RSNA Press Release

AI Performs Comparably to Human Readers of Mammograms

Released: September 5, 2023

At A Glance

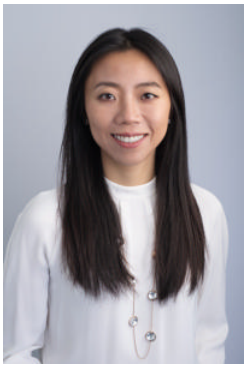
- Researchers compared the performance of a commercially available AI algorithm with human readers of screening mammograms.
- AI test scores were compared with the scores of 552 readers, including 315 board-certified radiologists and 237 non-radiologist readers.
- AI showed comparable sensitivity and specificity to human readers.

OAK BROOK, Ill. — Using a standardized assessment, researchers in the UK compared the performance of a commercially available artificial intelligence (AI) algorithm with human readers of screening mammograms. Results of their findings were published in *Radiology*, a journal of the Radiological Society of North America (RSNA).

Mammographic screening does not detect every breast cancer. False-positive interpretations can result in women without cancer undergoing unnecessary imaging and biopsy. To improve the sensitivity and specificity of screening mammography, one solution is to have two readers interpret every mammogram.

According to the researchers, double reading increases cancer detection rates by 6 to 15% and keeps recall rates low. However, this strategy is labor-intensive and difficult to achieve during reader shortages.

download photo



Yan Chen, Ph.D.

"There is a lot of pressure to deploy AI quickly to solve these problems, but we need to get it right to protect women's health," said Yan Chen, Ph.D., professor of digital screening at the University of Nottingham, United Kingdom.

Prof. Chen and her research team used test sets from the Personal Performance in Mammographic Screening, or PERFORMS, quality assurance assessment utilized by the UK's National Health Service Breast Screening Program (NHSBSP), to compare the performance of human readers with AI. A single PERFORMS test consists of 60 challenging exams from the NHSBSP with abnormal, benign and normal findings. For each test mammogram, the reader's score is compared to the ground truth of the AI results.

"It's really important that human readers working in breast cancer screening demonstrate satisfactory performance," she said. "The same will be true for AI once it enters clinical practice."

The research team used data from two consecutive PERFORMS test sets, or 120 screening mammograms, and the same two sets to evaluate the performance of the AI algorithm. The researchers compared the AI test scores with the scores of the 552 human readers, including 315 (57%) board-certified radiologists and 237 non-radiologist readers consisting of 206 radiographers and 31 breast clinicians.

"The 552 readers in our study represent 68% of readers in the NHSBSP, so this provides a robust performance comparison between human readers and AI," Prof. Chen said.

Treating each breast separately, there were 161/240 (67%) normal breasts, 70/240 (29%) breasts with malignancies, and 9/240 (4%) benign breasts. Masses were the most common malignant mammographic feature (45/70 or 64.3%), followed by calcifications (9/70 or 12.9%), asymmetries (8/70 or 11.4%), and architectural distortions (8/70 or 11.4%). The mean size of malignant lesions was 15.5 mm.

No difference in performance was observed between AI and human readers in the detection of breast cancer in 120 exams. Human reader performance demonstrated mean 90% sensitivity and 76% specificity. AI was comparable in sensitivity (91%) and specificity (77%) compared to human readers.

"The results of this study provide strong supporting evidence that AI for breast cancer screening can perform as well as human readers," Prof. Chen said.

Prof. Chen said more research is needed before AI can be used as a second reader in clinical practice.

"I think it is too early to say precisely how we will ultimately use AI in breast screening," she said. "The large prospective clinical trials that are ongoing will tell us more. But no matter how we use AI, the ability to provide ongoing performance monitoring will be crucial to its success."

Prof. Chen said it's important to recognize that AI performance can drift over time, and algorithms can be affected by changes in the operating environment.

"It's vital that imaging centers have a process in place to provide ongoing monitoring of AI once it becomes part of clinical practice," she said. "There are no other studies to date that have compared such a large number of human reader performance in routine quality assurance test sets to AI, so this study may provide a model for assessing AI performance in a real-world setting."

"Performance of a Breast Cancer Detection AI Algorithm Using the Personal Performance in Mammographic Screening Scheme." Collaborating with Dr. Chen were Adnan G. Taib, B.M.B.S., Iain T. Darker, Ph.D., and Jonathan J. James, FRCR.

In 2023, *Radiology* is celebrating its 100th anniversary with 12 centennial issues, highlighting *Radiology*'s legacy of publishing exceptional and practical science to improve patient care.

Radiology is edited by Linda Moy, M.D., New York University, New York, N.Y., and owned and published by the Radiological Society of North America, Inc. (https://pubs.rsna.org/journal/radiology)

RSNA is an association of radiologists, radiation oncologists, medical physicists and related scientists promoting excellence in patient care and health care delivery through education, research and technologic innovation. The Society is based in Oak Brook, Illinois. (RSNA.org)

For patient-friendly information on breast cancer screening, visit RadiologyInfo.org.

Images (JPG, TIF):

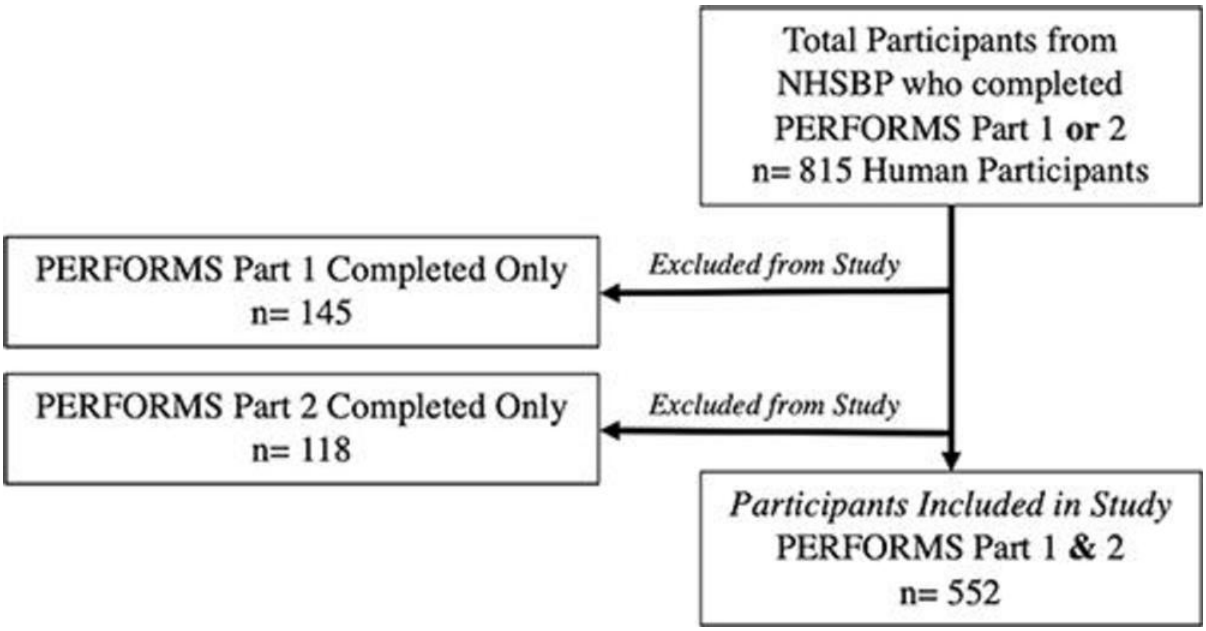


Figure 1. Flow diagram shows human reader inclusion and exclusion criteria. NHSBSP = National Health Service Breast Screening Programme, PERFORMS = Personal Performance in Mammographic Screening.
High-res (TIF) version

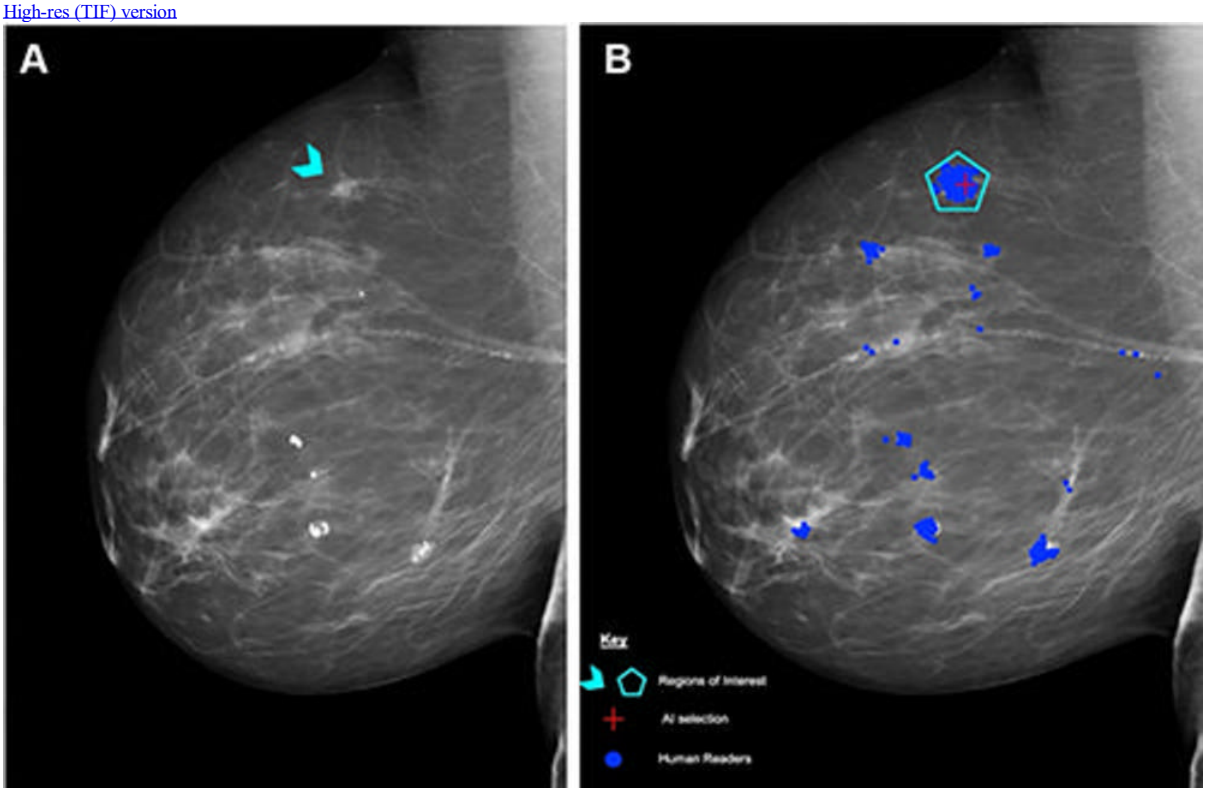


Figure 2. (A) Right mediolateral oblique unadulterated mammogram shows an 8-mm ill-defined mass (arrowhead), which, after biopsy, was determined to

be a histologic grade 2 ductal carcinoma of no special type. **(B)** Mammogram shows findings by human readers (blue areas) and the Lunit INSIGHT MMG artificial intelligence (AI) algorithm (red cross). Each blue dot is a mark placed by an individual human reader on a perceived abnormality when the Personal Performance in Mammographic Screening (PERFORMS) case was read. A region of interest (pentagon) has been annotated by the PERFORMS scheme organizers and their expert radiology panel. AI has correctly marked the region of interest in the right breast for recall. Source: PERFORMS via Yan Chen.
High-res (TIF) version

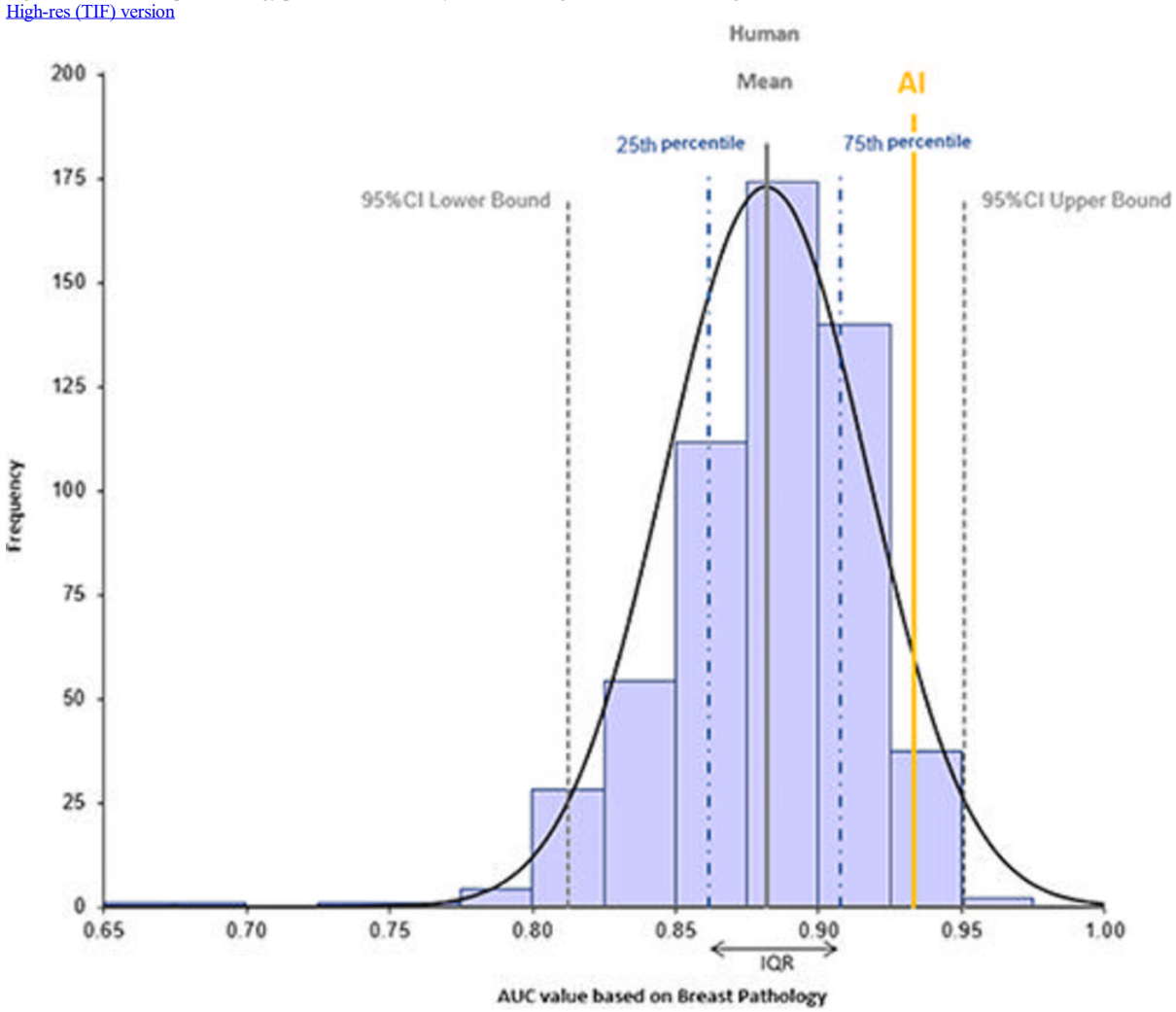


Figure 3. Histogram shows the distribution of area under the receiver operating characteristic curve (AUC) values for human readers (mean, solid gray line) and the Lunit INSIGHT MMG artificial intelligence (AI) algorithm (yellow line) interpreting mammograms at the breast level (n = 240) for recall or return to normal screening. Dashed gray lines represent lower and upper bounds of the AUC 95% CI, and dotted and dashed blue lines represent 25th and 75th percentiles./a>
High-res (TIF) version

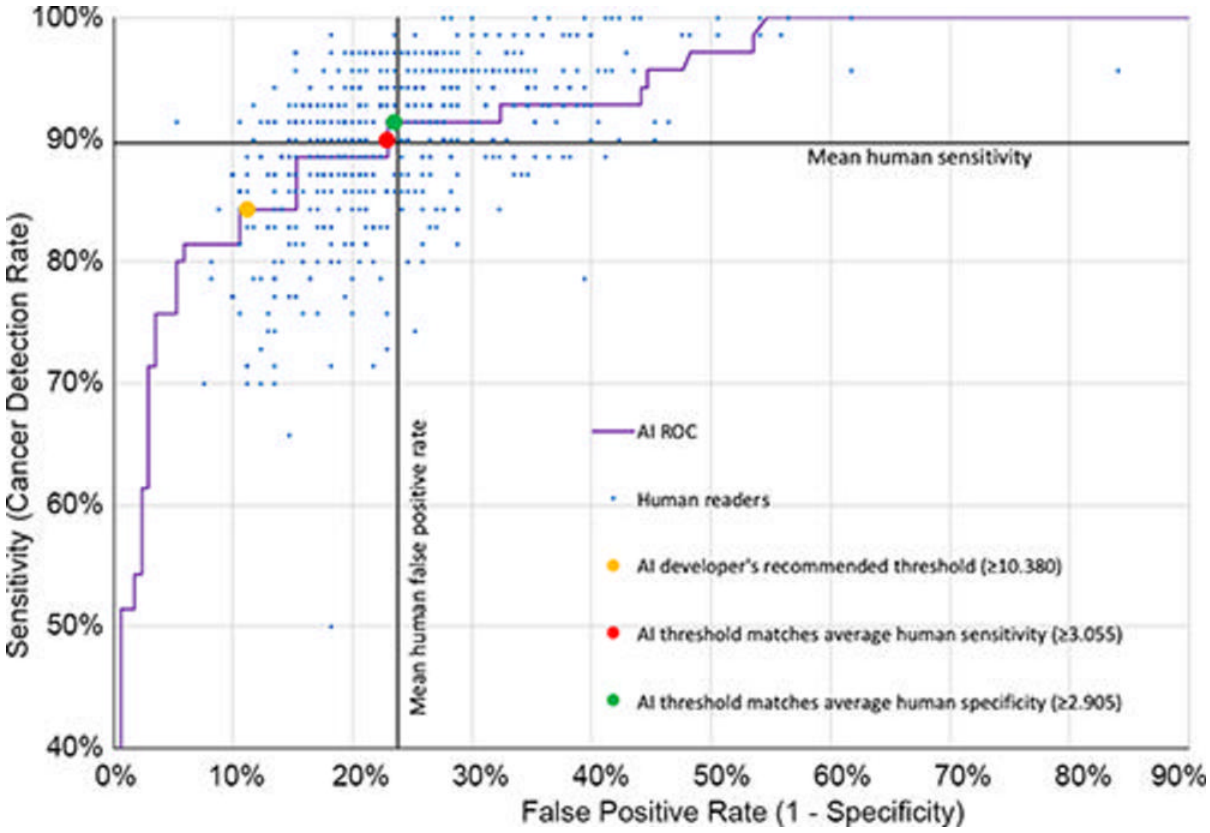


Figure 4. Receiver operating characteristic curve for the Lunit INSIGHT MMG artificial intelligence (AI) algorithm plotted with individual results of the 552 human readers (blue dots). The y-axis represents sensitivity (cancer detection rate), and the x-axis represents 1 – specificity (false-positive rate). The cancer detection and false-positive rates for AI at a recall threshold score of 10.38 or more (the developer's recommended threshold), 3.055 or more (the threshold matching average human sensitivity), and 2.905 or more (the threshold matching average human specificity) are represented by yellow, red, and green dots, respectively. ROC = receiver operating characteristic curve.
High-res (TIF) version

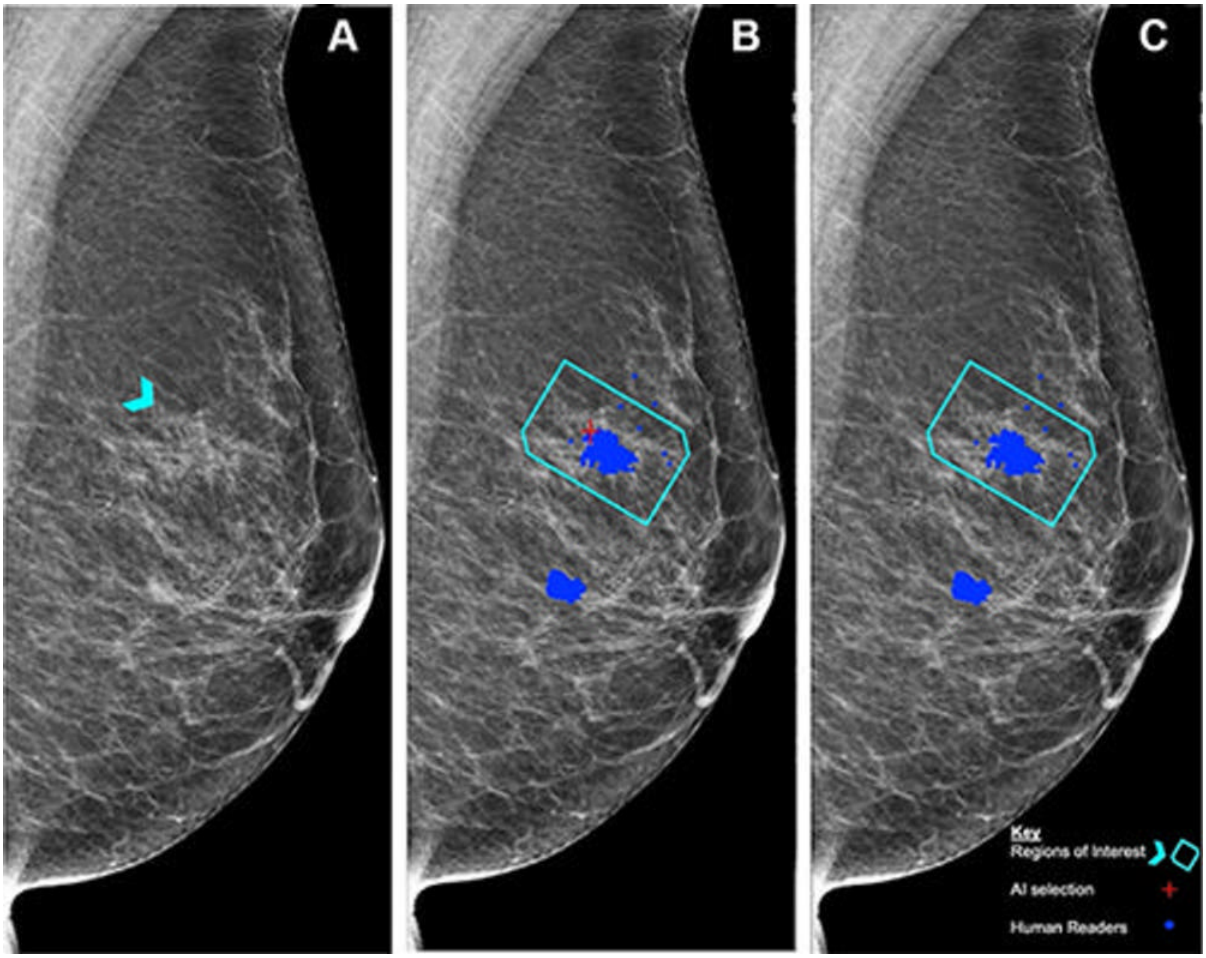


Figure 5. (A) Left mediolateral oblique mammogram. Unadulterated mammogram shows an asymmetric density (arrowhead) which, after biopsy, was determined to be a histologic grade 2 ductal carcinoma. **(B)** Artificial intelligence (AI) has correctly marked the region of interest in the left breast for recall (red cross) when set at a recall threshold of 2.91 or higher to match average human specificity, demonstrating a true-positive case. **(C)** AI has not marked the region of interest in the same breast when set at a recall threshold of 3.06 or higher, indicating a false-negative case. Blue dots indicate findings identified by the human readers. This shows how modifying the threshold for recall can impact the sensitivity of the AI model. Source: Personal Performance in Mammographic Screening via Yan Chen.
High-res (TIF) version

Resources:

Editorial
Study abstract