CURRENT TOPICS IN NUTRACEUTICAL RESEARCH Vol. 23, No. 2, pp. 1–8, 2025
doi: https://doi.org/10.37290/ctnr.v23i2.24
ISSN 1540-7535 print, ISSN 2641-452X online,

Research

Diagnostic Accuracy of Artificial Intelligence-Assisted Mammography Interpretation Vs. Radiologist Alone: A Systematic Review and Meta-Analysis

Loai Saleh Albinsaad, Eman Abdullah Almubarak, Raghad Mohammad Ahmed Balkhair, Khawlah Abdullah Ali Almana, Ammar Mohammed Ali Alamri, Ruba Mahmoud Abdullah Almuallim, Heba Yousef Habib Alkhamis, Mohammed Yousef Alessa

College of Medicine, King Faisal University, Al Ahsa, Saudi Arabia
Prince Saud Bin Jalawi Hospital, Alahsa, Saudi Arabia
Medical Intern, Faculty of Medicine, Taibah University, Al-Madinah, Saudi Arabia
Medical Student, Faculty of Medicine and Surgery, King Khalid University, Abha- Saudi Arabia
General Practitioner, Alnamas General Hospital, Alnamas-Aseer, Saudi Arabia
Medical Intern, Faculty of Medicine, Tabuk University, Tabuk - Saudi Arabia

Background: Breast cancer has the highest mortality rate among women worldwide, and early detection through mammography is critical for reducing breast cancer mortality. Artificial intelligence (AI) has become a promising tool to support radiologists in mammography interpretation. However, the diagnostic accuracy of AI-assisted interpretation needs to be evaluated and compared with current methods. **Objectives:** This systematic review and meta-analysis aimed to evaluate the diagnostic accuracy of AI-assisted mammography interpretation versus radiologist-alone interpretation in detecting breast cancer. **Methods:** A comprehensive literature search identified studies that compared AI-assisted mammography interpretation with radiologist-alone assessment. Twenty-five eligible studies were included, encompassing retrospective cohorts, prospective trials, and multi-reader studies across diverse healthcare settings and AI algorithms. Data were extracted for sensitivity, specificity, likelihood ratios, and diagnostic odds ratios. Pooled estimates were calculated using a bivariate I^2 statistic, and heterogeneity was assessed with variance measures, including median odds ratios. The risk of bias was evaluated using ROBINS-I and QUADAS-2 tools. Results: The pooled sensitivity of AI-assisted interpretation was 0.82 (95% CI: 0.77–0.85), and specificity was 0.90 (95% CI: 0.85–0.93). The area under the SROC curve (AUC) was 0.91 (95% CI: 0.88–0.93), demonstrating high overall diagnostic performance. The pooled Positive likelihood ratio (PLR) was 7.9 (95% CI: 5.2–11.8), the negative likelihood ratio (NLR) was 0.21 (95% CI: 0.16–0.26), and the diagnostic odds ratios (DOR) were 38 (95% CI: 22–66). These findings suggest that AI significantly enhances the sensitivity of breast cancer detection while maintaining high specificity. However, substantial heterogeneity was observed across studies, reflecting differences in populations, algorithms, and thresholds. **Conclusion:** AI-assisted mammography interpretation comparable to radiologists alone demonstrates high diagnostic accuracy, which may improve early detection rates and reduce false positives. It has the potential to augment radiologist performance in breast cancer screening and integrate into larger-scale breast cancer screening programs. While these results support integration of AI into clinical workflows, further prospective, real-world validation with standardized thresholds is required before widespread adoption.

Keywords: Breast cancer; Mammography; AI-assisted interpretation; Diagnostic accuracy; Breast cancer screening

Abbreviations Used: AI – Artificial intelligence; CI – Confidence interval; SROC – Summary receiver operating characteristic;AUC – Area under the curve; PLR – Positive likelihood ratio; NLR – Negative likelihood ratio;DOR – Diagnostic odds ratio; ROBINS-I – Risk of Bias in Non-randomized Studies of Interventions; QUADAS-2 – Quality Assessment of Diagnostic Accuracy Studies-2

Corresponding Author: Loai Saleh Albinsaad; College of Medicine, King Faisal University, Al Ahsa, Saudi Arabia. E-mail:

INTRODUCTION

Breast cancer is a major public health problem worldwide and a leading cause of cancer deaths among women (1,2,3). Mammography screening programs have been established to reduce breast cancer mortality through early detection (1,2,4). However, the value and sensitivity of mammography screening are still in dispute (5).

Radiologists play a critical role in breast cancer screening by interpreting mammograms to detect suspicious abnormalities that require additional work-up (6,7,3). The benefit of mammography lies in the qualitative visual assessment and subjective interpretation by radiologists, which generates diagnostic information from the images (6,3). This interpretation is subject to reader variability, and suboptimal clinical performance regarding radiologist-to-radiologist variation in diagnostic accuracy may result in unnecessary recalls and missed cancers (8,3). Tumoral architectural distortion is the most difficult sign to determine, given its subtle nature and variable characteristics (3). Novel approaches are required to address the growing shortage of radiologists and prevent missing lesions in mammography (9).

Artificial intelligence (AI) has shown promise in addressing challenges in mammography screening (8,5). AI algorithms may potentially alleviate workload related to false-positive recalls while increasing cancer detection rates and addressing workforce pressures in screening programs (10,11). A promising AI breakthrough for computer-aided diagnosis in mammography is deep learning (DL), a machine learning method that emphasizes the development of convolutional neural networks (12). AI systems can function at a radiologist-comparable level in the assessment of digital mammography (DM), allowing for more efficient and accurate breast cancer screening (13). Artificial intelligence (AI) algorithms can also recognize architectural distortion AD (12).

Despite the possible advantages, AI applied to breast cancer screening needs to be evaluated and compared with current methods (14,15). Prospective studies are required to learn how AI influences cancer detection and false-positive rates in clinical practice (16). Whether AI can compensate for weaknesses in human mammography interpretation and improve interpretative accuracy (17) is worth investigating. Research is needed to directly compare AI with single human interpretation and to assess cancer detection and recall rates for AI-human screen-reading (through simulation) against human double-reading (18). Validation from other external studies is critical for the assessment of AI in datasets other than the one on which the algorithm was trained (11). Given that AI performance continues to climb, the so-called decision-referral approach (19,20)— where AI algorithms predict, based on their quantification of uncertainty—invites closer examination for some high-risk cases, potentially making screening more accurate and less burdensome; however, the approach does not eliminate the need for a radiologist in the loop.

Our systematic review and meta-analysis aim to rigorously consolidate available evidence, offering clarity on the comparative effectiveness of AI versus radiologists. This work will directly inform clinical practice, healthcare policy, and future research, making it highly relevant and timely for the evolving landscape of breast cancer screening. It also includes the maximum quantity of AI models from 2006 to 2024, reflecting technological advancements that earlier reviews missed. We also focus on clinically important diagnostic accuracy outcomes (sensitivity, specificity, false positive rates, cancer detection rates) that are critical for real- world adoption.

MATERIALS AND METHODS

Strategy and Eligibility Criteria

This systematic review was derived from the protocol filed on PROSPERO (CRD420251043734) and was executed and documented in accordance with the Preferred Reporting Items for Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies guidelines (21).

A systematic review of articles assessing the clinical diagnostic accuracy of artificial intelligence-assisted mammography interpretation compared to radiologist interpretation was conducted on November 12, 2024, and subsequently updated on March 9, 2025, and March 29, 2025, utilizing MEDLINE (PubMed "Mesh"). The search was confined to articles published in peer-reviewed journals that included an accessible abstract, without restrictions on publication date. After combining the findings and eliminating redundant research, abstracts and titles were evaluated separately by three reviewers. After then, predetermined inclusion and exclusion criteria were used to evaluate entire texts. A fourth reviewer arbitrated disputes before they were settled by consensus. Four reviewers followed a standardized form to retrieve the data independently. Study design, demographics, diagnostic performance metrics, and results assessed by sensitivity, specificity, false positive rates, and cancer detection rates were among the characteristics that were extracted. Emails were sent to the authors of studies that lacked certain details.

Data Extraction and Quality Assessment

The review procedure outlined above was also used for data extraction and two reviewers independently evaluated each included study's methodologic quality using the ROBINS-I and Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tools, with disagreements being settled by consensus. The research with the highest sample size was kept when many reports were generated from the same cohort in order to prevent data duplication. Performance indicators were averaged for experiments with several readers. The most thorough threshold was chosen for the main analysis if more than one was provided.

Statistical analysis

Diagnostic performance data were obtained by extracting true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values from each study. These values were used to compute the primary accuracy metrics synthesized in the meta-analysis.

Sensitivity was defined as the fraction of diseased individuals who yielded positive results, reflecting the test’s ability to correctly identify true cases. Specificity referred to the percentage of non-diseased individuals who tested negative, representing the capacity to exclude unaffected cases.

The Positive Likelihood Ratio (PLR) expressed how much more probable a positive result was in subjects with the target condition compared to those without it, whereas the Negative Likelihood Ratio (NLR) indicated how much less probable a negative result was among affected individuals relative to the healthy population. An overall index of diagnostic accuracy was estimated using the Diagnostic Odds Ratio (DOR), calculated as the ratio between theodds of a positive test result among diseased versus non-diseased participants. Larger DOR values were interpreted as evidence of stronger discriminative ability (22).

To visually illustrate diagnostic accuracy, a Summary Receiver Operating Characteristic (SROC) curve was plotted, displaying the balance between sensitivity (y-axis) and specificity (x-axis). The diagonal line represented the theoretical threshold of no discrimination. Additionally, forest plots were constructed to show pooled point estimates and confidence intervals for both sensitivity and specificity. Potential publication bias was examined through Deeks' funnel plot asymmetry test, while clinical interpretability was explored using the Fagan nomogram, which integrates pre-test probability with the calculated likelihood ratios to estimate post-test probabilities(23). All analyses were performed using Meta-DiSc 2.0 and Stata 17.0, employing the "MIDAS" analytical framework.

Heterogeneity Analysis

To thoroughly assess heterogeneity within the diagnostic data, several statistical methods were employed. The variance in logit-transformed sensitivity and specificity was calculated, where lower values indicate less inconsistency across study results. The median odds ratio (MOR) was also used to quantify between-study heterogeneity; an MOR of 1 suggests no variability, whereas higher values imply increasing heterogeneity (23). Furthermore, the bivariate I^2 statistic, which ranges from 0 to 1, was used to reflect the extent of between-study variation, with higher scores denoting greater heterogeneity (24,25).

RESULTS

Study selection

Initially, 486 records were retrieved. After excluding zero duplicates, 486 unique studies remained for screening. Title and abstract evaluation resulted in 71 potentially eligible articles, which were subsequently subjected to full-text review. Ultimately, 25 studies met all inclusion criteria and were incorporated into both the qualitative synthesis and the quantitative meta-analysis. The detailed selection process is summarized in Figure 1.

Study Characteristics and Meta-analysis of Diagnostic Test Accuracy

The diagnostic accuracy of AI-assisted mammography interpretation versus radiologist-alone assessment was evaluated across the included studies. The pooled sensitivity, as shown in Figure 6, was 0.82 (95% CI: 0.77–0.85), demonstrating a strong capacity to accurately detect genuine positive cases, while the specificity, demonstrated in Figure 7, was 0.90 (95% CI: 0.85–0.93), reflecting strong capability to exclude non-disease cases. The SROC curve illustrated in Figure 8 provides a 95% confidence contour along with a thorough analysis of sensitivity and specificity across the listed studies. The AI model's high diagnostic accuracy was highlighted by the computed area under the curve (AUC), which was 91% (95% CI: 88%–93%).

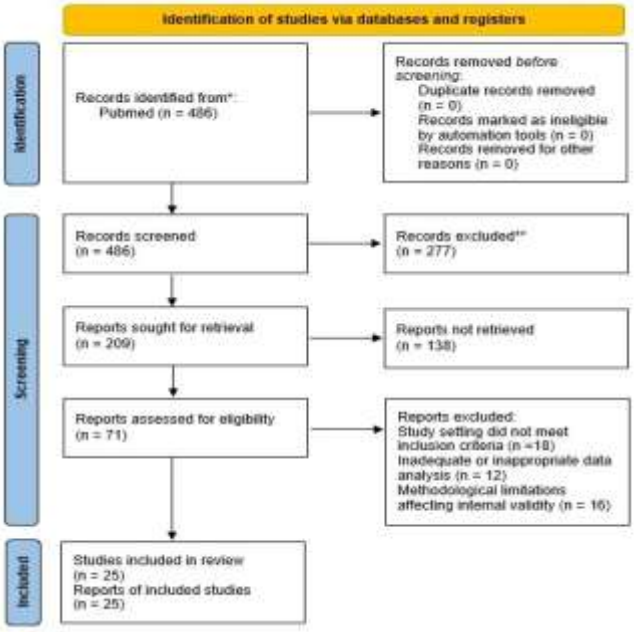

Figure 1. PRISMA flow diagram of study selection

Further analysis of diagnostic accuracy metrics included DOR, PLR, and NLR. The pooled PLR was 7.9 (95% CI: 5.2–11.8), suggesting that individuals with positive AI-assisted results were nearly eight times more likely to truly have breast cancer. The NLR was 0.21 (95% CI: 0.16–0.26), suggesting a significant decline in the probability of disease after a negative result. Besides, the DOR reached 38 (95% CI: 22–66), highlighting the robust discriminative performance of AI tools.

Heterogeneity assessment revealed significant variability across studies. The variance in logit sensitivity was 0.45, and for logit specificity it was 1.211. The bivariate I^2 statistic was notably high at 0.918, suggesting substantial heterogeneity in both sensitivity and specificity. The median odds ratio (MOR) for sensitivity was 1.896, while for specificity it was higher at 2.856, indicating notable between-study differences in diagnostic performance, especially in specificity. These findings underscore the promising diagnostic potential of AI-assisted mammography, though variability across studies warrants cautious interpretation and consideration of contextual clinical factors.

Methodologic Quality

Our meta-analysis included studies that assessed the diagnostic performance of artificial intelligence (AI)- assisted mammography interpretation compared with standard radiologist-alone reading. A total of 25 studies were analyzed, comprising hundreds of mammography examinations across diverse populations and healthcare settings. The included studies spanned various geographic regions, including the USA, Denmark, Australia, South Korea, the UK, Norway, and Turkey. The study designs included retrospective cohort studies, simulation models, multi-reader studies, and prospective trials. Sample sizes varied widely, from smaller cohorts of fewer than 300 patients to large-scale national screening datasets involving thousands of

mammograms. AI tools evaluated in these studies included a range of commercial and in-house algorithms such as Lunit INSIGHT, Transpara, and DeepHealth, often compared against single or double reading by radiologists. Across studies, AI performance was commonly reported in terms of AUC, sensitivity, specificity, and odds ratios. In general, AI algorithms demonstrated comparable or improved diagnostic accuracy relative to radiologists, particularly when used in hybrid or triage workflows. Further details regarding demographic characteristics, AI software versions, study settings, and performance metrics are demonstrated in Table 1.

Dual independent review evaluated the bias using the ROBINS-I and QUADAS-2 tools. These tools were chosen based on the study designs included in this review. The ROBINS- I tool assesses risk of bias in specific areas, including confounding, participant selection, intervention classification, deviations from intended interventions, missing data, outcome measurement, and reporting bias. In contrast, the QUADAS-2 tool looks at risk of bias of four main areas: patient selection, the index test, the reference standard, and flow and timing. Each area was rated as low, moderate, serious, or unclear risk of bias, following the guidance from the official manual of the tools. Any disagreements were resolved through joint decision, and a third reviewer provided arbitration when needed. The overall risk of bias using the ROBINS-I tool was seen as moderate to serious across the included studies. The most common sources of bias included a lack of control for confounding factors and significant missing data in some studies. Meanwhile, the overall risk of bias using the QUADAS-2 tool was high in 5 out of 8 studies, primarily due to using a case-control design, inappropriate exclusions, and lack of blinding to the reference standard result.

A visualization of risk of bias judgments across all studies is provided in Figures 2-5

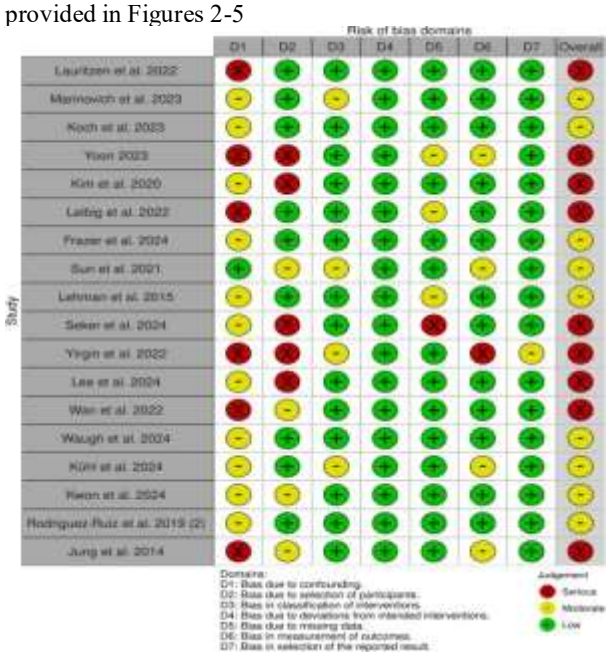

Figure 2. Study-level risk of bias (ROBINS-I) across all included studies.

Figure 3. Proportion of studies judged at low, moderate, or high risk of bias across ROBINS-I domains

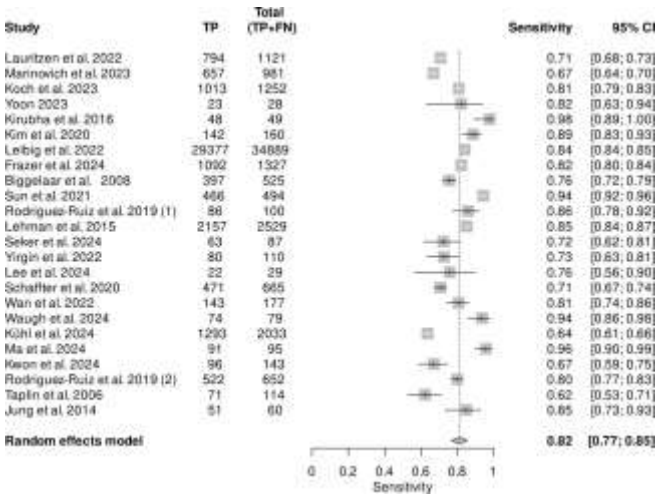

Figure 4. Summary of risk of bias in diagnostic accuracy studies using QUADAS-2 domains.

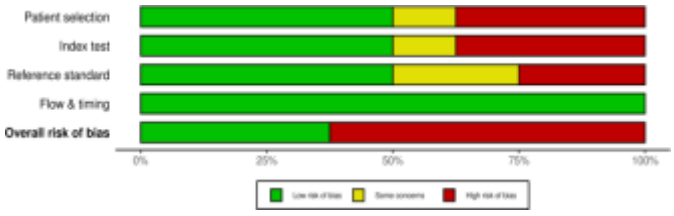

Figure 5. Study-level risk of bias (QUADAS-2) across diagnostic accuracy studies.

Meta-analysis of diagnostic test accuracy

The diagnostic accuracy of AI-assisted mammography interpretation versus radiologist-alone assessment was evaluated across the included studies. The pooled sensitivity, as shown in Figure 6, was 0.82 (95% CI: 0.77–0.85), demonstrating a strong capacity to accurately detect genuine positive cases, while the specificity, demonstrated in Figure 7, was 0.90 (95% CI: 0.85–0.93), reflecting strong capability to exclude

non-disease cases. The SROC curve illustrated in Figure 8 provides a 95% confidence contour along with a thorough analysis of sensitivity and specificity across the listed studies. The AI model's high diagnostic accuracy was highlighted by the computed area under the curve (AUC), which was 91% (95% CI: 88%–93%).

Further analysis of diagnostic accuracy metrics included DOR, PLR, and NLR. The pooled PLR was 7.9 (95% CI: 5.2–11.8), suggesting that individuals with positive AI-assisted results were nearly eight times more likely to truly have breast cancer. The NLR was 0.21 (95% CI: 0.16–0.26), suggesting a significant decline in the probability of disease after a negative result. Besides, the DOR reached 38 (95% CI: 22–66), highlighting the robust discriminative performance of AI tools.

Heterogeneity assessment revealed significant variability across studies. The variance in logit sensitivity was 0.45, and for logit specificity it was 1.211. The bivariate I^2 statistic was notably high at 0.918, suggesting substantial heterogeneity in both sensitivity and specificity. The median odds ratio (MOR) for sensitivity was 1.896, while for specificity it was higher at 2.856, indicating notable between-study differences in diagnostic performance, especially in specificity. These findings underscore the promising diagnostic potential of AI-assisted mammography, though variability across studies warrants cautious interpretation and consideration of contextual clinical factors.

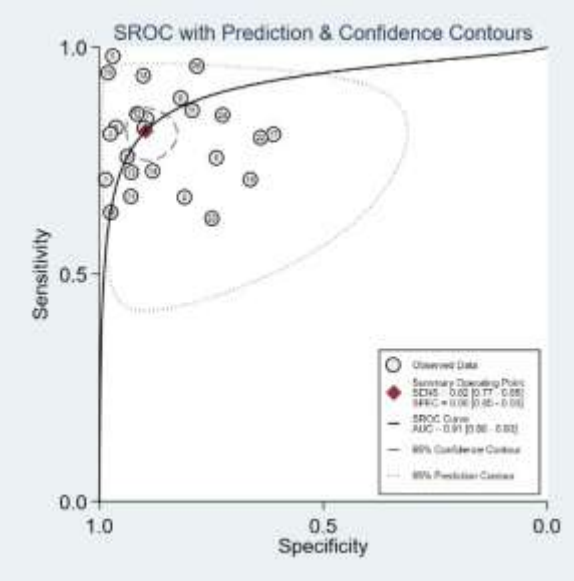

Figure 6: Pooled estimate of sensitivity for AI-assisted mammography interpretation.

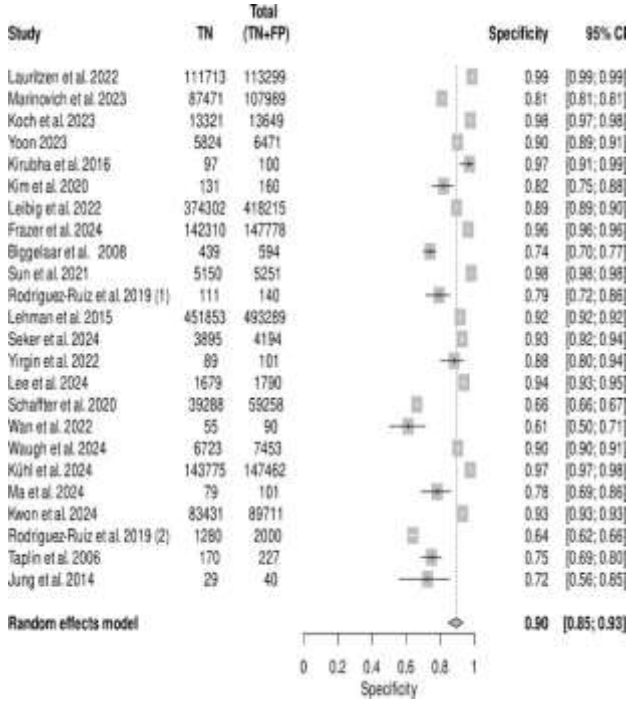

Figure 7: Pooled estimate of specificity for AI-assisted mammography interpretation.

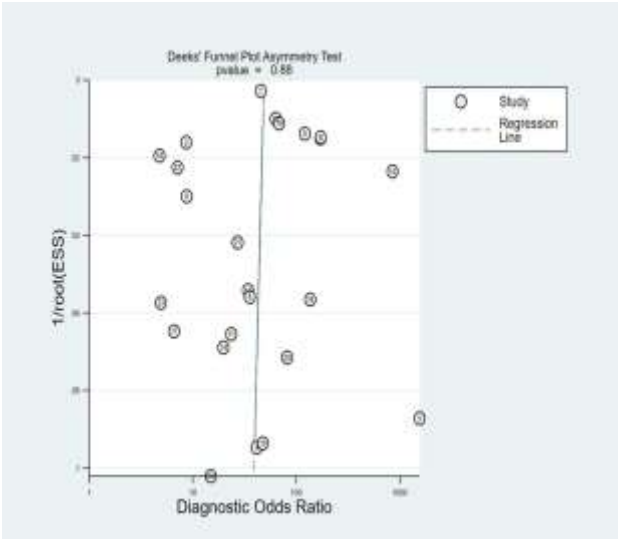

Figure 8: Summary receiver operating characteristic (SROC) curve of AI-assisted mammography interpretation.

Publication bias and clinical utility

We assessed the presence of publication bias using Deeks' funnel plot asymmetry test, as depicted in Figure 9. The test revealed no significant asymmetry (P-value = 0.88), indicating a low likelihood of publication bias across the included studies. Furthermore, the clinical applicability of the diagnostic test was evaluated through Fagan's nomogram, presented in Figure 10. Assuming a pre-test probability of 3%, a positive result increased the post-test probability to 20%, corresponding to a PLR of 8. In contrast, a negative result reduced the post-test probability to 1%, with an NLR of 0.21.

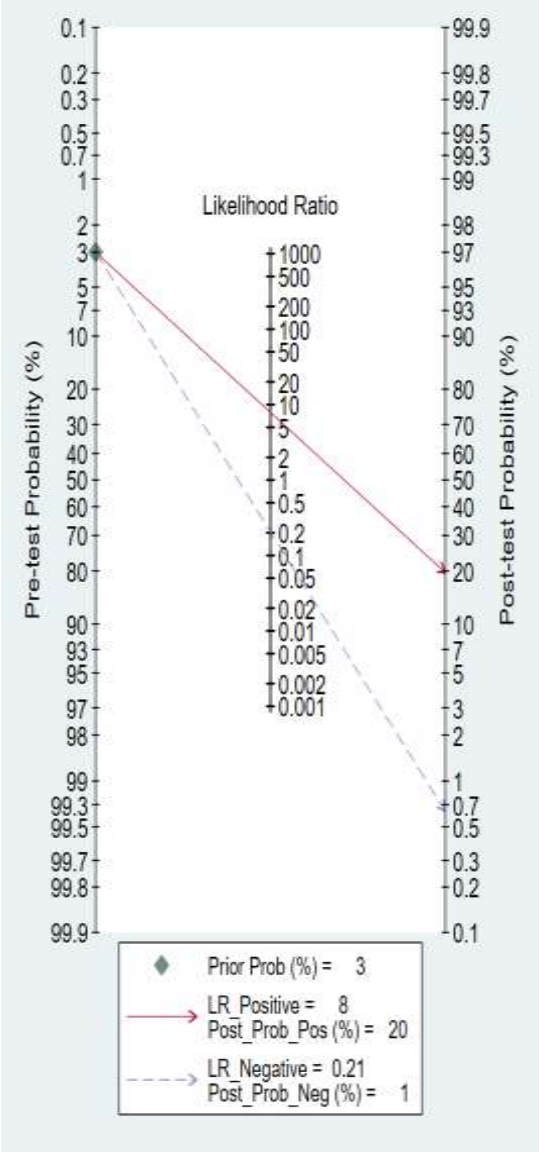

Figure 10: Fagan's Nomogram for clinical utility.

DISCUSSION

The current meta-analysis provides the most comprehensive synthesis to date on the diagnostic performance of AI-assisted mammography interpretation compared to radiologist-alone assessment, revealing diagnostic utility for AI tools in breast cancer detection. Our collective analysis brought to light a high sensitivity of 0.82 (95% CI: 0.77– 0.85), indicating the AI-assisted approach's strong ability to correctly identify true positive cases. In parallel, the overall specificity reached 0.90 (95% CI: 0.85–0.93), reflecting a robust capability to accurately exclude non-cancerous cases. The high Area Under the Curve (AUC) of 91% (95% CI: 88%–93%) on the SROC curve strengthens the overall diagnostic accuracy, emphasizing the strong discriminative performance of AI-assisted mammography.

These results suggest that, radiologist performance could be augmented by it, especially with the high pooled sensitivity and specificity. Therefore, AI-assisted mammography could significantly improve early detection rates and reduce false positives, providing a timely and appropriate patient management. The calculated pooled Positive Likelihood Ratio (PLR) of 7.9 (95% CI: 5.2–11.8) further strengthens this assertion, suggesting that individuals with a positive AI-assisted result are nearly eight times more likely to truly have breast cancer. On the contrary, the Negative Likelihood Ratio (NLR) of 0.21 (95% CI: 0.16–0.26) implies a substantial decrease in the probability of disease following a negative AI-assisted result, offering reassurance in ruling out the condition.

Other than augmentation of the radiologist's performance, there is a substantial clinical implications of these findings, suggesting that there is a strong potential with the discriminative performance (underscored by a high DOR and AUC) for integration into larger scale breast cancer screening programs, as well as for contributing to standardized screening practices.

Despite the promising findings, this study is not without limitations. Among the 14 included studies, nine were rated as having a high risk of bias, as assessed using the ROBINS-I tool, and five using the QUADAS-2 tool. These biases stemmed from issues such as non-randomized designs, lack of blinding, inappropriate exclusions, and the use of retrospective data. Moreover, there was considerable heterogeneity in study populations, AI algorithms, and diagnostic thresholds, which may affect the generalizability of the findings.

An additional limitation is that full texts for some potentially relevant studies were not able to be obtained, which may have influenced comprehensive coverage of the review and precluded more in-depth exploration of their background and conclusions. Finally, most of the studies were not based on real-world clinical practice, being instead essentially retrospective or simulation-based.

To support future research and clinical implementation, prospective studies with blinded readers and standardized reporting thresholds are urgently needed. Establishing large, shared databases and fostering collaboration between institutions will accelerate clinical validation. Most importantly, AI should not be viewed as a replacement for radiologists but as a tool that enhances their expertise and integrates smoothly into screening workflows for optimal patient outcomes.

CONCLUSION

This systematic review and meta-analysis thoroughly evaluated the diagnostic performance of artificial intelligence (AI)– assisted mammography in comparison to radiologist interpretation. Across 25 studies, AI integration demonstrated robust pooled sensitivity (0.82) and specificity (0.90), with an AUC of 0.91, demonstrating its robust diagnostic performance. The diagnostic odds ratio of 38 and favourable likelihood ratios further attest to the reliability of AI

systems in differentiating malignant from non-malignant cases. In combination, these results demonstrate artificial intelligence's capability to increase the accuracy of breast cancer diagnosis, while reducing false positives, and therefore provide strong support for population-based screening programs involving radiologists.

In spite of such promising results, appreciable levels of heterogeneity became apparent, reflecting differences in patient demographics, algorithmic decisions, and criteria for interpretation. Additionally, the prevalence of retrospective and simulation-based designs, along with moderate to significant risks of bias in most studies, limits the generalizability of the results. In daily practice, artificial intelligence also needs to be considered as an ancillary tool, not intended to replace radiologists, with capabilities to improve diagnostic uniformity, workload reduction, and facilitate earlier disease recognition.

Prospective multi-center trials with standardized measures of reporting will need to be included in future studies to allow for successful and safe integration of this approach as a regular part of care. In brief, AI-assisted mammography is a promising supplement to optimal breast cancer screening; however, its large-scale deployment in everyday use requires rigorous verification and validation.

REFERENCES

Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019;111(9):916-922. doi:10.1093/jnci/djy222

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

Wan Y, Tong Y, Liu Y, et al. Evaluation of the Combination of Artificial Intelligence and Radiologist Assessments to Interpret Malignant Architectural Distortion on Mammography. Front Oncol. 2022;12:880150. Published 2022 Apr 20. doi:10.3389/fonc.2022.880150

Marinovich ML, Wylie E, Lotter W, et al. Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. EBioMedicine. 2023;90:104498. doi:10.1016/j.ebiom.2023.104498

Kizildag Yirgin I, Koyluoglu YO, Seker ME, et al. Diagnostic Performance of AI for Cancers Registered in A Mammography Screening Program: A Retrospective Analysis. Technol Cancer Res Treat. 2022;21:15330338221075172. doi:10.1177/15330338221075172

Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms [published correction appears in JAMA Netw Open. 2020 Mar 2;3(3):e204429. doi: 10.1001/jamanetworkopen.2020.4429.]. JAMA Netw Open. 2020;3(3):e200265. Published 2020 Mar 2. doi:10.1001/jamanetworkopen.2020.0265

Marinovich ML, Wylie E, Lotter W, et al. Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. EBioMedicine. 2023;90:104498. doi:10.1016/j.ebiom.2023.104498

Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study Dembrower, Karin et al. The Lancet Digital Health, Volume 5, Issue 10, e703 - e711

Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019;111(9):916-922. doi:10.1093/jnci/djy222

Marinovich ML, Wylie E, Lotter W, et al. Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. EBioMedicine. 2023;90:104498. doi:10.1016/j.ebiom.2023.104498

Marinovich ML, Wylie E, Lotter W, et al. Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. EBioMedicine. 2023;90:104498. doi:10.1016/j.ebiom.2023.104498

Wan Y, Tong Y, Liu Y, et al. Evaluation of the Combination of Artificial Intelligence and Radiologist Assessments to Interpret Malignant Architectural Distortion on Mammography. Front Oncol. 2022;12:880150. Published 2022 Apr20. doi:10.3389/fonc.2022.880150

Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019;111(9):916-922. doi:10.1093/jnci/djy222

Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study Dembrower, Karin et al. The Lancet Digital Health, Volume 5, Issue 10, e703 - e711

Marinovich ML, Wylie E, Lotter W, et al. Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. EBioMedicine. 2023;90:104498. doi:10.1016/j.ebiom.2023.104498

Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study Dembrower, Karin et al. The Lancet Digital Health, Volume 5, Issue 10, e703 - e711

Wan Y, Tong Y, Liu Y, et al. Evaluation of the Combination of Artificial Intelligence and Radiologist Assessments to Interpret Malignant Architectural Distortion on Mammography. Front Oncol. 2022;12:880150. Published 2022 Apr 20. doi:10.3389/fonc.2022.880150

Marinovich ML, Wylie E, Lotter W, et al. Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. EBioMedicine. 2023;90:104498. doi:10.1016/j.ebiom.2023.104498

Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. 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

Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. 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

Salameh JP, Bossuyt PM, McGrath TA, Thombs BD, Hyde CJ, Macaskill P, Deeks JJ, Leeflang M, Korevaar DA, Whiting P, Takwoingi Y, Reitsma JB, Cohen JF, Frank RA, Hunt HA, Hooft L, Rutjes AWS, Willis BH, Gatsonis C, Levis B, Moher D, McInnes MDF. Preferred reporting items for systematic review and meta- analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ. 2020 Aug 14;370:m2632. doi: 10.1136/bmj.m2632. PMID: 32816740.\

A.S. Glas, J.G. Lijmer, M.H. Prins, et al., The diagnostic odds ratio: a singleindicator of test performance, J. Clin. Epidemiol. 56 (2003) 1129–1135, https://doi.org/10.1016/s0895-4356(03)00177-x.

C.G.B. Caraguel, R. Vanderstichel, The two-step Fagan's nomogram: ad hoc interpretation of a diagnostic test result without calculation, Evid. Based Med. 18 (2013) 125–128, https://doi.org/10.1136/eb-2013-101243.

M.N. Plana, T. P´erez, J. Zamora, New measures improved the reporting of heterogeneity in diagnostic test accuracy reviews: a metaepidemiological study, J. Clin. Epidemiol. 131 (2021) 101–112, https://doi.org/10.1016/j.jclinepi.2020.11.011

Y. Zhou, N. Dendukuri, Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy, Stat. Med. 33 (2014) 2701–2717, https://doi.org/10.1002/sim.6115

Table 1. Characteristics of included studies

Study	Type of study	Sample Size	Demographics (Age-Gender-Inclusion/Exclusion Criteria)	Type of Intervention	Comparison	Effect Sizes (Cohen's d,Odds ratio (OR), risk ratio (RR), or hazard ratio (HR))
Lauritzen et al. 2022	Retrospective Simulation Study	114421	Women aged 50–69 in Denmark's Capital Region breast cancer screening program	AI (Transpara v1.7.0) excluded normal cases, sent suspicious directly to recall; moderate risk read by radiologists	Standard double reading by radiologists	AI-based sensitivity: 69.7%, radiologists: 70.8% (noninferior); AI specificity: 98.6% vs 98.1%
Marinovich et al. 2023	Retrospective cohort study	108,970 women with consecutive screening mammograms.	Mean age: 61.0 years (SD 6.9); Gender: Female; Inclusion: Women aged 50–74 years from the BreastScreen Western Australia (BSWA) program.	AI algorithm (DeepHealth Saige-Q v2.0.0) for automated reading of mammograms.	AI performance vs. radiologists in a simulated AI-radiologist reading workflow with arbitration.	AI-radiologist reading resulted in lower recall rates (3.14% vs. 3.38% for standard double-reading) and a small reduction in CDR (6.37 vs. 6.97 per 1000). AI detected interval cancers that were missed by radiologists. Effect Sizes: AI standalone AUC: 0.83 (95% CI: 0.80–0.86); Radiologists: AUC 0.93 (95% CI: 0.90–0.96). AI sensitivity: 0.67 (95% CI: 0.64–0.70); Radiologists: 0.68 (95% CI: 0.66–0.71)
Koch et al. 2023	Retrospective, registry study	14,900 examinations, including 1254 breast cancer cases, 12,642 negative controls, and prior exams for 1004 women with cancer.	Age: Mean age was 58 years (SD = 6). Gender: Female Inclusion/Exclusion Criteria: All participants were part of the BreastScreen Norway program, aged 50–69 years, with data from 2010– 2018. Exclusions included cases with incomplete data.	AI system (Transpara version 1.7.0) processed the mammograms and assigned risk scores from 1 to 10. These scores were compared with the independent double reading results by radiologists.	Results of AI scoring were compared to the traditional double reading by two radiologists.	Sensitivity: Radiologists (double reading): 62.8% (dense breasts). AI (score 10 threshold): 80.9% (dense breasts). Specificity: Radiologists: 97.6% (first reader). AI: Matched to 97.6% specificity
Yoon 2023	Retrospective Study	6499 mammograms from 5228 women	Age: Not specified; Gender: Female; Inclusion: Women undergoing routine mammography screening	AI-CAD software (Lunit INSIGHT for Mammography, version 1.1.0.1)	Radiologists' interpretations	AI-CAD detected 17.9% of cancers missed by radiologists but had high false-positive rates.- Effect Sizes: AI-CAD detected 17.9% of cancers overlooked by radiologists

Kirubha et al. 2016	The study developed an automated computer-aided diagnosis (CAD) system to segment breast masses from mammograms with high accuracy, using a multimodal segmentation algorithm 1. The algorithm combines dual-tree complex wavelet transform (DTCWT), watershed segmentation, and K-means clustering 2 3. The mammograms were processed using MATLAB 7 software	A total of 150 mammograms were used in the study after excluding 8 subjects with micro-calcification from an initial download of 158 subjects' data from the DDSM database	The data consisted of majorly Caucasian women who underwent breast examinations in the United States 6. Age: The studied data were classified into three groups: Group I (Normal): n = 50, mean age = 55 ± 8 years 1 7 Group II (Benign): n = 50, mean age = 58 ± 11 years 1 7 Group III (Malignant): n = 50, mean age = 58 ± 9 years 1 7 Gender: The study focused on female subjects. Inclusion/Exclusion Criteria: Inclusion: Mammograms with normal, benign breast cancer, and malignant breast cancer were included 1 7. Exclusion: Subjects with micro-calcification in any one side of the breast were excluded from the study	The primary intervention was the application of an automated multimodal segmentation algorithm for breast mass detection in mammograms 1 2. This algorithm combines DTCWT, watershed segmentation, and K-means clustering to automatically segment abnormal breast masses	The results of the automated segmentation were compared to the abnormal region outlined in mammograms by radiologists of the American College of Radiology (ACR), which was considered the "standard"	Performance Metrics: The study reported the performance of the multimodal segmentation algorithm in terms of sensitivity, specificity, accuracy, false positive rate, precision, and Kappa coefficient . Statistical Data: The segmented benign and malignant breast mass regions were well-matched with the abnormality regions outlined by radiologists of ACR 1. In normal cases, the algorithm segmented different breast regions like the pectoral muscle and breast region based on gray scale texture features and pixel density 1. The measured breast mass size in benign cases ranged from 126 to 2243 pixels, whereas in malignant cases, it ranged from 225 to 2786 pixels 2. Effect Sizes, P-values, and Confidence Intervals: The document does not explicitly provide effect sizes such as Cohen's d, Odds Ratio (OR), Risk Ratio (RR), or Hazard Ratio (HR), nor does it list specific p-values or confidence intervals . However, it provides accuracy measurements using the Kappa statistic .
Kim et al. 2020	Retrospective, multireader study	170,230 mammography examinations	Mean age: 50.3 years (SD 10.0); Gender: Female; Inclusion:	AI algorithm (Lunit INSIGHT MMG) for	Radiologists without AI assistance vs. radiologists with AI assistance.	AI significantly improved radiologists' diagnostic performance, particularly in

		(development dataset); 320 mammograms (reader study)	Mammograms from five institutions in South Korea, the USA, and the UK.	breast cancer detection in mammography.		detecting early-stage cancers. Effect Sizes: AI standalone AUROC: 0.959 (95% CI: 0.952–0.966); Radiologists without AI: AUROC 0.810 (95% CI: 0.770–0.850); Radiologists with AI: AUROC 0.881 (95% CI: 0.850–0.911). and cancers in dense breasts.
Leibig et al. 2022	Retrospective Analysis	1193197	453,104 women screened in eight German screening centers from 2007–2020	Standalone AI vs. Decision-referral AI system (normal triage + safety net)	Compared with unaided radiologist reading ahead of consensus conference	Standalone AI slightly worse than radiologist; decision- referral approach improved sensitivity and specificity
Frazer et al. 2024	Simulation study (retrospective)	149,105 screening episodes from 92,839 clients	Age: Women aged 40–74 years ,Gender: Female only ,Inclusion Criteria: All women who underwent screening mammograms in the BreastScreen Victoria program from 2016 to 2019 ,Exclusion Criteria: Missing clinical data, incomplete image data, and patients who underwent multiple screening attempts	Various AI-integrated pathways (AI as a standalone reader, as a second reader, and other combinations with human involvement)	Standard two-reader system with arbitration	AI standalone achieved higher sensitivity (75.0%) and specificity (96.0%) than individual human readers
Choi et al. 2023	Retrospective diagnostic accuracy study.	Total Patients: 392 women.	Average Age: 57.3 ± 12.1 years (range: 30–94 years). Gender: Female. Inclusion Criteria: Women diagnosed with malignancy who underwent digital mammography prior to biopsy. Exclusion Criteria: Patients without mammography, those with film mammography, or those with computed radiography.	Use of Lunit INSIGHT MMG (AI software) for breast cancer detection in mammograms.	Two radiologists with access to clinical symptoms and prior mammography.	Kappa Coefficient: AI: 0.698. Odds Ratio (OR): Prior mammography influence: OR = 8.55. Clinical symptoms: OR = 5.49. Fatty breast density: OR = 5.18.
Biggelaar et al. 2008	prospictive	1,048 patients (51 cancers).	79% diagnostic, 21% screening; mean age 51	SecondLook CAD (CAD).	Radiologist vs. technologist performance with/without CAD.	CAD improved technologists' sensitivity but lowered specificity. Effect: CAD sensitivity: 78% (40/51 cancers). - Technologists' mean sensitivity: 92% vs. radiologist: 84%.

Sun et al. 2021	Multicenter clinical study (retrospective and prospective)	Retrospective training and validation: 4,113 patients (2,454 malignant, 1,665 benign). Prospective clinical application: 5,746 patients (495 malignant,	Age: Mean age ranged from 49.99 to 51.50 years across centers. Gender: Female (implied by breast cancer focus). Inclusion Criteria: Patients with complete clinical data, mammogram data, and pathological diagnosis or more than 2 years of follow-up. Exclusion Criteria: Unqualified images or inconsistency in lesion location between mammograms and pathological results.	AI-assisted diagnosis model based on deep learning for mammography.	Comparison of radiologists' performance with and without the AI model.	Sensitivity: 0.908 (after matching) at a false positive rate of 0.25 per image. Radiologists' AUC with AI model: 0.852. Prospective sensitivity: 0.887 at a false positive rate of 0.25.

		337 benign, 4,914 negative).				
Rodriguez-Ruiz et al. 2019 (1)	Retrospective Study	240 women (100 with cancer, 40 false positives, 100 normal)	women (median age 62 years; range 39–89 years)	Single Reading vs Double Reading vs CAD		AI support improved cancer detection (AUC increased, sensitivity improved, specificity trended toward- Effect Sizes: AUC increased from 0.87 to 0.89 with AI support (P = .002) improvement)
Lehman et al. 2015	Retrospective Observational Study	625625	323,973 women aged 40–89 years undergoing digital screening mammography	Computer-aided detection (CAD) during mammogram interpretation	Comparison between mammograms interpreted with CAD vs without CAD	No significant improvement with CAD in any metric
Seker et al. 2024	Retrospective observational study	5,136 mammograms from 4,282 women, with 105 diagnosed with breast cancer	Age: Women aged 40–69 years Gender: Female Inclusion/Exclusion Criteria: Mammograms from a population- based screening program in Bahcesehir, Istanbul, Turkey. Excluded if mammogram data was missing or of poor quality.	AI-based mammography interpretation (Lunit INSIGHT MMG), which assigned a risk score from 1 to 100 to each mammogram.	Comparison between radiologists' BI-RADS assessments and AI's risk score. Different workflow scenarios were also simulated (AI as a second reader, hybrid triage, etc.).	AI identified 72.38% of cancers correctly, with high specificity (92.86%). AI also detected 23% of cancers earlier than radiologists, with a mean detection time 29.92 months earlier.
Yirgin et al. 2022	Retrospective Study	211	Women aged 40-69 from Bahcesehir screening program (Turkey)	AI algorithm (Lunit INSIGHT MMG) evaluated as third reader	Comparison with radiologists and combined performance	AI AUC=0.853, Sensitivity=72.8%, Specificity=88.3%, better than radiologists

Lee et al. 2024	Retrospective Study	1819	Women (mean age 50.8 ± 9.4) undergoing screening mammography + ultrasound	AI-CAD using Lunit INSIGHT MMG integrated with PACS	Comparison: radiologists with vs. without AI-CAD, and stand-alone AI	AI-CAD improved specificity, accuracy, and reduced recall rate vs radiologists
Schaffter et al. 2020	Diagnostic accuracy study	144,231 screening mammograms from 85,580 US women and 166,578 examinations from 68,008 Swedish women	Women from the US and Sweden undergoing mammography screening	AI algorithms for mammography interpretation, combined with radiologist assessment	AI algorithms vs. radiologists, standalone AI vs. AI combined with radiologists	Combining AI with radiologist assessment improved specificity and overall accuracy. Effect Sizes: Top-performing AI algorithm AUC: 0.858 (US), 0.903 (Sweden); combined AI and radiologist AUC: 0.942
Wan et al. 2022	Retrospective case-control study.	Total Patients: 267 (177 malignant, 90 benign). Mammograms: Bilateral mammograms with two views (craniocaudal and mediolateral oblique).	Age: Malignant group: Mean = 49.51 ± 9.12 years. Benign group: Mean = 48.18 ± 7.65 years. Gender: Female. Inclusion Criteria: Patients with architectural distortion on mammography, confirmed by histopathology or follow-up. Exclusion Criteria: History of breast cancer or prior surgery.	Use of an AI system to assist radiologists in detecting malignant AD.	Junior radiologists (3 first-readers with 2–4 years of experience). Senior radiologists (3 second-readers with 8–12 years of experience).	AUC: AI alone: 0.792. AI + Junior radiologist (Reader First-1): 0.880. AI + Senior radiologist (Reader Second-1): 0.893. AI + Consensus: 0.908.
Waugh et al. 2024	Retrospective, observational study	7533 women screened in 2017.	Age: Participants were primarily women aged 50–74, although ages ranged from 40 to 85 years. Gender: Female Inclusion/Exclusion Criteria: Women attending for their prevalent mammogram through the BreastScreen Australia program in 2017, with 96.2% of mammograms included in the analysis.	Digital mammograms were obtained using Siemens DR, Mammomat and Inspiration, Sectra DR, and Hologic Dimensions units. The AI software used was Transpara (version 1.7.0),	AI scores were compared with the results of the two independent radiologists who assessed the mammograms.	Sensitivity: Radiologists (2017 round): 100% (67/67) for screen-detected cancers. AI (score 10 threshold): 94% (63/67) for screen-detected cancers. 88.1% (59/67) when including interval cancers and 2019 round minimal signs (vs. radiologists' 80.6%, p=0.24). Specificity: Radiologists: 91.2% (6,805/7,466). AI (score 10 threshold): 90.2% (6,731/7,466).

Kühl et al. 2024	Retrospective Cohort Study	249,402 screenings (149,495 women)	Women aged 50–69 years; 2,033 breast cancers (1,475 screen-detected, 558 interval cancers)	AI-based mammogram interpretation	AI system vs. first-reading radiologists	AI system showed comparable accuracy to first readers when matched at specificity. Effect Sizes: AI sensitivity: 63.6% (AIsens), 62.6% (AIspec); specificity: 97.5% (AIsens), 97.7% (AIspec)
Ma et al. 2024	Prospective study	196 patients with 202 breast masses	Age: Patients with malignant masses were significantly older (59.69 ± 13.18 years) than those with benign masses (44.03 ± 14.09 years) 7. Gender: Female 3 8 Inclusion Criteria: (1) Breast masses were incidentally detected during ultrasound scanning as part of a routine health examination, with no prior history of detected masses; (2) The identified breast masses were categorized as BI-RADS 3-5 upon initial examination; (3) The ultrasound images of the masses were of high quality; (4) The masses were identifiable by AI software; (5) The masses were ultimately confirmed through surgical excision or core needle biopsy 8. Exclusion Criteria: (1) Patients with masses larger than 40 mm; (2) Breast masses with unclear ultrasound images; (3) Non-mass lesions in the breast; (4) Breast masses with inconclusive pathological findings	AI-assisted diagnosis using the AI-SONIC Breast intelligent assisted diagnosis system developed by Zhejiang Deshan Yunxing Company	Comparison of diagnostic performance between junior radiologists, senior radiologists, and AI, with and without AI assistance	AI BI-RADS classification had a high sensitivity (95.79%) compared to radiologists . AI had the highest diagnostic efficiency Q3 an AUC of 0.950 (p = 0.000) The integration of AI software enhanced the diagnostic sensitivity and negative predictive value of all three radiologists (all $p < 0.05$)

Kwon et al. 2024	Retrospective study	89,855 women	Age: Mean age 43.5 ± 8.7 years • Gender: Female Inclusion Criteria: Korean women aged 34 years or older who underwent initial screening digital mammography between 2009 and 2020 Exclusion Criteria: Women with follow-up durations of less than 12 months, history of breast cancer, previous breast surgery, or those who underwent simultaneous breast ultrasound or PET-CT	Standalone AI algorithm for breast cancer detection (Lunit Inc., INSIGHT MMG, version 1.1.7.2)	Radiologists' interpretation of the mammograms	AI algorithm showed higher specificity (93.0% vs. 77.6%) and PPV (1.5% vs. 0.5%) compared to radiologists.
Rodriguez-Ruiz et al. 2019 (2)	Retrospective multi-reader, multi-case (MRMC) study.	Total Exams: 2652 (653 malignant, 768 benign, 1233 normal). Radiologists: 101.	Age: Varied across datasets (e.g., 40– 80 years in dataset A, 51–86 years in dataset B). Gender: Female (implied, as the study focuses on breast cancer). Inclusion Criteria: DM exams from screening and clinical settings, with ground truth verified by histopathology or follow-up. Exclusion Criteria: Exams with implants or missing data.	Use of an AI system (Transpara 1.4.0) for breast cancer detection in DM.	101 radiologists with varying experience levels (1–44 years).	AUC: AI system: 0.840 (95% CI: 0.820–0.860). Sensitivity: AI system: Higher sensitivity than 55 of 95 radiologists (57.9%). Specificity: AI system: Comparable to radiologists at their average specificity.
Taplin et al. 2006	Experimental Study	341	Women with fatty and dense breasts; stratified random sample with oversampling of difficult cases	ImageChecker M2 1000 system (version 2.2, R2 Technology)	Interpretation with and without CAD	CAD did not improve overall sensitivity but increased specificity; better sensitivity for marked visible lesions
Jung et al. 2014	Retrospective study	100 image sets from mammographies performed between June 2008 and June 2010.	Age: Not specified in the article. Gender: Female (mammograms). Inclusion/Exclusion Criteria: Image sets included 30 masses (15 benign and 15 malignant), 30 microcalcifications (15 benign and 15 malignant), and 40 normal mammography images. Excluded were cases without biopsy results or follow-up for at least two years.	Computer-aided detection (CAD) system (Image Checker, R2, software).	The diagnostic performance of the mammograms was compared with and without the use of CAD by two groups: breast radiologists (BR) and radiology residents (RR).	CAD use improved sensitivity in both BR and RR groups. For BRs, sensitivity increased from 81.10% to 84.29%, and for RRs, it increased from 75.38% to 77.95%.

Abbreviations: AI = Artifical Intelagiance; SD = standard deviation ;CDR = cancer detection rate ;AUROC = Area Under the Receiver Operating Characteristic curve; AUC = Area Under the Curve; CI= confidence interval ;BI-RADS = Breast Imaging Reporting and Data System; AD= Architectural Distortion; Q3 = third quartile; DM = digital mammography .