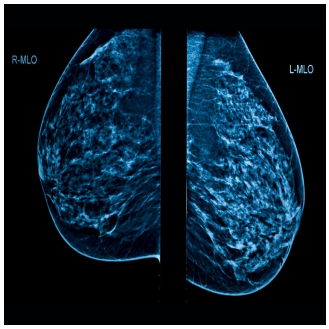
HealthManagement
Promoting Management and Leadership

AI in Diagnostic Mammography Matches Radiologist Accuracy



Breast cancer remains the most common cancer in women. While incidence has been rising, mortality has declined, supported by earlier diagnosis through mammography screening. Yet performance varies, with sensitivity reported at 80–98% and falling below 70% in women with dense breast tissue. Double reading can increase detection but is challenged by workforce constraints and concerns about higher recall. Artificial intelligence (AI) has emerged as an adjunct in breast imaging, with retrospective and screening-focused work suggesting potential benefits. Prospective evidence from diagnostic settings has been limited. A single-centre prospective evaluation conducted in a tertiary breast clinic assessed a commercially available AI system against radiologists with varying experience and explored how AI outputs might influence reassessment, while using histopathology and two-year follow-up as ground truth.

Prospective Diagnostic Setting and Workflow

The evaluation included 1063 women aged over 40 who underwent two-view full-field digital mammography between April and July 2022. Mediolateral oblique and craniocaudal views were acquired on a single vendor platform. Five radiologists participated, including three with 5–10 years of experience, one resident and one with more than 20 years in breast radiology. Each breast was analysed independently, yielding 2126 examinations. Radiologists recorded Breast Imaging-Reporting and Data System (BI-RADS) assessments using the fifth edition lexicon, dichotomised as negative (BI-RADS 1–3) or positive (BI-RADS 0, 4–5). Decisions were made from the current mammograms alone without prior imaging, clinical data or adjunct modalities. Majority voting resolved discordant cases.

Must read:AI-Guided Hybrid Reading Maintains Mammography Accuracy

A standalone AI system (Lunit INSIGHT MMG version 1.1.7.1) analysed the same images, generated heatmaps and assigned lesion-level risk scores from 1 to 100. A threshold of 30.44 defined a positive AI result. The system also provided a quantitative breast density score, while radiologists classified density visually from A to D. Ground truth incorporated histopathology for cancers and a two-year negative follow-up for non-cancers. As a second step after a one-year washout, radiologists re-evaluated only AI-positive cases with AI markings available, enabling comparison of initial reads, AI performance and reassessment.

The cohort was enriched for dense breasts. Of 1063 women, 6.6% had type A density, 26.2% type B, 57.9% type C and 9.3% type D. Radiologist and AI density assessments showed moderate agreement with a kappa of 0.602. Across the study, 29 cancers were found in 28 women, and no interval cancers were observed during follow-up. The median tumour diameter was 21 mm.

Comparable Accuracy with Different Trade-offs

AI returned a positive result in 2.44% of examinations (52/2126), and 46.15% of these positives were true cancers (24/52). Standalone AI detected 82.75% of cancers (24/29), comprising 70.83% invasive cancers and 29.16% ductal carcinoma in situ (DCIS). By comparison, majority voting among radiologists flagged 8% of cases as positive (172/2126) and detected 89.65% of cancers (26/29), with 73.07% invasive and 26.92% DCIS. Receiver operating characteristic analysis showed close performance: the area under the curve (AUC) was 94.4% for AI and 94.7% for majority voting, a non-significant difference. When radiologists re-evaluated AI-positive mammograms, the AUC reached 94.8%, significantly higher than the initial evaluation but not different from AI.

Across diagnostic metrics, AI favoured specificity and positive predictive value (PPV), while radiologists showed higher sensitivity. Accuracy was 98% for AI and 92% for both initial and reassessment reads. Sensitivity was 83% for AI compared with 90% for radiologists before and after reassessment. Specificity was 99% for AI versus 92% initially and 93% at reassessment. PPV was 45% for AI and 14% for radiologists both before and after reassessment. Negative predictive value (NPV) was 100% across approaches. The recall rate, reflecting the proportion of

examinations flagged positive, was lower with AI at 2.5%, compared with 8.9% at initial read and 8.6% after reassessment.

Inter-reader agreement among radiologists ranged from poor to good. Fleiss' kappa was 45.8 prior to reassessment and 46 after, with no significant difference between the two time points. The radiologist with the highest AUC had the greatest experience in breast imaging. These results indicate that the standalone AI system achieved accuracy similar to experienced human reading, while operating at a lower recall rate and higher specificity but lower sensitivity compared with radiologists' consensus.

Dense Breasts, Missed Lesions and Reader Variability

Dense parenchyma limits mammographic sensitivity by masking lesions. In this cohort, two cancers were detected on ultrasound and one on MRI but were not visible on mammography, and all three were missed by both radiologists and AI. Two additional invasive ductal carcinomas presented as focal asymmetries in type C density; these were missed by AI, aligning with the system's inability to perform side-to-side comparisons or incorporate prior examinations. One cancer was marked by AI with a risk score of 13, below the positivity threshold. Analysis of AI false positives highlighted algorithmic limitations when prior images, patient history and bilateral comparisons are not leveraged. Examples included vascular and skin calcifications, asymmetries symmetric with the contralateral breast, postoperative changes and microcalcifications stable on prior imaging.

These case patterns underline differing strengths. AI's higher specificity and lower recall suggest potential to reduce unnecessary work-ups, whereas radiologists' higher sensitivity supports detection, including in complex density patterns. The second-look reassessment of AI-positive cases improved AUC over the initial read without altering recall materially, and inter-reader agreement remained unchanged. The dataset's density profile, the single-vendor imaging platform and the relatively low number of cancers are important context for interpreting these findings. The threshold for AI positivity, set from retrospective data, may warrant calibration, although only one missed cancer had a score close to but below the threshold.

In a prospective diagnostic clinic setting, a standalone AI system demonstrated diagnostic accuracy comparable to radiologists' majority voting, with higher specificity, lower recall and lower sensitivity. Reassessment of AI-flagged examinations improved overall discrimination relative to the initial read without surpassing AI's performance. The cohort's high proportion of dense breasts illustrated shared challenges for AI and readers, including masked lesions and focal asymmetries. These results support the role of AI as an effective computer-aided diagnosis tool that can help reduce inter-reader variability and may be integrated as a second reader within diagnostic workflows. Generalisability, device dependence, the cancer prevalence and the AI threshold choice should be considered, and larger cohorts are needed to refine implementation in routine practice.

Source: **Academic Radiology**

Image Credit: **iStock**

Published on : Mon, 29 Sep 2025