ORIGINAL RESEARCH ARTICLE

CENTRAL ASIAN CANCER SCIENCES

Selective artificial intelligence arbitration by BI-RADS category: an algorithm to reduce overdiagnosis in mammographic breast cancer screening

Askar Adilkhodjaev[1] ORCID (https://orcid.org/0000-0001-7292-3910) · **Umid Tokhtamuratov**[1,2*] ORCID (https://orcid.org/0009-0004-7285-8070)

[1] Republican Specialized Scientific and Practical Medical Center of Oncology and Radiology, 14a Avtomobil Halqa Yo'li Street, Olmazor District, Tashkent, Uzbekistan

[2] Ministry of Health of the Republic of Uzbekistan, Tashkent, Uzbekistan

ABSTRACT

Background. Mammography screening in Uzbekistan is moving to hybrid reading by a radiologist and an artificial intelligence (AI) system. How the two should be combined in a setting where radiologist sensitivity is already high but false-positive recalls are frequent has not been studied.

Aim. To develop and evaluate an algorithm in which the AI system reviews the radiologist's assessment only in the Breast Imaging Reporting and Data System (BI-RADS) categories where that assessment is least reliable.

Materials and methods. Retrospective diagnostic accuracy study of 1,759 digital mammography examinations from Tashkent, Nukus and Samarkand with a verified outcome (316 breast cancers, 18.0%). Radiologists' BI-RADS assessments and the continuous risk score of a national AI system were obtained independently and compared with histopathology or 6–24-month follow-up. The algorithm was derived from the structure of radiologist errors and reader–AI disagreement: BI-RADS 4B–5 assessments are kept, the AI arbitrates BI-RADS 0 and 4A (threshold 0.30) and acts as a safety net for BI-RADS 1–3 (threshold 0.65). Paired comparisons used McNemar's test.

Results. Of 302 false-positive radiologist assessments, 277 (91.7%) fell into BI-RADS 0, 4A and 4B. All 6 cancers missed by radiologists scored above the AI operating threshold, and no cancer occurred among 982 examinations rated negative by both. The algorithm removed 70 unjustified positive assessments; specificity rose from 79.1% to 83.9% and positive predictive value from 50.7% to 57.3%, while

sensitivity was unchanged (98.1% vs 98.4%; p = 1.000). The error rate fell from 17.5% to 13.5% (difference 4.0 percentage points, 95% CI 2.8–5.3; p < 0.001), and the reduction was seen in every density, age and regional subgroup.

Conclusion. Using AI only where radiologist uncertainty is highest reduces overdiagnosis without loss of cancer detection. The algorithm needs prospective validation in a population screening cohort.

Keywords: breast cancer; screening mammography; artificial intelligence; BI-RADS; false-positive results; clinical decision support

Introduction

Breast cancer is the most common malignancy in women, with more than 2.3 million new cases worldwide in 2020 [1], and it holds a leading place in the cancer incidence of women in Uzbekistan [2]. The World Health Organization Global Breast Cancer Initiative places early diagnosis first among the measures needed to reduce mortality [3], and Presidential Resolution PP-402 of 22 November 2024 introduced mammography screening with AI-supported reading in the country [4].

The weak point of screening has long been known: interpretation is subjective. Elmore et al. and Kerlikowske et al. showed in the 1990s how widely radiologists disagree on the same images [5, 6]. The US national benchmarks put screening sensitivity at about 87% and specificity at about 89% [7]. European programmes control quality with independent double reading [8], which doubles the demand for trained readers – precisely the resource that is scarce in Central Asia.

Deep learning systems have matched radiologists in retrospective studies [9–11]. In the randomised MASAI trial, AI-supported reading detected 20% more cancers and cut the screen-reading workload by 44% [12]; a Swedish prospective study and the nationwide German implementation reported similar gains [13, 14]. Almost everywhere, however, the benefit is measured as higher detection and lower workload [15, 16]. That is natural where radiologists already read with high specificity. Where the starting point is different – high sensitivity bought with many false alarms – AI has a different job to do, and foreign implementation schemes cannot simply be copied.

We have previously shown that a national AI system reaches an area under the ROC curve of 0.914 on Uzbek mammograms, being less sensitive but more specific than the radiologist [17]. Here the question was not which of the two is more accurate, but a practical one: in which examinations does

the AI score improve the radiologist's decision, and in which does it only get in the way?

Aim

To derive, from the structure of radiologist errors and reader–AI disagreement, an algorithm for selective use of AI according to BI-RADS category, and to evaluate its effect on the diagnostic performance of mammography.

Materials and methods

Study design and ethics.

Retrospective multicentre diagnostic accuracy study based at Republican Specialized Scientific and Practical Medical Center of Oncology and Radiology, with data from Tashkent, Nukus and Samarkand. The study was approved by the Institutional Review Board of the institution Republican Specialized Scientific and Practical Medical Center of Oncology and Radiology; informed consent was waived because only fully anonymised data were used. The study complied with the Declaration of Helsinki.

Study population.

The cohort has been described in detail previously [17]. From 33,599 full-field digital mammography examinations performed between March 2022 and April 2023, records with technical failures, incomplete views, postoperative changes, implants or an unverified outcome were excluded, leaving 1,759 examinations of women aged 35–80 years (mean 54.5 ± 7.6 years). The sample was enriched: diagnostic as well as screening examinations were included to obtain enough confirmed cancers.

Mammography and radiologist assessment.

Craniocaudal and mediolateral oblique views were acquired on FUJIFILM AMULET systems. The images were reported in routine practice by 12 staff radiologists with 5–15 years of experience in mammography, using the ACR BI-RADS lexicon [18]; 768 examinations (43.7%) were double read. An assessment was considered positive if it led to recall or biopsy, i.e. BI-RADS 0, 4A, 4B, 4C or 5.

AI system.

A national detection model based on the YOLO11 architecture was trained on public datasets only; the Uzbek data were used solely for testing [17]. For each examination the model outputs a continuous risk score from 0 to 1, with an operating threshold of 0.495 (maximum Youden index). Radiologist and AI assessments were obtained independently; the radiologist did not see the AI output.

Reference standard.

Cancer was confirmed by histopathology (n = 297) or clinically and on follow-up (n = 19); absence of cancer was established by

6–24 months of negative follow-up. Breast cancer was verified in 316 women (18.0%), benign lesions in 351 (20.0%), and no abnormality in 1,092 (62.1%).

Algorithm development.

The algorithm was built in three steps: (1) distribution of radiologists' false-positive and false-negative assessments across BI-RADS categories; (2) analysis of disagreement between radiologist and AI on the same examinations; (3) choice of a rule and threshold for each group of categories. Thresholds were set on clinical grounds – no missed cancers in the arbitration zone, a limited number of new recalls in the large group of negative assessments – rather than by a search over values.

Statistical analysis.

Sensitivity, specificity, positive (PPV) and negative (NPV) predictive values, accuracy and error rate were calculated with Wilson 95% confidence intervals (CI). Because both readings concerned the same images, paired proportions were compared with McNemar's test; ROC curves were compared with DeLong's method, and the CI of the difference in error rates was obtained by bootstrap. A two-sided $p < 0.05$ was considered significant. Analyses were performed in Python 3.13 and checked in IBM SPSS Statistics 26.0.

Results

Where is radiologists' error?

Radiologists made 308 errors (17.5%): 302 false-positive and 6 false-negative assessments. Sensitivity was 98.1% (95% CI 95.9–99.1), specificity 79.1% (76.9–81.1) and PPV 50.7%. The errors were far from evenly distributed across BI-RADS categories (Table 1).

Table 1. Verified cancers and radiologists' false-positive assessments by BI-RADS category (n = 1,759)

BI-RADS category	Total, n	Cancer, n (%)	False positives, n	Share of all false positives, %
0	70	4 (5.7)	66	21.9
1	635	3 (0.5)	–	–
2	460	2 (0.4)	–	–
3	52	1 (1.9)	–	–
4A	195	59 (30.3)	136	45.0
4B	108	33 (30.6)	75	24.8
4C	96	79 (82.3)	17	5.6
5	143	135 (94.4)	8	2.6
Total	1,759	316 (18.0)	302	100.0

Note: in BI-RADS 1–3, a cancer is a false-negative radiologist assessment (6 cases).

Thus 277 of the 302 false positives (91.7%) were concentrated in just three categories – BI-RADS 0, 4A and 4B – where cancer was confirmed in only 5.7–30.6% of cases. In categories 4C and 5, by contrast, the radiologist's assessment was highly reliable. All 6 missed cancers occurred in dense or heterogeneous breasts (B – 1, C – 3, D – 2); in three of them the mammogram had been read as normal, and 5 of the 6 later presented as interval cancers.

Where radiologist and AI disagree.

Used on its own, the AI was less sensitive than the radiologist (82.9% vs 98.1%; $p < 0.001$) but more specific (85.2% vs 79.1%; $p < 0.001$), and the error rates did not differ significantly (15.2% vs 17.5%; $p = 0.071$). The joint distribution of the two readings is shown in Table 2.

Table 2. Agreement between radiologist and AI (threshold 0.495) and proportion of verified cancer (n = 1,759)

Radiologist / AI	n	Cancer, n	No cancer, n	Cancer, %
Both positive	311	256	55	82.3
Radiologist +, AI –	301	54	247	17.9
Radiologist –, AI +	165	6	159	3.6
Both negative	982	0	982	0

Table 2 contains the key observation of the study. Not a single cancer was found among the 982 examinations rated negative by both readers. The quadrant “radiologist negative, AI positive” holds exactly 6 cancers – all of those missed by radiologists; their scores were 0.500, 0.579, 0.593, 0.597, 0.662 and 0.989, against a mean of 0.298 in women without cancer. In the quadrant “radiologist positive, AI negative”, 247 of 301 examinations (82.1%) were radiologist false alarms, but the same quadrant also contained 54 cancers underrated by the AI, mostly in high-suspicion categories. Simple combination rules therefore fail. The “either positive” rule gives 100% sensitivity but raises the error rate to 26.2%; the “both positive” rule lowers it to 6.5% at the cost of 60 missed cancers instead of 6.

Selective arbitration algorithm.

These findings lead to a three-branch design (Fig. 1). In BI-RADS 4B, 4C and 5 (n = 347) the radiologist’s assessment stands without review: PPV is high, and this is where the AI tends to underrate tumours. We deliberately left category 4B to the radiologist, although it accounts for a quarter of false alarms. In BI-RADS 0 and 4A (n = 265) the AI acts as arbiter: the examination remains positive if the score is at least 0.30 and is otherwise closed as negative. This threshold is well below the operating one, to keep the risk of dismissing a true cancer to a minimum. In BI-RADS 1–3 (n = 1,147) the AI provides a safety net: a score of 0.65 or higher sends the examination for a second look. Here the threshold is high on purpose – this is the largest group, and a softer rule would generate many new recalls.

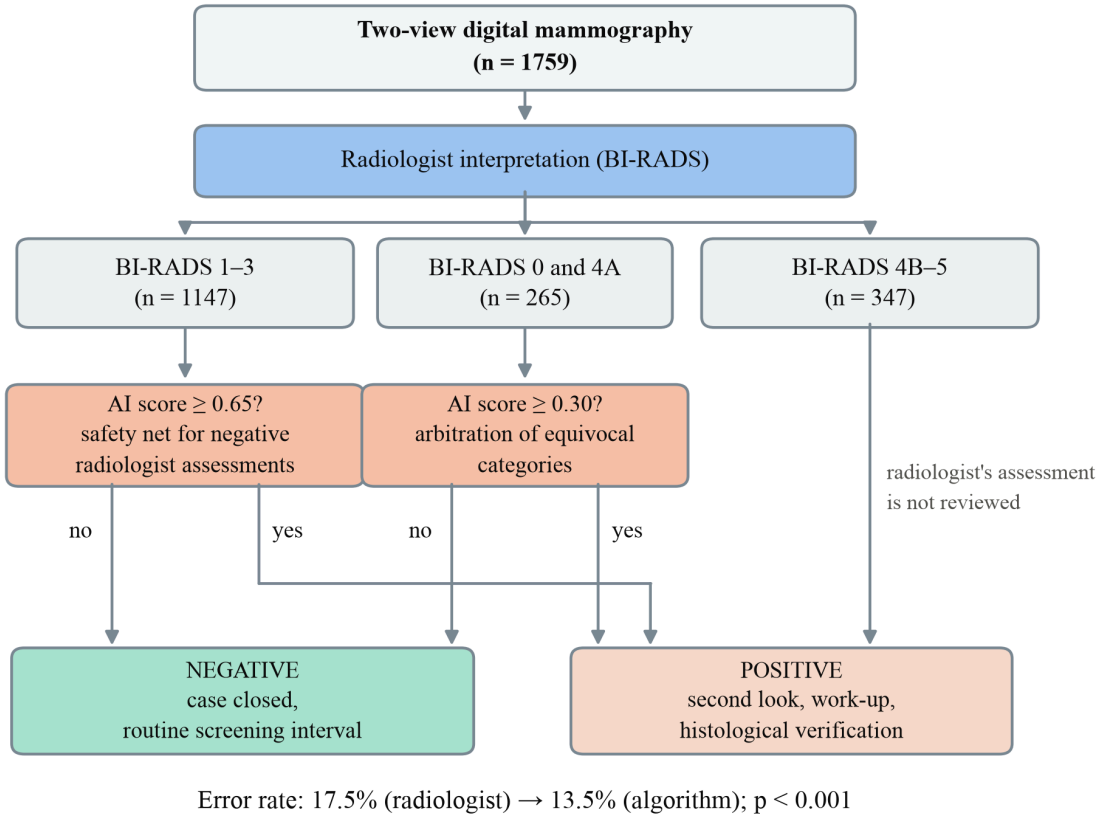


Figure 1. Mammography screening algorithm with selective use of the AI system according to BI-RADS category. BI-RADS 4B, 4C and 5: the radiologist's assessment stands without review. BI-RADS 0 and 4A (arbitration): the examination remains positive if the AI risk score is 0.30 or higher and is otherwise closed as negative. BI-RADS 1–3 (safety net): an AI risk score of 0.65 or higher sends the examination for a second look by the radiologist. AI – artificial intelligence; BI-RADS – Breast Imaging Reporting and Data System.

Table 3. Contribution of the algorithm branches (n = 1,759)

Algorithm branch	Examinations, n	False positives removed, n	Cancers added, n	Cancers lost, n
BI-RADS 4B–5: no review	347	0	0	0
BI-RADS 0 and 4A: arbitration (≥ 0.30)	265	100	0	1
BI-RADS 1–3: safety net (≥ 0.65)	1,147	–30	2	0
Total	1,759	70	2	1

Note: the negative value in the safety-net branch denotes new positive assessments added by the AI.

Arbitration in categories 0 and 4A removed 100 false alarms at the price of one missed cancer; the safety net added 30 second looks but recovered 2 of the 6 cancers missed by radiologists. Overall, the algorithm reduced the number of errors from

308 to 237 (Table 4).

Table 4. Diagnostic performance of radiologist reading and of the radiologist + AI algorithm (n = 1,759)

Measure	Radiologist	Algorithm	Difference
True positives, n	310	311	+1
False positives, n	302	232	–70
False negatives, n	6	5	–1
Sensitivity, % (95% CI)	98.1 (95.9–99.1)	98.4 (96.4–99.3)	+0.3 pp
Specificity, % (95% CI)	79.1 (76.9–81.1)	83.9 (81.9–85.7)	+4.8 pp
PPV, % (95% CI)	50.7 (46.7–54.6)	57.3 (53.1–61.4)	+6.6 pp
NPV, %	99.5	99.6	+0.1 pp
Accuracy, % (95% CI)	82.5 (80.6–84.2)	86.5 (84.9–88.0)	+4.0 pp
Error rate, %	17.5	13.5	–4.0 pp
Positive assessments per cancer detected	1.97	1.75	–0.22

pp – percentage points.

The 4.0-percentage-point reduction in the error rate (95% CI 2.8–5.3) was statistically significant: 102 discordant pairs favoured the algorithm and 31 favoured the radiologist ($\chi^2 = 36.84$; $p < 0.001$). The algorithm was also significantly more specific (100 vs 30 pairs; $p < 0.001$), whereas sensitivity did not differ (2 vs 1 pairs; $p = 1.000$).

Robustness across subgroups.

The algorithm was tested in subgroups defined by breast density, age and region (Table 5).

Table 5. Sensitivity, specificity and error rate by subgroup: radiologist / algorithm

Subgroup	n	Sensitivity, %	Specificity, %	Error rate, %
Density A	168	100.0 / 100.0	80.1 / 86.8	17.9 / 11.9
Density B	738	99.3 / 98.5	79.4 / 83.5	16.9 / 13.7
Density C	638	97.4 / 99.1	76.8 / 82.6	19.4 / 14.4
Density D	196	95.6 / 95.6	82.8 / 86.1	14.3 / 11.7
Age 35–44 years	172	100.0 / 100.0	81.9 / 84.2	18.0 / 15.7
Age 45–59 years	1,145	98.4 / 99.0	77.9 / 83.8	18.7 / 13.7
Age 60–74 years	436	97.5 / 97.5	81.2 / 84.4	14.2 / 11.9
Tashkent	899	98.1 / 98.7	82.1 / 86.4	15.1 / 11.5
Nukus	557	100.0 / 100.0	76.5 / 81.9	18.9 / 14.5
Samarkand	303	94.1 / 94.1	74.6 / 80.2	22.1 / 17.5

The error rate fell in every subgroup, and sensitivity nowhere dropped by more than 0.8 percentage points. Two findings stand out. In density C – the largest of the dense categories and the hardest for the radiologist – sensitivity rose from 97.4% to 99.1% while the error rate fell from 19.4% to 14.4%. And the algorithm helped most where baseline performance was worst: the error rate decreased by 4.6 percentage points in Samarkand, 4.4 in Nukus and 3.6 in Tashkent, narrowing the regional spread from 7.0 to 6.0 points.

Discussion

Our main conclusion is that the value of AI depends less on the accuracy of the model itself than on where in the diagnostic pathway it is applied. On its own the system did not differ significantly from the radiologist in error rate, as in our earlier report on the same cohort [17], and simple combination rules either multiplied overdiagnosis or lost dozens of tumours. A gain appeared only once the AI was confined to the categories in which the radiologist's assessment is objectively least reliable.

Leibig et al.

tested a related idea: in their decision-referral approach the AI confidently sorts part of the examinations and passes uncertain ones to the radiologist, and this division of labour outperformed either reader alone [19]. Our approach works the other way round – the radiologist's assessment is the starting point, and the AI is brought in only for equivocal BI-RADS categories. Where a radiologist reports every examination and carries legal responsibility for it, this seems

to us the more natural design: it keeps the familiar workflow and does not hand any final decision to the machine.

The nature of the gain also differs from that reported abroad. In MASAI and in the German programme, AI mainly increased detection and reduced workload [12, 14]; in the Danish study by Lauritzen et al. the reading volume fell by 62.6% [15]. In our setting the reserve lay on the other side: radiologist sensitivity was already 98.1%, while specificity was only 79.1%, compared with about 89% in the US benchmarks [7]. The task for AI was therefore to remove unnecessary alarms without touching detection. Among 1,759 examinations this meant 70 women spared further work-up and, often, a core biopsy; at the scale of national screening, with tens of thousands of examinations a year, the figure runs into thousands of avoidable procedures.

Dense breasts deserve separate mention.

Dense tissue masks tumours, and all radiologist misses in our material occurred there. Recent work shows that stand-alone AI retains discriminative ability at high density [20] – in our cohort its AUC ranged from 0.886 in density C to 0.935 in density B [17] – and that AI support improves radiologists' accuracy [21]. In our data the algorithm both raised sensitivity and lowered the error rate in density C – something neither arbitration nor the safety net achieves alone. We consider the narrowing of the regional gap equally important: the system performs the same way in the capital and in the regions and partly levels reporting quality where experienced breast radiologists are fewer.

The limitations should be stated plainly.

The study is retrospective, and the clinical benefit of AI can be confirmed only in prospective trials [22]. The sample is enriched with cancer (18.0% versus less than 1% in population screening), so predictive values will have to be recalculated for each population, whereas sensitivity and specificity do not depend on prevalence. The thresholds of 0.30 and 0.65 were chosen on the same data on which the effect was measured, so the effect may be somewhat overestimated. Finally, the radiologist did not see the AI output: this is right for a clean comparison, but it means that the influence of AI prompts on the reader, described by Al-Bazzaz et al. [23], could not be assessed.

In practice the algorithm requires no reorganisation of the service. The AI system is already integrated into the national digital health platform DMED and receives images automatically, and the choice of branch depends only on the BI-RADS category that the radiologist records anyway. The method has been implemented in two regional branches of the institution Republican Specialized Scientific and Practical Medical Center of Oncology and Radiology; prospective evaluation in a population cohort is the next necessary step.

Conclusions

1. Overdiagnosis in mammography is unevenly distributed: 91.7% of radiologists' false-positive assessments fall into BI-RADS 0, 4A and 4B, whereas BI-RADS 4C and 5 assessments are highly reliable (cancer in 82.3% and 94.4%).

2. All cancers missed by radiologists received AI scores above the operating threshold, and no cancer occurred when both readers were negative; simple rules for combining the two readings are clinically unacceptable.

3. Selective use of AI – arbitration of BI-RADS 0 and 4A (threshold 0.30) and a safety net for BI-RADS 1–3 (threshold 0.65), with no review of 4B–5 – reduces the error rate from 17.5% to 13.5% ($p < 0.001$) and raises specificity from 79.1% to 83.9% while preserving sensitivity at 98.4%.

4. The effect is consistent across subgroups and largest in density C and in regions with the weakest baseline performance; prospective validation is required before wide implementation.

Declarations

Funding. The study received no external funding.

Conflict of interest. The evaluated model is owned by Uzinfocom, a state-owned company, and was developed within the national programme established by Presidential Resolution PP-402. The study received no funding from Uzinfocom, and the authors have no financial or commercial interest in the model; they declare no conflict of interest.

Authors' contributions. Askar Adilkhodjaev: Conceptualization, Methodology, Supervision; Umid Tokhtamuratov: Investigation, Data Curation, Writing – Original Draft.

Data availability. The Uzbek clinical dataset is not public; anonymised data are available from the corresponding author on reasonable request, subject to approval by the Ministry of Health.

Declaration of Generative AI Use. The authors used ChatGPT 4.0 to improve language coherence and double-check the analysis.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–249. doi:10.3322/caac.21660 doi:10.3322/caac.21660 (https://doi.org/10.3322/caac.21660)
2. Tillyashaykhov MN, Djanklich SM, Ibragimov SN, Imamov OA. Analysis of cancer incidence structure in the Republic of Uzbekistan. Oncology and Radiology of Kazakhstan. 2021;3(61):4–8. doi:10.52532/2663-4864-2021-3-61-4-8 doi:10.52532/2663-4864-2021-3-61-4-8 (https://doi.org/10.52532/2663-4864-2021-3-61-4-8)
3. World Health Organization. Global Breast Cancer Initiative implementation framework: assessing, strengthening and scaling-up of services for the early detection and management of breast cancer. Geneva: WHO; 2023.
4. Resolution of the President of the Republic of Uzbekistan No. PP-402 of 22 November 2024 on measures for the early detection of breast and cervical cancer.
5. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med. 1994;331(22):1493–1499. doi:10.1056/NEJM199412013312206 doi:10.1056/NEJM199412013312206 (https://doi.org/10.1056/NEJM199412013312206)
6. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst. 1998;90(23):1801–1809. doi:10.1093/jnci/90.23.1801 doi:10.1093/jnci/90.23.1801 (https://doi.org/10.1093/jnci/90.23.1801)
7. Lehman CD, Arao RF, Sprague BL, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. Radiology. 2017;283(1):49–58. doi:10.1148/radiol.2016161174 doi:10.1148/radiol.2016161174 (https://doi.org/10.1148/radiol.2016161174)
8. Perry N, Broeders M, de Wolf C, et al. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition – summary document. Ann Oncol. 2008;19(4):614–622. doi:10.1093/annonc/mdm481 doi:10.1093/annonc/mdm481 (https://doi.org/10.1093/annonc/mdm481)
9. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94. doi:10.1038/s41586-019-1799-6 doi:10.1038/s41586-019-1799-6 (https://doi.org/10.1038/s41586-019-1799-6)

10. Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol. 2020;6(10):1581–1588. doi:10.1001/jamaoncol.2020.3321 doi:10.1001/jamaoncol.2020.3321 (https://doi.org/10.1001/jamaoncol.2020.3321)

11. Yoon JH, Strand F, Baltzer PAT, et al. Standalone AI for breast cancer detection at screening digital mammography and digital breast tomosynthesis: a systematic review and meta-analysis. Radiology. 2023;307(5):e222639. doi:10.1148/radiol.222639 doi:10.1148/radiol.222639 (https://doi.org/10.1148/radiol.222639)

12. Lång K, Josefsson V, Larsson AM, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. 2023;24(8):936–944. doi:10.1016/S1470-2045(23)00298-X doi:10.1016/S1470-2045(23)00298-X (https://doi.org/10.1016/S1470-2045(23)00298-X)

13. Dembrower K, Crippa A, Colón E, Eklund M, Strand F. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health. 2023;5(10):e703–e711. doi:10.1016/S2589-7500(23)00153-X doi:10.1016/S2589-7500(23)00153-X (https://doi.org/10.1016/S2589-7500(23)00153-X)

14. Eisemann N, Bunk S, Mukama T, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. Nat Med. 2025;31(3):917–924. doi:10.1038/s41591-024-03408-6 doi:10.1038/s41591-024-03408-6 (https://doi.org/10.1038/s41591-024-03408-6)

15. Lauritzen AD, Rodríguez-Ruiz A, von Euler-Chelpin MC, et al. An artificial intelligence-based mammography screening protocol for breast cancer: outcome and radiologist workload. Radiology. 2022;304(1):41–49. doi:10.1148/radiol.210948 doi:10.1148/radiol.210948 (https://doi.org/10.1148/radiol.210948)

16. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol. 2019;29(9):4825–4832. doi:10.1007/s00330-019-06186-9 doi:10.1007/s00330-019-06186-9 (https://doi.org/10.1007/s00330-019-06186-9)

17. Adilkhodjaev A, Ibragimov I, Kasparova G, Kurbanova S, Ososkov A, Seytmuratova A, Tokhtamuratov U. Deep learning-based breast cancer detection on mammography in Uzbekistan: a multicentre retrospective study. Clinical Oncology. 2026;16(4):1–8. doi:10.32471/clinicaloncology.2663-466X.36794 doi:10.32471/clinicaloncology.2663-466X.36794 (https://doi.org/10.32471/clinicaloncology.2663-466X.36794)

18. American College of Radiology. ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. 5th ed. Reston, VA: American College of Radiology; 2013.

19. Leibig C, Brehmer M, Bunk S, et al. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. Lancet Digit Health. 2022;4(7):e507–e519. doi:10.1016/S2589-7500(22)00070-X doi:10.1016/S2589-7500(22)00070-X (https://doi.org/10.1016/S2589-7500(22)00070-X)

20. Kwon MR, Chang Y, Ham SY, et al. Screening mammography performance according to breast density: a comparison between radiologists versus standalone intelligence detection. Breast Cancer Res. 2024;26(1):68. doi:10.1186/s13058-024-01821-w doi:10.1186/s13058-024-01821-w (https://doi.org/10.1186/s13058-024-01821-w)

21. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health. 2020;2(3):e138–e148. doi:10.1016/S2589-7500(20)30003-0 doi:10.1016/S2589-7500(20)30003-0 (https://doi.org/10.1016/S2589-7500(20)30003-0)

22. Houssami N, Marinovich ML. AI for mammography screening: enter evidence from prospective trials. Lancet Digit Health. 2023;5(10):e641–e642. doi:10.1016/S2589-7500(23)00176-0 doi:10.1016/S2589-7500(23)00176-0 (https://doi.org/10.1016/S2589-7500(23)00176-0)

23. Al-Bazzaz H, Janicijevic M, Strand F. Reader bias in breast cancer screening related to cancer prevalence and artificial intelligence decision support – a reader study. Eur Radiol. 2024;34(8):5415–5424. doi:10.1007/s00330-023-10514-5 doi:10.1007/s00330-023-10514-5 (https://doi.org/10.1007/s00330-023-10514-5)