HealthManagement
Promoting Management and Leadership

Clinical Value of AI in Mammogram Assessment

Breast cancer remains the most commonly diagnosed cancer among women worldwide. While mortality rates have declined due to better treatments and early detection, mammography continues to play a central role in reducing breast cancer deaths. Despite its value, mammography has well-documented limitations, particularly in women with dense breast tissue where sensitivity can fall significantly. Diagnostic accuracy is further affected by human error, often stemming from fatigue, inexperience or subtle lesion presentation.

As radiology services face increasing pressure due to workforce shortages, artificial intelligence has gained attention as a potential solution. Although several studies have assessed AI in screening environments, its standalone use in clinical diagnostics has yet to be validated extensively. A recent prospective study aimed to evaluate the performance of standalone AI against radiologists of varying experience in a clinical setting.

Comparative Performance of AI and Radiologists

The study enrolled 1,063 women aged over 40 who presented for diagnostic or screening mammography. Each underwent two-view digital mammography, generating a total of 2,126 breast images. A commercially available AI tool independently analysed each image, assigning a malignancy risk score on a scale from 1 to 100. A score above 30.44 was deemed positive. Simultaneously, five radiologists, including one with over two decades of experience and others with five to ten years or less, assessed the same images using the BI-RADS classification system. The diagnoses were confirmed via histopathology and a two-year clinical follow-up.

The AI system detected 24 out of 29 cancers, including invasive and in situ cases, resulting in a sensitivity of 83 percent. In comparison, radiologists using majority voting identified 26 cancers with a sensitivity of 90 percent. Both AI and radiologists demonstrated excellent diagnostic performance, with the AI achieving an area under the curve (AUC) of 94.4 percent, very close to the 94.7 percent AUC recorded by radiologists. Specificity, however, was significantly higher for AI at 99 percent, compared with 92 percent for radiologists. This translated into a much lower recall rate for AI at 2.5 percent, compared to the 8.9 percent recall rate for human readers. Positive predictive value was also notably higher for AI, while both AI and radiologists achieved a perfect negative predictive value.

Second-Look Evaluation and Reader Variability

To further assess AI's contribution, radiologists re-evaluated all AI-flagged cases following a one-year washout period. This second-look process, informed by AI's output, resulted in an AUC of 94.8 percent. Although statistically superior to the initial radiologist-only evaluation, the improvement was marginal and not significantly different from AI's standalone performance. Sensitivity and specificity remained similar to the first round, while the recall rate showed only a slight decrease.

Must Read: ***Cost-Efficient Mammography: Sharing Tasks Between AI and Radiologists***

The study also explored inter-reader variability. Performance among radiologists varied, with AUC scores ranging from 91.6 percent to 95.4 percent. Unsurprisingly, the most experienced radiologist achieved the highest accuracy. Inter-reader agreement, measured through Kappa statistics, ranged from poor to good depending on the case and stage of evaluation. These discrepancies highlight the inconsistent nature of manual interpretation, reinforcing the potential value of AI in standardising assessments.

Limitations and Clinical Considerations

While AI performed well overall, it did fail to identify five cancers. Three of these were also missed by all radiologists and were later detected by ultrasound or MRI. The remaining two, visible only as focal asymmetries on dense tissue, were missed due to AI's inability to compare both breasts or access prior imaging. This reflects a key limitation: AI algorithms currently lack the contextual and comparative capabilities that radiologists often rely upon.

False positives were also a consideration. Most AI false positives stemmed from benign findings such as vascular or postoperative calcifications, which the algorithm flagged as suspicious. These were resolved on radiologist review, suggesting that a hybrid workflow could balance AI's high specificity with human clinical judgement. Furthermore, the study's single-centre design and use of a single mammography system may limit the generalisability of its findings. With only 29 cancer cases detected, further large-scale studies are warranted.

Despite these challenges, the data suggest that AI can play a valuable role in diagnostic workflows, especially in environments with limited radiology staffing. Its capacity to reduce recall rates and maintain high specificity makes it a compelling adjunct to human expertise. Importantly, AI did not increase unnecessary follow-ups, and when paired with radiologist re-evaluation, false positives were effectively managed.

The study demonstrated that standalone AI can achieve diagnostic performance on par with experienced radiologists in mammography interpretation. While AI may not yet replace human expertise, it clearly offers substantial support in reducing variability and improving efficiency. Its lower recall rate and high specificity make it particularly useful in minimising unnecessary interventions. With continued refinement and broader validation, AI holds promise as a second reader in clinical mammography, helping to improve consistency and maintain high standards in breast cancer diagnostics.

Source: **Academic Radiology**

Image Credit: **iStock**

Published on : Sun, 15 Jun 2025