AI Dual Read Breast Cancer Screening

Screening mammography guidelines recommend that scans be reviewed by multiple physicians. While mammography has significantly improved breast cancer outcomes, screening programmes add a substantial workload to healthcare workers who are already facing shortages. There is considerable interest in using artificial intelligence (AI) to address this issue. AI has been shown to outperform individual mammographers in retrospective studies and has potential as a triage tool and as a second reader (AISR). However, few AI models have received clinical approval, and those approved are mostly used for assisted read computer-aided detection, which can increase rather than decrease the burden on healthcare systems. To demonstrate AI's impact in breast cancer screening, evaluations must focus on clinical pathways relevant to the specific populations where AI will be used.

This retrospective study published in Radiology Advances investigates AI performance in an AISR setting, where AI with arbitration replaces the second reader in a double-read setting, aiming to reduce clinical burden without compromising performance. The study builds on a previously developed breast cancer detection model based on US and UK data, which had shown favourable performance compared to radiologists. The current assessment examines the model's external validity using a breast cancer screening cohort from Japan, a population with different breast density and size compared to the original training data. The study reports on model performance before and after fine-tuning and includes a reader study comparing the model to local physicians both individually and in a double-read setting, reflecting potential clinical deployment.

Study Cohort and Data Annotation Methods for AI-Assisted Mammography Screening

Data for this study were retrospectively collected from a cancer screening centre in Japan. Eligible mammograms included those from women of all ages who visited the centre between 2005 and 2021 and had at least 12 months of follow-up data. Patients were excluded if they were symptomatic outside the screening setting, recalled for reasons other than possible malignancy, had previous bilateral mastectomy, opted out of sharing data for research, lacked the standard four mammographic views, or had nondiagnostic image quality. The digital mammograms were acquired using equipment from Hologic (26%) and Fujifilm (74%).

Malignant cases were defined as those confirmed by biopsy within 12 months following screening. Normal cases required at least 12 months of follow-up to rule out interval or missed cancers, while benign cases were either biopsy negative or, in the absence of biopsy data, reported as benign with 12 months of follow-up.

The total eligible cohort consisted of 10,340 patients (43,166 cases), which were randomly assigned into training, validation, or test sets. The selection was based on power estimates for sensitivity, ensuring a clinically representative cohort for Japanese screening centres. The test set included 4,059 patients (17,265 cases), and a subset of 278 cases from the test set was chosen for a reader study to power statistical calculations on sensitivity. The reader study dataset had an approximately equal distribution of normal, benign, and malignant cases, though there were slightly fewer benign cases available.

Annotations for positive cases were performed by an experienced breast imaging reader, who had 26 years of experience and was not involved in the reader study. These annotations were made with access to follow-up imaging, additional views, ultrasound images, biopsy reports, and other metadata, using irregular polygon boxes to outline lesions. An average of 1.89 annotations were made per image for the 326 positive cases out of 394 that were reliably annotated.

Evaluation of AI-Enhanced Mammography Workflows

Three workflows were tested with five senior and five junior board-certified readers: (1) routine read, (2) AI as a second reader (AISR), and (3) AI-assisted junior reader. Initially, each reader evaluated cases individually. After a 3-month washout period to avoid memory effects, in the routine arm, senior readers paired with juniors reviewed the cases again, this time with access to the junior reads. In the AISR setting, AI interpretations were used to filter cases where the AI's reading differed from the junior reader's. For these cases, the senior readers provided two reads—one with only the junior read and another with both the junior and AI reads. In the AI-assisted junior setting, juniors reread the scans with access to AI reads after a 3-month washout period.

For all workflows, readers first made recall decisions individually, as they would in clinical practice, and then assigned a JRADS score. When possible, prior imaging and associated reports were included. Readers also identified suspected lesions in cases chosen for recall.

Performance Comparison of AI-Enhanced Breast Cancer Screening Workflows

The AI system achieved an AUC of 0.84 (95% CI, 0.80-0.88) on the test set, compared to 0.81 (95% CI, 0.78-0.84) for US datasets and 0.96 (95% CI, 0.931-0.980) for UK datasets. After fine-tuning, there was no significant difference between the model AUC and consensus JRADS thresholds, particularly for JRADS scores ≥3-1 and ≥3-2, which reflect low and medium risk of malignancy and approximate recall decisions in clinical practice.

In a reader study with a subset of the test set, the model's standalone AUC was 0.76 (95% CI, 0.70-0.82) and its area under the precision-recall curve was 0.687 (95% CI, 0.595-0.756). The model showed significantly higher specificity while maintaining similar sensitivity compared to 9 of the 10 readers (all 5 senior, 4 junior). The model's lesion localization sensitivity was 0.8 (95% CI, 0.654-0.840), outperforming all 10 readers ($P < .00001$).

In the AISR scenario, where AI output was used alongside junior readers' interpretations, the AI increased sensitivity but reduced specificity. In the routine double-read arm, the model's sensitivity of 0.80 and false-positive rate of 0.26 outperformed 2 of the 5 pairs of readers. In the AISR arm, average sensitivity improved by 7.6% (95% CI, 3.80–11.4) (P = .00004) and specificity dropped by 3.4% (95% CI, 1.42–5.43) (P = .0016). Reader performance in the AISR arm closely matched the model's performance, with the model outperforming only 1 of the 5 pairs. The positive predictive value did not significantly improve, but the negative predictive value improved by 2.32% (95% CI, 0.595–4.13) (P = .00902). Individual reader pairs showed varied improvements, with some pairs significantly increasing sensitivity.

Individual reader decision consistency had a moderate Cohen kappa of 0.54 (95% CI, 0.52-0.56), which improved to 0.65 (95% CI, 0.61-0.68) in the routine double-read arm and 0.74 (95% CI, 0.71-0.77) in the AISR setting. In AISR, only 29% of cases required arbitration, potentially saving 34 clinical person-hours per week. When arbitration was necessary, the senior reader agreed with the junior reader 63% of the time, resulting in more false negatives, and with the AI 37% of the time, resulting in slightly more false positives.

With AI assistance, junior reader consistency in recall decisions improved, but there was no significant improvement in sensitivity or specificity. Average reporting time per case decreased from 2.20 minutes to 1.04 minutes with AI assistance. The correlation between AI decisions and junior readers changing their decisions was moderate (r = 0.41), even when the AI correctly identified lesions.

Impact of AI as a Second Reader in Mammography Screening

This study demonstrates that using AI as a second reader (AISR) in mammography screening improves dual read sensitivity by 7.6%, with minimal decrease in specificity. It reduces the caseload for second readers by 71% and enhances consistency between reads. Given the global shortage of radiologists, AI integration can alleviate the burden on cancer screening services and increase access by improving throughput. Most developed countries recommend or require double reading of mammograms. This study’s findings are broadly applicable, but some considerations are specific to different countries. In Japan, arbitration is typically done by the second reader rather than a third, potentially increasing economic impact in countries involving more mammographers. In the US, where single reading predominates, this approach could enhance detection rates akin to dual screening without adding to physician workload.

Generalizability and Potential of AI in Mammography Screening

The AI model, when generalised to Japan, performed comparably to the JRADS consensus score and maintained similar performance to US datasets after fine-tuning. The higher performance in the UK dataset reflects different screening populations, with UK patients screened every three years, allowing more time for lesions to develop and become noticeable. The model also improved recall consistency among individual physicians, potentially reducing biases and standardising outcomes. Training on model interpretation could further improve human performance and reduce inter-reader variance. Future research should explore personalized AI assistance based on a reader’s baseline performance.

Limitations and Considerations in AI-Assisted Mammography Screening

The study has several limitations. The dataset consisted entirely of East Asian women, with higher average breast density than Western datasets. Younger women under 40, commonly screened in Japan, were included, whereas women aged 40 are rarely screened routinely in Western countries. Higher breast density can reduce reader sensitivity, making the model’s benefits potentially greater in the Japanese population. The reader study dataset was enriched for benign lesions, which are harder to differentiate from malignancies, resulting in lower performance across the board compared to the overall test set. The study was conducted in a well-resourced Japanese institute that often includes ultrasound in screenings, potentially making reader performance less representative of those relying solely on mammography. Prospective studies are needed to determine performance in other practice settings.

The study shows that AISR can increase sensitivity and reduce human reader burden in a double-read scenario, illustrating how AI can enhance the efficacy and scalability of double reading in breast screening.

Source: Radiology Advances

Image Credit: iStock

Published on : Wed, 12 Jun 2024