ARTICLES
https://doi.org/10.1038/s41551-021-00711-2

nature biomedical engineering

Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning

Xuejun Qian[1,2,10] ✉, Jing Pei[3,4,10], Hui Zheng[5,10], Xinxin Xie[5], Lin Yan[6], Hao Zhang[7], Chunguang Han[3,4], Xiang Gao[8], Hanqi Zhang[5], Weiwei Zheng[9], Qiang Sun[3,4], Lu Lu[8] and K. Kirk Shung[1]

The clinical application of breast ultrasound for the assessment of cancer risk and of deep learning for the classification of breast-ultrasound images has been hindered by inter-grader variability and high false positive rates and by deep-learning models that do not follow Breast Imaging Reporting and Data System (BI-RADS) standards, lack explainability features and have not been tested prospectively. Here, we show that an explainable deep-learning system trained on 10,815 multimodal breast-ultrasound images of 721 biopsy-confirmed lesions from 634 patients across two hospitals and prospectively tested on 912 additional images of 152 lesions from 141 patients predicts BI-RADS scores for breast cancer as accurately as experienced radiologists, with areas under the receiver operating curve of 0.922 (95% confidence interval (CI) = 0.868–0.959) for bimodal images and 0.955 (95% CI = 0.909–0.982) for multimodal images. Multimodal multiview breast-ultrasound images augmented with heatmaps for malignancy risk predicted via deep learning may facilitate the adoption of ultrasound imaging in screening mammography workflows.

Breast cancer is the most frequently diagnosed cancer and the second leading cause of cancer death among women worldwide[1]. The incidence of breast cancer is increasing, with over 1.6 million cases in 2010 and projections of 2.1 million by 2030[2,3]. Mammography is the recommended screening test to reduce breast cancer-related mortality. However, mammography has low sensitivity in dense breast parenchyma[4] and is not widely available in all countries, which may cause delayed diagnosis and worse outcomes[5]. Because of its low cost, wide availability and lack of ionizing radiation, the application of ultrasound (US) has been proposed to be expanded from the differentiation of cysts from solid masses to screening for breast cancer, especially for women with dense breasts[6]. The American College of Radiology published Breast Imaging Reporting and Data System (BI-RADS) lexicon guidelines for breast cancer screening, to standardize image interpretation by radiologists and dictate management recommendations. Despite improved consistency, subjective characterization of imaging findings and persistent intra- and inter-observer variability in medical image interpretation still remain as limitations[7,8]. In addition, there still exists a shortage of human experts to provide timely diagnosis and refer patients to the appropriate clinical care.

Machine learning has been leveraged for many years in computer-aided diagnosis in various types of cancer[9], including breast cancer[10]. Much of the previous work has focused on hand-engineered features, which involve computing explicit features selected by domain experts, resulting in algorithms designed for certain texture features or specific imaging modes. With the development of artificial intelligence (AI), deep convolutional neural networks—a special type of deep-learning technique[11] allowing an algorithm to learn the appropriate predictive features on the basis of imaging examples—have repeatedly been shown to be superior to hand-engineered features in the field of computer vision[12]. At the same time, deep learning has gained traction in advanced biomedical image analysis as a powerful tool to increase efficiency and reproducibility[13]. Recent advances in deep-learning studies have led to comparable sensitivity and specificity to readings by board-certified medical experts, as demonstrated through the use of hundreds of thousands of images[14–17]. Notably, deep-learning architectures have demonstrated ophthalmologist-level performance on optical coherence tomography images[14] and have achieved dermatologist-level classification of skin cancer[15].

Despite the fact that feasibility of applying deep learning on the classification of breast-US images has been demonstrated[18,19] and recently improved using ensemble[20,21] or modified[22,23] deep-learning architectures, different US modalities[24–26] or integration of BI-RADS features into the framework of AI[27,28], these studies were designed to neither demonstrate real-world clinical relevance nor illuminate potential guidance to clinicians in making clinical decisions. In other words, the clinical applicability of this technology has remained unresolved due to four key challenges. First, the design of previous AI studies deviated from existing BI-RADS reporting standards[29], resulting in potential issues relevant to clinical acceptance and diagnostic accuracy. In routine clinical workflows where either fourth-edition BI-RADS (via the assessment of US B-mode

[1]Department of Biomedical Engineering, University of Southern California, Los Angeles, CA, USA. [2]Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [3]Department of Breast Surgery, The First Affiliated Hospital of Anhui Medical University, Hefei, China. [4]Department of General Surgery, The First Affiliated Hospital of Anhui Medical University, Hefei, China. [5]Department of Ultrasound, The First Affiliated Hospital of Anhui Medical University, Hefei, China. [6]School of Computer Science and Technology, Xidian University, Xi'an, China. [7]Department of Neurosurgery, University Hospital Heidelberg, Heidelberg, Germany. [8]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. [9]Department of Ultrasound, Xuancheng People's Hospital, Xuancheng, China. [10]These authors contributed equally: Xuejun Qian, Jing Pei, Hui Zheng. ✉e-mail: xuejunqi@usc.edu

and US colour Doppler images) or fifth-edition BI-RADS (via the assessment of US B-mode, US colour Doppler and US elastography images) lexicon is adopted, multimodal multiview US images should be considered as part of the design of AI tools as the workflow standard. Second, previous AI studies were not evaluated in a prospective setting and had small or even no human comparison groups; therefore, they were at high risk of bias[30]. To address this issue, a prospective setting and an appropriately large human sample for fair comparison with AI are essential for ensuring reliability. Third, to gain trust from clinicians, the inner workings of medical AI systems, which were opaque in former AI studies, should be understandable. Lastly, previous AI studies failed to demonstrate whether the AI system can improve and/or guide clinical outcomes such as clinicians' decision-making on BI-RADS categorization.

To move beyond the limitations of previous AI approaches and accelerate a broader adoption of deep-learning technology by clinicians, we aim to develop a clinically applicable AI system for automatic breast cancer risk prediction using a deep-learning architecture, which creates an opportunity to mimic the routine clinical workflow by utilizing multimodal (that is, US B-mode, US colour Doppler and US elastography) and multiview (that is, transverse and longitudinal) breast-US images. The most important contribution of this work is the enhancement of clinical applicability. Specifically, the performance of the AI system is compared with that of seven human experts in a setting of prospectively consecutive patients. In addition, we create numerical heatmaps that emphasize the importance of the feature map relevant to the predictions of the AI system. Such explainability features allow assessment of the region of interest (ROI), with potential clinical value for each imaging mode (that is, they guide the clinicians to investigate the corresponding regions in the original US images and then re-evaluate their clinical value). As a consequence, combining such auxiliary information with the original clinicians' assessments could potentially increase confidence levels of clinicians when making final informed decisions such as BI-RADS categorization and routine referral biopsy/follow-up suggestions. We demonstrate that AI systems have the potential to offer reliable diagnoses, good generalizability and efficient deployment, all of which will greatly improve routine breast-US examinations and facilitate their development in clinical settings.

Results

With this work, we proposed a multipathway deep-learning architecture, as shown in Fig. 1. Deep-learning models to predict the risk of breast cancer were developed from a retrospectively collected development dataset consisting of 10,815 US images from 721 lesions of 634 patients who underwent breast-US examinations between October 2016 and December 2018. The retrospective workflow followed the BI-RADS guidelines (see Supplementary Table 1), including the inclusion and exclusion criteria, and an overview is depicted in Fig. 2. Out of the 721 lesions used, 556 lesions were benign and 165 lesions were malignant, based on biopsy-confirmed pathology results. Detailed patient demographics and breast lesion characteristics for the development dataset are summarized in Table 1. Lesions in the development dataset were randomly assigned to one of two sets: a training set (70%) and a validation set (30%).

Performance of deep-learning models. The multipathway deep convolutional neural network was trained using view-level multimodal US images and biopsy-confirmed labels (Supplementary Fig. 1). A convolution operator and SENet module[31] enabled the network to extract informative features through investigating both the spatial relationship and the channel relationship in each pathway (see Methods). Fully connected layers distilled meaningful representations to perform decision-making. To enhance the interpretability of the proposed deep-learning model on breast cancer risk prediction, features from the final convolutional layer in each pathway were extracted to generate a heatmap of each imaging modality via the Grad-CAM technique[32], which can aid human experts to understand highlighting decisions made by the AI system.

The effectiveness of our proposed multipathway deep-learning model was validated using three different aspects. First, ResNet-18 combined with the SENet backbone showed better performance than basic ResNet-18, as shown in Supplementary Fig. 2a. Second, to find the most suitable base model for predicting the risk of breast cancer, we compared our backbone network with other prevalent models, including VGG19 (ref. [33]), ResNet-50 (ref. [34]) and Inception-v3 (ref. [35]), all of which were integrated with the SENet block to ensure fair comparison. As displayed in Supplementary Fig. 2b, our model maintained the best performance with less complexity (for example, a complicated training model may cause overfitting in the condition of insufficient training data size) compared with VGG19 with SENet, ResNet-50 with SENet and Inception-v3 with SENet. Lastly, in Supplementary Fig. 2c, we show the efficacy of our proposed model via a cross-validation strategy on the development dataset.

Two deep-learning models were involved in our AI system—namely, a bimodal model (inputs with US B-mode and US colour Doppler images) and a multimodal model (inputs with US B-mode, US colour Doppler and US elastography images). As indicated in Supplementary Fig. 3, the bimodal and multimodal models were designed to mimic the routine clinical workflow of clinicians who use fourth- and fifth-edition BI-RADS lexicon, respectively. In the validation set, we evaluated the performance of the proposed AI system using view-level US images in terms of the area under the curve (AUC) of the receiver operating characteristic curve (ROC). The bimodal model achieved an AUC score of 0.877 (95% confidence interval (CI) = 0.830–0.914). With the additional US elastography information, the multimodal model attained a significantly better AUC score of 0.923 (95% CI = 0.883–0.953) with $P < 0.05$ (Supplementary Fig. 4).

To strengthen the generalizability of the AI system, it must not only be used to evaluate retrospective data, but also it should be assessed in a prospective study setting[36,37]. Therefore, we used one prospective clinical test set amassed from two different hospitals. On the test set of 152 lesions from 141 patients (Supplementary Fig. 5), the bimodal model obtained an AUC of 0.890 (95% CI = 0.829–0.935) on transverse-view US images and an AUC of 0.898 (95% CI = 0.839–0.942) on longitudinal-view US images. In contrast, the multimodal model accomplished superior performance to the bimodal model, reaching a significantly higher AUC of 0.936 (95% CI = 0.885–0.970) on transverse-view US images and an AUC of 0.935 (95% CI = 0.883–0.968) on longitudinal-view US images ($P < 0.05$). Supplementary Table 2 summarizes the statistical comparisons among various views and models. In terms of view-level US images, the performance of the AI system in the prospective test set was consistent with that in the retrospectively collected validation set, which demonstrates the generalizability of the AI system.

To mimic the workflow of radiologists who take multiview US images into consideration in clinical settings, we combined the malignancy risk probabilities from both views to produce an overall probability for the clinical test set, as a lesion-level US imaging evaluation. Varying the threshold of the overall probability in the interval 0–1 generated new ROCs of sensitivities and specificities. To summarize, the bimodal model attained an overall AUC of 0.922 (95% CI = 0.868–0.959). In comparison, the multimodal model fulfilled the best overall AUC of 0.955 (95% CI = 0.909–0.982), which was significantly higher than that of the bimodal model ($P < 0.05$), as shown in Fig. 3. On both the bimodal model and the multimodal model, the AUC evaluated on lesion-level US images showed better performance than that on view-level US images.

For comparison with radiologists on sensitivity and specificity, we varied the operating threshold to produce three different breast

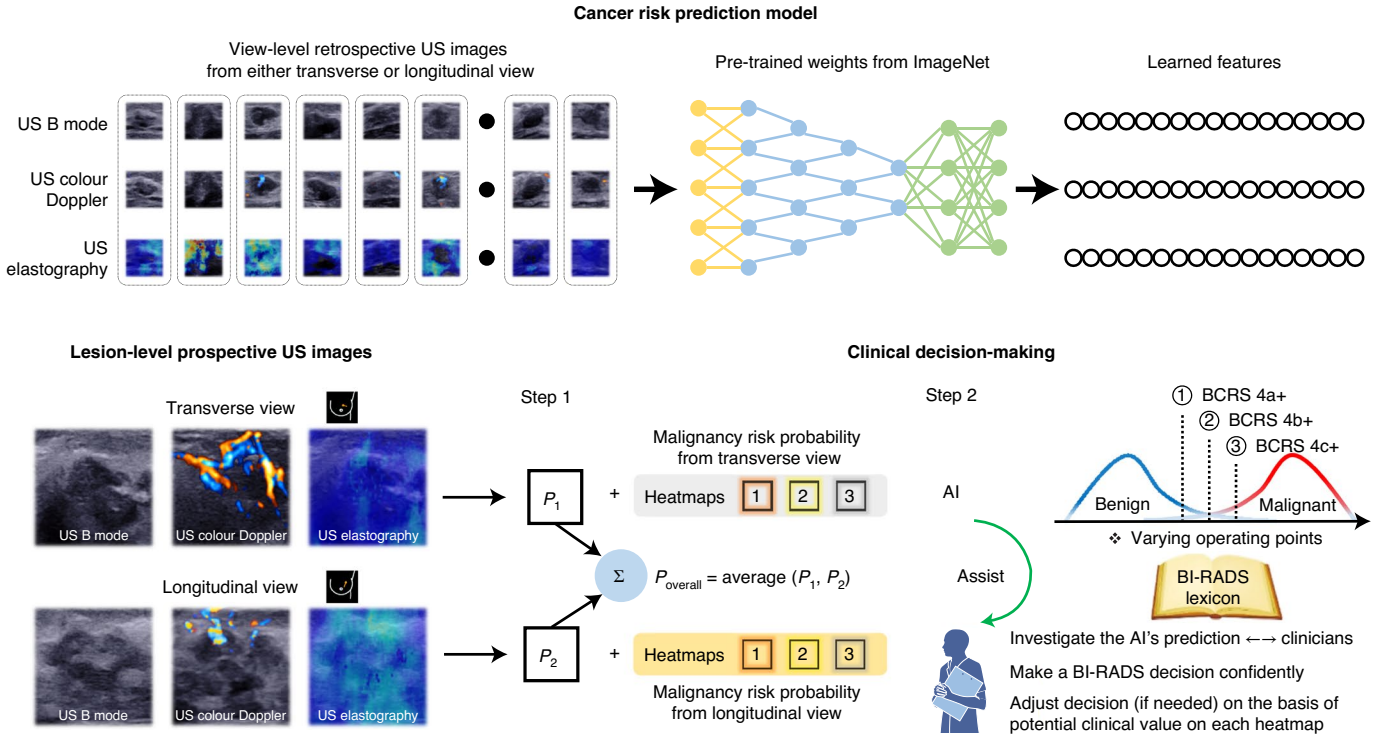

Fig. 1 | Overall AI system for breast cancer risk prediction. The model was developed on view-level multimodal US images (that is, US B-mode, US colour Doppler and US elastography images) using the deep-learning framework (see details in Supplementary Fig. 1). For each prospective clinical test lesion, the AI system utilizes one-view multimodal US images as inputs each time, evaluates the suspicious lesion from multiple views (that is, transverse and longitudinal views) and outputs an overall malignancy probability. Three different BCRSs were proposed in the AI system by varying the operating threshold to compare with and assist clinicians.

cancer risk scores (BCRSs)—namely, BCRS 4a+, 4b+ and 4c+. These BCRS modes corresponded to the decision-making modes of human experts in BI-RADS 3 versus 4a+, BI-RADS 3,4a versus 4b+ and BI-RADS 3,4a,4b versus 4c+, respectively (see Methods). At the BCRS 4a+ mode, the bimodal model achieved a sensitivity of 88.6%, which was higher than the 75.0 and 47.7% observed in the other two BCRS modes. The best specificity of 97.2% was obtained for the BCRS 4c+ mode while balanced sensitivity and specificity were achieved at the BCRS 4b+ mode (Supplementary Table 3). With respect to the multimodal model, improved sensitivity of breast-US lesion assessment without loss of specificity was observed at the BCRS 4b+ and 4c+ modes (Supplementary Table 4). However, at the BCRS 4a+ mode, the main diagnostic improvement was specificity rather than sensitivity.

AI system performance compared with radiologists. We conducted a two-part reader study with seven experienced radiologists with a range of experience of 5–30 years (average = 14 years) on the same test set as a way to compare with the AI system. In part 1 of the reader study, readers determined BI-RADS scores for the bimodal multiview US images. We compared breast cancer risk prediction of the bimodal model with that of the readers in two ways. One way was to compare the bimodal model with each individual reader's performance by measuring the sensitivity and specificity at three BCRS modes. The AI system achieved significantly better specificity ($P < 0.05$ for six of the seven readers) at the BCRS 4a+ mode (Fig. 3a and Supplementary Table 3). In other BCRS modes, the bimodal model's sensitivities and specificities showed comparable results with those of the readers. The other way was to compare the performance of the bimodal model with that of the average reader. Three averaged readers' assessments were located slightly below the bimodal model's ROC curve (Fig. 3a). Supplementary Table 5 lists the sensitivity and specificity results of the bimodal model and the averaged readers for three BCRS modes.

In part 2 of the reader study, in addition to bimodal US images, the multiview US elastography images were provided for review by the same readers. We performed the same comparison as in the previous reader study. On the assessment of multimodal multiview US images, the sensitivities and specificities of the multimodal model were on par with each reader's predictions, except for improved specificity ($P < 0.05$ for five of the seven readers) for the BCRS 4a+ mode (Fig. 3b and Supplementary Table 4). Figure 3b shows that the AUC of the multimodal model outperformed or matched most individual readers, as well as the averaged readers, since the sensitivity–specificity points of readers lie below the ROC curve. Compared with evaluation on bimodal US images, reduction of false negatives and false positives in the assessment of multimodal US images was observed with the AI system and for most readers.

A major challenge for clinicians in decision-making is to upgrade or downgrade BI-RADS 4 lesions since the error rate between BI-RADS 2, BI-RADS 3 and BI-RADS 5 lesions and biopsy results is minimal. In Fig. 3c,d, we have shown a comparison between the models and readers for the assessment of BI-RADS 4 lesions only in the clinical test set. The multimodal model achieved an AUC of 0.920 (95% CI = 0.840–0.969), indicating an improvement over the bimodal model, which had an AUC of 0.880 (95% CI = 0.790–0.940). Overall, the results from both the whole clinical test set and the subset of suspicious malignancy BI-RADS 4 lesions show that the AI system has reached the level of experienced radiologists with high reliability and accuracy, especially for the multimodal model.

Heatmaps for understanding AI decision-making. To further enhance the AI system's likelihood of clinical applicability, the decisions made by the AI system should be understandable to clinicians.

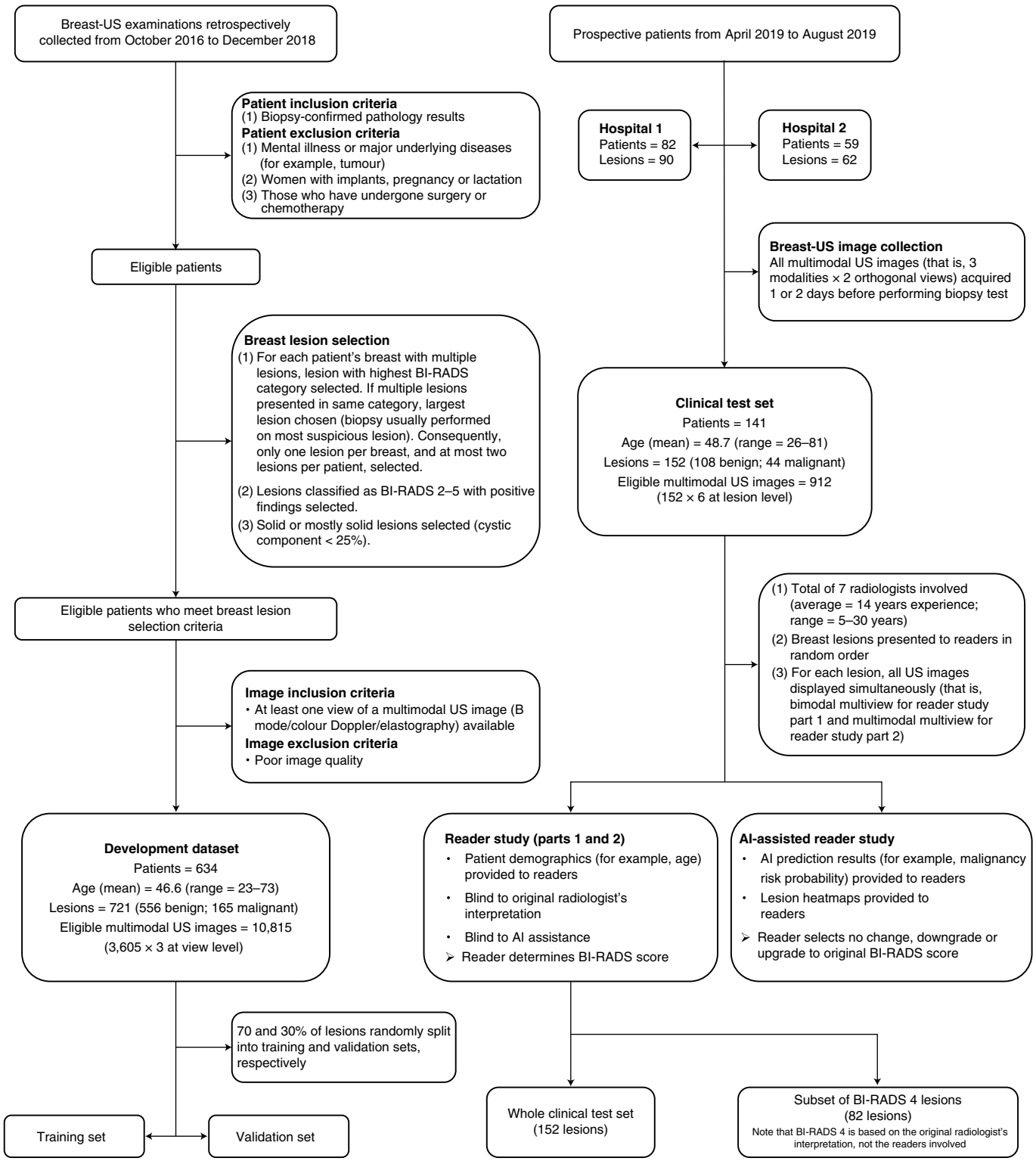

Fig. 2 | Overview of the retrospective and prospective workflow. Owing to the nature of retrospective investigation, multiview US images are not completely preserved and/or view descriptions are not clearly labelled in some lesions. To utilize the existing large multimodal US imaging dataset, the AI system was developed based on view-level multimodal US images (transverse or longitudinal view not distinguished). Lesion-level multimodal US images with explicitly labelled orthogonal views were collected in a prospective setting. It should be pointed out that all BI-RADS categories in this study were determined on US imaging exclusively.

Therefore, we incorporated a heatmap that highlights the important regions relevant to AI predictions to aid radiologists in their assessment of US images (see examples of AI prediction basis in Fig. 4).

Heatmaps clearly facilitated assessment of the ROI of each US imaging mode with potential clinical value (see more heatmap interpretations with various lesion types in Supplementary Figs. 6

Table 1 | Patient demographics and breast lesion characteristics

Characteristics	Retrospective dataset	Prospective dataset
Number of patients	634	141
Age (years) (mean)	46.6 (23–73)	48.7 (26–81)
Number of lesions	721	152
Lesion size (mm)		
<10	199 (27.6%)	36 (23.7%)
10–19.9	334 (46.3%)	69 (45.4%)
20–29.9	135 (18.7%)	37 (24.3%)
>30	53 (7.4%)	10 (6.6%)
Lesion depth (mm)		
<5	343 (47.6%)	64 (42.1%)
5–9.9	282 (39.1%)	68 (44.7%)
10–14.9	86 (11.9%)	16 (10.5%)
>15	10 (1.4%)	4 (2.7%)
BI-RADS category[a]		
2	3 (0.4%)	1 (0.7%)
3	326 (45.2%)	58 (38.2%)
4a	232 (32.2%)	49 (32.2%)
4b	84 (11.7%)	19 (12.5%)
4c	44 (6.1%)	14 (9.2%)
5	32 (4.4%)	11 (7.2%)
Lesion type		
Invasive ductal carcinoma	128 (17.8%)	37 (24.3%)
Invasive lobular carcinoma	4 (0.6%)	1 (0.7%)
Ductal carcinoma in situ	9 (1.2%)	2 (1.3%)
Other malignant[b]	24 (3.3%)	4 (2.6%)
Fibroadenoma	227 (31.5%)	46 (30.3%)
Other benign[c]	329 (45.6%)	62 (40.8%)

[a]The BI-RADS category is based on the interpretation of the radiologist who originally performed the US examinations before the biopsy test, not the radiologists involved in the reader study. It should be noted that all BI-RADS categories involved in this study were determined on breast-US images only. [b]Includes non-specific malignant results. [c]Includes adenosis, hyperplasia, benign phyllodes tumours and papillomas.

and 7). To be specific, there were three locations valuable for predicting the risk of breast cancer on US B-mode images (namely, the lesion margin, the echo pattern inside the lesion and the presence/absence of calcifications). With regard to US colour Doppler images, the emphasis of the heatmaps focused on the locations of the vasculature, especially for malignant lesions with abundant angiogenesis. In regard to heatmaps derived from US elastography, we found two meaningful locations: a high stiffness region and a marginal region. To some extent, the explainability features (that is, the heatmap used in this study) provided potential insight that could assist clinicians in their exploration and visualization of the suspicious lesions when there is inconsistency between the original assessment of clinicians and AI BCRS prediction.

For the purpose of investigating the advantageous use of explainability features by radiologists, we designed another AI-assisted reader study using the original readers' BI-RADS scores in reader study part 2 as the baseline for each reader (Fig. 5). In addition to the original multimodal multiview US images, corresponding heatmaps and AI predictions of malignancy risk probability were both presented to the same readers to help them understand the justification of the AI system. According to the feedback of the readers, the AI system could potentially guide them to make a better clinical BI-RADS decision. Specifically, if the AI BCRS prediction is in line with the clinician's original assessment, it would increase the confidence levels of clinicians for making BI-RADS decisions. Otherwise, if discrepancy exists, the heatmaps could aid clinicians in their understanding of the basis of the AI prediction decision, then guide them in analysing features of the respective highlighted regions.

Table 2 shows each reader's BI-RADS adjustments (no change, upgrade or downgrade) with the assistance of our AI system. After re-evaluating the updated BI-RADS categorization, we found that the readers preferred to downgrade more BI-RADS categories in benign lesions and upgrade more BI-RADS categories in malignant lesions, in accordance with our expectations. Moreover, with respect to the optimization of binary decision on follow-up/biopsy (that is, BI-RADS 3 versus 4a+), we observed that three more biopsies on average were avoided from 108 benign lesions and one more biopsy on average was recommended from 44 malignant lesions.

Discussion

A lack of consensus (that is, variability in inter- and intra-reader reproducibility) and high false positive rates[6] in breast-US examinations have been widely recognized; thus, both clinicians and engineers have put a substantial amount of effort towards improving the diagnostic output of breast-US examinations. An AI system that combines high sensitivity and specificity with consistency of interpretation could serve as an assistance for clinicians. Here, we developed a deep-learning-based breast cancer risk prediction AI system using over 10,000 US images, including US B-mode, US colour Doppler and US elastography images. Such an AI system using multimodal multiview US images closely matches routine clinical breast-US scanning and decision-making. It is worth noting that US elastography is a very recently developed technique that is not yet widely available in existing US scanning machines in the clinic. To make the AI system more applicable to clinical use, our proposed AI system incorporates both bimodal and multimodal models, and therefore has the ability to meet diagnostic requirements in accordance with either the fourth- or fifth-edition BI-RADS lexicon in clinical settings. We report that our AI system can be used to assess breast lesions at a level comparable to that of experienced human experts in a prospective setting. We have demonstrated that the incorporation of explainability features enhanced the AI system's clinical applicability, and further investigated the potential advantages of the understandable AI system in guiding clinicians' decision-making.

The majority of existing work in which deep learning was used for automated diagnoses of breast-US imaging was exclusively developed on retrospectively collected single-view US B-mode images, which deviated from existing BI-RADS lexicon reporting standards[29]. More specifically, in a study of 520 sonograms[18], deep-learning architecture was proposed to assess US breast lesions with an AUC of 0.896 ± 0.039. Later, in a larger dataset of 7,408 breast-US B-mode images[19], a deep-learning framework reported an AUC of >0.9. In addition, the performance of previous AI studies was not convincing due to a lack of large human comparator groups in a prospective setting. Particularly, generic deep-learning analysis software[38] has been implemented to classify breast cancer with an accuracy comparable to two radiologists and one medical student. Subsequently, a deep convolution neural network was proposed to mimic human decision-making in the classification of US breast lesions; however, its performance was compared with that of only two human radiologists[39]. Most recently, although various strategies[20–28] have been further investigated to strengthen the diagnosis performance of AI algorithms, a lack of evidence of interpretability for deep-learning systems, as well as a lack of guidance for relevant clinical decision-making, was identified as an impediment to the likelihood of clinical applicability.

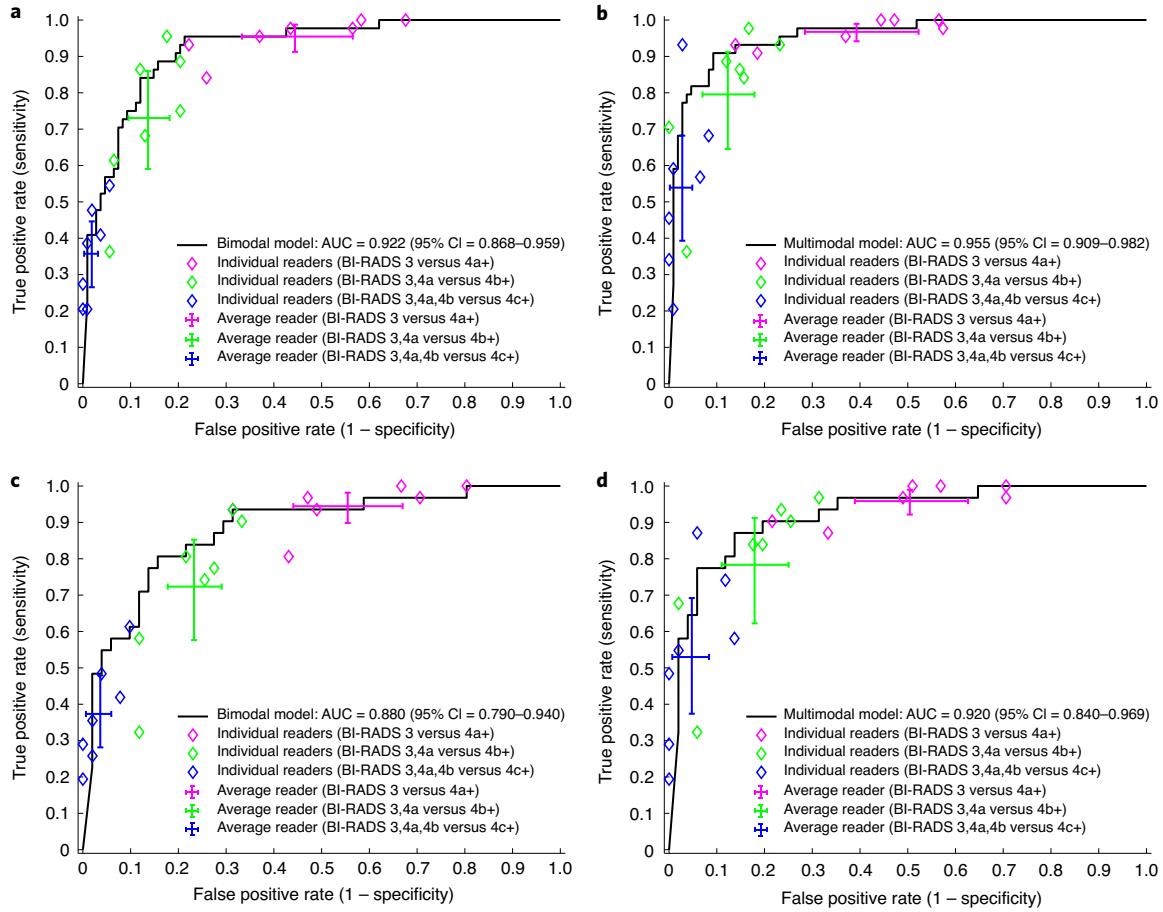

Fig. 3 | Performance of the AI system and readers in predicting the risk of breast cancer on the prospective clinical test set using lesion-level US images. **a**–**d**, The results correspond to bimodal (**a**,**c**) and multimodal (**b**,**d**) images of the whole set (**a**,**b**) and BI-RADS 4 subset (**c**,**d**). The performance of our AI system was compared with each of the seven readers and the averaged performance of the seven readers at three decision modes. Error bars represent the 95% CIs, which were calculated based on 1,000 bootstraps of the data.

In contrast, we have shown that the proposed AI system has high overall performance with the independent clinical test set used. Under the presence of US B-mode and US colour Doppler images, the bimodal model accomplished an AUC of 0.890 (95% CI = 0.829–0.935) for transverse views, an AUC of 0.898 (95% CI = 0.839–0.942) for longitudinal views and an overall AUC of 0.922 (95% CI = 0.868–0.959) for the combination of both views. With the comprehensive information of US B-mode, US colour Doppler and US elastography images, the AUC of the multimodal model was notably higher, reaching 0.936 (95% CI = 0.885–0.970) for transverse views, 0.935 (95% CI = 0.883–0.968) for longitudinal views and 0.955 (95% CI = 0.909–0.982) for the integration of two candidates. We observed that the multimodal model achieved substantially superior performance to the bimodal model, which implied the clinical value of US elastography for predicting breast cancer risk in the fifth edition of BI-RADS lexicon.

Taking both transverse and longitudinal views into clinical consideration is needed to ensure the quality of breast-US examinations in clinical practice[40]. In other words, the standard breast-US examination protocol requires radiologists to scan and analyse both views. Our study demonstrated that performance on lesion-level US images was in general better than on view-level US images. However, performance improvement was not remarkable owing to the fact that the corresponding diagnostic results evaluated from transverse and longitudinal views might be overlapped (Fig. 4a). As stated by clinicians, the information extracted from the second view predominantly serves as an auxiliary diagnosis to that from the first view in clinical decisions. The clinical value of evaluating breast cancer risk through a second view by both clinicians and the AI system is to diminish the risk of bias (Fig. 4b) and ensure clinical outcomes. As a consequence, to conform to the routine workflow of clinicians, our AI system was designed and evaluated on lesion-level US images.

To ensure a fair comparison between AI and clinicians, the AI system should be implemented in a prospective setting[30]. Therefore, our AI system was tested and compared with seven clinicians in an independent clinical dataset collected after development of the algorithm and from patients who underwent US examination with orthogonal views in two hospitals. Our AI system showed the ability to predict risk of breast cancer at a level comparable to that of the experienced radiologists in the study in terms of accuracy, sensitivity and specificity, as determined from ROC curves. With the additional information of US elastography in part 2 of the reader study, the AI system showed improved specificity of breast lesion assessment without loss of sensitivity for the BRCS 4a+ mode. These results were consistent with both our reader studies and previous clinical studies[41–43] via considering BI-RADS category 4a or higher as test positive for malignancy (BI-RADS 3 versus 4a+). Notably, despite the fact that both the AI system and reader performance in part 2 of the reader study improved relative to part 1, the AI system

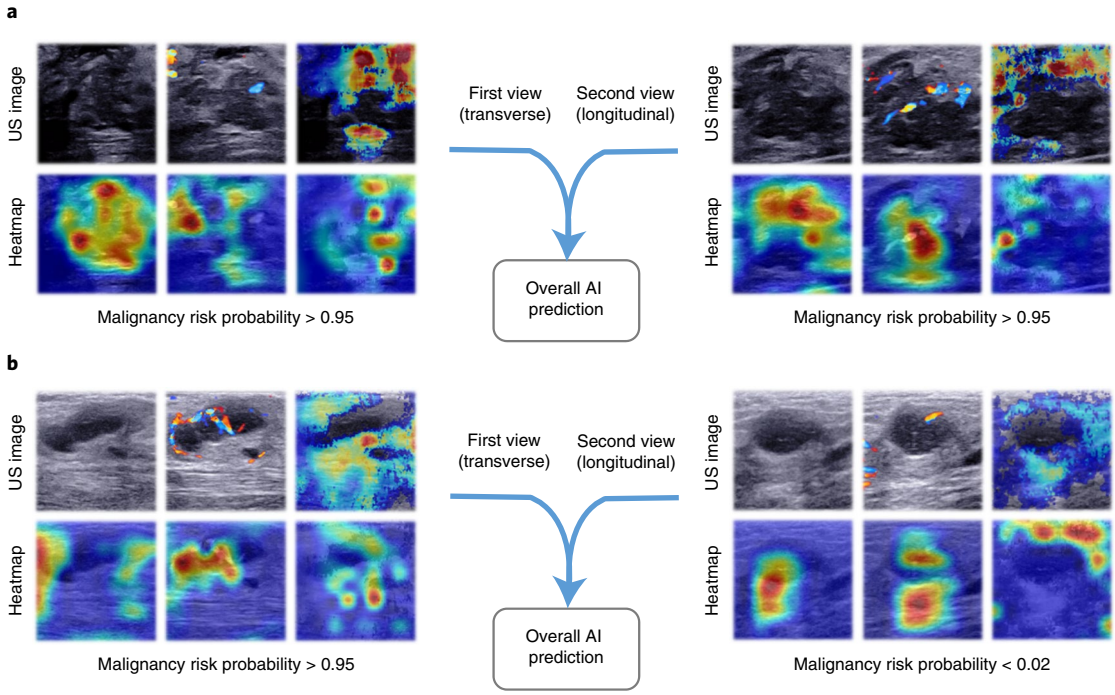

Fig. 4 | Examples of AI prediction basis. a,**b**, Colour-coded heatmaps overlaid with the corresponding US images were generated from the final convolution layer using the Grad-CAM approach. **a**, Examples of transverse (first) and longitudinal (second) views of a malignant lesion. Combining two malignancy probability scores, the AI system correctly classified the true positive lesion. **b**, Example of a benign lesion. Despite the fact that there is a disagreement between the false positive from the transverse view and the true negative from the longitudinal view, combing orthogonal views may mitigate overall prediction bias. The prediction basis can assist human experts to understand the justification of the decision made by the AI system.

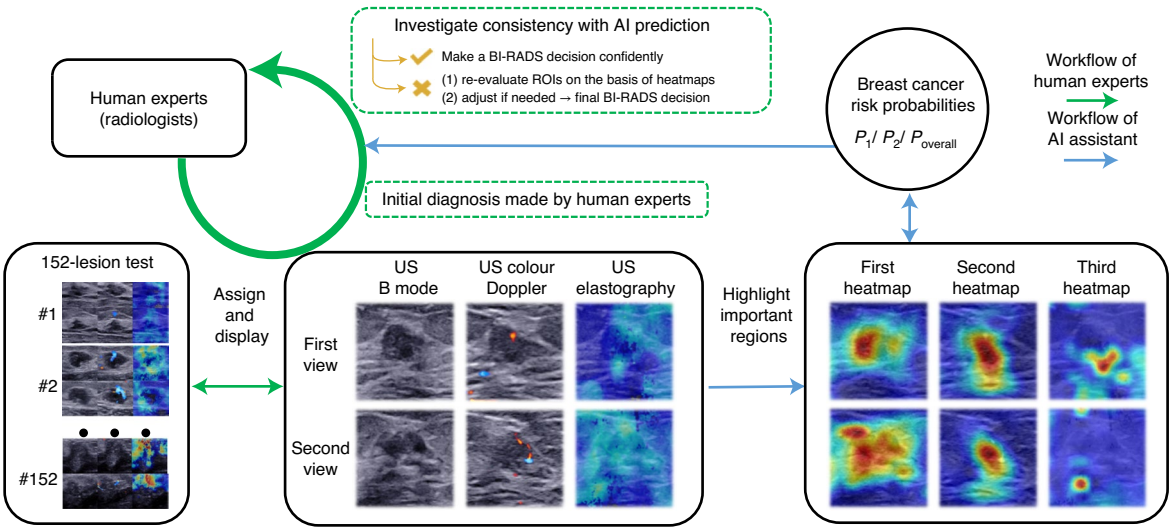

Fig. 5 | Understandable AI system potentially guides human experts to make better clinical decisions. The workflow of the AI-assisted reader study is shown. For each randomly assigned and displayed test lesion, the reader examined information from three aspects, including original multimodal multiview US images, the corresponding highlighted heatmaps and the AI prediction on malignancy risk probability, then finally determined the BI-RADS categorization.

showed a more favourable improvement than that of human experts, demonstrating the AI system's potential advantages for comprehensive interpretation of multimodal multiview US images.

It should be noted that the decision to upgrade or downgrade BI-RADS 4 lesions is a major challenge for clinicians. According to the American College of Radiology BI-RADS Atlas[29], diagnoses of BI-RADS 4 are confirmed by the presence or absence of breast cancer via biopsy. With the aim of reducing unnecessary biopsy for patients classified as BI-RADS 4, an AI diagnostic tool that provides deeper clinical insight, particularly for patients in the BI-RADS 4 category, is highly desirable[44]. On the test subset with only BI-RADS 4 lesions, both reader and AI system performance dropped relative to that in the whole test set as a result of only suspicious malignancy lesions being involved. However, the AI system still achieved

Table 2 | Summary of the changes of BI-RADS and biopsy decisions made by radiologists R1–R7 in completing the AI-assisted reader study

	Adjustment	Benign lesions ($n=108$)					Malignant lesions ($n=44$)				
			3 versus 4a+[a]		Biopsy			3 versus 4a+[a]		Biopsy	
			+	−	Solo	+AI		+	−	Solo	+AI
R1	No change	102	1	0	51	52	42	0	0	44	44
	Downgrade	5					1				
	Upgrade	1					1				
R2	No change	96	7	4	20	23	30	2	0	40	42
	Downgrade	5					2				
	Upgrade	7					12				
R3	No change	70	6	11	61	56	23	0	0	44	44
	Downgrade	28					5				
	Upgrade	10					16				
R4	No change	98	2	4	15	13	25	1	0	41	42
	Downgrade	4					0				
	Upgrade	6					19				
R5	No change	78	4	16	48	36	42	0	0	44	44
	Downgrade	26					1				
	Upgrade	4					1				
R6	No change	90	1	6	40	35	33	2	0	42	44
	Downgrade	12					5				
	Upgrade	6					6				
R7	No change	103	1	2	62	61	41	1	0	43	44
	Downgrade	4					0				
	Upgrade	1					3				

[a]Number of lesions that the radiologists altered from BI-RADS 3 to 4a+ (+) or from BI-RADS 4a+ to 3 (−) using computer assistance. Considering BI-RADS 4a+ as test positive for malignancy, a recommendation for biopsy was made by the readers without (solo) or with (+AI) computer assistance.

an AUC of 0.880 on bimodal US images and an AUC of 0.920 on multimodal US images, reaching the same high level attained by human experts.

The black box nature of deep learning has been identified as an obstacle to establishing trust with human experts in clinical practice[45]. The United States Food and Drug Administration requires that any AI system that supports clinical decision-making must explain the rationale for its decisions to enable the clinicians to review the recommendation basis[46]. As indicated by a previous study[47], an understandable AI tool not only increases confidence levels of clinicians in making or excluding a diagnosis but also provides educational feedback that will benefit non-experts such as non-radiologist clinicians or general radiologists. Therefore, to enhance the AI system's likelihood of clinical applicability, we incorporated explainability features (that is, heatmaps that make the AI decisions understandable to human experts).

As opposed to a malignancy risk score, such heatmaps allowed clinicians in our studies to visualize the basis of the AI prediction—information that could then be used to help guide their clinical decisions. In particular, the heatmap displays depicted highly relevant and informative features such as the value of the lesion margin, the interior echo pattern in US B-mode images, the presence of vasculature in US colour Doppler images and the stiffness distribution in US elastography images. Since malignant breast lesions were typically associated with distinct features, such as irregular shapes, non-circumscribed margins (for example, indistinct margins or lobulated margins), heterogeneous echo areas inside the lesions, abundant vessel signals and high stiffness regions, the highlighted regions in the heatmaps were helpful to identify these representative characteristics of malignant lesions. In contrast, most benign breast lesions rarely have detectable vasculature and have relatively uniform stiffness mapping. Therefore, the entire region in the neighbourhood of benign lesions is of importance in AI interpretation.

It must be pointed out that it is the clinicians who make the final clinical BI-RADS decisions. The heatmaps were proposed to aid clinicians in their understanding of the prediction basis of the AI decisions, and then to guide them in analysing/re-evaluating the highlighted regions in each imaging mode if necessary, resulting in the potential to make a better clinical BI-RADS decision. In the AI-assisted reader study, we demonstrated the potential advantages of an understandable AI system in improving the radiologists' clinical outcomes (that is, an average decrease of 7% biopsy in benign lesions and an average increase of 2% biopsy in malignant lesions). Such an observation implied that our AI system was more effective in improving the specificity rather than sensitivity of the clinicians' decisions, which could further diminish the false positive rate of breast-US examinations. Future work can now directly augment the size of prospective datasets for comprehensively evaluating the advantages of understandable AI systems. Despite the fact that the interpretability of deep learning is still at an early stage, our AI-assisted workflow has demonstrated potential insights for clinicians' decision-making.

There are several limitations to our study. As we collected data from two hospitals in China, the proposed deep-learning system may only apply to the Asian population in current form. All patient data were acquired using Aixplorer US scanners while excluding the variability generated from different US machines. Therefore, there is a limit to the conclusions that can be drawn about generalizability.

Future validation of the AI system should include data from several scanner manufacturers. Another limitation is that the patient population in the utilized test set is not representative of the natural distribution of patients with cancer in the screening population. Since only patients with biopsy-confirmed results were included in this study, the dataset lacked information for patients who underwent follow-up procedures. Moreover, further development of our system should include patients' medical histories, which would be relevant and could improve the performance of the AI decision.

Lastly, although data augmentation was implemented here to alleviate the influence of unavailable data and overfitting, such a methodology did not mimic true physical phenomena (for example, bad gel coupling, acoustic phase reverberation or US attenuation) and instrumentation settings (for example, time-gain compensation, dynamic range compression or nonlinear filtering) impacting the appearance of US images. In other words, the data-augmentation procedures implemented in this study have less capability to reflect the variance attributed to US physics and inter-subject variability of clinical US scans. Future studies will require the broader variability of clinical multimodal US images encountered in practice to develop the AI system.

In conclusion, these results represent a step towards automated breast cancer risk prediction through the use of a deep-learning-based AI system. The combination of multimodal and multiview US images in the workflow matches clinical US examination/screening reporting standards. The heatmap display and breast cancer risk probability provided by this understandable AI system has the potential to guide clinical decisions in a prospective setting. Such a clinically applicable AI system may be incorporated into future breast cancer US screening, as well as support assisted or second-read workflows.

Methods

Ethical approval. Our retrospective study was approved by the institutional review board of the hospitals, with a waiver granted for the requirement of informed consent. The prospective study was not an interventional prospective trial and was performed under guidelines approved by the institutional review board. All of the participants were informed about all aspects of the prospective study and gave informed, written consent. All images processed for this investigation were de-identified in accordance with the Health Insurance Portability and Accountability Act before transfer to study investigators.

Development dataset and clinical test dataset. We retrospectively collected breast diagnostic reports from patients who underwent US examination between October 2016 and December 2018 at The First Affiliated Hospital of Anhui Medical University (hospital 1) and Xuancheng People's Hospital of China (hospital 2). The breast-US examinations/scans were performed using an Aixplorer US system (SuperSonic Imagine) equipped with either an SL15-4 or an SL10-2 linear array transducer under the preset settings. Examinations/scans were performed by one of eight radiologists, each of whom had over 10 years of experience in breast US. According to the availability of multimodal US images (that is, US B-mode images for the assessment of morphology, US colour Doppler images for the assessment of vascularity and US elastography images for the assessment of tissue elasticity), as well as corresponding pathology results, a total of 10,815 US images from 721 breast lesions (165 positive for cancer) of 634 patients were selected and then assigned to the development dataset. Details of the inclusion and exclusion criteria of the development dataset are shown in Fig. 2. As part of the clinical routine in the hospitals, between 10 and 20 images were taken and saved for each breast lesion, including three to six images each for US B-mode, US colour Doppler or US elastography imaging (US elastography in this study refers to shear wave elastography, a quantitative measurement of tissue stiffness expressed in kPa). Owing to the nature of retrospective study, multiview US image data were not completely preserved in all patients, and in some cases the corresponding view descriptions were not clearly labelled. As a result, the retrospectively collected multimodal US images were from either the transverse view, the longitudinal view or both. To utilize the existing large multimodal US imaging dataset for model development, the development dataset included 3,605 view-level multimodal US images (transverse or longitudinal view not distinguished), and biopsy-confirmed pathology results served as the ground truth. Lesions in the development dataset were randomly assigned to a training set (70%) or a validation set (30%).

To further evaluate the model's performance and compare it with that of clinicians, we prospectively collected 90 lesions from 82 consecutive patients in hospital 1 and 62 lesions from 59 consecutive patients in hospital 2. These patients underwent multimodal multiview breast-US scans from April 2019 to August 2019. The breast-US examinations were performed by one of three radiologists who had 12, 9 and 14 years of experience in breast US using the Aixplorer US system (SuperSonic Imagine) with the preset instrument settings. To distinguish from the view-level US images in the retrospective dataset, we defined the collected US images from each lesion in the prospective setting as lesion-level data, which represented multimodal US images with two orthogonal views. These data were collected after development of the AI system and were thus not used for training of the model. The ground truth for the clinical test set was labelled based on core-needle biopsy or surgical confirmation of cancer via manual review of the pathology note.

Table 1 lists the patient demographics and breast lesion characteristics in this study. It should be emphasized that all BI-RADS categories included in this study were determined on breast-US imaging only. Pathology results were available for the patients or lesions classified as BI-RADS 2 or 3 following breast US due to either classification as BI-RADS 4a or higher following mammography or magnetic resonance imaging or requests from patients themselves.

Imaging preprocessing. Before the introduction of US images into the deep-learning network, a custom annotation tool (written in JavaScript using Electron 5.2 as its User Interfaces framework) was utilized by the radiologists to eliminate irrelevant information, such as text and instrument settings, in raw US images via a square segmentation mask. More specifically, during the US examinations, the radiologists were required to manually place a sampling box region where the corresponding vascularity (via US colour Doppler image) and elasticity (via US elastography image) measurements could be performed. Since the US B-mode and US elastography images were displayed on the screen simultaneously, there was a corresponding box region for US B-mode image as well. As a consequence, such box regions in each of the three US images provided guidance to the radiologists to eliminate irrelevant information.

The size of the square segmentation masks was first adjusted to the maximum, to maintain a sufficient margin (the margin is defined as the distance between the lesion boundary and the boundary of the segmentation mask itself) while not exceeding the sampling box boundary. Then, each of the segmentation masks was further altered by experienced radiologists to ensure similar lesion-to-mask proportions in each imaging mode. Finally, all of the cropped US images were resized to a 300×300 aspect ratio for quality control and fed as the network inputs. In summary, only multimodal US images of the lesions were retained for model development.

Model development. Supplementary Fig. 1 illustrates the detailed architecture of our proposed multipathway deep-learning network for multimodal US image classification. The front part of our network processed each imaging modality exclusively. Feature maps were first extracted by a convolutional neural network ResNet-18 model[34] with pre-trained weights from the ImageNet dataset[12] via transfer learning[48]. A SENet module[31] was then implemented to adaptively recalibrate channel-wise features by explicitly modelling interdependencies between channels (that is, feature channel relationships) in each pathway. Next, several convolutional layers followed by BatchNorm and ReLU as the activation functions were used. At the end of each pathway, the 1×1 two-dimensional average pooling layers aggregated features from the spatial domain. To obtain the global multimodal feature vector, the outputs of these pathways were concatenated by merging features from each modality. To evaluate breast malignancy risk, probability was predicted by two fully connected layers and a softmax function. The learning object was to minimize the sum of the softmax cross entropy loss. To avoid overfitting, dropout with a probability of 0.5 was implemented throughout fully connected layers. With regard to the bimodal model, the deep-learning architecture only retained two pathways to match the inputs of US B-mode and US colour Doppler images.

We implemented our models on the PyTorch (version 3.7.4; pytorch.org) deep-learning framework with a Manjaro Linux 18.04 computer equipped with two Intel Xeon central processing units and two NVIDIA RTX 2080 Ti graphics processing units for training, validation and clinical testing. Optimization of the model was performed using an adaptive moment estimation (ADAM) optimizer in a batch size of 20 with an initial learning rate of 0.0001, which then decayed every 50 epochs with a decay factor of 0.5. The maximum iteration was set to 13,000 steps and an early stopping criterion was used to terminate training due to the absence of further improvement in both loss and accuracy. During model development, we augmented our training data by applying colour jitter and geometric transformations. More specifically, brightness levels of 0–2, contrasts of 0–2, saturation levels of 0.3–1.7, hues of −0.5–0.5 (since US colour Doppler and US elastography images have adopted the default colour bar in clinical settings, hue was therefore not applied to these two imaging modes) and entire image rotation angles of −8°–8° were used for augmentation. All of these parameters were randomly and uniformly selected in the predefined ranges.

The bimodal model was trained to perform breast cancer risk prediction using both US B-mode and US colour Doppler images. The multimodal model was developed by taking US B-mode, US colour Doppler and US elastography imaging into consideration. Overall, the AI system was trained to take view-level US imaging candidates and automatically produced one malignancy risk probability score.

To predict the breast cancer risk of patients corresponding to lesion-level data with orthogonal-view candidates in a prospective setting, we averaged the malignancy probability from the transverse view (P_1) and longitudinal view (P_2) to form a final breast malignancy probability score ($P_{overall}$), as shown in Fig. 1. The ROC of the model was plotted by varying the threshold of overall probability ($P_{overall}$) in the interval 0–1. The AUC is the model's measure of performance, with a maximum value of 1.

Interpretability of the AI system. To assure trust by human experts, an understandable decision-making process is desired in clinical practice. A class activation mapping (CAM)-based approach[49], which uses a global average pooling layer at the end of neural networks instead of a fully connected layer, was recently proposed to highlight the class-specific discriminative regions. Despite the fact that excellent heatmaps of attention can be generated, the CAM approach has restrictions using global average pooling, and can only be applied to visualize the final-layer heatmaps.

The gradient-weighted CAM (Grad-CAM) technique[32], as a generalization to CAM, is applicable to a broader range of convolutional neural network model families without architectural changes or retraining, resulting in promising textual explanations for model decisions. Given the model prediction to a target class c, we defined the neuron importance weights α_k^c via global average pooling using equation (1):

$$\alpha_k^c = \frac{1}{S}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

where y^c is the gradient of the score for class c before being fed through the softmax function, A_{ij}^k corresponds to feature maps of a convolutional layer and S is the size of each feature map from the last block. The heatmap M^c is generated by a weighted combination of forward activation maps followed by a ReLU using equation (2):

$$M^c = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k\right) \quad (2)$$

where ReLU represents a rectified linear unit.

Supplementary Fig. 8 shows heatmap examples of biopsy-confirmed breast-US lesions generated by both CAM and Grad-CAM techniques. It was observed that CAM and Grad-CAM can identify similar ROIs with subtle discrepancies. Therefore, as a more generalized approach, Grad-CAM was adopted here to create the heatmap from the final convolutional layer in a test image, which could aid human experts in their understanding of the justification of the AI system for breast cancer risk prediction. Since the coarse heatmap was the same size as the convolutional feature map, we interpolated the grid size to the original size of the US image and then overlaid the corresponding image patches.

Operating point selection. We defined three operating points that represented, respectively, the decision-making modes of BCRS 4a+, BCRS 4b+ and BCRS 4c+, as a way to compare the model with reader performance. The three operating points were chosen by thresholding the overall malignancy risk probability to match a likelihood of 10, 50 or 90%. It is important to note that the proposed three operating points were primarily used for comparison between the AI system and the readers who used BI-RADS categorization in clinical practice. Selection of optimal parameters of operating points with respect to the proper trade-off between sensitivity and specificity is still under investigation.

Reader study. A two-part reader study was conducted to compare the performance of the model with that of seven radiologists with an average of 14 years of clinical experience (range = 5–30 years). A total of 152 breast lesions (44 positive for cancer) from 141 patients in the clinical test set were presented to readers and the model in random order. The readers were blinded to each other, to the original radiologist's interpretation and to the deep-learning model assessment. Each reader reviewed the same set of lesions independently and applied the BI-RADS criteria to determine a BI-RADS score[29,50]. The main features considered by readers on assessment of US images were summarized as follows. Lesion margin, lesion morphology, aspect ratio, interior acoustic echo and micro-calcification were mainly considered for US B-mode images. For US colour Doppler images, the readers focused on the position of vascularity and the number of detected blood signals. With respect to US elastography images, both stiffness distribution and maximum elasticity were considered by the readers.

Reader study part 1 included both transverse and longitudinal views of US B-mode and US colour Doppler images for each breast lesion. After the readers completed part 1, the same lesions were re-presented to the same readers, now with additional information of US elastography, as reader study part 2. For each lesion, readers were given access to associated patient demographics (mainly age information), whereas the deep-learning model did not have access to this information.

Performance comparison between the model and readers was evaluated from two aspects. First, we computed the sensitivity and specificity of the model at each of the three BCRS modes (4a+, 4b+ and 4c+). Next, we compared the model's BCRS results with each reader using corresponding BI-RADS 3 versus 4a+, BI-RADS 3,4a versus 4b+ and BI-RADS 3,4a,4b versus 4c+ modes, respectively. Second, comparisons for cancer risk prediction were made between the model and the average reader. We computed the average reader sensitivity and specificity by averaging the seven individual reader sensitivities and specificities at three modes, respectively. We adjusted the threshold of the model's sensitivity to match the average reader sensitivity and then compared the specificity. To compare sensitivity, the model's specificity was set to match the average reader specificity.

AI-assisted reader study. To evaluate the advantageous use of explainability features (that is, heatmaps) by the radiologists in guiding clinical decisions, an AI-assisted reader study was carried out. The same 152 breast lesions in the clinical test set were presented to the same seven readers again. For each lesion, together with the original multimodal multiview US images, the corresponding lesion heatmaps and AI prediction on malignancy risk probability were provided to the readers at the same time. Each reader was given the opportunity to upgrade or downgrade their original BI-RADS decisions, where the baseline BI-RADS scores were the results in reader study part 2.

The heatmap allowed the readers to assess the ROI with potential clinical value in each imaging mode. The malignancy risk prediction results from each individual view (P_1 and P_2) and the overall malignancy risk probability ($P_{overall}$) provided quantitative estimations to the readers, which might assist the readers to select the view with the highest probability of malignancy to review.

To evaluate the relevance of the heatmap to the clinician, we requested that all radiologists involved in our AI-assisted reader study summarize the information they obtained from the heatmaps over the whole clinical test set. Specifically, the readers were required to select the existing terms defined in BI-RADS lexicon to match the information they observed from the heatmaps. After collecting the information individually from the readers, we determined the relevant features of the heatmap based on the majority of the readers' observations (that is, at least four readers identified the informative features).

Statistical analysis. The diagnostic performances of the bimodal and multimodal models on the basis of view- and lesion-level US images were expressed as AUC values and compared using Delong's test[51] with binomial exact confidence intervals. The confidence intervals in sensitivity and specificity were computed based on 1,000 bootstraps of the data. McNemar's test was used to calculate two-sided P values for the sensitivity and specificity between the models and human readers. $P < 0.05$ was considered to be the threshold for a statistically significant difference. All statistical analyses were performed using standard statistical software (SPSS (version 22.0; IBM) or MedCalc (version 19.0.7; MedCalc software)).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The authors declare that the main data supporting the results of this study are available within the paper and its Supplementary Information. The raw US datasets from The First Affiliated Hospital of Anhui Medical University and Xuancheng People's Hospital of China cannot be made available for public release because of patient privacy. However, some data can be made available for academic purposes from the corresponding author on reasonable request, subject to permission from the institutional review boards of the hospitals.

Code availability

The deep-learning models were developed using standard libraries and scripts available in PyTorch. The pre-trained weights for ResNet were obtained from the torchvision library. Custom codes and the annotation tool for the deployment of the system are available for research purposes from the corresponding author upon reasonable request.

Received: 29 October 2019; Accepted: 8 March 2021;
Published online: 19 April 2021

References

1. Jemal, A. et al. Global cancer statistics. *CA Cancer J. Clin.* **61**, 69–90 (2011).
2. Bray, F., Jemal, A., Grey, N., Ferlay, J. & Forman, D. Global cancer transitions according to the Human Development Index (2008–2030): a population-based study. *Lancet Oncol.* **13**, 790–801 (2012).
3. Forouzanfar, M. H. et al. Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. *Lancet* **378**, 1461–1484 (2011).
4. Berg, W. A. 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. *J. Am. Med. Assoc.* **307**, 1394–1404 (2012).
5. Ohuchi, N. et al. Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial. *Lancet* **387**, 341–348 (2016).

6. Berg, W. A. et al. Ultrasound as the primary screening test for breast cancer: analysis from ACRIN 6666. *J. Natl Cancer Inst.* **108**, djv367 (2015).
7. Lee, H.-J. et al. Observer variability of Breast Imaging Reporting and Data System (BI-RADS) for breast ultrasound. *Eur. J. Radiol.* **65**, 293–298 (2008).
8. Abdullah, N., Mesurolle, B., El-Khoury, M. & Kao, E. Breast Imaging Reporting and Data System lexicon for US: interobserver agreement for assessment of breast masses. *Radiology* **252**, 665–672 (2009).
9. Doi, K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput. Med. Imaging Graph.* **31**, 198–211 (2007).
10. Yassin, N. I., Omran, S., El Houby, E. M. & Allam, H. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: a systematic review. *Comput. Methods Programs Biomed.* **156**, 25–45 (2018).
11. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
12. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
13. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
14. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J. Am. Med. Assoc.* **316**, 2402–2410 (2016).
15. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
16. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
17. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
18. Cheng, J.-Z. et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci. Rep.* **6**, 24454 (2016).
19. Han, S. et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Phys. Med. Biol.* **62**, 7714–7728 (2017).
20. Tanaka, H., Chiu, S.-W., Watanabe, T., Kaoku, S. & Yamaguchi, T. Computer-aided diagnosis system for breast ultrasound images using deep learning. *Phys. Med. Biol.* **64**, 235013 (2019).
21. Moon, W. K. et al. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Comput. Methods Prog. Biomed.* **190**, 105361 (2020).
22. Shin, S. Y., Lee, S., Yun, I. D., Kim, S. M. & Lee, K. M. Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. *IEEE Trans. Med. Imaging* **38**, 762–774 (2018).
23. Qi, X. et al. Automated diagnosis of breast ultrasonography images using deep neural networks. *Med. Image Anal.* **52**, 185–198 (2019).
24. Zhang, Q. et al. Deep learning based classification of breast tumors with shear-wave elastography. *Ultrasonics* **72**, 150–157 (2016).
25. Qian, X. et al. A combined ultrasonic B-mode and color Doppler system for the classification of breast masses using neural network. *Eur. Radiol.* **30**, 3023–3033 (2020).
26. Zhou, Y. et al. A radiomics approach with CNN for shear-wave elastography breast tumor classification. *IEEE Trans. Biomed. Eng.* **65**, 1935–1942 (2018).
27. Shan, J., Alam, S. K., Garra, B., Zhang, Y. & Ahmed, T. Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods. *Ultrasound Med. Biol.* **42**, 980–988 (2016).
28. Zhang, E., Seiler, S., Chen, M., Lu, W. & Gu, X. BIRADS features-oriented semi-supervised deep learning for breast ultrasound computer-aided diagnosis. *Phys. Med. Biol.* **65**, 125005 (2020).
29. Mendelson, E. B. et al. ACR BI-RADS® Ultrasound In *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System* (American College of Radiology, 2013).
30. Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *Br. Med. J.* **368**, m689 (2020).
31. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/CVPR.2018.00745 (IEEE, 2018).
32. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)* https://doi.org/10.1109/ICCV.2017.74 (IEEE, 2017).
33. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations* (2015); preprint at https://arxiv.org/abs/1409.1556 (2014).
34. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/CVPR.2016.90 (IEEE, 2016).
35. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking theInception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/CVPR.2016.308 (IEEE, 2016).
36. Chen, J. H. & Asch, S. M.Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N. Engl. J. Med.* **376**, 2507–2509 (2017).
37. Park, S. H. & Han, K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* **286**, 800–809 (2018).
38. Becker, A. S. et al. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *Br. J. Radiol.* **91**, 20170576 (2018).
39. Ciritsis, A. et al. Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making. *Eur. Radiol.* **29**, 5458–5468 (2019).
40. Wang, Y. et al. Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning. *Ultrasound Med. Biol.* **46**, 1119–1132 (2020).
41. Berg, W. A. et al. Shear-wave elastography improves the specificity of breast US: the BE1 multinational study of 939 masses. *Radiology* **262**, 435–449 (2012).
42. Lee, S. H. et al. Evaluation of screening US-detected breast masses by combined use of elastography and color Doppler US with B-mode US in women with dense breasts: a multicenter prospective study. *Radiology* **285**, 660–669 (2017).
43. Cho, N. et al. Distinguishing benign from malignant masses at breast US: combined US elastography and color Doppler US—influence on radiologist accuracy. *Radiology* **262**, 80–90 (2012).
44. Destrempes, F. et al. Added value of quantitative ultrasound and machine learning in BI-RADS 4–5 assessment of solid breast lesions. *Ultrasound Med. Biol.* **46**, 436–444 (2020).
45. Castelvecchi, D.Can we open the black box of AI? *Nature* **538**, 20–23 (2016).
46. *Clinical and Patient Decision Support Software (Draft Guidance)* (Food and Drug Administration, 2018).
47. Lee, H. et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat. Biomed. Eng.* **3**, 173–182 (2019).
48. Shin, H.-C. et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).
49. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/CVPR.2016.319 (IEEE, 2016).
50. Barr, R. G. et al. WFUMB guidelines and recommendations for clinical use of ultrasound elastography: part 2: breast. *Ultrasound Med. Biol.* **41**, 1148–1160 (2015).
51. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L.Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).

Acknowledgements

We are grateful to H. Zheng, Y. Liu, X. Shuai, G. Zhang, J.-S. Zhang, W. Yao and J.-X. Zhang for participation in the reader studies. We acknowledge help from Y. Lu and R. Wodnicki in manuscript revision and editing.

Author contributions

X.Q. conceived of, designed and supervised the project. J.P. and H. Zheng provided clinical expertise and guidance on the study design. X.Q., L.Y. and X.G. developed the deep-learning framework and software tools necessary for the experiments. X.Q., J.P., H. Zheng, X.X., Hao Zhang and C.H. created the datasets, interpreted the data and defined the clinical labels. X.X., C.H., Hanqi Zhang, W.Z. and Q.S. collected the raw US images and patients' pathology results in the clinic. X.Q., L.Y., X.G., Hao Zhang and L.L. executed the research and performed the statistical analysis. K.K.S. advised on the US imaging techniques. X.Q., J.P. and H. Zheng wrote the manuscript. All authors contributed to reviewing and editing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-021-00711-2.

Correspondence and requests for materials should be addressed to X.Q.

Peer review information *Nature Biomedical Engineering* thanks Guy Cloutier and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Xuejun Qian

Last updated by author(s): Feb 14, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All patient data were acquired with Aixplorer ultrasound machine (SuperSonic Imagine, Aix-en-Provence, France). A custom annotation tool (written in JavaScript using Electron 5.2 as its UI framework) was used to eliminate irrelevant information such as text and instrument settings, and for labeling the multimodal ultrasound images.
Data analysis	Python (version 3.7.4) and PyTorch (version 1.3.1): used to train, validate and test the deep-learning models. MATLAB (version 2018a): used to plot ROC curves and for the data analysis. MedCalc (version 19.0.7) and SPSS (version 22.0): used for statistical analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The authors declare that the main data supporting the results in this study are available within the paper and its supplementary information. The raw ultrasound datasets from the First Affiliated Hospital of Anhui Medical University and Xuancheng People's Hospital of China cannot be made available for public release because of patient privacy. However, some data can be made available for academic purposes from the corresponding author on reasonable request, subject to permission from the Institutional Review Boards of the hospitals.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	On the basis of published literature, it is generally agreed that a deep-learning system requires on the order of tens of thousands of examples. Thus, we collected as much available data as possible, according to the inclusion criteria.
Data exclusions	The exclusion criteria were: 1. Patients have mental illness, major underlying diseases (that is, have had a tumour). 2. Done surgery or chemotherapy. 3. Women with implants, pregnancy or lactation. 4. Poor image quality. 5. Images with breast lesions that were not confirmed by histology. For more details, please refer to Methods and to Fig. 2.
Replication	We tested the models using a prospectively collected clinical dataset from two hospitals. This dataset was not used for the training and validation of the models.
Randomization	Samples meeting the inclusion criteria were randomly allocated to the training and validation datasets.
Blinding	The radiologists who participated in the clinical evaluation were blinded to the ground truth and were not involved in dataset collection.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Patient data (100% female) were gathered retrospectively from Oct 2016 to Dec 2018 and prospectively from Apr 2019 to Aug 2019 when carrying out breast-ultrasound examinations in the two hospitals.
Recruitment	The retrospective study was approved by the institutional review board (IRB) of the hospitals with a waiver granted for the requirement of informed consent. For the prospective study, all participants signed an informed consent approved by the IRB. We collected breast-ultrasound images according to predetermined criteria.
Ethics oversight	The First Affiliated Hospital of Anhui Medical University Ethics Committee and Xuancheng People's Hospital Ethics Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	This study is not an interventional prospective trial.
Study protocol	This study is not an interventional prospective trial.
Data collection	All data were collected using the Aixplorer ultrasound system (SuperSonic Imagine, Aix-en-Provence, France) under the preset settings in the Department of ultrasound in two hospitals.
Outcomes	This study is not an interventional prospective trial.