Artificial Intelligence System Reduces False-Positive Findings in the Interpretation of Breast Ultrasound Exams

Yiqiu Shen[1,*], Farah E. Shamout[2,*], Jamie R. Oliver[3,*], Jan Witowski[3], Kawshik Kannan[4], Jungkyu Park[5], Nan Wu[1], Connor Huddleston[3], Stacey Wolfson[3], Alexandra Millet[3], Robin Ehrenpreis[3], Divya Awal[3], Cathy Tyma[3], Naziya Samreen[3], Yiming Gao[3], Chloe Chhor[3], Stacey Gandhi[3], Cindy Lee[3], Sheila Kumari-Subaiya[3], Cindy Leonard[3], Reyhan Mohammed[3], Christopher Moczulski[3], Jaime Altabet[3], James Babb[3], Alana Lewin[3], Beatriu Reig[3], Linda Moy[3,5], Laura Heacock[3], Krzysztof J. Geras[3,5,1,✉]

[1]Center for Data Science, New York University
[2]Engineering Division, NYU Abu Dhabi
[3]Department of Radiology, NYU Grossman School of Medicine
[4]Department of Computer Science, Courant Institute, New York University
[5]Vilcek Institute of Graduate Biomedical Sciences, NYU Grossman School of Medicine
[*]Equal contribution
[✉]k.j.geras@nyu.edu

Abstract

Ultrasound is an important imaging modality for the detection and characterization of breast cancer. Though consistently shown to detect mammographically occult cancers, especially in women with dense breasts, breast ultrasound has been noted to have high false-positive rates. In this work, we present an artificial intelligence (AI) system that achieves radiologist-level accuracy in identifying breast cancer in ultrasound images. To develop and validate this system, we curated a dataset consisting of 288,767 ultrasound exams from 143,203 patients examined at NYU Langone Health, between 2012 and 2019. On a test set consisting of 44,755 exams, the AI system achieved an area under the receiver operating characteristic curve (AUROC) of 0.976. In a reader study, the AI system achieved a higher AUROC than the average of ten board-certified breast radiologists (AUROC: 0.962 AI, 0.924±0.02 radiologists). With the help of the AI, radiologists decreased their false positive rates by 37.4% and reduced the number of requested biopsies by 27.8%, while maintaining the same level of sensitivity. To confirm its generalizability, we evaluated our system on an independent external test dataset where it achieved an AUROC of 0.911. This highlights the potential of AI in improving the accuracy, consistency, and efficiency of breast ultrasound diagnosis worldwide.

Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer-related deaths among women worldwide [1]. It is estimated that 281,550 new cases of invasive breast cancer will be diagnosed among women in the United States in 2021, eventually leading to approximately 43,600 deaths [2]. Identifying breast cancer at an early stage before metastasis enables more effective treatments and therefore significantly improves survival rates [3, 4]. Mammography has long been the most widely utilized imaging technique for screening and early detection of breast cancer, but it is not without limitations. In particular, for women with dense breast tissue, the sensitivity of mammography drops from 85% to 48-64% [5]. This is a significant drawback, as women with dense breasts have a 4-fold increased risk of developing breast cancer [6]. Moreover, mammography is not always accessible, especially in limited-resources settings, where the high cost of equipment is prohibitive and skilled technologists and radiologists are not available [7].

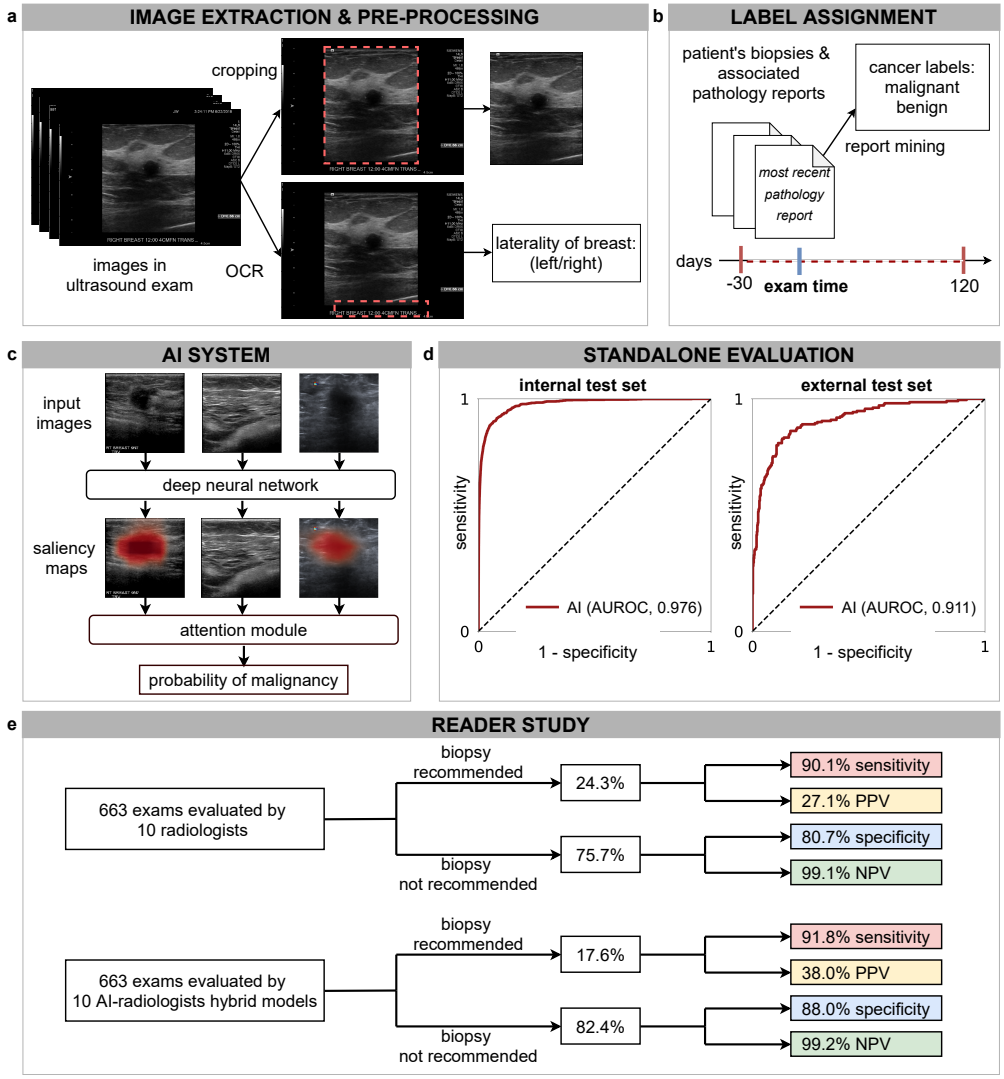

Figure 1: **Overview of the system's pipeline.** **a**, US images were pre-processed to extract the breast laterality (i.e., left or right breast) and to include only the part of the image which shows the breast (cropping out the image periphery which typically contains textual metadeta about the patient and US acquisition technique). **b**, For each breast, we assigned a cancer label using the recorded pathology reports for the respective patient within -30 and 120 days from the time of the US examination. We applied additional filtering on the internal test set to ensure that cancers in positive exams are visible in the US images and negative exams have at least one cancer-negative follow-up (see Methods section 'Additional filtering of the test'). **c**, The AI system processes all US images acquired from one breast to compute probabilistic predictions for the presence of malignant lesions. The AI system also generates saliency maps that indicate the informative regions in each image. **d**, We evaluated the system on an internal test set (AUROC: 0.976, 95% CI: 0.972, 0.980, $n = 79{,}156$ breasts) and an external test set (AUROC: 0.911, 95% CI: 0.885, 0.933, $n = 780$ images). **e**, In a reader study consisting of 663 exams ($n = 1{,}024$ breasts), we showed that the AI system can improve the specificity and positive predictive value (PPV) for 10 attending radiologists while maintaining the same level of sensitivity and negative predictive value (NPV).

Given the limitations of mammography, ultrasound (US) plays an important role in breast cancer diagnosis. It often serves as a supplementary modality to mammography in screening settings [8] and as the primary

imaging modality in many diagnostic settings, including the evaluation of palpable breast abnormalities [9]. Moreover, US can help further evaluate and characterize breast masses and is therefore frequently used for performing image guided breast biopsies [10]. Breast US has several advantages compared to other imaging modalities, including relatively lower cost, lack of ionizing radiation, and the ability to evaluate images in real time [4]. In particular, US is especially effective at distinguishing solid breast masses from fluid-filled cystic lesions. In addition, breast US is able to detect cancers obscured on mammography, making it particularly useful in diagnosing cancers in women with mammographically dense breast tissue [11].

Despite these advantages, interpreting breast US is a challenging task. Radiologists evaluate US images using different features including lesion size, shape, margin, echogenicity, posterior acoustic features, and orientation, which vary significantly across patients [12]. Ultimately, they determine if the imaged findings are benign, need short-term follow-up imaging, or require a biopsy based on their suspicion of malignancy. There is considerable intra-reader variability in these recommendations and breast US has been criticized for increasing the number of false-positive findings [13, 14]. Compared to mammography alone, the addition of US in breast cancer screening leads to an additional 5-15% of patients being recalled for further imaging and an additional 4-8% of patients undergoing biopsy [15, 16, 17]. However, only 7-8% of biopsies prompted by screening US are found to identify cancers [15, 17].

Computer-aided diagnosis (CAD) systems have been proposed to assist radiologists in the interpretation of breast US exams over a decade ago [18]. Early CAD systems often relied on handcrafted visual features that are difficult to generalize across US images that were acquired using different protocols and US units [19, 20, 21, 22, 23, 24]. Recent advances in deep learning have facilitated the development of AI systems for the automated diagnosis of breast cancer from US images [25, 26, 27]. However, the majority of these efforts rely on image-level or pixel-level labels, which require experts to manually mark images containing visible lesions within each exam or annotate lesions in each image, respectively [28, 29, 30, 31, 32, 33]. As a result, existing studies have been based on small datasets consisting of several hundreds or thousands of US images. Deep learning models trained on those datasets might not sufficiently learn the diverse characteristics of US images observed in clinical practice. This is especially important for US imaging as lesion appearance can vary substantially depending on the imaging technique and the manufacturer of the US unit system. Moreover, prior research has primarily focused on differentiating between benign and malignant breast lesions, hence evaluating AI systems only on the images which contain either benign or malignant lesions [34, 35, 36]. In contrast, the majority of breast cancer screening exams are negative (no lesions are present) [7, 11]. In addition, most AI systems in previous studies do not interpret the model's predictions, resulting with "black-box" models [28, 29, 30, 31, 32, 33, 34, 35, 36]. So far, there has been little work on interpretable AI systems for breast US.

In this work, we present an AI system (Figure 1) to identify malignant lesions in breast US images with the primary goal of reducing the frequency of false positive findings. The AI system was trained to perform classification and localization in a weakly supervised manner [37, 38, 39]. That is, our AI system is able to explain its predictions by indicating locations of malignant lesions even though it is trained with binary breast-level cancer labels only (see Methods section 'Breast-level cancer labels'), which were automatically extracted from pathology reports. The explainability of our system enables clinicians to develop trust and better understand its strengths and limitations.

The proposed system provides several advances relative to previous work. First, to the best of our knowledge, the dataset used to train and evaluate this AI system is larger than any prior dataset used for this application [29, 40]. Second, to understand the potential value of this AI system in clinical practice, we conducted a reader study to compare its diagnostic accuracy with ten board-certified breast radiologists. The AI system achieved a higher area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) than the ten radiologists on average. Moreover, we showed that the hybrid model, which aggregates the predictions of the AI system and radiologists, improved radiologists' specificity and decreased biopsy rate while maintaining the same level of sensitivity. In addition, we showed that the performance of the AI system remained robust across patients from different age groups and mammographic breast densities. Accuracy of our system also remained high when tested on an external data set [40].

Results

Datasets. The AI system was developed and evaluated using the NYU Breast Ultrasound Dataset [41] consisting of 5,442,907 images within 288,767 breast US exams (including both screening and diagnostic exams) collected from 143,203 patients examined between 2012 and 2019 at NYU Langone Health in New York, USA. The NYU Langone hospital system spans multiple sites across New York City and Long Island, allowing the inclusion of a diverse patient population. The dataset included 28,914 exams associated with a biopsy, and among those, the biopsy yielded benign and malignant results for 26,843 and 5,593 breasts, respectively. Patients in the dataset were randomly divided into a training set (60%) that was used for model training, a validation set (10%) that was used for hyperparameter tuning, and an internal test set (30%) that was used for model evaluation. Each patient was included in only one of the three sets. We used a subset of the internal test set for the reader study. The statistics of the overall dataset, the internal test set, and the reader study set are summarized in Table 1.

Each breast within an exam was assigned a label indicating the presence of cancer using pathology results. The pathology examinations were conducted on tissues obtained using image-guided biopsy or surgical excision. As shown in Figure 1b, all cancer-positive exams were accompanied by at least one pathology report indicating malignancy collected either 30 days prior or 120 days after the US examination. This time frame was chosen to maximize the inclusion of both lesions found at primary screening US and lesions found during targeted US after an initial imaging workup with a different modality. We filtered the internal test set to ensure that cancers were visible on positive exams and that negative exams had either cancer-negative biopsy or at least one negative follow-up US exam (see Methods section 'Additional filtering of the test set'). Studies with neither a biopsy nor any negative follow-up were included in the training and validation set but excluded from the internal test set.

To assess the ability of the AI system to generalize across patient populations and image acquisition protocols, we further evaluated it on the public Breast Ultrasound Images (BUSI) dataset collected at the Hospital for Early Detection and Treatment of Women's Cancer in Cairo, Egypt [40]. This external test set consisted of 780 images, of which 437 were benign, 210 were malignant, and 133 were negative (no lesion present). These images were collected from 600 patients. Of note, the BUSI dataset was acquired using different US machines and was collected from patients with contrasting demographic backgrounds compared to the NYU dataset. Each image in the BUSI dataset was associated with a label indicating the presence of any malignant lesions.

AI system performance. On the internal test set of 44,755 US exams (25,003 patients, 79,156 breasts), the AI system achieved an AUROC of 0.976 (95% CI: 0.972, 0.980) in identifying breasts with malignant lesions. Additionally, we stratified patients by age, mammographic breast density, US machine manufacturer, and evaluated AI model performance across these sub-populations (Table 2). The AI system maintained high diagnostic accuracy among all age groups (AUROC: 0.969-0.981), mammographic breast densities (AUROC: 0.964-0.979), and US device manufacturers (AUROC: 0.974-0.990). In addition, we evaluated the AI system on the external test set (BUSI dataset) [40]. Even though the AI system was not trained on any images of the external test set, it maintained a high level of diagnostic accuracy (0.911 AUROC, 95% CI: 0.885, 0.933).

Reader study. To compare the performance of the AI system with that of breast radiologists, we constructed a reader study subset by selecting 663 exams (644 patients, 1,024 breasts) from the internal test set. Among the exams selected for this study, 73 breasts had biopsy-proven cancer, 535 breasts had a biopsy yielding exclusively benign findings, and 416 breasts were not biopsied but were evaluated by radiologists as likely benign and had a follow-up benign evaluation at 1-2 years. These proportions were chosen to increase the difficulty of the interpretation task and increase statistical power. Readers were informed that the study dataset was enriched with cancers but were not informed of the enrichment level.

Ten board-certified breast radiologists rated each breast according to the Breast Imaging Reporting and Data System (BI-RADS) [12]. Radiologists' experience is described in Table A.1. Readers were provided with contextual information typically available in the clinical setting, including the patient's age, burnt-in

Table 1: **Statistics of the overall NYU Breast Ultrasound Dataset, internal test set, and reader study set.** This dataset was collected from NYU Langone Health over an eight-year period. Exam-level BI-RADS were issued by radiologists based on patients' breast US exams. Breast densities were determined using existing screening and diagnostic mammography reports. Patients who were not matched with any mammograms were assigned "unknown" for breast density. Abbreviations: *N*, number; *SD*, standard deviation.

Characteristic, unit	**Overall**	**Internal test set**	**Reader study**
Patients, *N*	143,203	25,003	644
Age, mean years (*SD*)	53.7 (13.7)	55.5 (12.7)	52.8 (14.0)
<40 yrs old, *N* (%)	18,218 (12.7)	1,857 (7.4)	90 (14.0)
$40-49$ years old, *N* (%)	33,955 (23.7)	5,811 (23.2)	175 (27.2)
$50-59$ years old, *N* (%)	34,942 (24.4)	6,567 (26.3)	146 (22.7)
$60-69$ years old, *N* (%)	26,671 (18.6)	5,198 (20.8)	104 (16.1)
≥ 70 years old, *N* (%)	17,703 (12.4)	3,359 (13.4)	81 (12.6)
Exams, *N*	288,767	44,755	663
Images, *N*	5,442,907	858,636	13,582
Average no. of images per exam, *N*	18	19	20
Exams associated with biopsy, *N* (%)	28,914 (10.0)	8,337 (18.6)	587 (88.5)
Breasts, *N*	510,271	79,156	1,024
Breasts with benign findings, *N*	26,843	7,879	567
Breasts with malignant findings, *N*	5,593	1,324	73
Exam-level BI-RADS			
BI-RADS 0, *N* (%)	14,078 (4.9)	1,092 (2.4)	80 (12.1)
BI-RADS 1, *N* (%)	86,347 (29.9)	12,374 (27.6)	56 (8.4)
BI-RADS 2, *N* (%)	136,322 (47.2)	21,675 (48.4)	80 (12.1)
BI-RADS 3, *N* (%)	27,711 (9.6)	3,586 (8.0)	25 (3.8)
BI-RADS 4, *N* (%)	22,133 (7.7)	5,578 (12.5)	391 (59.0)
BI-RADS 5, *N* (%)	1,348 (0.5)	338 (0.8)	22 (3.3)
BI-RADS 6, *N* (%)	518 (0.2)	69 (0.2)	3 (0.5)
Unknown BI-RADS, *N* (%)	310 (0.1)	43 (0.1)	6 (0.9)
Exam-level mammographic density			
A (breasts are almost entirely fatty), *N* (%)	5,384 (1.9)	695 (1.6)	13 (2.0)
B (scattered areas of fibroglandular density), *N* (%)	69,948 (24.2)	11,048 (24.7)	143 (21.6)
C (breasts are heterogeneously dense), *N* (%)	165,855 (57.4)	26,509 (59.2)	376 (56.7)
D (the breasts are extremely dense), *N* (%)	31,829 (11.0)	5,189 (11.6)	76 (11.5)
Unknown density, *N* (%)	15,751 (5.5)	1,314 (2.9)	55 (8.3)

annotations showing measurements of suspicious findings, and notes from the technologist, such as specifying any region of palpable concern or pain. In contrast, the AI system was not provided any contextual information.

For each reader, we computed a receiver operating characteristic (ROC) curve and a precision-recall curve by comparing their BI-RADS scores to the ground-truth outcomes (see Methods section 'Statistical analysis'). The ten radiologists achieved an average AUROC of 0.924 (SD: 0.020, 95% CI: 0.905, 0.944) and an average AUPRC of 0.565 (SD: 0.072, 95% CI: 0.465, 0.625) (Figure A.1). Compared to the average radiologist in this study, the AI system achieved a higher AUROC of 0.962 (95% CI: 0.943, 0.979) with an AUROC improvement of 0.038 (95% CI: 0.028, 0.052, P<0.001) and a higher AUPRC of 0.752 (95% CI: 0.675, 0.849) with an AUPRC improvement of 0.187 (95% CI: 0.140, 0.256, P<0.001) (Figure 2). In addition, we also compared the specificity and sensitivity achieved by the AI system and radiologists. We assigned a positive prediction to any breast a radiologist gave a BI-RADS score of $\geq$4, and a negative prediction to any breast that was given a BI-RADS score of 1-3. A BI-RADS score of $\geq$4 is an assessment that indicates a radiologist thinks an exam is suspicious for malignancy. This was selected as the threshold for positive predictions since this is the score above which a patient will typically undergo an invasive procedure (biopsy or surgical excision) to definitively determine whether they have cancer [12]. With this methodology, the ten radiologists achieved an

Table 2: **AI performance on the internal test set across different sub-populations.** We reported the AUROC of the AI system with 95% confidence intervals on the internal test set. The biopsied population only includes exams where at least one biopsy was recommended. We stratified exams based on patient age, mammographic breast density, and the manufacturer of the US devices. Mammographic breast density was categorized based on the BI-RADS standards [42].

Population	AUROC (95% CI)	No. of breasts	No. of cancers
Overall population	0.976 (0.972, 0.980)	79,078	1,248
Biopsied population	0.940 (0.933, 0.945)	12,973	1,248
Age			
<40 yrs old	0.969 (0.955, 0.978)	5,176	72
$40-49$ yrs old	0.970 (0.960, 0.984)	19,677	160
$50-59$ yrs old	0.981 (0.970, 0.986)	24,142	292
$60-69$ yrs old	0.980 (0.975, 0.984)	19,039	326
≥ 70 yrs old	0.969 (0.963, 0.976)	11,044	398
Breast density			
Entirely fatty	0.964 (0.944, 0.978)	1,157	54
Scattered fibroglandular densities	0.975 (0.969, 0.984)	19,199	441
Heterogeneously dense	0.979 (0.977, 0.986)	47,255	610
Extremely dense	0.964 (0.938, 0.984)	9,398	90
Unkown	0.970 (0.953, 0.976)	2,069	53
Manufacturer			
General Electric	0.984 (0.974, 0.978)	5,708	47
Medison	0.990 (0.982, 0.984)	2,673	13
Philips	0.977 (0.972, 0.986)	28,943	412
Siemens	0.974 (0.965, 0.984)	37,572	699
Toshiba	0.986 (0.982, 0.976)	4,180	77
Other	-	2	0

average specificity of 80.7% (SD: 4.7%, 95% CI: 78.9%, 82.6%) and an average sensitivity of 90.1% (SD: 4.3%, 95% CI: 86.4%, 93.8%). At the average radiologist's specificity, the AI system achieved a sensitivity of 94.5% (95% CI: 89.4%, 100.0%) and an improvement in sensitivity of 4.4% (95% CI: -0.3%, 7.5%, P=0.0278). At the average radiologist's sensitivity, the AI system achieved a higher specificity of 85.6% (95% CI: 83.9%, 88.0%) with an absolute increase in specificity of 4.9% (95% CI: 3.0%, 7.1%; P<0.001). At the average radiologist's sensitivity, the AI system recommended tissue biopsies on 19.8% (95% CI: 17.9%, 22.1%) of breasts and 32.5% (95% CI: 26.9%, 39.2%) of these biopsies were for breasts ultimately found to have cancer. Compared to the average reader's biopsy rate of 24.3% (SD: 4.5%, 95% CI: 22.0%, 26.5%) and average PPV of 27.1% (SD: 4.1%, 95% CI: 22.9%, 33.1%), the AI system achieved an absolute reduction in biopsy rate of 4.5% (95% CI: 2.9%, 6.5%, P<0.001) which corresponds to 18.6% of all biopsies recommended by the average radiologist and achieved an absolute improvement in PPV of 5.4% (95% CI: 2.4%, 8.9%, P<0.001). The performance of the AI system and readers is summarized in Table A.2.

Subgroup analysis on the biopsied population. We conducted additional analyses on two clinically relevant subgroups in the reader study to understand the relative strengths of the AI system and radiologists. The first analysis examined diagnostic accuracy exclusively amongst breasts with lesions that had undergone biopsy evaluation (73 breasts with biopsy-confirmed malignant lesions and 535 breasts with exclusively biopsy-confirmed benign lesions). Breasts that yielded normal findings were not included. As expected, compared to the overall reader study population, AUROC (mean: 0.896, SD: 0.024, 95% CI: 0.874, 0.929) and specificity (mean: 69.8%, SD: 6.9%, 95% CI: 67.7%, 73.6%) of radiologists declined in this sub-population. Additionally, the average biopsy rate of radiologists increased to 37.4% (SD: 6.4%, 95% CI: 33.1%, 39.8%). On this subgroup, the AI system achieved an AUROC of 0.941 (95% CI: 0.922, 0.968). Compared to radiologists, the AI system demonstrated an absolute improvement of 8.5% (95% CI: 5.3%, 11.1%; P<0.001) in specificity,

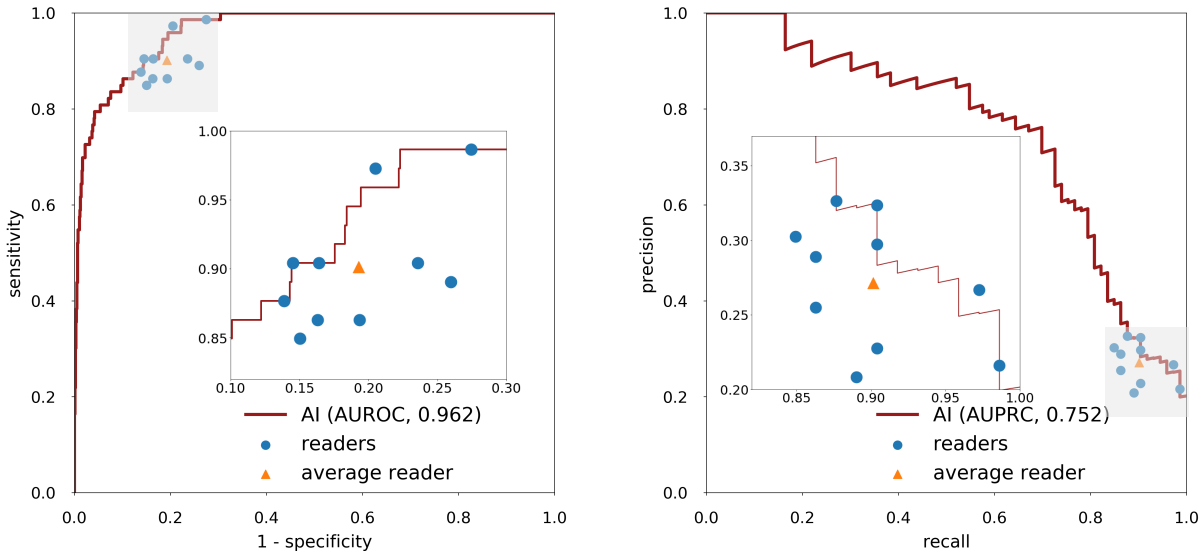

Figure 2: **Reader study results.** The performance of the AI system on the reader study population ($n =$ 1,024 breasts) using ROC curve (left) and precision-recall curve (right). The AI achieved 0.962 (95% CI: 0.943, 0.979) AUROC and 0.752 (95% CI: 0.675, 0.849) AUPRC. Each data point represents a single reader and the triangles correspond to the average reader performance. The inset shows a magnification of the grey shaded region.

an absolute reduction of 7.5% (95% CI: 4.4%, 9.6%, P<0.001) in biopsy rate, and an absolute improvement in PPV of 6.7% (95% CI: 3.0%, 9.8%, P<0.001), while matching the average radiologist's sensitivity. The performance of each reader is shown in Table A.3.

Next, we evaluated the accuracy of readers and the AI system exclusively amongst breasts with biopsy-confirmed cancers (97 malignant lesions across 73 breasts). As shown in Table A.4, we stratified malignant lesions by cancer subtype, histologic grade, and biomarker profile. This was done to further investigate the AI system's ability to discriminate between benign and malignant lesions. Certain types of breast cancers (such as high grade, triple biomarker negative cancers) may closely resemble benign masses (more likely to have oval/round shape and circumscribed margins, less likely to have posterior attenuation compared to other cancers) and are considered particularly difficult to characterize [43]. This analysis demonstrated that the sensitivity of the AI system was similar to that of the readers across all stratification categories. There were no significant differences in sub-populations of patients where the AI system had inferior performance.

Qualitative analysis of saliency maps. In an attempt to understand the AI system's potential utility as a decision support tool, we qualitatively assessed six studies using the AI's saliency maps. These saliency maps indicated where the system identified potentially benign and malignant lesions, and represent data that could be made available to radiologists (in addition to breast level predictions of malignancy) if the AI system were integrated into clinical practice. Figure 3a,b shows two 1.5cm irregularly shaped hypoechoic masses with indistinct margins, that ultimately underwent biopsy and were found to be invasive ductal carcinoma. All readers as well as the AI system correctly identified these lesions as being suspicious for malignancy. Figure 3c displays a small 7mm complicated cystic/solid nodule with a microlobulated contour, which 7 out of 10 readers as well as the AI system thought appeared benign. However, this lesion ultimately underwent biopsy and was found to be invasive ductal carcinoma. Figure 3d displays a 7mm superficial and palpable hypoechoic mass with surrounding echogenicity, that underwent biopsy and was found to be benign fat necrosis. However, the AI system as well as 9 out of 10 readers incorrectly thought this lesion was suspicious for malignancy, and recommended it undergo biopsy. Lastly, Figure 3e shows a small 7mm

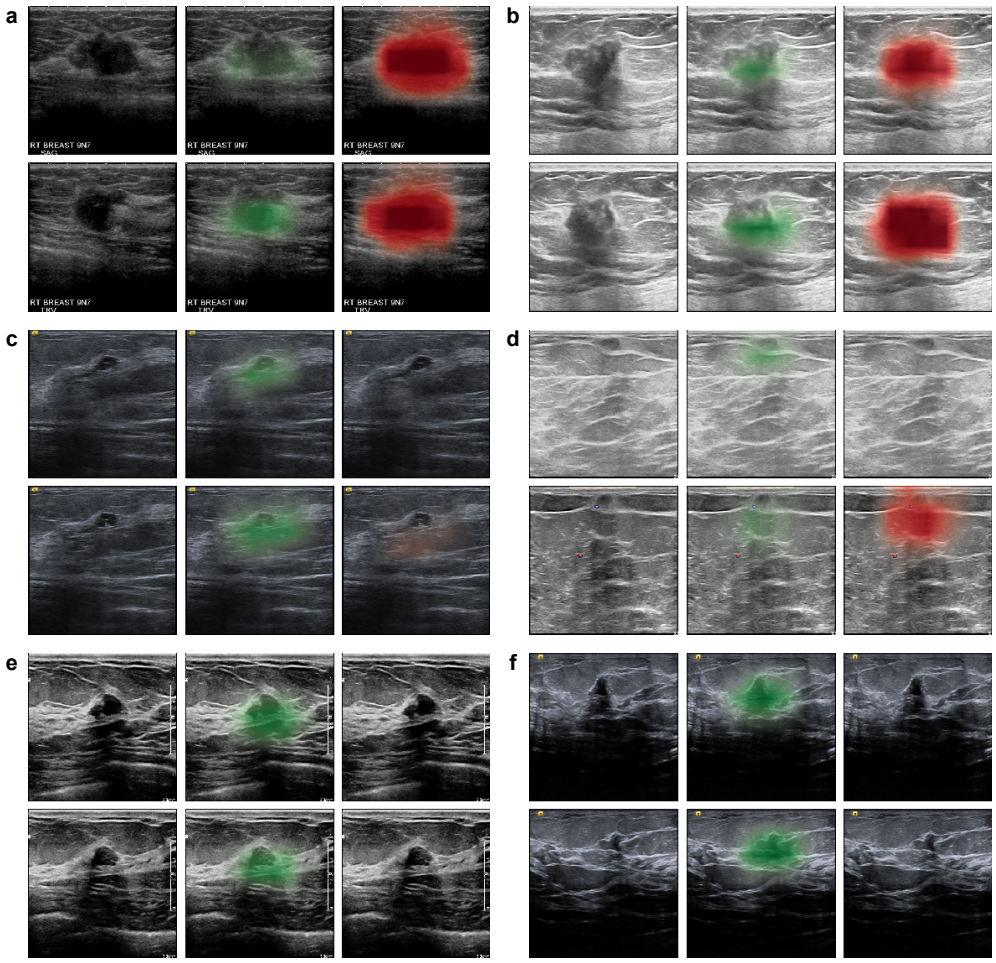

Figure 3: **Qualitative analysis of saliency maps**. In each of the six cases (**a-f**) from the reader study, we visualized the sagittal and transverse views of the lesion (left) and the AI's saliency maps indicating the predicted locations of benign (middle) and malignant (right) findings (see Methods section 'Deep neural network architecture'). Exams **a-c** display lesions that were ultimately biopsied and found to be malignant. All readers and the AI system correctly classified exams **a-b** as suspicious for malignancy. However, the majority of readers (7/10) and the AI system incorrectly classified case **c** as benign. Cases **d-f** display lesions that were biopsied and found to be benign. The majority of readers incorrectly classified exams **d** (9/10), **e** (10/10), and **f** (10/10) as suspicious for malignancy and recommended the lesions undergo biopsy. In contrast, the AI system classified exam **d** as malignant, but correctly identified exams **e-f** as being benign.

ill-defined area and Figure 3f displays a 9mm mildly heterogenous lobulated solid nodule. All 10 radiologists thought these two lesions appeared suspicious and recommended they undergo biopsy. In contrast, the AI system correctly classified the exams as benign, and the lesions were ultimately found to be benign fibrofatty tissue (Figure 3e) and a fibroadenoma (Figure 3f). Although we were unable to determine clear patterns among these US exams, the presence of cases where the AI system correctly contradicted the majority of readers and produced appropriate localization information underscores the potential complementary role the AI system might play in helping human readers more frequently reach accurate diagnoses.

Potential clinical applications. To evaluate the potential of our AI system to augment radiologists' diagnosis, we created hybrid models of the AI system and the readers. The predictions of each hybrid model

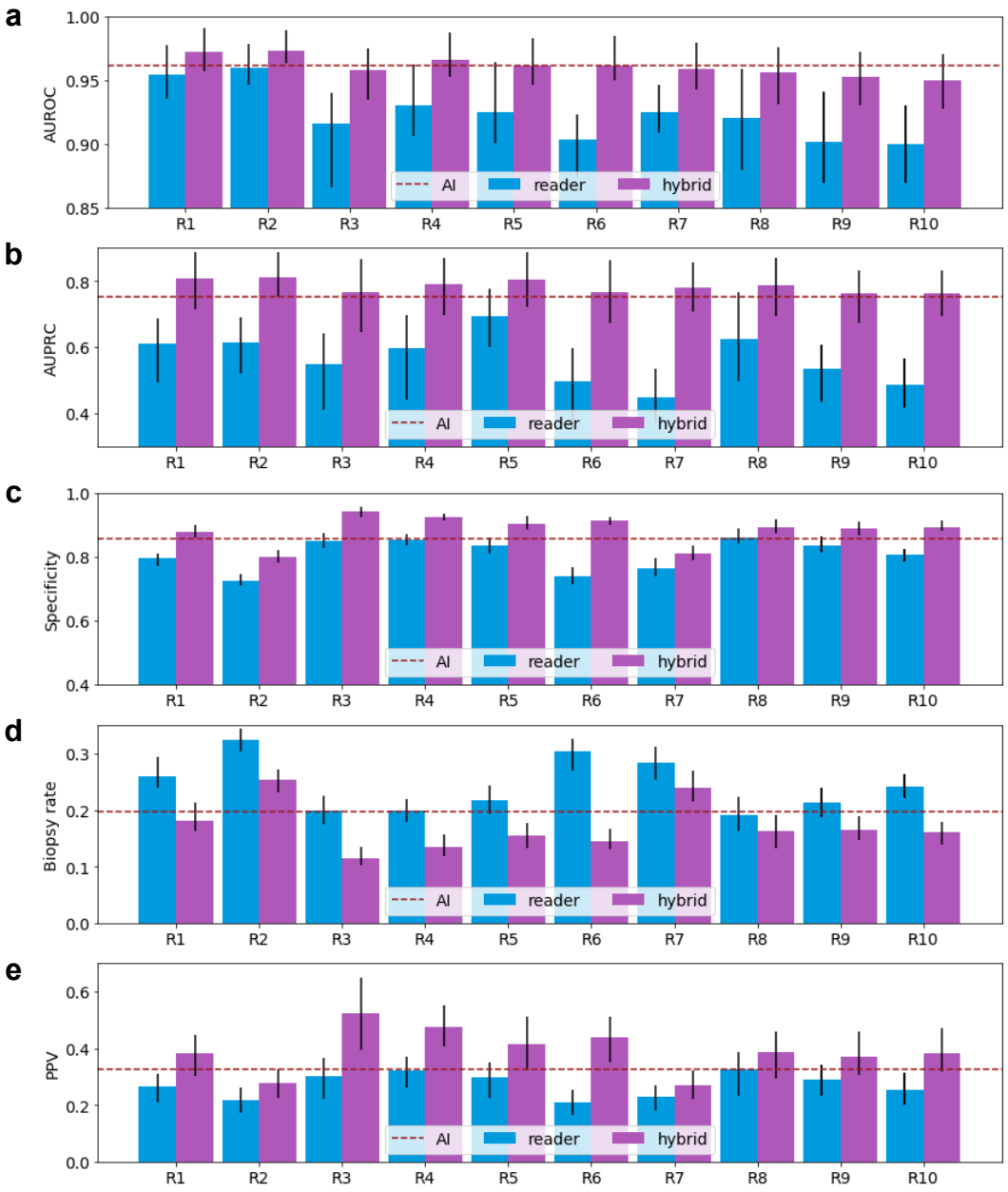

Figure 4: **Performance of readers, AI, and hybrid models.** We reported the values and 95% confidence intervals of AUROC (**a**), AUPRC (**b**), specificity (**c**), biopsy rate (**d**), and PPV (**e**) of ten radiologists (R1-R10), AI, and the hybrid models on the reader study set (n = 1,024 breasts). The predictions of each hybrid model are a weighted average of each reader's BI-RADS scores and the AI's probablistic predictions (see Methods section 'Hybrid model'). We dichotomized each hybrid model's probabilistic predictions to match the sensitivity of its respective reader. We dichotomized the AI's predictions to match the average radiologists' sensitivity. The collaboration between AI and readers improves readers' AUROC, AUPRC, specificity, and PPV, while reducing biopsy rate.

were computed as an equally weighted average between the AI system and each reader (see Methods section

'Hybrid model'). This analysis revealed that the performance of all readers was improved by incorporating the predictions of the AI system (Figure 4, Table A.5). On average, the hybrid models improved radiologists' AUROC by 0.037 (SD: 0.013, 95% CI: 0.011, 0.070, P<0.001) and improved their AUPRC by 0.219 (SD: 0.060, 95% CI: 0.089, 0.372, P<0.001). At the radiologists' sensitivity levels, the hybrid models increased their average specificity from 80.7% to 88.0% (average increase 7.3%, SD: 3.8%, 95% CI: 2.7%, 18.5%, P<0.001), increased their PPV from 27.1% to 38.0% (average increase 10.8%, SD: 5.3%, 95% CI: 3.7%, 25.0%, P<0.001), and decreased their average biopsy rate from 24.3% to 17.6% (average decrease, 6.8%, SD: 3.5%, 95% CI: 2.3%, 17.1%, P<0.001). The reduction in biopsies achieved by the hybrid model represented 27.8% of all biopsies recommended by radiologists.

In addition, the AI system could also be used to assist radiologists to triage US exams. To evaluate the potential of the AI system in identifying cancer-negative cases with high confidence, we selected a very low decision threshold to triage women into a no-radiologist work stream. On the reader study subset, using this triage paradigm, the AI system achieved an NPV of 99.86% while retaining a specificity of 77.7%. This result suggests that it may be feasible to dismiss 77.7% of normal/benign cases and skip radiologist review if we accept missing one cancer in every 740 negative predictions, which is less than 1/6 of the false negative rate observed among radiologists in the reader study (one missed cancer for every 109 negative evaluations). To evaluate the potential of the AI system in triaging patients into an enhanced assessment work stream, we used a very high decision threshold. In this enhanced assessment work stream, the AI system achieved a PPV of 84.4% while retaining a sensitivity of 52.1%. These results suggest that it may be feasible to rapidly prioritize more than half of cancer cases, with approximately five out of six biopsies leading to a diagnosis of cancer. For comparison, only 27.1% biopsies that the radiologists recommended were diagnosed with cancer. While we demonstrated the potential of AI in automatically triaging breast US exams, confirmation of these performance estimates would require extensive validation in a clinical setting.

Discussion

In this work, we present a radiologist-level AI system that is capable of automatically identifying malignant lesions in breast US images. Trained and evaluated on a large dataset collected from 20 imaging sites affiliated with a large medical center, the AI system maintained a high level of diagnostic accuracy across a diverse range of patients whose images were acquired using a variety of US units. By validating its performance on an external dataset, we produced preliminary results substantiating its ability to generalize across a patient cohort with different demographic composition and image acquisition protocols.

Our study has several strengths. First, in the reader study subset, we found that the AI system performed comparably to board-certified breast radiologists. The ten radiologists achieved an average sensitivity of 90.1% (SD: 4.3%, 95% CI: 86.4%, 93.8%) and an average specificity of 80.7% (SD: 4.7%, 95% CI: 78.9%, 82.6%). The sensitivity of radiologists in our study is consistent with the results reported in other breast US reader studies [10, 44], as well as the sensitivity of breast radiologists observed in clinical practice, despite the fact that radiologists in our study did not have access to the patient's medical record or prior breast imaging [15, 45]. Compared to radiologists in our reader study, the AI system was able to detect cancers with the same sensitivity, while obtaining a higher specificity (85.6%, 95% CI: 83.9%, 88%), a higher PPV (32.5%, 95% CI: 26.9%, 39.2%), and a lower biopsy rate (19.8%, 95% CI: 17.9%, 22.1%). Moreover, the AI system achieved a higher AUROC (0.962, 95% CI: 0.943, 0.979) and AUPRC (0.752, 95% CI: 0.675, 0.849) than all ten radiologists. This trend was confirmed in our subgroup analysis which showed that the system could accurately interpret US exams that are deemed difficult by radiologists.

Another strength of this study is that we explored the benefits of collaboration between radiologists and AI. We proposed and evaluated a hybrid diagnostic model that combined the predictions from radiologists and the AI system. The results from our reader study suggest that such collaboration improves the diagnostic accuracy and reduces false positive biopsies for all ten radiologists (Table A.5). In fact, breast US has come under criticism for having a high false positive rate [13, 14]. As reported by multiple clinical studies, only 7-8% of breast biopsies performed under US guidance are found to yield cancers [15, 17]. Indeed, for the ten radiologists in our cancer-enriched reader study subset, on average 19.3% (SD: 4.7%, 95% CI: 17.7%, 20.6%)

of cancer-negative exams were falsely diagnosed as positive and only 27.1% (SD: 4.1%, 95% CI: 22.9%, 33.1%) of the exams that they recommended to undergo biopsy actually had cancer. In this study, we showed that the hybrid models reduced the average radiologist's false positive rate to 12.0% (SD: 3.9%, 95% CI: 7.6%, 21.0%), representing a 37.4% (SD: 13.0%, 95% CI: 34.1%, 40.0%) relative reduction. The hybrid models also increased the average radiologist's PPV to 38.0% (SD: 6.0%, 95% CI: 24.1%, 50.0%). These results indicate that our AI system has the potential to aid radiologists in their interpretation of breast US exams to reduce the number of false positive interpretations and benign biopsies performed.

Beyond improving radiologists' performance, we also explored how AI systems could be utilized to assist radiologists to triage US exams. We showed that high-confidence operating points provided by the AI system can be used to automatically dismiss the majority of low-risk benign exams and escalate high-risk cases to an enhanced assessment stream (Table A.6). Prospective clinical studies will be required to understand the full extent to which this technology can benefit US reading.

Finally, we have made technical contributions to the methodology of deep learning for medical image analysis. Prior work on AI systems for interpreting breast US exams, and other similar applications, rely on manually collected image-level or pixel-level labels [28, 29, 30, 31, 32, 33]. In contrast, our AI system was trained using breast-level labels which were automatically extracted from pathology reports. This is an important difference, as developing a reliable AI system for clinical use requires training and validation on large-scale datasets to ensure the network will function well across the broad spectrum of cases encountered in clinical practice. At such a scale, it is impractical to collect labels manually. We address this issue by adopting the weakly supervised learning paradigm to train models at scale without the need for image-level or pixel-level labels. This paradigm enables the model to generate interpretable saliency maps that highlight informative regions in each image. With the saliency maps, researchers can perform qualitative error analysis and understand the strength and limitations of the AI system. Furthermore, an interpretable AI system trained with such a large dataset could help discover novel data-driven imaging biomarkers, leading to a better understanding of breast cancer.

Despite the contributions of our study in advancing breast cancer diagnosis, it has some limitations. We focused on the evaluation of an AI system that detects breast cancer only using US imaging. In clinical practice, US imaging is often used as a complementary modality to mammography. One promising research direction is to utilize multimodal learning [46, 47] to combine information from other imaging modalities. Moreover, the diagnosis produced by our AI system is based only on a single US exam, while breast radiologists often refer to patients' prior imaging to evaluate the morphological changes of suspicious findings over time. Future research could focus on augmenting AI systems to extract relevant information from past US exams.

Another limitation of this work is the design of reader study. To provide a fair comparison with the AI system, readers in our study were only provided with US images, patients' ages, and notes from the operating technician. In clinical practice, breast radiologists also have access to other information such as patients' prior breast imaging and their electronic medical records. Additionally, in the breast cancer screening setting, a screening US examination is typically accompanied by a screening mammogram. Even if prior US exams are not available, radiologists can typically refer to the mammogram for additional information, which can also influence the way that an US exam is interpreted. Finally, the qualitative analysis presented in this study was conducted over a limited set of exams. A systematic study on the differences between the AI system and the perception of radiologists in sonography interpretation is required to understand the limitations of such systems.

Despite these limitations, we believe this study is a meaningful contribution to the emerging field of AI-based decision support systems for interpreting breast US exams. On a clinically realistic population, our AI system achieved a higher diagnostic accuracy (AUROC: 0.976, 95% CI: 0.972, 0.980) than prior AI systems for breast US lesion classification (AUROC: 0.82-0.96) [32, 34, 48, 49, 50, 51, 52], though we acknowledge these systems can be compared only approximately as they were evaluated on different datasets. Key features that contributed to our AI system's high level performance were the large dataset used in training, along with utilization of the weakly supervised learning paradigm that enables the system to learn from automatically extracted labels. Furthermore, as our AI system was evaluated on a large test set (>44,000 US exams) acquired from a diverse range of US units and patients of diverse demographics, we are optimistic of its

ability to perform well prospectively, in the hands of radiologists. A few recent studies have demonstrated in retrospective reader studies that AI systems can improve the performance of radiologists when they have access to the decision support tool while reviewing US exams [48, 52]. However, these studies utilized an AI system that required radiologists to localize lesions by manually drawing bounding boxes. Moreover, these studies used small datasets and did not evaluate the AI's performance on sub-populations stratified by age and breast density. This makes it hard to determine if the system would maintain performance across the broad range of US exams that a radiologist might encounter in different clinical settings. Regardless of these limitations, these studies demonstrate that an AI system with a relatively low AUROC of 0.86-0.88 can substantially improve the diagnostic accuracy of radiologists. Based on these results, we are optimistic that our AI system, which does not require radiologists to localize lesions and achieved a higher diagnostic accuracy (AUROC: 0.976) on a larger diverse patient population, could enable radiologists to achieve even greater levels of performance. As a next step, our system requires prospective validation before it can be widely deployed in clinical practice. The potential impact that such a system could have on women's imaging is immense, given the enormous volume of women who undergo breast US exams each year.

In conclusion, we examined the potential of AI in US exam evaluation. We demonstrated in a reader study that deep learning models trained with a sufficiently large amount of data are able to produce diagnosis as accurate as experienced radiologists. We further showed that the collaboration between AI and radiologists can significantly improve their specificity and obviate 27.8% of requested biopsies. We believe this research could supplement future approaches to breast cancer diagnosis. In addition, the general approach employed in our work, mainly the framework for weakly supervised classification and localization, may enable utilization of deep learning in similar medical image analysis tasks.

References

[1] H. Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a Cancer Journal for Clinicians* (2021).

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal. "Cancer Statistics, 2021." In: *CA: a Cancer Journal for Clinicians* 71.1 (2021), pp. 7–33.

[3] E. K. Arleo, R. E. Hendrick, M. A. Helvie, and E. A. Sickles. "Comparison of recommendations for screening mammography using CISNET models". In: *Cancer* 123.19 (2017), pp. 3673–3680.

[4] S. Feig. "Cost-effectiveness of mammography, MRI, and ultrasonography for breast cancer screening". In: *Radiologic Clinics* 48.5 (2010), pp. 879–891.

[5] T. M. Kolb, J. Lichy, and J. H. Newhouse. "Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations". In: *Radiology* 225.1 (2002), pp. 165–175.

[6] N. F. Boyd et al. "Mammographic density and the risk and detection of breast cancer". In: *New England Journal of Medicine* 356.3 (2007), pp. 227–236.

[7] W. A. Berg et al. "Ultrasound as the primary screening test for breast cancer: analysis from ACRIN 6666". In: *Journal of the National Cancer Institute* 108.4 (2016), djv367.

[8] P. J. Dempsey. "The history of breast ultrasound". In: *Journal of Ultrasound in Medicine* 23.7 (2004), pp. 887–894.

[9] M. Chung et al. "US as the primary imaging modality in the evaluation of palpable breast masses in breastfeeding women, including those of advanced maternal age". In: *Radiology* 297.2 (2020), pp. 316–324.

[10] R. Sood et al. "Ultrasound for breast cancer detection globally: a systematic review and meta-analysis". In: *Journal of Global Oncology* 5 (2019), pp. 1–17.

[11] W. A. Berg et al. "Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer". In: *JAMA* 299.18 (2008), pp. 2151–2163.

[12] E. A. Sickles et al. "ACR BI-RADS® Atlas, Breast imaging reporting and data system". In: *Reston, VA: American College of Radiology* (2013), pp. 39–48.

[13] P. Crystal, S. D. Strano, S. Shcharynski, and M. J. Koretz. "Using sonography to screen women with mammographically dense breasts". In: *American Journal of Roentgenology* 181.1 (2003), pp. 177–182.

[14] E. Lazarus, M. B. Mainiero, B. Schepps, S. L. Koelliker, and L. S. Livingston. "BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value". In: *Radiology* 239.2 (2006), pp. 385–391.

[15] L. Yang et al. "Performance of ultrasonography screening for breast cancer: a systematic review and meta-analysis". In: *BMC Cancer* 20.1 (2020), pp. 1–15.

[16] W. A. Berg et al. "Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk". In: *JAMA* 307.13 (2012), pp. 1394–1404.

[17] V. Corsetti et al. "Evidence of the effect of adjunct ultrasound screening in women with mammography-negative dense breasts: interval breast cancers at 1 year follow-up". In: *European Journal of Cancer* 47.7 (2011), pp. 1021–1026.

[18] D.-R. Chen and Y.-H. Hsiao. "Computer-aided diagnosis in breast ultrasound". In: *Journal of Medical Ultrasound* 16.1 (2008), pp. 46–56.

[19] W.-C. Shen, R.-F. Chang, W. K. Moon, Y.-H. Chou, and C.-S. Huang. "Breast ultrasound computer-aided diagnosis using BI-RADS features". In: *Academic Radiology* 14.8 (2007), pp. 928–939.

[20] J.-H. Lee et al. "Fourier-based shape feature extraction technique for computer-aided b-mode ultrasound diagnosis of breast tumor". In: *Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2012, pp. 6551–6554.

[21] J. Ding, H.-D. Cheng, J. Huang, J. Liu, and Y. Zhang. "Breast ultrasound image classification based on multiple-instance learning". In: *Journal of Digital Imaging* 25.5 (2012), pp. 620–627.

[22] L. Bing and W. Wang. "Sparse representation based Multi-Instance learning for breast ultrasound image classification". In: *Computational and Mathematical Methods in Medicine* 2017 (2017).

[23] T. Prabhakar and S. Poonguzhali. "Automatic detection and classification of benign and malignant lesions in breast ultrasound images using texture morphological and fractal features". In: *2017 10th Biomedical Engineering International Conference (BMEiCON)*. IEEE. 2017, pp. 1–5.

[24] Q. Zhang, J. Suo, W. Chang, J. Shi, and M. Chen. "Dual-modal computer-assisted evaluation of axillary lymph node metastasis in breast cancer patients on both real-time elastography and B-mode ultrasound". In: *European Journal of Radiology* 95 (2017), pp. 66–74.

[25] Y. Gao, K. J. Geras, A. A. Lewin, and L. Moy. "New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence". In: *American Journal of Roentgenology* 212.2 (2019), pp. 300–307.

[26] K. J. Geras, R. M. Mann, and L. Moy. "Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives". In: *Radiology* 293.2 (2019), pp. 246–259.

[27] T. Fujioka et al. "The utility of deep learning in breast ultrasonic imaging: a review". In: *Diagnostics* 10.12 (2020), p. 1055.

[28] J.-Z. Cheng et al. "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans". In: *Scientific Reports* 6.1 (2016), pp. 1–13.

[29] M. H. Yap et al. "Automated breast ultrasound lesions detection using convolutional neural networks". In: *IEEE Journal of Biomedical and Health Informatics* 22.4 (2017), pp. 1218–1226.

[30] W. Al-Dhabyani, M. Gomaa, H. Khaled, and F. Aly. "Deep learning approaches for data augmentation and classification of breast masses using ultrasound images". In: *International Journal of Advanced Computer Science and Applications* 10.5 (2019).

[31] E. Fleury and K. Marcomini. "Performance of machine learning software to classify breast lesions using BI-RADS radiomic features on ultrasound images". In: *European Radiology Experimental* 3.1 (2019), p. 34.

[32] H. Tanaka, S.-W. Chiu, T. Watanabe, S. Kaoku, and T. Yamaguchi. "Computer-aided diagnosis system for breast ultrasound images using deep learning". In: *Physics in Medicine & Biology* 64.23 (2019), p. 235013.

[33] Z. Cao, L. Duan, G. Yang, T. Yue, and Q. Chen. "An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures". In: *BMC Medical Imaging* 19.1 (2019), p. 51.

[34] S. Han et al. "A deep learning framework for supporting the classification of breast lesions in ultrasound images". In: *Physics in Medicine & Biology* 62.19 (2017), p. 7714.

[35] A. S. Becker et al. "Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study". In: *British Journal of Radiology* 91 (2018), p. 20170576.

[36] T. Xiao et al. "Comparison of transferred deep neural networks in ultrasonic breast masses discrimination". In: *BioMed Research International* 2018 (2018).

[37] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. "Is object localization for free?-weakly-supervised learning with convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 685–694.

[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2921–2929.

[39] Z.-H. Zhou. "A brief introduction to weakly supervised learning". In: *National Science Review* 5.1 (2018), pp. 44–53.

[40] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. "Dataset of breast ultrasound images". In: *Data in Brief* 28 (2020), p. 104863.

[41] F. Shamout et al. *The NYU Breast Ultrasound Dataset v1.0*. Tech. rep. Available at `https://cs.nyu.edu/~kgeras/reports/ultrasound_datav1.0.pdf`. 2021.

[42] L. Liberman and J. H. Menell. "Breast imaging reporting and data system (BI-RADS)". In: *Radiologic Clinics* 40.3 (2002), pp. 409–430.

[43] H.-Y. Du, B.-R. Lin, and D.-P. Huang. "Ultrasonographic findings of triple-negative breast cancer". In: *International Journal of Clinical and Experimental Medicine* 8.6 (2015), p. 10040.

[44] A. Ciritsis et al. "Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making". In: *European Radiology* 29.10 (2019), pp. 5458–5468.

[45] N. Houssami, S. Ciatto, L. Irwig, J. Simpson, and P. Macaskill. "The comparative sensitivity of mammography and ultrasound in women with breast symptoms: an age-specific analysis". In: *The Breast* 11.2 (2002), pp. 125–130.

[46] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. "Multimodal machine learning: A survey and taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2018), pp. 423–443.

[47] T. Zhou, S. Ruan, and S. Canu. "A review: Deep learning for medical image segmentation using multi-modality fusion". In: *Array* 3 (2019), p. 100004.

[48] L. Barinov et al. "Impact of data presentation on physician performance utilizing artificial intelligence-based computer-aided diagnosis and decision support systems". In: *Journal of Digital Imaging* 32.3 (2019), pp. 408–416.

[49] F. Dong et al. "One step further into the blackbox: a pilot study of how to build more confidence around an AI-based decision system of breast nodule assessment in 2D ultrasound". In: *European Radiology* (2021), pp. 1–10.

[50] C. Zhao et al. "Reducing the number of unnecessary biopsies of US-BI-RADS 4a lesions through a deep learning method for residents-in-training: a cross-sectional study". In: *BMJ Open* 10.6 (2020), e035757.

[51] T. Fujioka et al. "Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network". In: *Japanese Journal of Radiology* 37.6 (2019), pp. 466–472.

[52] V. L. Mango, M. Sun, R. T. Wynn, and R. Ha. "Should We Ignore, Follow, or Biopsy? Impact of Artificial Intelligence Decision Support on Breast Ultrasound Lesion Assessment". In: *American Journal of Roentgenology* 214.6 (2020), pp. 1445–1452.

Methods

Ethical approval. This retrospective study was approved by the NYU Langone Health Institutional Review Board (ID#i18-00712_CR3) and is compliant with the Health Insurance Portability and Accountability Act. Informed consent was waived since the study presents no more than minimal risk. This study is reported following the TRIPOD guidelines [53].

NYU breast ultrasound dataset. The dataset used in this study was collected from NYU Langone Health system (New York, USA) across 20 imaging sites. The final dataset contained 288,767 exams (5,442,907 images) acquired from 143,203 patients imaged between January 2012 and September 2019. Each US exam included between 4 and 70 images with 18.8 images per exam on average (Figure A.2a). The images had an average resolution of 665×603 pixels in width and height, respectively (Figure A.2b). A summary of the acquisition devices is shown in Table A.7. Each exam was associated with additional patient metadata as well as a radiology report summarizing the findings. We extracted breast tissue density from the patients' past mammography reports and assigned "unknown" to patients who did not have any mammography exams. Both screening and diagnostic US exams were included. Screening exams are performed for women who have no symptoms or signs of breast cancer while diagnostic US exams can be used to evaluate women who present with symptoms such as a new lump or pain in the breast or can be used to further evaluate abnormalities detected on a screening examination. While screening exams are typically comprehensive and image both breasts, diagnostic US exams vary in terms of how targeted they are, and might image both breasts, one breast, or sometimes just a single lesion. The dataset was filtered as described in the next section. Further details can be found in the technical report [41].

Filtering of the dataset. We initially extracted a dataset of 425,506 breast US exams consisting of 8,448,978 images collected from 212,716 unique patients. We then applied a few levels of filtering to obtain the final dataset for training and evaluating the neural network. This entailed the exclusion of exams with invalid patient identifiers, exams collected before 2012, exams collected from patients younger than 16 years of age, duplicate images, exams from non-female patients, and invalid images based on the `ImageType` attribute, which consisted of non-US images such as reports or demographic data screenshots. We further excluded images that were collected during biopsy procedures based on the `PerformedProcedureStepDescription`, `StudyDescription` & `RequestedProcedureDescription` attributes of the image metadata, in that order, images with missing metadata information relating to the type of procedure, images with more than 80% zero pixels, exams with multiple patient identifiers or study dates, exams with an extreme number of images, and exams with missing image laterality.

Patients were then randomly split among training (60%), validation (10%) and test (30%) sets. After splitting, each patient appeared in only one of the training, validation, and test sets. The training set consisted of 3,930,347 images within 209,162 exams collected from 101,493 patients. The validation set consisted of 653,924 images within 34,850 exams collected from 16,707 patients. The test set consisted of 858,636 images within 44,755 exams collected from 25,003 patients. The training set was used to optimize learnable parameters in the models. The validation set was used to tune the hyperparameters and select the best models. The test set was used to evaluate the performance of the models selected using the validation set. We applied additional filtering on the test set as described in the next section.

Additional filtering of the test set. To provide a clinically realistic evaluation of the AI system, we additionally refined the test set using the steps summarized in Figure A.3. First, we ensured that each non-biopsied exam was followed with a subsequent cancer-negative exam. Non-biopsied patients who had a negative (BI-RADS 1) or benign (BI-RADS 2) US exams were only included in the test set if they did not have any malignant breast pathology found within 0-15 months following their US exam, and had follow up imaging between 6 and 24 months that was also negative or benign (BI-RADS 1-2). Patients who did not undergo biopsy and had probably benign US exams (BI-RADS 3) were included in the test set if they did not have any malignant breast pathology found within 0-15 months following their exam, and met one of two

additional criteria: all of their subsequent US exams in the 4-36 months following their initial US exam were BI-RADS 1-2, or they had at least one follow-up US exam at 24-36 months which was evaluated as BI-RADS 1-3.

Next, we refined exams with biopsy-proven benign findings to determine if the pathology results were deemed by the radiologist to be concordant or discordant with the imaging features of the breast lesion. Patients with biopsy reports that confirmed a discordant benign finding were only included in the test set if they received a subsequent biopsy (that was not discordant) or breast surgery within the 6 months following the initial discordant biopsy. Patients with benign discordant biopsies that did not receive subsequent pathological evaluation were excluded.

Lastly, we ensured that exams with biopsy-proven cancers contained images of these cancers. Since breast US produces small images which do not comprehensively capture the entire breast, a proportion of patients diagnosed with breast cancer did not have images of the cancer in any of their US images. US exams with a label indicating malignancy and a BI-RADS score of 1-2 were excluded as these exams typically did not contain images of the cancer. Additionally, patients diagnosed with breast cancer who did not have any breast pathology obtained using US-guided biopsy were also excluded, since the majority of patients diagnosed using MRI and stereotactic-guided biopsies had malignancies that were sonographically occult. US exams that received a BI-RADS score of 0, 3, and 6, as well as patients who had breast pathology obtained using multi-modal image guidance (US plus stereotactic and/or MRI guided biopsies) had their cases manually reviewed to confirm that breast cancer was visible on the US exam. Patients who were given a BI-RADS score of 4-5 and had all their breast pathology obtained using US-guided biopsy were presumed to have visible cancers and were not manually reviewed.

Breast-level cancer labels. Among all the exams in the dataset, 28,914 exams (approximately 10%) were associated with at least one biopsy performed within 30 days prior or 120 days after the US examination. The cancer labels of biopsies were determined using their associated pathology reports. In cases where there were multiple pathology reports recorded within the considered time window, all of these reports were evaluated. Malignant findings included primary breast cancers: invasive ductal carcinoma, invasive lobular carcinoma, special-type invasive carcinoma (including tubular, mucinous and cribriform carcinomas), inflammatory carcinoma, intraductal papillary carcinoma, microinvasive carcinoma, ductal carcinoma in situ, as well as non-primary breast cancers: lymphoma and phyllodes. Benign findings included cyst, fibroadenoma, scar, sclerosing adenosis, lobular carcinoma in situ, columnar cell changes, atypical lobular hyperplasia, atypical ductal hyperplasia, papilloma, periductal mastitis and usual ductal hyperplasia. The labels were automatically extracted from the corresponding pathology reports using a natural language processing pipeline developed earlier [41]. Of note, patients with multiple pathology reports could be assigned both malignant and benign labels if their exam contained both types of lesions.

Breast Ultrasound Images Dataset. This external dataset was collected in 2018 from Baheya Hospital for Early Detection and Treatment of Women's Cancer (Cairo, Egypt) with the LOGIQ E9 ultrasound system and the LOGIQ E9 Agile ultrasound. It included 780 breast US images, with an average resolution of 500×500 pixels, acquired from 600 female patients whose ages ranged between 25 and 75 years old. Among these 780 images, 133 were normal images without cancerous masses, 437 were images containing malignant masses and 210 were images with benign masses. We refer the reader to the original paper for more information about this public dataset [40].

Deep neural network architecture. We present a deep learning model (DLM) whose architecture is shown in Figure A.4. To explain the mechanics of this model, we need to introduce some notation. Let $\mathbf{x} \in \mathbb{R}^{H,W,3}$ denote an RGB US image with a resolution of $H \times W$ pixels and let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K\}$ denote an *image set* that contains all images acquired from the patient during an US exam from one breast. This DLM is trained to process the image set $\mathbf{X}$, which may vary in number of the images it contains (Figure A.2), and generate two probability estimates $\hat{y}^b$, $\hat{y}^m \in [0, 1]$ that indicate the predicted probability of the presence of benign and malignant lesions in the patient's breast, respectively. The DLM is designed to resemble the

diagnostic procedure performed by radiologists. First, it generates saliency maps and probability estimates for each image $\mathbf{x}_k$ in the image set. This step is similar to a radiologist roughly scanning through each US image and looking for abnormal findings. Then it computes a set of attentions scores which indicate the importance of each image to the cancer diagnosis task. This procedure can be seen as an analogue to a radiologist concentrating on images that contain suspicious lesions. Finally, it forms a breast-level cancer diagnosis by combining information collected from all images. This is analogous to modelling a radiologist comprehensively considering signals in all images to render a full diagnosis. Below we describe each step in detail.

1) *Saliency maps.* The DLM first utilizes a convolutional neural network [54] f_g (parameterized as ResNet-18 [55]) to extract a representation of each image $\mathbf{x}_k$, in an image set $\mathbf{X}$, denoted by $\mathbf{h}_k \in \mathbb{R}^{h,w,C}$. The height, the width, and the number of channels are denoted by h, w, and C, respectively. Inspired by Zhou et al. [38], we then apply a convolutional layer with 1×1 convolutional filters followed by sigmoid non-linearity to transform $\mathbf{h}_k$ into two saliency maps $\mathbf{A}_k^b \in \mathbb{R}^{h,w}$ and $\mathbf{A}_k^m \in \mathbb{R}^{h,w}$. These saliency maps highlight approximate locations of benign and malignant lesions in each image. Each element $\mathbf{A}_k^b[i,j], \mathbf{A}_k^m[i,j] \in [0,1]$ denotes the contribution of spatial location (i,j) towards predicting the presence of benign/malignant lesions. The resolutions of the saliency maps (h,w) depends on the implementation of f_g. The sizes (h,w) are usually smaller than the resolution of the input image (H,W). In this work, we set $h = w = 8$, $C = 512$, and $H = W = 256$.

2) *Attention scores.* The images in the image set $\mathbf{X}$ might significantly differ in how relevant each of them is to the classification task. To address this issue, we utilize the Gated Attention Mechanism [56], allowing the model to select which information to incorporate from all images. Specifically, we first apply global max pooling to transform the representation $\mathbf{h}_k$ computed for the image $\mathbf{x}_k$ into a vector $\mathbf{v}_k \in \mathbb{R}^C$. Two attention scores α_k^b and $\alpha_k^m \in [0,1]$ that indicate the importance of each image $\mathbf{x}_k$ to the estimation of the probability of the presence of benign and malignant findings in the breast are computed as

$$\boldsymbol{\alpha}_k = \frac{\exp\{\mathbf{W}^\intercal(\tanh(\mathbf{V}\mathbf{v}_k^\intercal) \odot \mathrm{sigm}(\mathbf{U}\mathbf{v}_k^\intercal))\}}{\sum_{j=1}^{K} \exp\{\mathbf{W}^\intercal(\tanh(\mathbf{V}\mathbf{v}_j^\intercal) \odot \mathrm{sigm}(\mathbf{U}\mathbf{v}_j^\intercal))\}}, \tag{1}$$

where $\boldsymbol{\alpha}_k = \begin{bmatrix} \alpha_k^b \\ \alpha_k^m \end{bmatrix}$ denotes the concatenation of attention scores for both benign and malignant findings, $\odot$ denotes an element-wise multiplication, and $\mathbf{W} \in \mathbb{R}^{L,2}$, $\mathbf{V} \in \mathbb{R}^{L \times M}$ and $\mathbf{U} \in \mathbb{R}^{L \times M}$ are matrices of learnable parameters. In all experiments, we set $L = 512$ and $M = 128$.

3) *Cancer diagnosis.* Lastly, the DLM aggregates the information from all US images in the image set $\mathbf{X}$ and generates the final diagnosis using the attention scores and saliency maps. We first use an aggregation function $f_{\mathrm{agg}}(\mathbf{A}) : \mathbb{R}^{h,w} \mapsto [0,1]$ to transform the saliency maps into image-level predictions:

$$\hat{y}_k^b = f_{\mathrm{agg}}(\mathbf{A}_k^b) \quad \hat{y}_k^m = f_{\mathrm{agg}}(\mathbf{A}_k^m). \tag{2}$$

In our work, we parameterize f_{agg} as the *top* t% *pooling* proposed by Shen et al. [57]. Namely, we define the aggregation function as

$$f_{\mathrm{agg}}(\mathbf{A}) = \frac{1}{|H^+|} \sum_{(i,j) \in H^+} \mathbf{A}_{i,j}, \tag{3}$$

where H^+ denotes the set containing locations of top t% values in $\mathbf{A}$, and t is a hyperparameter. The breast-level cancer prediction $\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}^b \\ \hat{y}^m \end{bmatrix}$ is then defined as the average of all image-level cancer predictions weighted by the attention scores:

$$\hat{y}^b = \sum_k^K \alpha_k^b \hat{y}_k^b, \quad \hat{y}^m = \sum_k^K \alpha_k^m \hat{y}_k^m. \tag{4}$$

Training details. In order to constrain the saliency maps to only highlight important regions, we impose the L_1 regularization on $\mathbf{A}$ which penalizes the DLM for highlighting irrelevant pixels:

$$L_{\text{reg}}(\mathbf{A}) = \sum_{(i,j)} |\mathbf{A}[i,j]|. \tag{5}$$

Despite the relative complexity of our proposed framework, this DLM can be trained end-to-end using stochastic gradient descent with the following loss function, defined for a single training example (i.e. one breast) as

$$L(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{c \in \{b,m\}} \text{BCE}(y^c, \hat{y}^c) + \beta \sum_{k=1}^{K} L_{\text{reg}}(\mathbf{A}_k^c), \tag{6}$$

where BCE is the binary cross-entropy and β is a hyperparameter. For all experiments, the training loss is optimized using Adam [58]. Of note, labels indicating the presence of benign lesions (y^b) were also used during training to regularize the network through multi-task learning [59]. On the test set, we focus on evaluating predictions of malignancy since it is a more clinically relevant task: identification of malignant lesions has an immediate and significant impact on patient management (biopsy, potential surgery), whereas identification of a benign breast lesions typically does not alter management compared to patients without breast lesions [12].

We optimized the hyperparameters with random search [60]. Specifically, we searched for the learning rate $\eta \in 10^{[-5.5,-4]}$ on a logarithmic scale, regularization hyperparameter $\beta \in 10^{[-3,0.5]}$ on a logarithmic scale, weight decay hyperparameter $\lambda \in 10^{[-6,-3.5]}$ on a logarithmic scale, and the pooling threshold $t \in [0.1, 0.5]$ on a linear scale. We trained 30 separate models using hyperparameters uniformly sampled from the ranges above. Each model was trained for 50 epochs. We saved the model weights from the training epoch that achieves the highest AUROC on the validation set. To further improve our results, we used model ensembling [61]. Specifically, we average the breast-level predictions of the top 3 models that achieved the highest AUROC on the validation set to produce the overall prediction of the ensemble.

During training, we adopt image augmentation including random horizontal flipping (p=0.5), random rotation (-45° to 45°), random translation in both horizontal and vertical directions (up to 10% of the image size), scaling by a random factor between 0.7 and 1.5, and random shearing (-25° to 25°). The resulting image was then resized to 256×256 pixels using bilinear interpolation and normalized. During the validation and test stages, the original image was resized and normalized without any augmentation.

Reader study. We performed a reader study to compare the performance of the proposed DLM with breast radiologists. This study included ten board-certified breast radiologists with an average of 15 years of clinical experience (Table A.1). Their experience ranged from 3 to 40 years. Nine of the ten radiologists were fellowship-trained in breast imaging. The one radiologist who did not receive formal fellowship training (R10) worked as a sub-specialized breast radiologist and had over 30 years of breast imaging experience. The readers were provided with US images including metadata (breast laterality, position of the probe, notes from the sonographer) and the age of the patient. For each breast in all exams, the readers were then asked to provide a diagnostic BI-RADS score using the values 1, 2, 3, 4A, 4B, 4C or 5. A score of 0 was not permitted.

Hybrid model. To explore the potential benefit that the AI system might be able to provide, we created a hybrid model for each radiologist, whose predictions were created by averaging the predictions of the respective radiologist and the AI model: $\hat{\mathbf{y}}_{\text{hybrid}} = \lambda \hat{\mathbf{y}}_{\text{expert}} + (1-\lambda)\hat{\mathbf{y}}_{\text{AI}}$. The BI-RADS scores of radiologists were used as their predictions. Both $\hat{\mathbf{y}}_{\text{AI}}$ and $\hat{\mathbf{y}}_{\text{expert}}$ were standardized to have zero mean and unit variance. In this study, we set $\lambda = 0.5$. We note that $\lambda = 0.5$ is not the optimal value. On the other hand, the performance obtained by retroactively fine-tuning λ on the reader study is not transferable to realistic clinical settings. Therefore, we chose $\lambda = 0.5$ as the most natural way of aggregating two predictions without prior knowledge of their quality.

Statistical analysis. In this study, we evaluated the performance of the AI system, radiologists, and the hybrid models using the following evaluation metrics: area under receiver operating characteristic curves (AUROC), area under precision-recall curve (AUPRC), sensitivity, specificity, biopsy rate, negative predictive value (NPV), and positive predictive value (PPV). AUROC and AUPRC were used to assess the diagnostic accuracy of the probabilistic predictions generated by the AI system/hybrid models and the BI-RADS scores of the readers. The BI-RADS scores were treated as a 6-point index of suspicion for malignancy: scores of 1 and 2 were collapsed into the lowest category of suspicion; scores 3, 4A, 4B, 4C and 5 were treated independently as increasing levels of suspicion. AUROC avoids the subjectivity in selecting the thresholds to dichotomize continuous predictions, since it compares performance across all possible recall rates. However, AUROC weights omission and commission errors equally and therefore could provide excessively optimistic estimates in extremely imbalanced classification tasks such as cancer diagnosis where the negative cases often overwhelm the positive cases [62]. Therefore, to complement AUROC, we also reported AUPRC which solely evaluates the ability to correctly identify the positive cases. We calculated both AUROC and AUPRC using the Python Scikit-learn API [63].

In addition, we also evaluated the binary predictions of the AI system, the hybrid models, and the readers using sensitivity, specificity, biopsy rate, NPV, and PPV. These metrics are commonly used to assess the diagnostic accuracy in clinical studies [7, 11, 15]. The PPV reported in this study corresponds to PPV_2, which is defined as the number of breasts with cancer that were recommended to undergo biopsy divided by the total number of breast biopsies recommended [12]. For each breast, the AI system and the hybrid models produced a probabilistic score that represents the likelihood of cancer being present. We dichotomized these scores to produce binary predictions by selecting a score threshold that separates positive and negative decisions. To compute sensitivity, we dichotomized the AI system's probabilistic predictions to match average reader's specificity. To calculate the specificity, biopsy rate, PPV and NPV, we dichotomized the AI system's probabilistic predictions by matching the average reader's sensitivity. We similarly dichotomized the predictions of each hybrid model using the sensitivity/specificity of its respective reader. For all evaluation metrics, we estimated the confidence intervals at 95% by 1,000 iterations of the bootstrap method [64].

In the reader study, we compared the AUROC, AUPRC, sensitivity, specificity, PPV, and biopsy rate of the AI system and hybrid models with those of the average radiologists. The confidence interval for these differences was obtained through 1,000 iterations of bootstrap method [64]. The p-values were computed using one-tailed permutation test [65]. In each of 10,000 trials, we randomly swapped the AI/hybrid model's score with one of the comparator reader's score for each case, yielding a reader–AI difference sampled from the null distribution. A one-sided p-value was computed by comparing the observed statistic to the empirical quantiles of the null distribution. We used a statistical significance threshold of 0.05.

Data availability

The external test dataset is publicly available at `https://scholar.cu.edu.eg/?q=afahmy/pages/dataset`. The NYU Breast Ultrasound Dataset is not currently permitted for public release by the institutional review board. We published the following report explaining how the dataset was created for reproducibility: `https://cs.nyu.edu/~kgeras/reports/ultrasound_datav1.0.pdf`.

Code availability

The neural networks used in our AI system were developed in PyTorch [66]. Code for preprocessing the data and running the inference, sufficient to evaluate our system on other datasets, is available for research purposes upon a reasonable request made to the corresponding author.

References

[7] W. A. Berg et al. "Ultrasound as the primary screening test for breast cancer: analysis from ACRIN 6666". In: *Journal of the National Cancer Institute* 108.4 (2016), djv367.

[11] W. A. Berg et al. "Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer". In: *JAMA* 299.18 (2008), pp. 2151–2163.

[12] E. A. Sickles et al. "ACR BI-RADS® Atlas, Breast imaging reporting and data system". In: *Reston, VA: American College of Radiology* (2013), pp. 39–48.

[15] L. Yang et al. "Performance of ultrasonography screening for breast cancer: a systematic review and meta-analysis". In: *BMC Cancer* 20.1 (2020), pp. 1–15.

[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2921–2929.

[40] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. "Dataset of breast ultrasound images". In: *Data in Brief* 28 (2020), p. 104863.

[41] F. Shamout et al. *The NYU Breast Ultrasound Dataset v1.0*. Tech. rep. Available at `https://cs.nyu.edu/~kgeras/reports/ultrasound_datav1.0.pdf`. 2021.

[53] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons. "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement". In: *Circulation* 131.2 (2015), pp. 211–219.

[54] Y. LeCun, Y. Bengio, et al. "Convolutional networks for images, speech, and time series". In: *The Handbook of Brain Theory and Neural Networks* 3361.10 (1995), p. 1995.

[55] K. He, X. Zhang, S. Ren, and J. Sun. "Identity mappings in deep residual networks". In: *European Conference on Computer Vision*. Springer. 2016, pp. 630–645.

[56] M. Ilse, J. M. Tomczak, and M. Welling. "Attention-based deep multiple instance learning". In: *arXiv:1802.04712* (2018).

[57] Y. Shen et al. "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization". In: *Medical Image Analysis* 68 (2021), p. 101908.

[58] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *arXiv:1412.6980* (2014).

[59] R. Caruana. "Multitask learning: a knowledge-Based source of inductive bias". In: *Proceedings of the Tenth International Conference on Machine Learning*. Morgan Kaufmann, 1993, pp. 41–48.

[60] J. Bergstra and Y. Bengio. "Random search for hyper-parameter optimization". In: *Journal of Machine Learning Research* 13.Feb (2012).

[61] T. G. Dietterich. "Ensemble methods in machine learning". In: *International Workshop on Multiple Classifier Systems*. Springer. 2000, pp. 1–15.

[62] J. M. Lobo, A. Jiménez-Valverde, and R. Real. "AUC: a misleading measure of the performance of predictive distribution models". In: *Global Ecology and Biogeography* 17.2 (2008), pp. 145–151.

[63] F. Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[64] R. W. Johnson. "An introduction to the bootstrap". In: *Teaching Statistics* 23.2 (2001), pp. 49–54.

[65] L. Chihara and T. Hesterberg. *Mathematical Statistics with Resampling and R*. Wiley Online Library, 2011.

[66] A. Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Acknowledgements

The authors would like to thank Mario Videna, Abdul Khaja and Michael Costantino for supporting our computing environment, Benny Huang and Marc Parente for extracting the data, Yizhuo Ma for providing graphical design consultation, and Catriona C. Geras for proofreading the manuscript. We also gratefully acknowledge the support of Nvidia Corporation with the donation of some of the GPUs used in this research. This work was supported in part by grants from the National Institutes of Health (P41EB017183, R21CA225175), the National Science Foundation (1922658), the Gordon and Betty Moore Foundation (9683), the Polish National Agency for Academic Exchange (PPN/IWA/2019/1/00114/U/00001) and NYU Abu Dhabi.

Author contributions

YS, FES and JO are the co-first authors of this paper. YS, FES and KJG designed the experiments with neural networks. YS conducted the experiments with neural networks. YS, FES, JO, JW, KK, JP, NW and CH built the data preprocessing pipeline. YS and FES conducted the reader study and analyzed the data. YS, FES, and JO conducted literature search. YS and JB conducted the statistical analysis. JO, CH, SW, AM, RE, DA, CT, NS, YG, CC, SG, JA, CL, SKS, CL, RM, CM, AL, BR, LM and LH collected the data. LH analyzed the results from a clinical perspective. KJG and FES supervised the execution of all elements of the project. All authors provided critical feedback and helped shape the manuscript.

Competing interests

The authors declare no competing interests.

A Extended Data

Table A.1: **Experience of readers who participated in the reader study.** We summarized the experience of readers who participated in the reader study, in terms of the estimated number of breast ultrasound reads per year and the number of years of experience. All readers are attending radiologists who specialize in breast imaging.

Reader	Estimated reads per year	Years of experience
Reader 1	6500	6
Reader 2	2500	7
Reader 3	3000	4
Reader 4	750	32
Reader 5	1500	13
Reader 6	600	3
Reader 7	6000	35
Reader 8	6500	6
Reader 9	6000	7
Reader 10	3500	40

Table A.2: **Reader study performance.** We reported the values and 95% confidence intervals of AUROC, AUPRC, specificity, sensitivity, biopsy rate, PPV, and NPV achieved by the AI system and radiologists on the reader study set (n = 1,024 breasts). We also showed the mean and standard deviation of radiologists' performance. We calculated the specificity and sensitivity of the AI system by dichotomizing its probabilistic predictions to match the average reader's sensitivity and specificity respectively. We similarly calculated the biopsy rate, PPV, and NPV of the AI system by matching the average reader's sensitivity.

Reader	AUROC	AUPRC	Specificity(%)	Sensitivity(%)	Biopsy rate(%)	PPV(%)	NPV(%)
R1	0.955 (0.935, 0.978)	0.612 (0.492, 0.688)	79.5 (75.7, 81.5)	97.3 (93.4, 100.0)	26.0 (24.5, 30.7)	26.7 (22.3, 32.5)	99.7 (99.5, 100.0)
R2	0.960 (0.946, 0.978)	0.616 (0.522, 0.689)	72.6 (70.9, 75.1)	98.6 (96.1, 100.0)	32.5 (30.3, 34.3)	21.6 (18.4, 26.3)	99.9 (99.6, 100.0)
R3	0.916 (0.866, 0.940)	0.550 (0.411, 0.640)	85.0 (82.0, 87.9)	84.9 (73.8, 91.8)	20.0 (17.5, 23.2)	30.2 (23.8, 36.3)	98.7 (97.8, 99.4)
R4	0.930 (0.906, 0.962)	0.596 (0.441, 0.696)	85.5 (84.0, 87.1)	90.4 (84.3, 95.1)	19.9 (17.8, 22.6)	32.4 (24.5, 38.5)	99.1 (98.5, 99.6)
R5	0.924 (0.900, 0.964)	0.695 (0.599, 0.777)	83.6 (81.1, 86.2)	90.4 (86.9, 96.7)	21.7 (18.7, 24.2)	29.7 (23.6, 36.8)	99.1 (98.6, 99.8)
R6	0.904 (0.874, 0.923)	0.498 (0.354, 0.598)	74.0 (70.5, 77.6)	89.0 (83.3, 93.4)	30.5 (26.3, 33.3)	20.8 (17.0, 26.3)	98.9 (98.3, 99.3)
R7	0.925 (0.909, 0.947)	0.447 (0.371, 0.535)	76.4 (73.8, 78.7)	90.4 (84.3, 96.7)	28.3 (25.9, 31.2)	22.8 (18.6, 27.4)	99.0 (98.3, 99.7)
R8	0.920 (0.879, 0.959)	0.624 (0.496, 0.765)	86.1 (83.4, 89.6)	87.7 (80.0, 92.4)	19.1 (15.1, 22.2)	32.7 (25.9, 38.8)	98.9 (97.8, 99.3)
R9	0.902 (0.870, 0.941)	0.533 (0.435, 0.609)	83.7 (81.2, 85.7)	86.3 (79.2, 91.6)	21.3 (18.9, 24.1)	28.9 (25.0, 34.8)	98.8 (97.9, 99.3)
R10	0.900 (0.869, 0.930)	0.484 (0.417, 0.566)	80.7 (77.8, 83.6)	86.3 (78.9, 91.1)	24.1 (22.2, 27.1)	25.5 (20.2, 33.8)	98.7 (97.6, 99.2)
Avg	0.924 ± 0.020 (0.905, 0.944)	0.565 ± 0.072 (0.465, 0.625)	80.7 ± 4.7 (78.9, 82.6)	90.1 ± 4.3 (86.4, 93.8)	24.3 ± 4.5 (22.0, 26.5)	27.1 ± 4.1 (22.9, 33.1)	99.1 ± 0.4 (98.4, 99.5)
AI	0.962 (0.943, 0.979)	0.752 (0.675, 0.849)	85.6 (83.9, 88.0)	94.5 (89.4, 100.0)	19.8 (17.9, 22.1)	32.5 (26.9, 39.2)	99.1 (98.2, 99.6)

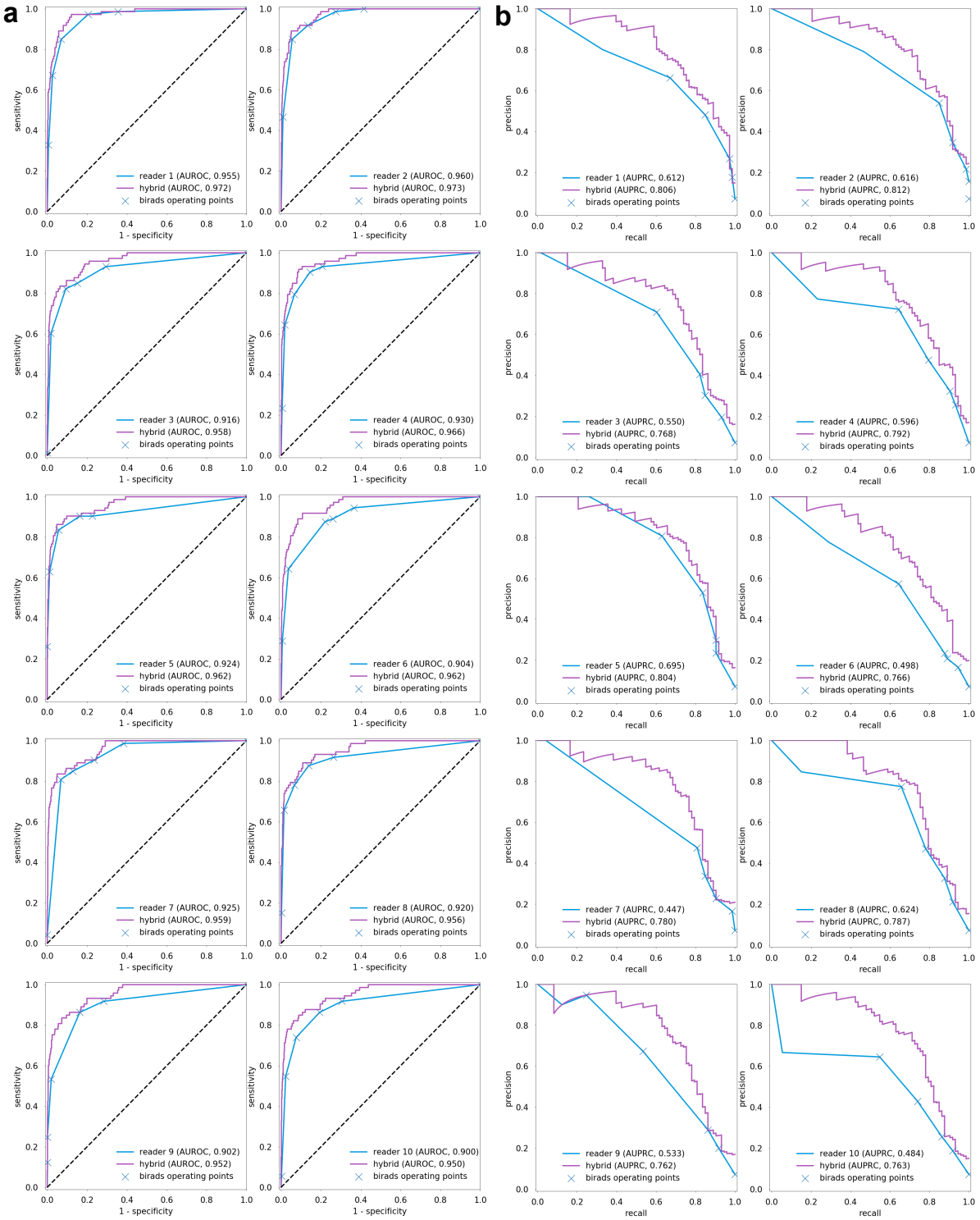

Figure A.1: **ROC and precision-recall curves for radiologists in the reader study.** We visualized the ROC (**a**) and precision-recall curves (**b**) derived from the predictions made by ten radiologists and their corresponding hybrid models (see Methods section 'Hybrid model') in the reader study (n = 1,024 breasts). For each reader, we highlight the operating points which correspond to the performance this radiologist achieved when dichotomizing the radiologist's predictions using a threshold of BI-RADS categories (see Methods section 'Statistical analysis').

Table A.3: **Subgroup analysis results: benign vs. malignant.** We reported the values and 95% confidence intervals of AUROC, AUPRC, specificity, sensitivity, biopsy rate, PPV, and NPV achieved by the AI system and radiologists on the subgroup analysis. In this analysis, we included 574 exams ($n = 608$ breasts) from the reader study that yielded biopsy-confirmed benign or malignant findings. We also showed the mean and standard deviation of radiologists' performance. We calculated the specificity and sensitivity of the AI system by dichotomizing its probabilistic predictions to match the average reader's sensitivity and specificity respectively. We similarly calculated the biopsy rate, PPV, and NPV of the AI system by matching the average reader's sensitivity.

Reader	**AUROC**	**AUPRC**	**Specificity(%)**	**Sensitivity(%)**	**Biopsy rate(%)**	**PPV(%)**	**NPV(%)**
R1	0.932 (0.913, 0.957)	0.635 (0.570, 0.717)	67.9 (64.5, 71.9)	97.3 (91.5, 100.0)	40.0 (36.8, 42.4)	29.2 (23.1, 33.5)	99.5 (98.6, 100.0)
R2	0.937 (0.923, 0.967)	0.636 (0.593, 0.758)	57.4 (52.7, 62.5)	98.6 (95.9, 100.0)	49.3 (44.1, 53.6)	24.0 (20.3, 27.7)	99.7 (99.0, 100.0)
R3	0.889 (0.855, 0.929)	0.576 (0.517, 0.722)	76.6 (72.6, 82.5)	84.9 (79.5, 90.6)	30.8 (25.2, 33.9)	33.2 (27.0, 39.9)	97.4 (96.0, 98.7)
R4	0.908 (0.879, 0.944)	0.619 (0.523, 0.748)	76.6 (73.2, 80.7)	90.4 (87.1, 96.1)	31.4 (26.8, 35.0)	34.6 (28.7, 40.7)	98.3 (97.6, 99.3)
R5	0.907 (0.878, 0.951)	0.709 (0.646, 0.794)	72.9 (67.2, 77.9)	90.4 (87.0, 95.2)	34.7 (29.4, 40.5)	31.3 (27.7, 37.3)	98.2 (97.4, 99.3)
R6	0.866 (0.831, 0.920)	0.525 (0.450, 0.605)	59.4 (56.1, 64.3)	89.0 (82.8, 93.8)	46.4 (41.1, 49.3)	23.0 (20.2, 25.9)	97.5 (96.4, 98.8)
R7	0.890 (0.859, 0.915)	0.478 (0.439, 0.534)	64.9 (60.2, 68.3)	90.4 (84.3, 95.5)	41.8 (38.2, 46.2)	26.0 (22.3, 29.1)	98.0 (96.2, 99.1)
R8	0.896 (0.850, 0.949)	0.639 (0.535, 0.735)	77.6 (75.1, 81.4)	87.7 (79.5, 95.3)	30.3 (25.8, 33.1)	34.8 (29.1, 41.9)	97.9 (96.4, 99.3)
R9	0.870 (0.821, 0.922)	0.555 (0.458, 0.663)	75.0 (70.6, 79.6)	86.3 (77.4, 95.3)	32.4 (26.8, 36.0)	32.0 (26.7, 35.9)	97.6 (96.0, 99.3)
R10	0.868 (0.836, 0.925)	0.516 (0.433, 0.662)	69.9 (66.5, 73.9)	86.3 (81.2, 93.8)	36.8 (32.6, 40.3)	28.1 (22.2, 32.9)	97.4 (96.4, 99.0)
Avg	0.896 ± 0.024 (0.874, 0.929)	0.589 ± 0.067 (0.557, 0.671)	69.8 ± 6.9 (67.7, 73.6)	90.1 ± 4.3 (86.8, 93.8)	37.4 ± 6.4 (33.1, 39.8)	29.6 ± 4.0 (25.2, 33.6)	98.1 ± 0.8 (97.3, 99.0)
AI	0.941 (0.922, 0.968)	0.762 (0.695, 0.841)	78.3 (74.7, 81.0)	95.9 (90.1, 98.6)	29.9 (26.8, 33.1)	36.3 (30.8, 40.7)	98.4 (97.1, 99.5)

Table A.4: **Subgroup analysis results: cancer subtypes.** We compared the number of correctly identified malignant lesions between the AI system and radiologists. In this analysis, we included 72 exams (73 breasts, 97 lesions) from the reader study with biopsy-confirmed malignant findings. We stratified the lesions by their cancer subtype, histological grade, and biomarker profile. For each stratification, we reported the total number of lesions (n), the number of lesions identified as malignant by the AI, and the number of lesions identified as malignant by radiologists. We dichotomized AI's probabilistic predictions by matching radiologists' average specificity in the reader study.

Lesion characteristics	n	**AI**	**radiologists (mean ± std)**
Cancer Subtype			
Invasive ductal carcinoma	75	72	70.7 ± 2
Invasive lobular carcinoma	9	9	8.2 ± 0.4
Other invasive carcinoma	8	8	5.9 ± 2.2
Ductal carcinoma in situ (DCIS)	5	4	4.1 ± 0.7
Histologic Grade (Invasive Cancers)			
Well differentiated	9	9	8.2 ± 0.4
Moderately differentiated	35	33	32.2 ± 1.2
Poorly differentiated	39	38	37.5 ± 1.5
Histologic Grade (DCIS)			
Well differentiated	0	-	-
Moderately differentiated	4	3	3.1 ± 0.7
Poorly differentiated	1	1	1 ± 0.0
Biomarkers of Invasive Cancers			
ER/PR-positive, HER2-negative	55	52	50.4 ± 1.9
HER2-positive	25	25	23.8 ± 1.2
ER/PR/HER2-negative	8	8	7.9 ± 0.3

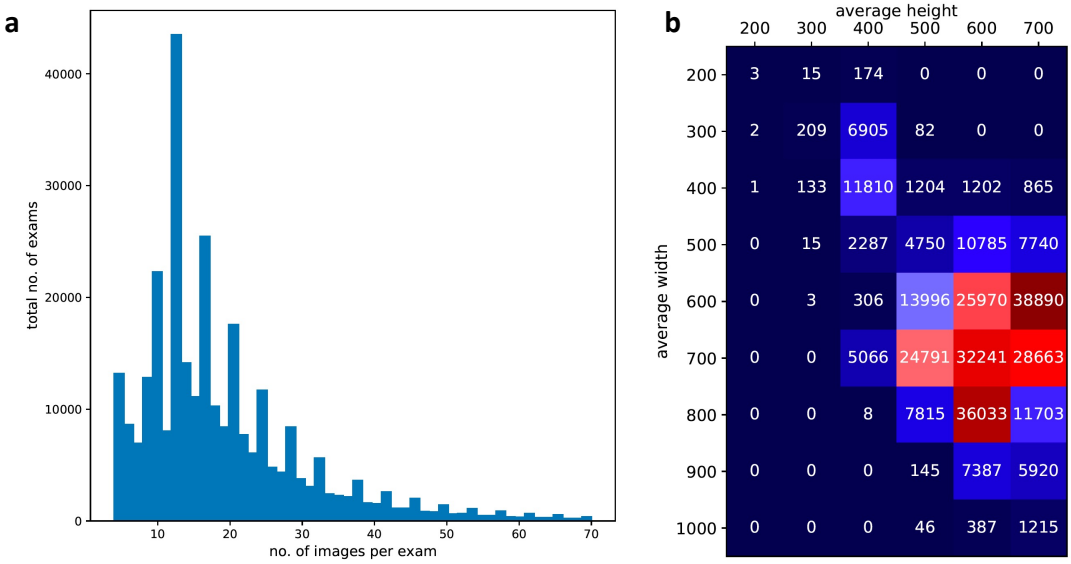

Figure A.2: **Number of images and resolution of images in the dataset. a**, The distribution of the total number of images per exam. On average, each exam contains 18.8 US images. **b**, The distribution of the average size of the images in each exam. The x-axis represents the average image height per exam while the y-axis represents the average image width per exam (rounded to the nearest hundredth). The height and width are measured in number of pixels. The average resolution of images in this dataset is 665 × 603 pixels.

Table A.5: **Performance of the hybrid models.** We reported the values and 95% confidence intervals of AUROC, AUPRC, specificity, sensitivity, biopsy rate, PPV, and NPV achieved by the hybrid models (see Methods section 'Hybrid model') that combine the predictions of AI with each of the ten radiologists (R1-R10) on the reader study set (n = 1,024 breasts). The delta values show the difference (hybrid model-radiologist) and 95% confidence intervals in each metrics between each hybrid model and its respective reader. We calculated the specificity and sensitivity of each hybrid model by dichotomizing its probabilistic predictions to match its respective reader's sensitivity and specificity respectively. We similarly calculated the biopsy rate, PPV, and NPV of each hybrid model by matching the its respective reader's sensitivity.

Reader	AUROC	AUPRC	Specificity (%)	Sensitivity (%)	Biopsy rate (%)	PPV (%)	NPV (%)
R1	0.972	0.806	87.9	97.3	18.2	38.2	99.8
	(0.957, 0.991)	(0.713, 0.887)	(85.0, 89.2)	(93.4, 100.0)	(16.4, 22.2)	(30.4, 45.5)	(99.5, 100.0)
Δ	0.017	0.195	8.4	0.0	-7.8	11.5	0.0
	(0.012, 0.025)	(0.124, 0.245)	(6.5, 10.7)	(0.0, 0.0)	(-9.9, -6.1)	(8.1, 13.2)	(0.0, 0.1)
R2	0.973	0.812	80.1	100.0	25.5	27.6	99.9
	(0.964, 0.990)	(0.754, 0.889)	(78.6, 82.1)	(100.0, 100.0)	(22.9, 27.4)	(24.1, 33.2)	(99.6, 100.0)
Δ	0.013	0.196	7.6	1.4	-7.0	6.0	0.0
	(0.009, 0.018)	(0.153, 0.263)	(6.2, 8.9)	(0.0, 3.9)	(-8.3, -5.7)	(4.5, 7.0)	(0.0, 0.0)
R3	0.958	0.768	90.4	87.7	15.0	40.9	98.9
	(0.935, 0.975)	(0.646, 0.867)	(88.3, 92.0)	(80.3, 93.5)	(13.0, 17.3)	(32.9, 48.3)	(98.1, 99.4)
Δ	0.042	0.218	5.5	2.7	-5.0	10.7	0.2
	(0.021, 0.068)	(0.157, 0.291)	(4.0, 7.1)	(0.0, 6.6)	(-6.5, -3.7)	(7.6, 14.5)	(0.0, 0.5)
R4	0.966	0.792	91.7	93.2	14.2	45.5	99.2
	(0.953, 0.987)	(0.699, 0.871)	(90.4, 93.0)	(84.6, 98.4)	(12.3, 16.1)	(38.7, 52.7)	(98.1, 99.8)
Δ	0.036	0.195	6.2	2.7	-5.8	13.2	0.1
	(0.021, 0.053)	(0.138, 0.272)	(4.5, 7.9)	(0.0, 7.7)	(-7.3, -4.2)	(9.5, 16.4)	(-0.4, 0.5)
R5	0.962	0.804	90.2	90.4	15.5	41.5	99.2
	(0.947, 0.983)	(0.721, 0.886)	(89.1, 92.0)	(86.9, 96.7)	(13.3, 17.7)	(33.8, 47.5)	(98.7, 99.8)
Δ	0.037	0.109	6.6	0.0	-6.2	11.8	0.1
	(0.019, 0.047)	(0.062, 0.169)	(5.5, 9.3)	(0.0, 0.0)	(-8.6, -5.1)	(9.2, 15.3)	(0.0, 0.1)
R6	0.962	0.766	91.6	95.9	14.2	44.8	99.1
	(0.950, 0.985)	(0.673, 0.864)	(90.3, 93.3)	(91.8, 100.0)	(12.1, 16.3)	(37.9, 50.0)	(98.5, 99.8)
Δ	0.058	0.268	17.6	6.8	-16.3	24.0	0.2
	(0.038, 0.099)	(0.162, 0.377)	(15.6, 20.4)	(2.7, 13.1)	(-18.8, -14.1)	(19.6, 28.3)	(-0.2, 0.8)
R7	0.959	0.780	81.1	90.4	24.0	26.8	99.1
	(0.942, 0.980)	(0.708, 0.857)	(78.8, 82.8)	(84.3, 96.7)	(21.9, 26.7)	(22.4, 32.7)	(98.4, 99.7)
Δ	0.034	0.333	4.6	0.0	-4.3	4.1	0.1
	(0.023, 0.046)	(0.291, 0.440)	(3.8, 5.7)	(0.0, 0.0)	(-5.4, -3.6)	(3.4, 5.4)	(0.0, 0.1)
R8	0.956	0.787	89.2	89.0	16.4	38.7	99.1
	(0.931, 0.976)	(0.693, 0.870)	(87.2, 92.4)	(82.2, 93.7)	(12.5, 18.7)	(31.8, 46.0)	(98.1, 99.4)
Δ	0.036	0.163	3.0	1.4	-2.7	6.0	0.2
	(0.016, 0.052)	(0.089, 0.232)	(2.2, 4.2)	(0.0, 3.3)	(-3.8, -2.0)	(4.2, 8.9)	(0.0, 0.3)
R9	0.952	0.762	88.9	86.3	16.5	37.3	98.8
	(0.931, 0.972)	(0.673, 0.834)	(86.6, 90.7)	(79.2, 91.6)	(14.2, 18.9)	(31.5, 43.2)	(98.0, 99.3)
Δ	0.051	0.229	5.2	0.0	-4.8	8.4	0.1
	(0.030, 0.068)	(0.175, 0.298)	(4.3, 6.7)	(0.0, 0.0)	(-6.2, -4.0)	(6.4, 11.5)	(0.0, 0.1)
R10	0.950	0.763	89.3	87.7	16.1	38.2	98.8
	(0.928, 0.970)	(0.693, 0.832)	(87.9, 91.0)	(80.0, 92.8)	(14.1, 18.3)	(32.5, 48.1)	(97.7, 99.3)
Δ	0.050	0.278	8.6	1.4	-8.0	12.7	0.1
	(0.034, 0.063)	(0.208, 0.391)	(6.7, 11.3)	(0.0, 3.3)	(-10.5, -6.3)	(8.8, 15.6)	(-0.2, 0.3)

Table A.6: **High-confidence triage analysis.** We experimented with varying the operating point to improve the confidence of the AI system. A very low threshold results in high NPV and enables the AI to confidently identify negative cases. On the other hand, a very high threshold results in high PPV and enables the AI to confidently prioritize cases that are highly suspicious of malignancy. For either triage scenario, we reported the values and 95% confidence intervals of sensitivity, specificity, and NPV/PPV, along with number of breasts (n) associated with each metrics.

Triage	**Sensitivity (%)**	**Specificity (%)**	**Reliability of triage decision**
negative	98.63% (95.35%, 100%) $n = 73$	77.71% (74.82%, 80.19%) $n = 951$	99.86% (NPV) (99.59%, 100%) $n = 740$
positive	52.05% (40.84% - 63.77%) $n = 73$	99.26% (98.73% - 99.67%) $n = 951$	84.44% (PPV) (72.97% - 93.76%) $n = 44$

Table A.7: **Distribution of ultrasound devices.** Breakdown of studies in the NYU Breast Ultrasound Dataset subsets by a ultrasound machine model. There was no bias in terms of device preference when splitting the studies into training, validation and test sets.

Device	**Training set**	**Validation set**	**Test set**
Affiniti 70G	79080	13329	14715
S1000	40097	6684	9148
S3000	29676	4937	5785
S2000	24701	4054	5655
LOGIQ7	6316	1035	1647
Xario	6029	947	1541
iU22	4988	803	1696
LOGIQ9	3659	585	544
TUS-A300	3540	618	954
Accuvix V10	2478	389	788
Antares	2468	395	709
LOGIQ5	2232	471	832
Sequoia	1868	263	28
Accuvix V20	1851	311	680
LOGIQE9	152	19	26
HDI 5000	10	6	1
LOGIQS8	8	1	5
Aixplorer	4	1	0
LOGIQS7	4	2	1
UGEO H60	1	0	0

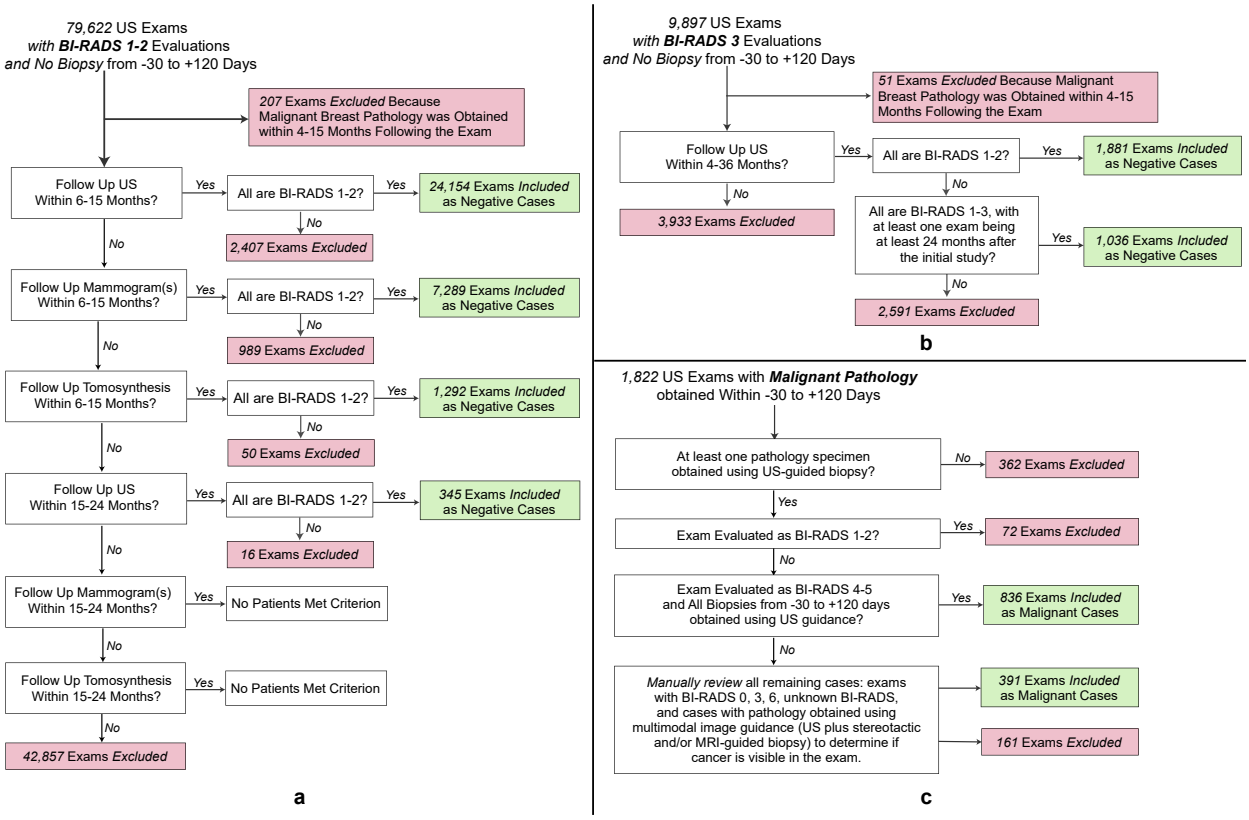

Figure A.3: **Filtering protocol applied on the internal test set.** Cancer-negative exams were filtered to ensure that they are associated with a negative pathology report or have at least one cancer-negative follow-up. The specific workup for BI-RADS 1&2 and BI-RADS 3 exams were illustrated in **a** and **b** respectively. **c**, Exams with biopsy-proven cancers were filtered to ensure that cancers were visible on the US images.

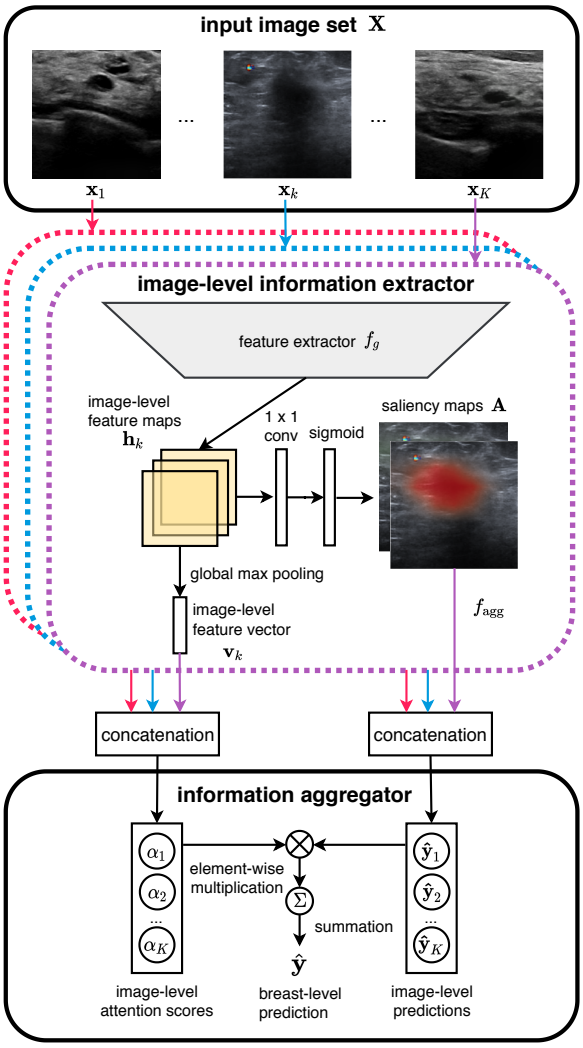

Figure A.4: **Overall structure of the deep neural network used in this study.** The image-level information extractor first independently processes each ultrasound image $\mathbf{x}_k$ in the image set $\mathbf{X}$ and generates two saliency maps ($\mathbf{A}_k^b$, $\mathbf{A}_k^m$) that indicate the informative regions in the image. The network then calculates two attention scores (α_k^b, α_k^m) which indicate the importance of $\mathbf{x}_k$ for the diagnosis of benign and malignant lesions respectively. Lastly, the information aggregator then combines classification signals from all images and yields a breast-level prediction $\hat{\mathbf{y}}$.