**

Concordance Measures

Table 1 shows estimated concordance rates and their 95% confidence intervals together with non-inferiority testing results using a non-inferiority margin of $\delta_0 = 0.1$, between S-Detect™ for Breast and the 10 readers (p_s), along with the concordance rate among the 10 readers (p_r).

We found that the concordance rate between S-Detect™ for Breast and the readers was significantly $(P < .05)$ non-inferior to the concordance rate among readers in shape, orientation, margin, and posterior classification. The echo pattern contributed with other feature categories in the assessment of the breast lesions, but is found to display low specificity alone.[20] Overall, the performance results show that the non-inferiority test, by combining the concordance scores of all five lexicons, was very significant $(P < .001)$.

Figure 2. ROC curves for radiologists compared with dichotomous S-Detect™ for Breast results. S-Detect™ for Breast (single square) and two reader groups (more and less experienced).



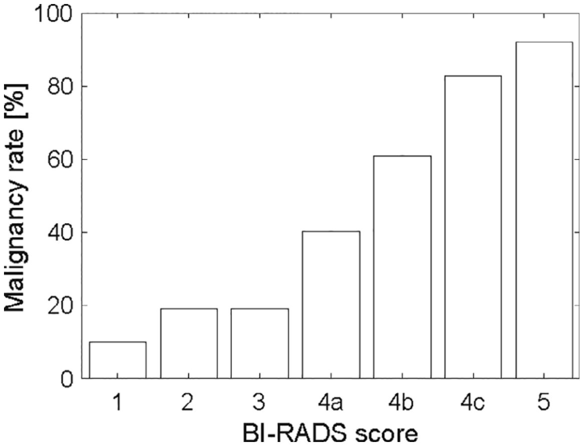
ROC Outcomes

Figure 2 represents the ROC curves showing the averaged performances of the experienced and less experienced groups, along with the dichotomous outcome of S-Detect™ for Breast. To obtain the two curves from the reader groups in Figure 2, all reading results from radiologists were collected for each group, resulting in 0.813 and 0.807 as the AUC (area under curve) for experienced and less experienced, respectively. These are empirical results, the issue of further statistical analysis is considered in the Discussion section. Table 2 also shows the combined cancer detection results for 10 radiologists and S-Detect™ for Breast. The AUC for the combined radiologists was 0.810. We found that the highest sensitivity and specificity values for the readers as reported in Table 2 were achieved when the ROC threshold is set such that BI-RADS scores of 4b and higher are considered as malignant. Since BI-RADS 4a represents a low suspicion of malignancy,[20] this subgroup contains a number of benign lesions which are accounted as false positives when 4a is used as a threshold. The overall percent of malignancy within each category is given in Figure 3, showing an increasing proportion of malignancy with increasing BI-RADS score. Empirically, the dichotomous S-Detect™ for Breast performance was comparable to the experienced group. For example, the sensitivity of S-Detect™ for Breast and the radiologists was 0.810 and 0.703, respectively. The specificity of S-Detect™ and the radiologists was 0.827 and 0.755, respectively. Thus, when considering sensitivity, specificity, and accuracy, S-Detect™ for Breast produced a favorable performance compared to the radiologists, as an empirical result in the study. Comparing the results of the more and less experienced readers, an ANOVA analysis found no statistically significant difference between the two groups for AUC, sensitivity, and accuracy. However, the less-experienced group has

Table 2. Performance for Radiologists and Initialized S-Detect™ Algorithm in Patients

	Initialized S-Detect™	Radiologists (N = 10)	Radiologists (Experienced, N = 5)	Radiologists (Less Experienced, N = 5)
AUC	Dichotomous	0.8097	0.8128	0.8066
Sensitivity	0.8095	0.7029	0.7810	0.6248
Specificity	0.8265	0.7554	0.6645	0.8463
Accuracy	0.8186	0.7310	0.7186	0.7434

Figure 3. Malignancy rate of lesions within each BI-RADS category as assigned by radiologists. The malignancy rate increases monotonically with increasing score.



higher specificity than the more-experienced readers, with a *P*-value of .044.

Discussion

Overall, our results supported the hypothesis that the concordance rate of BI-RADS descriptors between S-Detect™ for Breast and the readers was non-inferior to the concordance rate amongst readers. Similar concordance rates between S-Detect™ for Breast and the readers was found for descriptors including shape, orientation, margin, and posterior feature classification. Our performance test results have demonstrated that the non-inferiority test by combining the concordance scores for all five lexicons was very significant ($P < .001$).

Furthermore, a comparison of the readers' ROC curves showed favorable decisions when using the initialized S-Detect™ for Breast as compared to the reading by radiologists. However, the assessments, including AUC, sensitivity, and specificity, are comparable or lower than the S-Detect™ design study from Han et al,[17] which could be caused by the following differences between the previous and this study. The test data used in Han's study was obtained from the same hospital where the training data was originally obtained. In general, internal validation performance using training and testing data from the same hospital is higher than external validation performance. In fact, other studies of S-Detect™ for Breast conducted in other hospitals in Korea, Asian countries, or Western countries all showed results that are lower in performance than Han's paper.[17,21] Also, S-Detect™ was trained using dense breast cases from an Asian population, whereas this study included less than 3% of Asian patients. However, the three assessments of this study are over 0.8, which are comparable to typical performances from ultrasound.[22] Thus, S-Detect™ for Breast appears to be useful for more diverse populations, although their breasts have different characteristics, including density. Furthermore, this study included European and US breast radiologists, whereas the previous study was performed with Korean radiologists. It is known that there can be diagnostic trends depending on area, for example, the sensitivity of US doctors is higher than European doctors.[23] Thus, there can be assessment differences between regions, however the subset of major categories of benign and malignant cases (n = 226, see Table A2) was chosen to represent cases where the pathology diagnosis as a gold standard would be consistent across different international sites.

The original S-Detect™ for Breast training was performed using ultrasound images acquired from a iU22 system (Philips Healthcare) and an RS80A (Samsung Medison Co. Ltd.): 71 and 29% from iU22 and RS80A, respectively. The S-Detect™ for Breast image set of this study was acquired using Samsung RS80A and RS85 systems. However, S-Detect™ for Breast uses post-processed B-mode images as input images, which generally have different textures of images between companies. Therefore, we speculate that by adding more training set images from the RS85 ultrasound system, the accuracy of S-Detect™ for Breast from RS85 studies could increase.

Three limitations of the study results are related to the empirical nature of our ROC analyses. First, the pooled averaging of ROC results may not be strictly area-preserving and may in some cases bias to a lower AUC.[24] This may underrepresent the performance of the pooled groups. Second, a detailed statistical analysis of the S-Detect™ for breast ROC curve is left for further research. The statistical generalization of our results is complicated by the fact that our study does not include a standard MRMC (multiple reader, multiple cases) set of outputs,[25] and so a model-based statistical generalization of ROC curves remains for future study. Finally, the scoring by radiologists may not be indicative of their results in clinical practice since the constrained nature of the study did not allow the radiologists to access patient information such as age, history, cine loop sweeps, and mammogram images which would ordinarily inform their decisions.

Conclusion

Our performance results have demonstrated that the non-inferiority test by combining the concordance scores for all five lexicons was very significant (P <.001). Furthermore, the ROC analyses derived from this study show that the initialized S-Detect™ for Breast program can achieve a sensitivity, specificity, and accuracy greater than 0.8, commensurate with that of experienced radiologists in this study (albeit under restricted conditions) and international studies[23] using ultrasound for imaging breast cancer.

References

1. Wu WJ, Lin SW, Moon WK. Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. *Comp Med Imaging Graphics* 2012; 36:627–633. https://doi.org/10.1016/j.compmedimag.2012.07.004.
2. Liu B, Cheng HD, Huang J, Tian J, Tang X, Liu J. Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images. *Pattern Recogn* 2010; 43:280–298. https://doi.org/10.1016/j.patcog.2009.06.002.
3. Shan J, Cheng HD, Wang Y. Completely automated segmentation approach for breast ultrasound images using multiple-domain features. *Ultrasound Med Biol* 2012; 38:262–275. https://doi.org/10.1016/j.ultrasmedbio.2011.10.022.
4. Cheng HD, Shan J, Ju W, Guo Y, Zhang L. Automated breast cancer detection and classification using ultrasound images: a survey. *Pattern Recogn* 2010; 43:299–317. https://doi.org/10.1016/j.patcog.2009.05.012.
5. Becker AS, Mueller M, Stoffel E, Marcon M, Ghafoor S, Boss A. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *Br J Radiol* 2018; 91:20170576. https://doi.org/10.1259/bjr.20170576.
6. Singh BK, Verma K, Thoke AS. Fuzzy cluster based neural network classifier for classifying breast tumors in ultrasound images. *Expert Syst Appl* 2016; 66:114–123. https://doi.org/10.1016/j.eswa.2016.09.006.
7. Liu S, Wang Y, Yang X, et al. Deep learning in medical ultrasound analysis: a review. *Engineering* 2019; 5:261–275. https://doi.org/10.1016/j.eng.2018.11.020.
8. Zhang Q, Xiao Y, Dai W, et al. Deep learning based classification of breast tumors with shear-wave elastography. *Ultrasonics* 2016; 72:150–157. https://doi.org/10.1016/j.ultras.2016.08.004.
9. Choi JS, Han BK, Ko ES, et al. Effect of a deep learning framework-based computer-aided diagnosis system on the diagnostic performance of radiologists in differentiating between malignant and benign masses on breast ultrasonography. *Korean J Radiol* 2019; 20:749–758. https://doi.org/10.3348/kjr.2018.0530.
10. Bartolotta TV, Orlando A, Cantisani V, et al. Focal breast lesion characterization according to the BI-RADS US lexicon: role of a computer-aided decision-making support. *Radiol Med* 2018; 123:498–506. https://doi.org/10.1007/s11547-018-0874-7.
11. Lee SE, Moon JE, Rho YH, Kim EK, Yoon JH. Which supplementary imaging modality should be used for breast ultrasonography? Comparison of the diagnostic performance of elastography and computer-aided diagnosis. *Ultrasonography* 2017; 36:153–159. https://doi.org/10.14366/usg.16033.
12. Wu JY, Zhao ZZ, Zhang WY, et al. Computer-aided diagnosis of solid breast lesions with ultrasound: factors associated with false-negative and false-positive results. *J Ultrasound Med* 2019; 38:3193–3202. https://doi.org/10.1002/jum.15020.
13. Wu GG, Zhou LQ, Xu JW, et al. Artificial intelligence in breast ultrasound. *World J Radiol* 2019; 11:19–26. https://doi.org/10.4329/wjr.v11.i2.19.
14. Park HJ, Kim SM, La Yun B, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: added value for the inexperienced breast radiologist. *Medicine* 2019; 98:e14146. https://doi.org/10.1097/md.0000000000014146.
15. Zhang Z, Sejdic E. Radiological images and machine learning: trends, perspectives, and prospects. *Comp Biology and Med* 2019; 108:354–370. https://doi.org/10.1016/j.compbiomed.2019.02.017.
16. Choi JH, Kang BJ, Baek JE, Lee HS, Kim SH. Application of computer-aided diagnosis in breast ultrasound interpretation: improvements in diagnostic performance according to reader experience. *Ultrasonography* 2018; 37:217–225. https://doi.org/10.14366/usg.17046.
17. Han S, Kang HK, Jeong JY, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Phys Med Biol* 2017; 62:7714–7728. https://doi.org/10.1088/1361-6560/aa82ec.
18. Di Segni M, de Soccio V, Cantisani V, et al. Automated classification of focal breast lesions according to S-detect: validation and

role as a clinical and teaching tool. *J Ultras* 2018; 21:105–118. https://doi.org/10.1007/s40477-018-0297-2.

19. Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J Radiol* 2004; 5:11–18. https://doi.org/10.3348/kjr.2004.5.1.11.
20. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA. *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System.* Reston, VA: American College of Radiology; 2013.
21. Li J, Sang T, Yu WH, et al. The value of S-detect for the differential diagnosis of breast masses on ultrasound: a systematic review and pooled meta-analysis. *Med Ultrasonog* 2020; 22:211–219. https://doi.org/10.11152/mu-2402.
22. Sood R, Rositch AF, Shakoor D, et al. Ultrasound for breast cancer detection globally: a systematic review and meta-analysis. *J Glob Oncol* 2019; 5:1–17. https://doi.org/10.1200/jgo.19.00127.
23. Domingo L, Hofvind S, Hubbard RA, et al. Cross-national comparison of screening mammography accuracy measures in U.S., Norway, and Spain. *Eur Radiol* 2016; 26:2520–2528. https://doi.org/10.1007/s00330-015-4074-8.
24. Chen W, Samuelson FW. The average receiver operating characteristic curve in multireader multicase imaging studies. *Br J Radiol* 2014; 87:20140016. https://doi.org/10.1259/bjr.20140016.
25. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Investigative Radiol* 1992; 27:723–731.

Appendix

Table A1. Lesion Descriptors Used in this Study

Ground Truth	Lesion Type
Shape	Oval
	Round
	Irregular
Margin	Circumscribed
	Indistinct
	Angular
	Microlobulated
	Spiculated
Orientation	Parallel
	Not parallel
Echo pattern	Anechoic
	Hypoechoic
	Complex cystic and solid
	Isoechoic
	Hyperechoic
	Heterogeneous
Posterior features	No features
	Enhancement
	Shadowing
	Combined pattern

Table A2. List of Major Categories and Pathology Descriptors of Benign and Malignant Lesions Included in Readers' ROC Analyses

Ground Truth	Lesion Type
Benign (N = 105)	Stromal fibrosis
	Fibroadenoma
	Fibroadenomatoid changes (FAC)
	Fibrocystic changes with stromal fibrosis
	Intraductal papilloma
	Cyst (micocyst cluster, ruptured cyst, simple cyst)
	Follow-up stable mass
Malignant (N = 121)	Ductal carcinoma in situ (DCIS)
	Invasive ductal carcinoma (IDC)
	Invasive lobular carcinoma (ILC)
	Invasive ductal and lobular carcinoma
	Invasive ductal carcinoma with micropapillary features
	Invasive mammary carcinoma (IMC)