RSNA Press Release

AI Hybrid Strategy Improves Mammogram Interpretation

Released: August 19, 2025

OAK BROOK, Ill. — A hybrid reading strategy for screening mammography, developed by Dutch researchers and deployed retrospectively to more than 40,000 exams, reduced radiologist workload by 38% without changing recall or cancer detection rates. The study, which emphasizes AI confidence, was published today in *Radiology*, a journal of the Radiological Society of North America (RSNA).

"Although the overall performance of state-of-the-art AI models is very high, AI sometimes makes mistakes," said Sarah D. Verboom, M.Sc., a doctoral candidate in the Department of Medical Imaging at Radboud University Medical Center in the Netherlands. "Identifying exams in which AI interpretation is unreliable is crucial to allow for and optimize use of AI models in breast cancer screening programs."

hybrid reading strategy involves using a combination of radiologist readers and a stand-alone AI interpretation of cases in which the AI model performs as well as, or better than, the radiologist.

We can achieve this performance level if the AI model provides not only an assessment of the probability of malignancy (PoM) for a case but also a rating of its certainty of that assessment," Verboom said. "Unfortunately, the PoM itself is not always a good predictor of certainty because deep neural networks tend to be overconfident in their predictions."

download photo

Sarah D. Verboom, M.Sc.

To develop and evaluate a hybrid reading strategy, the researchers used a dataset of 41,469 screening mammography exams from 15,522 women (median age 59 years) with 332 screen-detected cancers and 34 interval cancers. The exams were performed between 2003 and 2018 in Utrecht, Netherlands, as part of the Dutch National Breast Cancer Screening Program.

The dataset was divided at the patient level into two equal groups with identical cancer detection, recall and interval cancer rates. The first group was used to determine the optimal thresholds for the hybrid reading strategy, while the second group was used to evaluate the reading strategies.

Of the uncertainty metrics evaluated by the researchers, the entropy of the mean PoM score of the most suspicious region produced a cancer detection rate of 6.6 per 1,000 cases and a recall rate of 23.7 per 1,000 cases, similar to rates of standard double-reading by radiologists.

The final hybrid reading strategy involved AI evaluating every screening mammogram to produce two outputs: the PoM and an uncertainty estimate of that prediction. When AI determined the PoM was below the established threshold with certainty, the case was considered normal. When AI detected a PoM above the established threshold, women were recalled for further testing, but only when that prediction was deemed confident. Otherwise, the exam was double-read by radiologists.

Although the majority of AI decisions were uncertain and deferred to a human reader, 38% were classified as certain and could be read solely by AI. Using the researchers' strategy reduced radiologist reading workload to 61.9% without changing recall (23.6‰ vs 23.9‰) or cancer detection (6.6‰ vs 6.7‰) rates, both of which are comparable to those of standard double-reading.

When the AI model was certain, the area under the curve (AUC) was higher (0.96 vs 0.87). Its sensitivity nearly matched that of double radiologist reading (85.4% vs 88.9%). Younger women with dense breasts were more likely to have an uncertain AI score.

"The key component of our study isn't necessarily that this is the best way to split the workload, but that it's helpful to have uncertainty quantification built into AI models," Verboom said. "I hope commercial products integrate this into their models, because I think it's a very useful metric."

Verboom noted that if the study results occurred in clinical practice, the decision to recall 19% of women would be made by AI without the intervention of a radiologist.

"Several studies have shown that women participating in breast cancer screening programs have positive attitudes about the use of AI," she said. "However, most women prefer their mammogram to be read by at least one radiologist."

She said it may be more acceptable for radiologists to review exams deemed uncertain by AI, as well as AI recall cases.

"The use of AI with uncertainty quantification can be a possible solution for workforce shortages and could help build trust in the implementation of AI," Verboom said.

Verboom said further research, ideally a prospective trial, is needed to determine how the workload reduction achieved by the hybrid reading strategy could decrease radiologist reading time.

"I think in the future, we could get to a point where a portion of women are sent home without ever having a radiologist look at their mammogram because AI will determine that their exam is normal," she said. "We're not there yet, but I think we could get there with this uncertainty metric and quality control."

This study is part of the aiREAD project, which is financed by the Dutch Research Council, Dutch Cancer Society and Health Holland.

"AI Should Read Mammograms Only When Confident: A Hybrid Breast Cancer Screening Reading Strategy." Collaborating with Verboom were Jaap

Kroes, Ph.D., Santiago Pires, M.Sc., Mireille J. M. Broeders, Ph.D., and Ioannis Sechopoulos, Ph.D.

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):

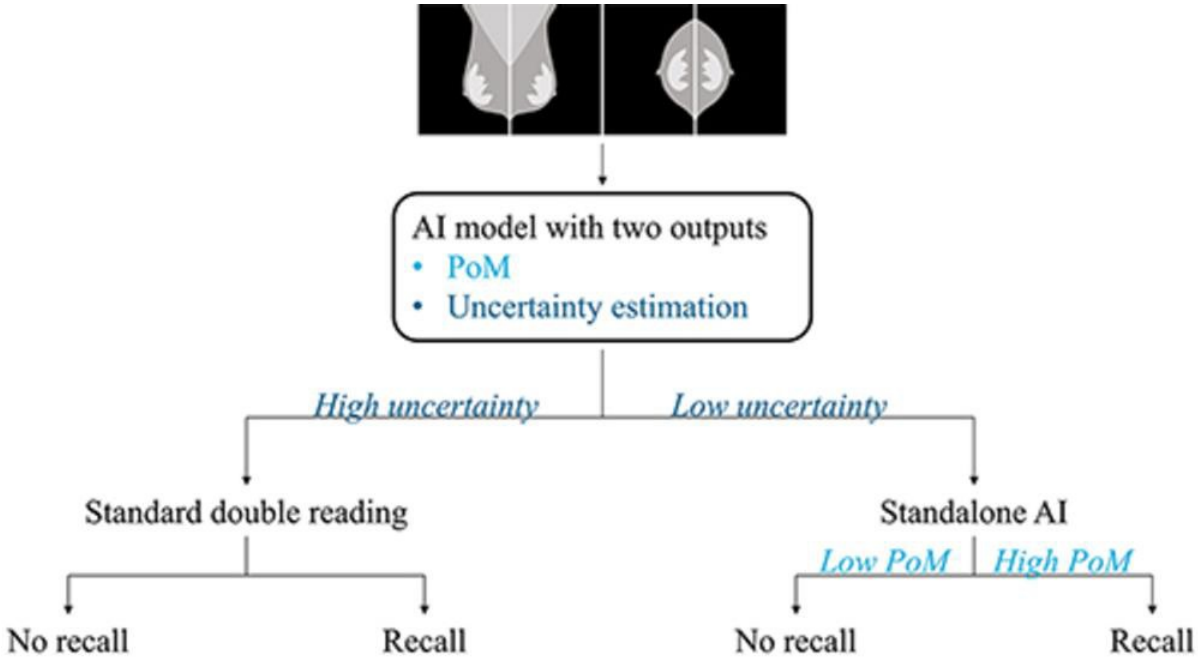


Figure 1. Schematic representation of the proposed hybrid reading strategy based on a model uncertainty quantification metric. All examinations are processed by an AI mammography interpretation model that outputs two metrics: a probability of malignancy (PoM) and an estimation of the uncertainty of that prediction. Examinations with a high uncertainty are referred for standard double reading by radiologists. Otherwise, the recall decision is solely based on the PoM.
High-res (TIF) version

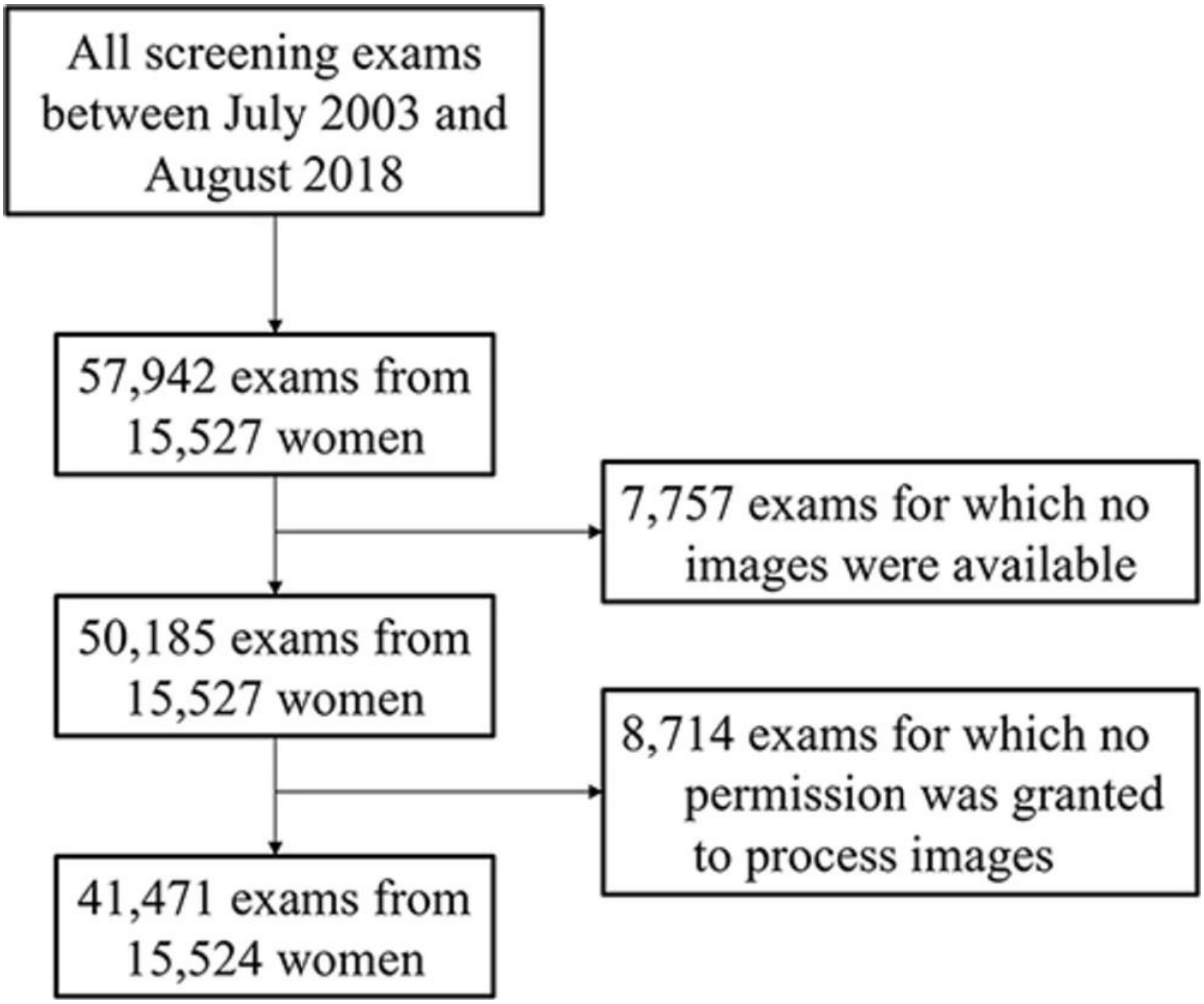


Figure 2. Flowchart of the inclusion of examinations and women from the breast cancer screening unit
High-res (TIF) version

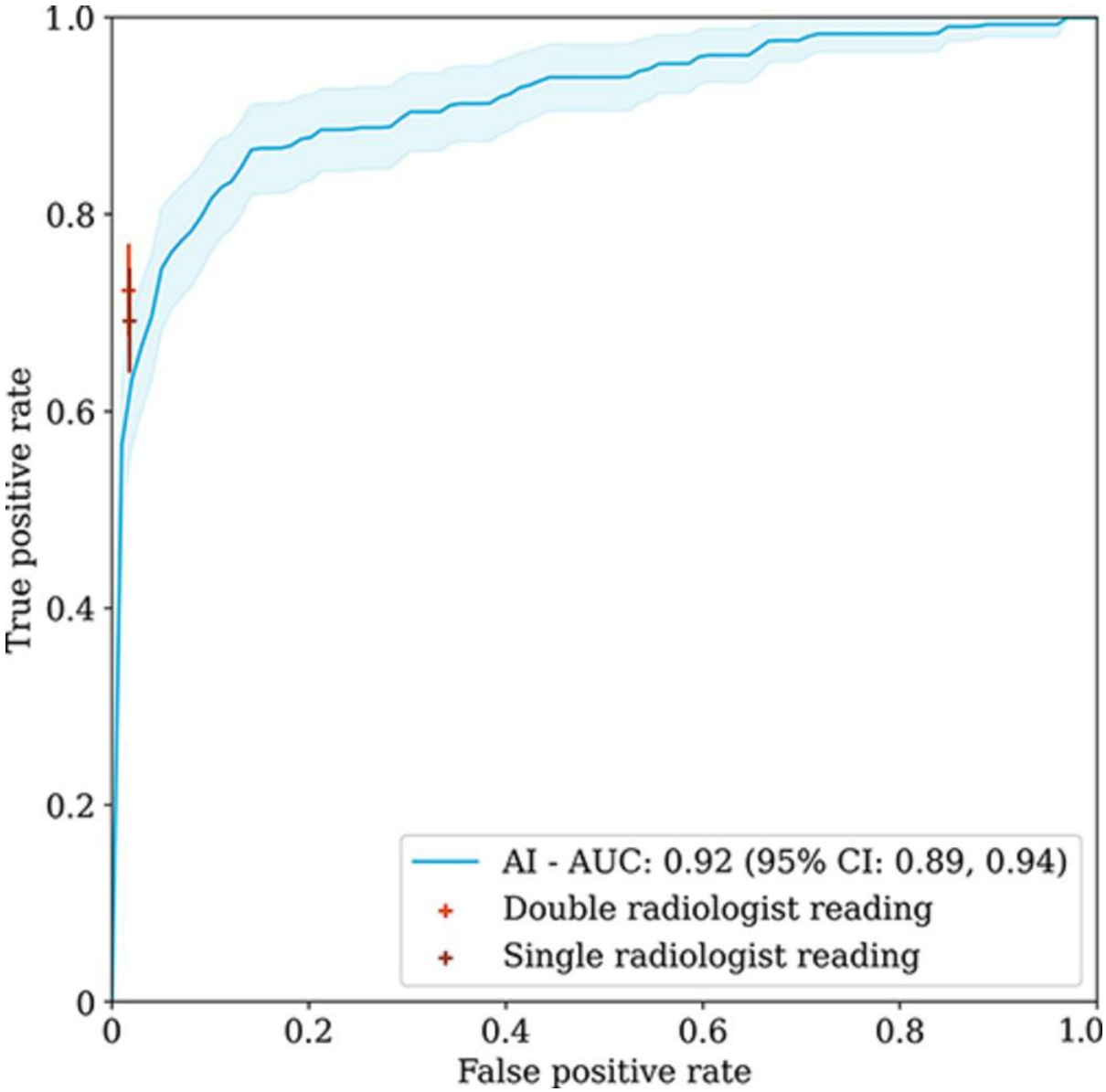


Figure 3. Receiver operating characteristic curve and its 95% CI, as shown by the shaded area, for all examinations of the stand-alone AI mammography interpretation model compared with the operating points of standard double radiologist reading and single radiologist reading. AUC = area under the receiver operating characteristic curve.
High-res (TIF) version

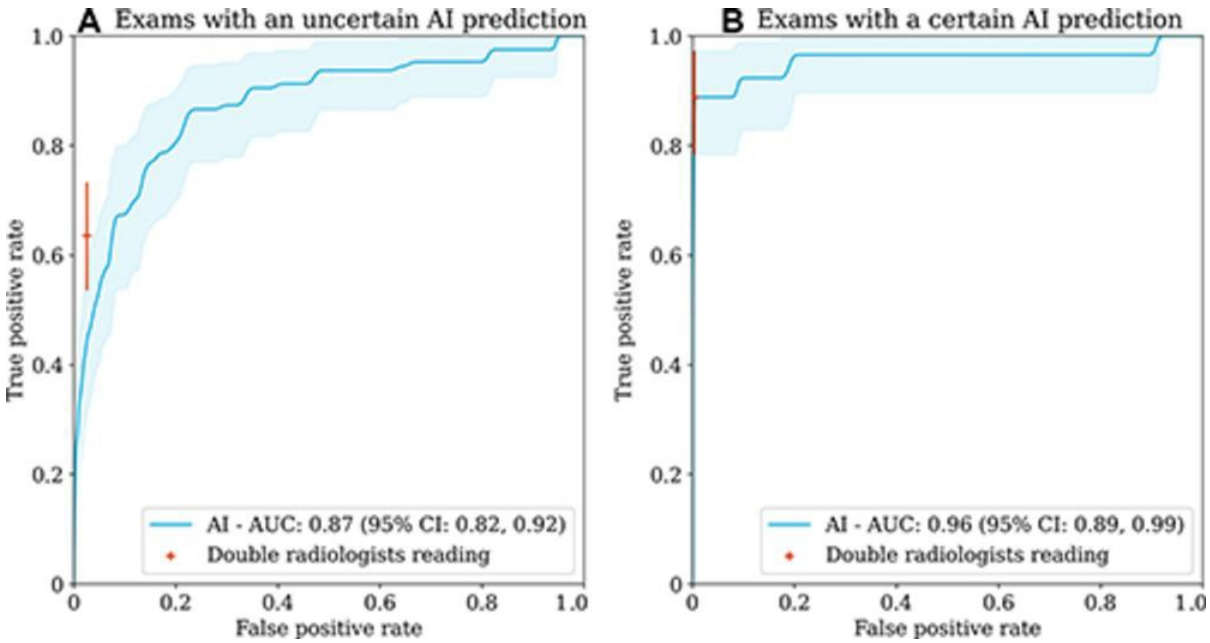
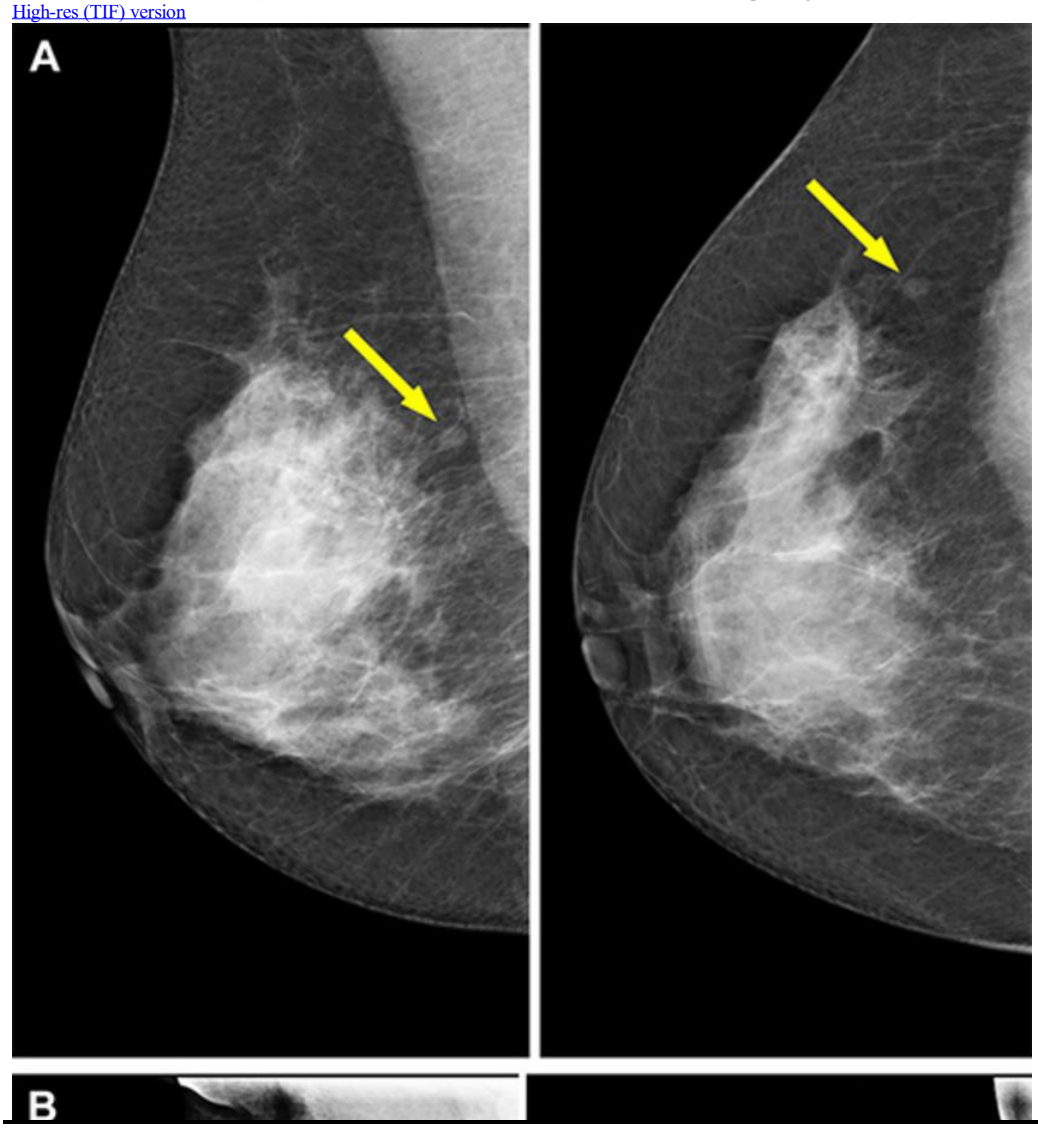


Figure 4. Receiver operating characteristic curves and their 95% CIs, as shown by the shaded area, for **(A)** the examinations where the AI model was uncertain and **(B)** the examinations where the AI model was certain. The operating point of double reading by radiologists for the corresponding examinations is shown in orange. The cancer prevalence is 8.6 per 1000 examinations (95% CI: 6.7, 10.5) for uncertain examinations **(A**and 9.8 per 1000 examinations (95% CI: 7.2, 12.6) for certain examinations **(B)** . AUC = area under the receiver operating characteristic curve.
High-res (TIF) version



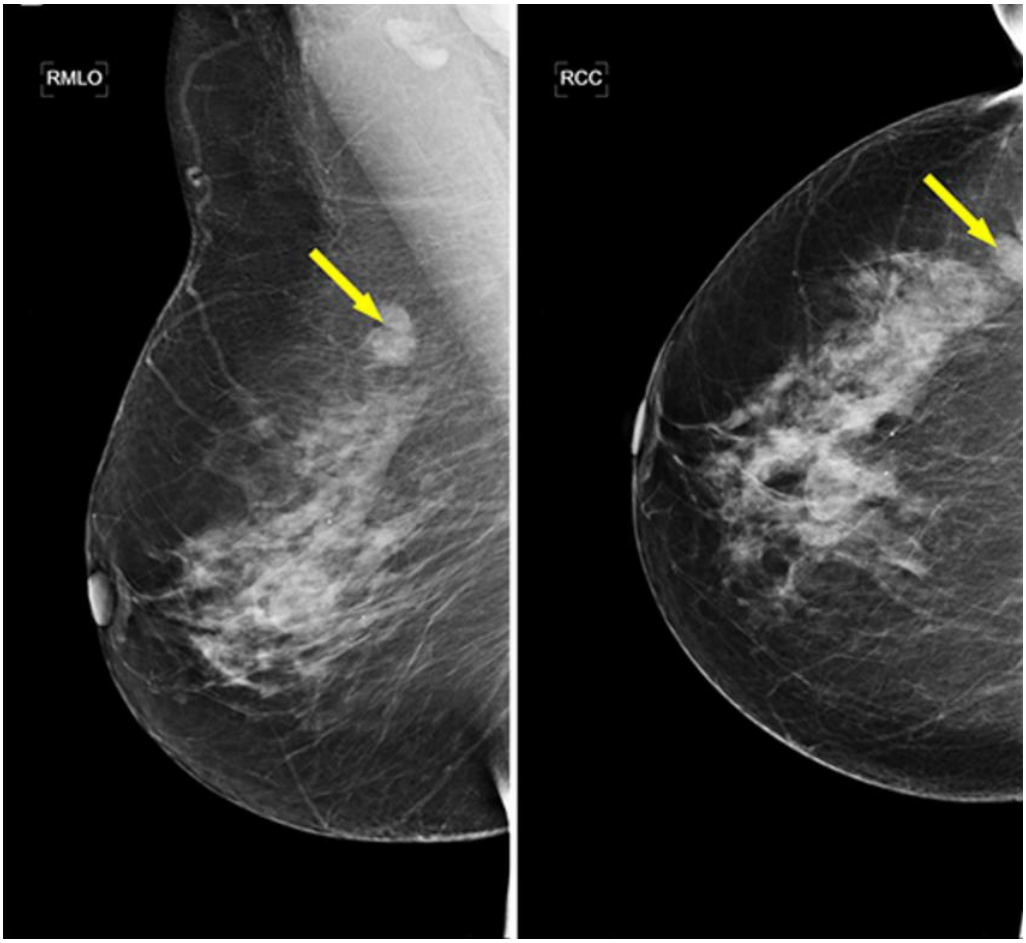


Figure 5. Example screening mammograms with an invasive ductal carcinoma (arrows) in which the women would not have been recalled with an AI–reading-only strategy. However, these examinations would have been shown to radiologists in a hybrid reading strategy based on the AI uncertainty score of the entropy of the mean probability of malignancy (PoM) score of the most suspicious region. For both examinations, mediolateral oblique (left) and craniocaudal (right) views of the affected breast are shown. **(A)** Images in a 67-year-old woman who was recalled because both radiologists scored the right breast as Breast Imaging Reporting and Data System (BI-RADS) 0. The woman would not have been recalled if the examination was read by the AI model, which assigned a PoM score of 40, but the prediction would have been classified as an uncertain prediction with an uncertainty quantification of 0.86. **(B)** Images in a 63-year-old woman who was recalled because both radiologists scored the right breast as BI-RADS 4. The woman would not have been recalled if the examination was read by the AI model, with a PoM score of 44, but the prediction would be classified as an uncertain prediction with an uncertainty quantification of 0.98.
High-res (TIF) version

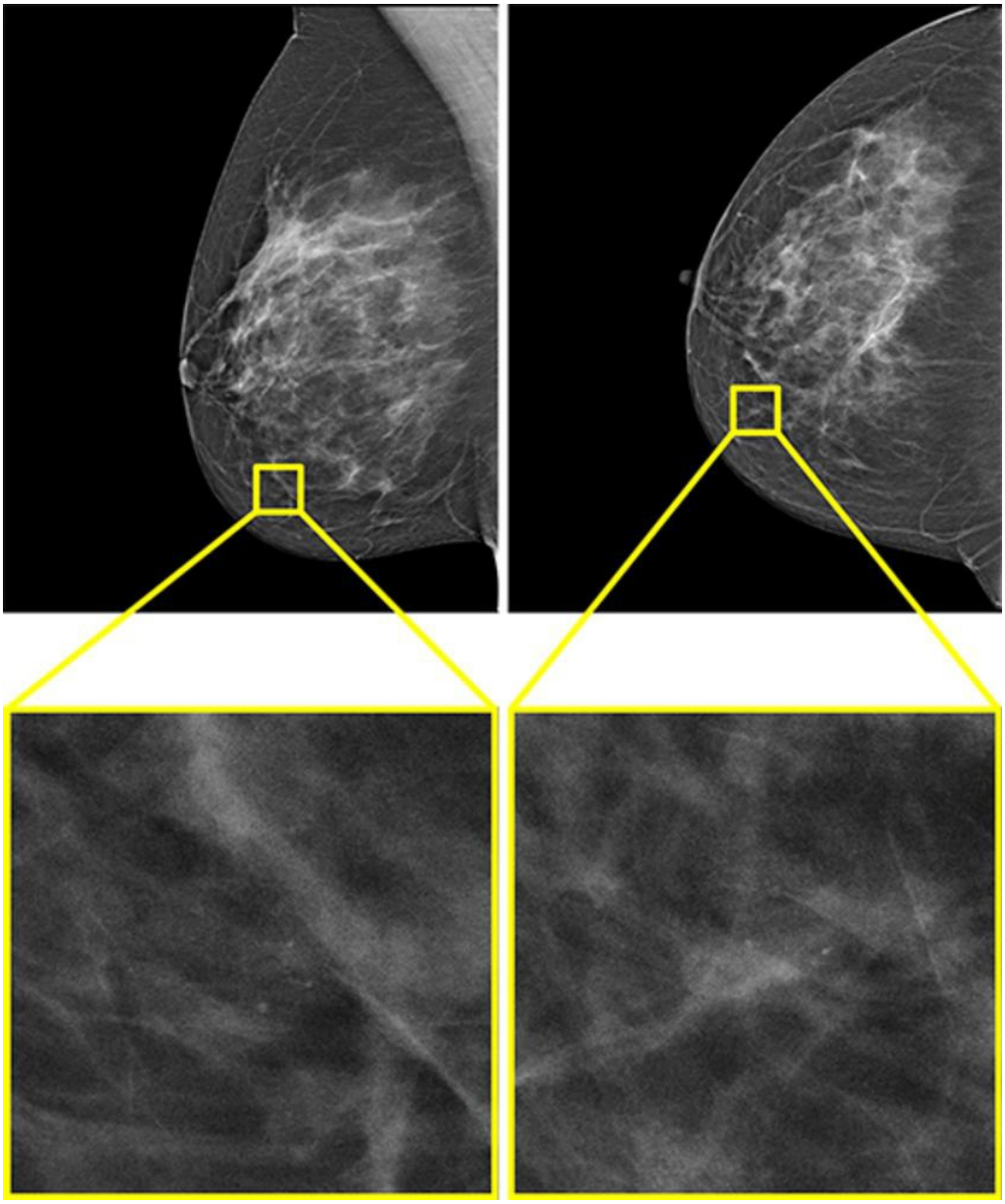


Figure 6. The only example of a screening examination with a screen-detected cancer that would have been missed by AI in a hybrid reading strategy based on the AI uncertainty score of the entropy of the mean probability of malignancy (PoM) score of the most suspicious region. During screening, a 52-year-old woman was recalled following arbitration scoring of the right breast as Breast Imaging Reporting and Data System (BI-RADS) 4 after the first and second radiologists scored the right breast as BI-RADS 1 and 4, respectively. This woman would not have been recalled if the examination was read by the AI model, which assigned a PoM score of 30, which would be classified as a certain prediction with an uncertainty quantification of 0.57. Both the mediolateral oblique (left) and craniocaudal (right) views of the affected breast are shown. The boxes indicate the calcifications found during screening, and the final diagnosis of this examination was ductal carcinoma in situ.
High-res (TIF) version

Resources:

Study abstract