rfm
Revista *de la* Facultad *de* Medicina

SYSTEMATIC REVIEW

Artificial intelligence algorithms versus conventional radiological interpretation in breast cancer screening and classification: Systematic review and meta-analysis

Algoritmos de inteligencia artificial versus interpretación radiológica convencional en tamizaje y clasificación del cáncer de mama: revisión sistemática con metaanálisis

Juan Pablo Alzate-Granados[1,2] María José Sotomayor-Ricardo[3] María Kamila Avella-Espinosa[2]

[1] Universidad Nacional de Colombia - Bogotá Campus - Faculty of Medicine - Department of Pathology - Doctoral Program in Oncology - Bogotá D.C. - Colombia.
[2] Centro de Investigaciones Oncológicas Clínica San Diego CIOSAD S.A.S. - Medical Coordination - Bogotá D.C. - Colombia.
[3] Universidad Libre - Barranquilla Campus - Faculty of Medicine - Medical Program - Barranquilla - Colombia.

Abstract

Introduction: Population-based breast cancer screening programs relying on direct interpretation of mammograms and ultrasounds by radiologists have limitations such as high workload, interobserver variability, false positives, and a shortage of specialists. Artificial intelligence (AI), through deep neural networks and other algorithms, could optimize early detection and characterization of breast lesions.
Objective: To evaluate the efficacy, in terms of operational characteristics (sensitivity, specificity, and area under the curve [AUC]), of AI algorithms in the interpretation of imaging studies (mammography, ultrasound, and magnetic resonance imaging) for (1) the detection of breast cancer in screening programs, (2) the classification of breast lesions (benign vs. malignant and BI-RADS categories), and (3) the stratification of molecular subtypes of breast cancer (e.g., triple-negative) compared to conventional interpretation by a radiologist (i.e., without AI assistance).
Materials and methods: Systematic review with meta-analysis of observational studies and clinical trials published up to May 29, 2024, in Medline, Embase, and LILACS. PROSPERO registration code: CRD42024507843. Studies comparing the sensitivity, specificity, and AUC of AI algorithms versus conventional radiological interpretation methods were included. Given the heterogeneity between studies, the meta-analysis included only the sensitivity and specificity of AI algorithms for detecting breast cancer (Dersimonian-Lard random effects model).
Results: Out of 1 156 records identified, 32 studies were included in the review (32 for qualitative analysis and 26 for meta-analysis). The AUC, sensitivity, and specificity ranges of the AI algorithms for detecting breast cancer, classifying breast lesions, and discriminating molecular subtypes of breast cancer were 0.706-0.98, 63.7-96.89%, and 45-97.9%, 0.838-0.978, 76-100%, and 78. 71-100%, and 0.535-0.978, 50-96.6%, and 52.4-95.56%, respectively. As for the meta-analysis, the pooled sensitivity and specificity of AI for breast cancer detection were 87% and 89%, respectively.
Conclusion: AI demonstrated a high diagnostic performance in breast cancer screening and classification, surpassing human performance in several studies. Therefore, AI may serve as a complementary tool to enhance the efficiency and quality of screening programs. However, prospective studies are needed to assess its clinical implementation, standardize protocols, and determine its impact on long-term health outcomes.

Resumen

Introducción. Los programas de tamizaje poblacional de cáncer de mama basados en la interpretación directa de mamografías y ecografías por radiólogos presentan limitaciones como carga de trabajo elevada, variabilidad interobservador, falsas alarmas y escasez de especialistas. La inteligencia artificial (IA), mediante redes neuronales profundas y otros algoritmos, podría optimizar la detección precoz y la caracterización de lesiones mamarias.
Objetivo. Evaluar la eficacia, en términos de características operativas (sensibilidad, especificidad y área bajo la curva [AUC]), del uso de algoritmos de IA en la interpretación de estudios de imagen (mamografía, ultrasonido y resonancia magnética) para 1) la detección de cáncer de mama en programas de cribado, 2) la clasificación de lesiones mamarias (benignas vs. malignas y clasificación BI-RADS) y 3) la diferenciación de subtipos moleculares de cáncer de mama (p. ej., triple negativo) en comparación con la interpretación convencional por radiólogos (i.e., sin asistencia de IA).
Materiales y métodos. Revisión sistemática con metaanálisis de estudios observacionales y ensayos clínicos publicados hasta el 29 de mayo de 2024 en Medline, Embase y LILACS. Código de registro en PROSPERO: CRD42024507843. Se incluyeron trabajos que compararan sensibilidad, especificidad y AUC de algoritmos de IA versus métodos radiológicos convencionales. Dada la heterogeneidad entre estudios, el metaanálisis incluyó únicamente la sensibilidad y especificidad de la IA para detectar cáncer de mama (modelo de efectos aleatorios de Dersimonian-Lard).
Resultados. De 1 156 registros identificados, 32 fueron incluidos en la revisión (32 para el análisis cualitativo y 26 para el metaanálisis). Los rangos de AUC, sensibilidad y especificidad de la IA para detectar cáncer de mama, clasificar lesiones mamarias y discriminar subtipos moleculares de cáncer fueron 0.706-0.98, 63.7-96.89% y 45-97.9%, 0.838-0.978, 76-100% y 78.71-100%, y 0.535-0.978, 50-96.6% y 52.4-95.56%, respectivamente. En cuanto al metaanálisis, la sensibilidad y especificidad combinadas de la IA para la detección de cáncer de mama fueron 87% y 89%, respectivamente.
Conclusión. La IA demostró un elevado desempeño diagnóstico en el tamizaje y la clasificación de lesiones mamarias, superponiéndose en varios estudios al rendimiento humano. Por lo tanto, la IA tiene el potencial de convertirse en una herramienta complementaria para mejorar la eficiencia y la calidad de los programas de cribado, aunque se requieren estudios prospectivos que evalúen su implementación clínica, estandaricen protocolos y determinen su impacto en resultados de salud a largo plazo.

Open access
Received: 09/05/2024
Accepted: 04/04/2025

Corresponding author: Juan Pablo Alzate-Granados. Doctorado de oncología, Departamento de Patología, Universidad Nacional de Colombia. Bogotá D.C. Colombia.
E-mail: jpalzategr@unal.edu.co.

Keywords: Breast neoplasms; Artificial Intelligence; Diagnosis, Computer-Assisted; Screening; Systematic Review; Meta-analysis (MeSH).

Palabras clave: Neoplasias de la Mama; Inteligencia Artificial; Diagnóstico por Computador; Tamizaje; Revisión Sistemática; Metaanálisis (DeCS).

How to cite: Alzate-Granados JP, Sotomayor-Ricardo MJ, Avella-Espinosa MK. Artificial intelligence algorithms versus conventional radiological interpretation in breast cancer screening and classification: Systematic review and meta-analysis. Rev. Fac. Med. 2025;73:e114297. English. doi: https://doi.org/10.15446/revfacmed.v73.114297.

Cómo citar: Alzate-Granados JP, Sotomayor-Ricardo MJ, Avella-Espinosa MK. [Algoritmos de inteligencia artificial versus interpretación radiológica convencional en tamizaje y clasificación del cáncer de mama: revisión sistemática con metaanálisis]. Rev. Fac. Med. 2025;73:e114297. English. doi: https://doi.org/10.15446/revfacmed.v73.114297.

Introduction

Breast cancer is the most common malignant neoplasm among women worldwide,[1-4] posing a serious threat to their health and lives. According to data from the Global Cancer Observatory[1] and the World Health Organization (WHO),[5] 2 296 840 new cases were diagnosed in 2022, with approximately 670 000 deaths attributable to this type of cancer. Although its incidence is relatively lower in low- and middle-income countries than in high-income countries, mortality is higher in the former,[5,6] which has been attributed, mainly in resource-limited countries in Africa and Central and South Asia, to factors such as the widespread absence of population screening strategies and programs, lack of access to diagnostic centers in rural areas, detection of the disease in advanced stages, and inability to access state-of-the-art multidisciplinary treatments.[6]

Early detection and treatment of breast cancer can significantly improve survival rates, with early detection being essential to improving prognosis.[5-7] For this reason, many developed countries have implemented large-scale screening programs based on the interpretation of mammograms.[7] Conventional breast cancer screening programs rely on radiologists' interpretation of mammograms and ultrasounds.[6,7] Such programs, although regarded as effective in reducing mortality associated with this disease, are highly labor-intensive due to the large number of women screened and the use of double reading.[8] Moreover, the interpretation of these images remains challenging as the accuracy achieved by experts in cancer detection varies widely,[7] which can lead to diagnostic delays and a high number of false positives and negatives.[7,8]

Since the 1990s, computer-aided detection (CAD) systems have been developed to automatically detect and classify breast cancer lesions in mammograms, but to date there is no evidence that traditional CAD systems directly improve screening performance or the cost-effectiveness of screening programs, mainly due to their low specificity.[8] However, progress in artificial intelligence (AI) has led to the creation of new algorithms based on deep convolutional neural networks (DCNNs), some of which have shown very promising results.[8,9]

AI, due to its ability to analyze large volumes of data and learn complex patterns, has the potential to improve diagnostic accuracy in the detection and classification of breast lesions, reducing false positive and false negative rates and facilitating earlier diagnoses.[4,8,10] In this regard, a recent study reported that the implementation of AI-assisted workflows in mammogram reading resulted in a significant improvement in radiologists' accuracy in detecting breast cancer.[11]

Notwithstanding the above, and although AI has shown potential for increasing the cancer detection rate in mammograms used in screening programs, either as stand-alone or in combination with assessment by a radiologist,[7,11,12] some studies have reported that the use of AI algorithms does not improve the diagnostic accuracy of mammograms.[10] It has also been reported that, despite their high performance when trained on large datasets, these algorithms, as stand-alone or in combination with reading by a radiologist, may not necessarily be generalizable to new populations.[13]

For example, Schaffer *et al.*[12] stated that while AI algorithms did not individually outperform radiologists in interpreting mammograms, the combined use of these algorithms together with a radiologist's evaluation improved overall accuracy in a single-reader interpretation setting (i.e., compared to interpretation by a radiologist without AI assistance). However, it has also been pointed out that AI models must be adjusted or calibrated to the characteristics of the target populations in order to achieve a better performance.[13]

Considering the foregoing, the objective of this systematic review was to evaluate the efficacy, in terms of operational characteristics (sensitivity, specificity, and area under the curve [AUC]), of AI algorithms in the interpretation of imaging studies (mammography, ultrasound, and magnetic resonance imaging) for (1) the detection of breast cancer in screening programs, (2) the classification of breast lesions (benign vs. malignant and BI-RADS categories), and (3) the stratification of molecular subtypes of breast cancer (e.g., triple-negative) compared to conventional interpretation by a radiologist (i.e., without AI assistance).

Materials and methods

Systematic review with meta-analysis conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines.[14] The review protocol was registered in PROSPERO (ID: CRD42024507843).

Inclusion criteria

In order to ensure consistency between the findings and the objective of the review in the three defined scenarios, only diagnostic accuracy studies with an appropriate design (cross-sectional diagnostic accuracy studies and diagnostic cohorts) were considered. Case-control studies were excluded, with the exception of those evaluating early stages (phase I-II) of diagnostic test development.

Participants

Screening/detection: asymptomatic women of screening age (e.g., 40-74 years).
Diagnosis/classification: women with suspicious findings or palpable masses referred for confirmation.
Classification of lesions/differentiation of molecular subtypes: patients with imaging and histopathological confirmation of benign or malignant lesions.

Intervention

Use/application of AI algorithms based on deep convolutional neural networks, support vector machines, or other machine learning models in the interpretation of mammograms (analog or digital), ultrasound, and/or MRI.

Comparator

Interpretation by radiologists without any AI assistance.

Outcomes

Sensitivity, specificity, and AUC. Data on positive and negative predictive values (PPV, NPV) were also described when available.

Search strategy

On May 29, 2024, systematic searches were conducted in Medline, Embase, and LILACS without restrictions on language, publication status (published, in press, and preprints), or publication period (start of the database - May 29, 2024). The search strategy combined controlled vocabulary (MeSH, Emtree, DeCS) and free terms for “breast cancer screening” OR “breast cancer diagnosis” AND “artificial intelligence” AND (“mammography” OR ‘ultrasound’ OR “MRI”), using Boolean and proximity operators in the title and abstract. Full details of each search are available in Annex 1. A manual search of reference lists of included studies was also performed to identify studies potentially relevant to the review objectives (snowball method).

Study selection

Once duplicates were removed, two authors (JPA and MJS) independently screened and selected studies. First, they read the titles and abstracts of the retrieved records to select studies potentially relevant to the review based on the defined inclusion criteria. Then, the full text of these studies was read to decide on their final inclusion in the review; studies published in languages other than English or Spanish were excluded due to limitations in understanding the information. Disagreements were resolved by consensus.

Data extraction

The data extraction process performed on the studies included in the review was carried out independently by two authors (JPA and MKA) using a pre-designed form for this purpose. Once this process was completed, both files were compared to identify possible inconsistencies, which, if found, were resolved by consensus, thus obtaining a final file with consolidated data.

The following data were extracted for each study: title, authors, year of publication, country where the study was conducted, objective, study design, sample size, characteristics of the participants and/or images used (mammography, ultrasound, etc.), characteristics of the AI or AI algorithms used, comparator used, results on the diagnostic performance of AI or AI algorithms for breast cancer detection (sensitivity, specificity, predictive values), and breast cancer detection rate of AI or AI algorithms.

Assessment of risk of bias

The risk of bias in the included studies was assessed independently by two authors (JPA and MKA) using the criteria described in the Cochrane Handbook for Systematic Reviews of Interventions.[15] Disagreements were resolved by consensus. In this regard, diagnostic accuracy studies were evaluated using the QUADAS-2 tool, cohort studies were evaluated using the Newcastle-Ottawa Scale (NOS), and cross-sectional studies were evaluated using a version of the NOS adapted for this type of study, a practice commonly used in systematic reviews[16,17] (evaluation of seven criteria in three main domains [selection, comparability, and outcomes] for a maximum score of eight, since one of the criteria [adjustment for confounding factors] could award up to two points). This ensured that the assessment was tailored to the specific characteristics of each design.

Meta-analysis

Although sensitivity, specificity, and AUC data were extracted for each clinical scenario, given the heterogeneity between studies, the meta-analysis included only the sensitivity and specificity of AI for detecting breast cancer (Dersimonian-Lard random effects model). Other diagnostic accuracy metrics such as PPV and NPV were also extracted and are presented descriptively in the study characterization tables but were not statistically pooled in the meta-analysis. In addition, a summary receiver operating characteristic (SROC) curve was constructed to estimate the AUC and Q index.

Finally, heterogeneity between studies was assessed using chi-square (x^2) statistics and the heterogeneity index (I^2). All analyses were performed using the STATA software (version 18.0).

Results

Studies identified and selected

The initial search yielded 1 156 records. After removing duplicates (n=20), 1 136 studies were screened by reading the title and abstract, and 1 094 of them were excluded because they did not meet the inclusion criteria. Of the 42 studies selected for full-text reading, 5 were discarded because the full document could not be accessed. Finally, of the 37 records that were read in full, 32 were included in the review (32 for qualitative analysis and 26 for quantitative analysis [meta-analysis]). The search and selection process for the studies is presented in Figure 1 (PRISMA flowchart).

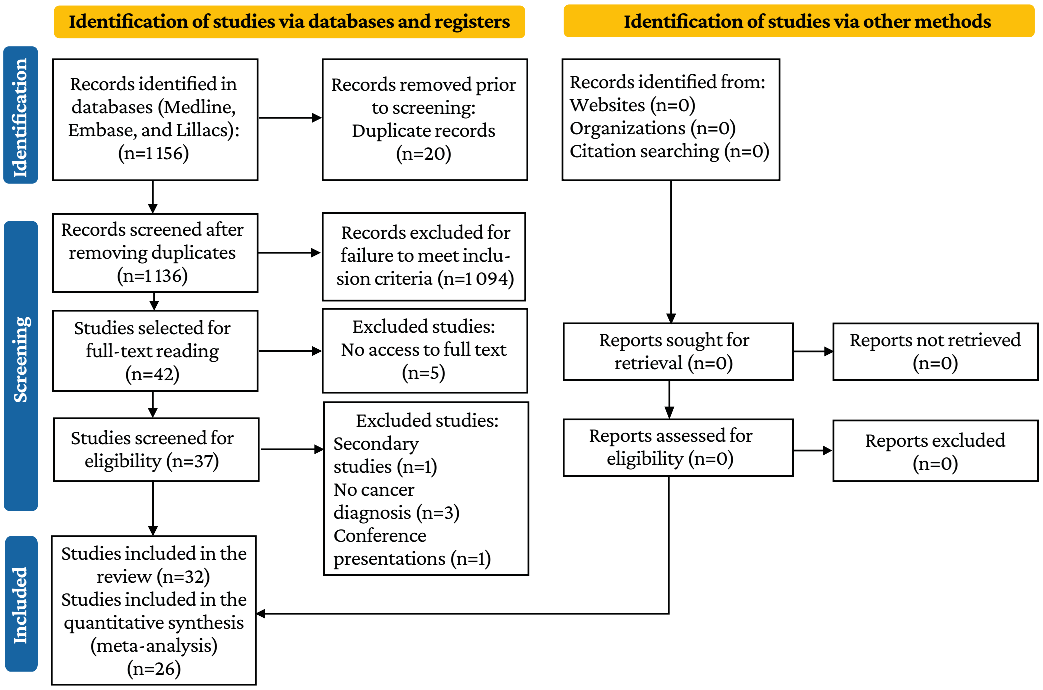
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified in databases (Medline, Embase, and Lillacs): (n=1 156)
Records removed prior to screening: Duplicate records (n=20)
Records identified from: Websites (n=0) Organizations (n=0) Citation searching (n=0)
Records screened after removing duplicates (n=1 136)
Records excluded for failure to meet inclusion criteria (n=1 094)
Studies selected for full-text reading (n=42)
Excluded studies: No access to full text (n=5)
Reports sought for retrieval (n=0)
Reports not retrieved (n=0)
Studies screened for eligibility (n=37)
Excluded studies: Secondary studies (n=1) No cancer diagnosis (n=3) Conference presentations (n=1)
Reports assessed for eligibility (n=0)
Reports excluded (n=0)
Studies included in the review (n=32) Studies included in the quantitative synthesis (meta-analysis) (n=26)

Figure 1. Flowchart for the study search, screening, and selection process.

Summary of results from the cohort studies included

Twelve cohort studies were included. They compared the diagnostic performance of AI algorithms to conventional interpretation by radiologists (i.e., without AI assistance) in breast cancer detection.[18-29] These studies were published between 2014 and 2023, and sample sizes ranged from 94 to 1 193 197 patients and/or ultrasound or mammography images, with ages ranging from 19 to 93 years. Among these 12 studies, 6 reported operational characteristics for breast cancer detection,[22-26,28] 4 for breast lesion classification,[19-21,29] and 2 for breast cancer molecular subtype differentiation.[18,27]

The AI algorithms used in these studies ranged from decision tree models to DCNNs and were evaluated using diagnostic performance metrics such as AUC, accuracy, sensitivity, and specificity. The comparators were diagnoses made by radiologists based on the interpretation of mammograms and ultrasounds. The results of these studies showed that AI algorithms can achieve high sensitivity and specificity rates, with AUC values >0.97 in some cases. Furthermore, reduced reading times and agreements ranging from moderate to perfect were observed across evaluations1863a1 performed by AI algorithms and those performed by radiologists. Overall, these studies indicate that AI has the potential to improve the accuracy of breast lesion detection and classification in general and can complement the assessment performed by radiologists (Table 1).

Table 1. Key characteristics and results of the included cohort studies comparing the performance of artificial intelligence algorithms with human diagnosis in breast cancer detection, breast lesion classification, and cancer subtype differentiation (n=12).

Author	Country	Year of publication	Study objective	Sample characteristics: images and/or participants	AI used	Comparator	Results	Other results
Ma *et al.*[18]	China	2022	To evaluate the performance of interpretable machine learning models in predicting breast cancer molecular subtypes.	600 patients (mean age: 47.59 years) with invasive breast cancer diagnosed between 2012 and 2019 with preoperative mammograms and ultrasounds. - Training set: 450 patients. - Testing set: 150 patients.	Decision tree model	Human diagnosis (4 radiologists)	Decision tree model performance for differentiating TNBC from other cancer subtypes: AUC: 0.97 Sensitivity: 0.90 Specificity: 0.94 Accuracy: 0.94	Out of the five interpretable ML models built, the decision tree model had the best performance in predicting breast cancer molecular subtypes.
Sun *et al.*[19]	China	2021	To develop and verify an AI-assisted diagnostic model for interpreting mammograms based on a CNN, evaluate its efficacy in improving radiologists' accuracy for detecting and classifying breast lesions compared to conventional interpretation (i.e., without AI assistance), and validate it in clinical practice.	- First part (model development and verification): 16 476 mammograms from 4 119 individuals (2 454 with malignant lesions and 1 665 with benign lesions): * Training set: 13 556 images from 3 389 patients (average age: 52.45 years; 2 001 with cancer) obtained between October 1, 2014, and May 31, 2016. *Validation set: 2 920 images from 730 patients (average age: 53.23 years; 453 with cancer) obtained between June 1 and September 30, 2016. - Second part (model efficacy evaluation): mammograms from 200 patients (average age: 59 years; 70 with cancer, 30 false positives, and 100 without cancer) obtained between October 1 and 31, 2015. Third part (model validation): 5 746 patients (age range: 49.99-51.50 years; 495 with malignant lesions, 337 with benign lesions, and 4,914 without lesions).	Deep learning model for automatic mammogram analysis based on three deep neural models: - Lesion detection module: use of Faster R-CNN with ResNet-50 to detect suspicious lesions. - Matching module: use of a neural network to match lesions. - Malignant degree assessment module: use of a ResNet-based CNN to estimate lesion malignancy degree.	Human diagnosis (12 radiologists)	Performance of radiologists using the deep learning model to differentiate benign from malignant lesions: AUC: 0.983, sensitivity: 94.36%, specificity: 98.07%, PPV: 87.76%, NPV: 99.09%.	Average reading time: 62.28 seconds with the model and 80.18 seconds without the model.
Gu *et al.*[20]	China	2023	To establish a breast injury risk stratification system using ultrasound images to predict breast malignancy and evaluate BI-RADS categories (2, 3, 4a, 4b, 4c, and 5) simultaneously.	Breast ultrasound images from 5 012 patients (3 220 with benign lesions and 1 792 with malignant lesions) obtained between December 2018 and December 2020: - Training set: 4 212 patients (average age: 44.05 years). - Internal test set: 416 patients (average age: 43.81 years). - External test set: 384 patients (average age: 40.06 years).	RepVGG-based deep learning models for binary classification (benign and malignant) and simultaneous BI-RADS classification of breast lesions.	Human diagnosis (3 radiologists and consensus among the 3 radiologists).	- Performance of the deep learning model in the external test set: AUC for binary categorization: 0.980 and AUC for BI-RADS categorization (6 categories): 0.945. - Performance of the deep learning model in the human interpreter study set in BI-RADS categorization (deep learning models vs. radiologist consensus): AUC: 0.901 vs. 0.933 (*p*=0.0632), sensitivity: 90.98% vs. 95.90% (*p*=0.1094), accuracy: 83.33% vs. 79.01%) (*p*=0.0541), specificity: 78.71% vs. 68.81% (*p*=0.0012).	- Binary categorization (benign vs. malignant): agreement was "almost perfect" in the external test cohort with a kappa value of 0.823 (95%CI: 0.760-0.886), and "substantial" in the internal testing cohort with a kappa value of 0.759 (95%CI: 0.693-0.824). - BI-RADS categorization: agreement was "substantial" in both the external test cohort and the internal test cohort, with kappa values of 0.669 (95%CI: 0.619-0.719) and 0.626 (95%CI: 0.575-0.674), respectively.
Schönenberger *et al.*[21]	Switzerland	2021	To investigate the potential of a DCNN to accurately classify microcalcifications in mammograms in order to create a standardized, observer-independent system based on the BI-RADS catalog.	- Test cohort: more than 56,000 images from 268 mammograms of 94 patients (age range: 35-89) classified according to the BI-RADS system. - Validation cohort: 141 images from 51 mammograms of 26 patients.	DCNN trained with preprocessed images for the classification of breast lesions according to the BI-RADS (3 models: BI-RADS 4 cohort, BI-RADS 5 cohort, and BI-RADS 4 + 5 cohort)	Human diagnosis	- Performance of the model for classifying breast lesions in the validation set: accuracy: 99.5%, 99.6%, and 98.1% for the BI-RADS 4, BI-RADS 5, and BI-RADS 4+5 cohorts, respectively.	- Performance of the model for classifying breast lesions in the test set: accuracy: 39.0%, 80.9%, and 76.6% for the BI-RADS 4, BI-RADS 5, and BI-RADS 4+5 cohorts, respectively.

Table 1. Key characteristics and results of the included cohort studies comparing the performance of artificial intelligence algorithms with human diagnosis in breast cancer detection, breast lesion classification, and cancer subtype differentiation (n=12). (Continued)

Author	Country	Year of publication	Study objective	Sample characteristics: images and/or participants	AI used	Comparator	Results	Other results
Leibig *et al.*[22]	Germany	2022	To assess the performance, in terms of sensitivity and specificity, of an AI system for breast cancer detection, both as a standalone system and as part of a decision-referral approach, compared to the original radiologist decision.	1 193 197 mammograms of 453 104 asymptomatic women (age range: 50-70 years) performed between January 1, 2007, and December 31, 2020, at 8 breast cancer screening centers. - Internal dataset: 524 143 mammograms (229 796 patients) out of 839 282 performed at 6 screening centers. Training set: 8 729 with malignant lesions and 367 088 without lesions; validation set: 1 666 with malignant lesions and 19 700 without lesions; internal test set: 1 670 with malignant lesions and 19 997 without lesions. - External dataset: 213 694 mammograms (92 585 patients) out of a total of 353 915 studies performed at 2 screening centers. External test set: 2 793 with malignant lesions and 80 058 without lesions.	Model based on a DCNN trained with mammography images using labels across different scales (patch, image, and study)	Human diagnosis	- Performance of the AI system alone: * Internal testing dataset: AUROC: 0.944, sensitivity: 84.2%, specificity: 89.5%. * External testing dataset: AUROC: 0.951, sensitivity: 84.6%, specificity: 91.3%. - Performance of the AI system as part of the decision-referral approach: * Internal testing dataset: sensitivity: 89.7%, specificity: 93.8% (a 4.0% improvement in sensitivity and a 0.5% improvement in specificity were observed compared to the radiologist's performance alone). * External testing dataset: sensitivity: 89.8%, specificity: 94.3% (a 2.6% improvement in sensitivity and a 1.0% improvement in specificity were observed compared to the performance of the radiologist alone).	The decision-referral approach leverages the strengths of both the radiologist and AI, demonstrating improvements in sensitivity and specificity that outperform those of the radiologist alone and those of the standalone AI system.
Lee *et al.*[23]	South Korea	2022	To evaluate and compare breast density categorization performed by an AI-based computer-assisted diagnosis program (AI-CAD) to radiologists alone and an automated density assessment system (Volpara®), using interobserver agreement.	488 mammograms from 488 women (average age: 56.2 years; 9 with cancer) collected between March and May 2020.	AI-CAD program (Lunit INSIGHT MMG version 1.1.4.3) to evaluate breast density categorization on a scale of 1-10.	Human diagnosis (radiologists alone) and Volpara® (automated density assessment software)	Inter-rater agreement between AI-CAD and radiologist consensus: kappa of 0.52 and agreement rate of 68.2%. - Agreement between Volpara® and radiologist consensus: kappa of 0.50 and agreement rate of 62.7%. - Concordance between AI-CAD and Volpara®: Kappa of 0.54 and coincidence rate of 61.5%.	Density assessments performed with AI-CAD showed acceptable agreement with the assessments performed by radiologists, comparable to the agreement between the commercial automated density assessment program and radiologists.
Sasaki *et al.*[24]	Japan	2020	To compare breast cancer detection performance in digital mammograms by a panel of three unaided radiologists versus a stand-alone AI system.	Digital mammograms of 310 patients (average age: 56.2 years, 69 with malignant lesions) performed between January and October 2018.	DCNN-based model: Transpara AI system (ScreenPoint Medical, 1.3.0) dedicated to interpreting mammography images.	Human diagnosis	- Transpara system performance: AUC: 0.706, sensitivity: 93% for a Transpara score ≥4 and 85% for a Transpara score ≥7, specificity: 45% for a Transpara score ≥4 and 67% for a Transpara score ≥7. - Performance of unaided radiologists: AUC: 0.816, sensitivity: 89%, specificity: 86%.	-
Elhakim *et al.*[25]	Denmark	2023	To evaluate the accuracy of breast cancer detection by a commercial AI system in a screening population (mammograms) through two simulated scenarios (standalone AI and AI-integrated screening replacing the first reader) compared to the interpretation of the first reader and double reading with arbitration.	257 671 mammograms from 153 372 women (2 041 with cancer) obtained between August 4, 2014, and August 15, 2018, from breast cancer screening centers in four cities of Southern Denmark.	DCNN-based model: Transpara AI system (ScreenPoint Medical BV, version 1.7.0)	Human diagnosis (22 certified breast radiologists)	AI system performance in different scenarios: sensitivity: 63.7-76.2%, specificity: 96.5-97.9%, PPV: 12.6-22.0%, NPV: 99.7-99.8%.	-

Table 1. Key characteristics and results of the included cohort studies comparing the performance of artificial intelligence algorithms with human diagnosis in breast cancer detection, breast lesion classification, and cancer subtype differentiation (n=12). (Continued)

Author	Country	Year of publication	Study objective	Sample characteristics: images and/or participants	AI used	Comparator	Results	Other results
Marinovich *et al.*[26]	Australia	2023	To evaluate the accuracy of an AI algorithm for reading digital mammograms compared to reading by a single radiologist and to compare AI-simulated AI-human screen-reading with human double-reading (standard practice in most breast screening programs) in terms of cancer detection and completeness.	Consecutive digital mammograms (Siemens systems) of 108,970 women (average age: 61.0 years) performed between November 1, 2015, and December 31, 2016, in the context of a population-based screening program (women in the breast cancer screening range in Australia).	DeepHealth deep AI model (Saige-Q v2.0.0) modified for Siemens image processing	Human diagnosis	- AI algorithm performance for reading digital mammograms: AUC: 0.83, sensitivity: 0.67, specificity: 0.81. - Radiologist performance for reading digital mammograms: AUC: 0.93, sensitivity: 0.68, specificity: 0.97.	The rates of completeness and cancer detection were significantly lower in the AI-radiologist reading than in the human double-reading: 3.14% vs. 3.38%. (*p*<0.001) and 6.37 vs. 6.97 per 1 000 (*p*<0.001).
Raafat *et al.*[27]	Egypt	2022	To compare the sensitivity of an AI system to that of digital mammogram interpretation by humans in detecting different types of breast cancer and to evaluate the sensitivity of the AI system in detecting different morphological characteristics of breast cancer (mass, calcifications, asymmetry, and distortion).	Digital mammograms of 123 women (average age: 53.6 years) with 134 pathologically proven malignant breast lesions between December 2020 and June 2021: 91 with invasive ductal carcinoma (IDC), 29 with invasive lobular carcinoma (ILC), 9 with ductal carcinoma in situ (DCIS), and 5 with other forms of cancer. Eleven patients had bilateral carcinomas.	Lunit INSIGHT MMG (2019 version) for Fujifilm digital mammography system. This AI system generates susceptibility of malignancy scores from <10% to 100%, where <10% is considered a low score.	Human diagnosis (two radiologists with at least 15 years of experience in breast radiology)	The sensitivity and false negative rate of the AI system for detecting different types of cancer were 96.6% and 3.4%, respectively, while the sensitivity and false negative rate for digital mammogram interpretation by radiologists were 87.3% and 12.7%, respectively. The AI system showed greater sensitivity in detecting IDC, ILC, and DCIS than manual interpretation of digital mammograms (96.7%, 96.6%, and 100% vs. 89%, 82.2%, and 88.9%, respectively); sensitivity for detecting other rare types of breast cancer was the same (80%).	AI showed greater sensitivity than digital mammogram interpretation in detecting the morphological characteristics of lesions: - Sensitivity for detecting suspicious masses: 95.2% vs. 75% - Sensitivity for detecting suspicious calcifications: 100% vs. 86.5% - Sensitivity for detecting asymmetry and distortion of the lesion: 100% vs. 84.6%.
Wang *et al.*[28]	China	2022	To develop a deep learning network incorporating an automatic segmentation network for the morphological analysis of breast tumors and determine its performance for the diagnosis of breast cancer ABUS.	ABUS images of 769 breast lesions (600 for the training set and 169 for the test set) from patients aged between 27 and 79 years old.	Deep learning networks (ResNet 34 v2, 50 v2, 101 v2) that incorporated an automatic segmentation network and extracted morphological information of breast tumors.	Human diagnosis (two radiologists with different levels of experience)	- Performance of the ResNet34 v2 model: AUC: 0.83, sensitivity: 74.00%, specificity: 76.81%, accuracy: 75.15%, PPV: 82.22, NPV: 67.08% - ResNet50 v2 model performance: AUC: 0.84, sensitivity: 81.00%, specificity: 73.91%, accuracy: 78.11%, PPV: 81.82%, NPV: 72.86% - ResNet101 v2 model performance: AUC: 0.85, sensitivity: 85.00%, specificity: 66.67%, accuracy: 77.51%, PPV: 78.70%, NPV: 71.88%.	The F1 score increased from 0.77 to 0.78, 0.81, and 0.82 with the three new deep learning network models, but their performance was inferior to that of the experienced radiologist. The three ResNet models developed performed better than the novice radiologist.
Jiang *et al.*[29]	United States	2021	To compare radiologist performance in differentiating cancerous from non-cancerous breast lesions in DCE-MRI, first incorporating a conventional decision support system with kinetic maps, and subsequently incorporating this support together with an advanced AI system for MRI interpretation.	111 DCE-MRIs from 111 patients (average age: 52 years, 54 with malignant lesions and 57 with benign lesions)	QuantX: AI system for analyzing breast images developed to assist radiologists in the evaluation and characterization of breast lesions; it includes image registration and automatic lesion segmentation.	Human diagnosis (19 radiologists specializing in breast imaging)	- Radiologist performance with the incorporation of the conventional decision support system: AUC: 0.71. - Radiologist performance with the incorporation of the conventional decision support system and the AI system: AUC: 0.76. The average sensitivity improved when BI-RADS category 3 was used as the cutoff point (from 90% to 94%) but not when category 4a was used (from 80% to 85%). The average specificity showed no difference when BI-RADS category 4a or category 3 was used as the cutoff point (52% and 52%, and 29% to 28%, respectively).	-

AI: artificial intelligence; US: ultrasound; AUC: area under curve; TNBC: triple negative breast cancer; CNN: convolutional neural network; DCNN: deep convolutional neural network; PPV: positive predictive value; NPV: negative predictive value; BI-RADS: breast imaging-reporting and data system; AI-CAD: artificial intelligence-computer-aided detection/diagnosis; DCE-MRI: dynamic contrast-enhanced magnetic resonance imaging; ABUS: automatic breast ultrasound; ABVS: automated breast volume scanner; HHUS: handheld ultrasound.

Summary of results from the cross-sectional studies included

Four cross-sectional studies (images from 7 052 participants) were included.[30-33] These studies evaluated the performance of various AI algorithms, such as convolutional neural networks and machine learning algorithms, in the classification of breast lesions (n=2) and the detection of breast cancer (n=2), compared to the performance of conventional methods (reading by radiologists). The studies were published between 2021 and 2023.

Classification of breast lesions

Shia *et al.*[30] and Shia & Cheng,[31] in studies published in 2021, reported that the AUC, sensitivity, specificity, PPV, and NPP of two AI algorithms [a model that combined pyramid histogram descriptors of oriented gradients (PHOG) with sequential minimum optimization (SMO) and a model that combined deep learning with transfer learning] for classifying malignant tumors were 0.847 and 0.938, 81.64% and 94.34%, 87.76% and 93.22%, 84.1% and 92.6%, and 85.8% and 94.8%, respectively (Table 2).

Breast cancer screening

Trang *et al.*,[32] in a study published in 2023, reported that the sensitivity, specificity, and AUC of a combined deep learning and machine learning model for breast cancer detection were 89.7%, 78.1%, and 0.88, respectively. Yoon *et al.*,[33] also reported in a study published in 2023 that the sensitivity and specificity of a DCNN-based diagnostic program for breast cancer detection were 82.1% and 90.3%, respectively (Table 2).

Table 2. Main characteristics and results of the cross-sectional studies included in the review comparing the performance of artificial intelligence algorithms to human diagnosis in breast cancer detection and breast lesion classification (n=4).

Author	Country	Year of publication	Study objective	Sample characteristics: images and/or participants	AI used	Comparator	Results	Other results
Shia *et al.*[30]	Taiwan	2021	To improve diagnostic performance associated with the classification of BI-RADS 4 malignant tumors in breast ultrasound without the need for tumor region preselection	Breast ultrasound images from 677 patients (age range: 35–75 years, 365 with malignant lesions and 370 with benign lesions) obtained between January 1, 2017, and December 31, 2018	Pyramidal histogram descriptor of oriented gradients (PHOG) and sequential minimum optimization (SMO)	Human diagnosis	Performance for classifying benign and malignant lesions: - Unsupervised ML: AUC: 0.847, sensitivity: 81.64%, specificity: 87.76%, PPV: 84.1%, NPV: 85.8%. - Radiologists: AUC: 0.574, sensitivity: 95.28%, specificity: 19.50%, PPV: 48.2%, NPV: 84.0%	-
Shia & Cheng[31]	Taiwan	2021	To evaluate diagnostic accuracy of breast tumor classification in bidirectional ultrasound images using a transfer learning method	2 099 breast ultrasound images from 543 patients (age range: 35–75 years, 241 with malignant lesions and 302 with benign lesions) obtained between January 1, 2017, and December 31, 2018	AI model that combines a pre-trained deep residual network model (ResNet-101) to extract high-level features and a linear SVM model to classify the extracted features	Human diagnosis	Performance of the AI model for classifying malignant lesions: AUC: 0.938, sensitivity: 94.34%, specificity: 93.22%, PPV: 92.6%, NPV: 94.8%	-

Table 2. Main characteristics and results of the cross-sectional studies included in the review comparing the performance of artificial intelligence algorithms to human diagnosis in breast cancer detection and breast lesion classification (n=4). (Continued)

Author	Country	Year of publication	Study objective	Sample characteristics: images and/or participants	AI used	Comparator	Results	Other results
Trang et al.[32]	Vietnam	2023	To evaluate the diagnostic performance of CNNs and ML methods in the detection of breast cancer in mammograms	731 mammograms performed between July and September 2017 on 357 women (average age: 48 years, 136 with malignant lesions) with at least one mammogram and clinical records for at least 6 months prior to the mammogram	Combined deep learning (X-ception, VGG16, ResNet-v2, ResNet50, CNN3) and ML model for breast cancer detection (k-nearest neighbor, SVM, random forest, artificial neural network, and gradient boosting machine)	Human diagnosis	Performance of the combined model for breast cancer detection: AUC: 0.88, sensitivity: 89.7%, specificity: 78.1%, accuracy: 84.5%	-
Yoon et al.[33]	South Korea	2023	To evaluate the diagnostic performance of an AI-based computer-aided diagnosis (AI-CAD) system for the detection of breast cancer in screening mammograms	6 499 mammograms (6 282 negative, 189 benign, and 28 malignant) from 5,228 women obtained between January 2016 and December 2017 at a screening center	Diagnostic support software (Lunit INSIGHT for Mammography, version 1.1.0.1) based on a DCNN trained with over 170 000 mammograms for breast cancer detection	Human diagnosis	Diagnostic performance for cancer detection: - AI-CAD system: sensitivity: 82.1%, specificity: 90.3%. - Radiologists: sensitivity: 67.9%, specificity: 96.9. The AI-CAD system had significantly lower specificity than conventional interpretation by radiologists; differences in sensitivity were not statistically significant	Cancer detection rate and recall rate: - IA-CAD system: 3.5x1 000 and 10.0%. - Radiologists: 2.9x1 000 and 3.4%.

AI: artificial intelligence; AUC: area under curve; PPV: positive predictive value; NPV: negative predictive value; ML: machine learning: AI-CAD: artificial intelligence-computer-aided detection/diagnosis; CNN: convolutional neural network; DCNN: deep convolutional neural network; SVM: support vector machine; VGG: visual geometry group.

Summary of the results of the diagnostic testing studies included

Sixteen diagnostic testing studies published between 2019 and 2023 were reviewed.[34-39] The study populations included participants in screening programs, high-risk cohorts, and patients who underwent biopsy or surgery. The studies evaluated the capacity of various AI algorithms to detect breast cancer, classify breast lesions, and differentiate molecular subtypes of cancer in mammograms, ultrasounds, and MRIs. It should be noted that the study conducted by Suh *et al.*[41] evaluated the capacity of a deep learning model to detect breast cancer and classify breast lesions, while the study conducted by Ye *et al.*[45] evaluated the performance of a DCNN for breast lesion classification and molecular cancer subtype discrimination.

Breast cancer detection

The capacity of AI (stand-alone or as an aid to radiologists) to detect breast cancer in screening programs or risk cohorts was evaluated in 5 studies,[35,39-41,43] in which the sensitivity of the models evaluated ranged from 80.9% to 100% and the specificity from 67.7% to 96.1%. For example, the sensitivity of the Transpara model ranged from 80.9%[35] to 86%[43] (Table 3).

Classification of breast lesions (benign vs. malignant and BI-RADS)

The capacity of AI (stand-alone or as an aid to radiologists) to discriminate between benign and malignant lesions or to categorize lesions based on the BI-RADS classification was analyzed in 10 studies,[34,36-38,41,44,45,47-49] in which the sensitivity of the models evaluated ranged from 65.9%[37] to 100%[38] and the specificity from 23.7%[37] to 100%.[48] For example, an AI system based on a ResNet-50 model and its modifications obtained sensitivities and specificities ranging from 91.67% to 99.24% and from 87.36% to 96.70%, respectively, for a CBIS-DDSM (curated breast imaging subset of digital database for screening mammography) dataset and between 97.14% and 100% respectively for an INbreast[38] dataset (Table 3).

Differentiation of molecular subtypes of breast cancer

The efficacy of AI (stand-alone or as an aid to radiologists) for distinguishing cancer subtypes, such as triple negative (TN) and human epidermal growth factor receptor 2 positive (HER2+), was evaluated in three studies.[42,45,46] The AUC, sensitivity, and specificity ranged from 0.535[46] to 0.9789,[45] 50.0%[46] to 92.5%,[45] and 52.4%[46] to 95.56%,[46] respectively. For example, a combined CNN-SVM classifier automatically segmented tumors on MRI with a sensitivity of 0.92, a PPV of 0.94, and a Dice similarity coefficient of 0.93,[42] while a model that included the logistic regression classifier achieved an AUC of 0.824 for TN, with a sensitivity of 81.8%, and a specificity of 74.2%, and 0.778 for HER2+, with a sensitivity of 71.4% and a specificity of 71.6% (Table 3).

Table 3. Key features and results of the included diagnostic test studies comparing the performance of artificial intelligence algorithms to human diagnosis in the detection and/or classification of breast cancer, breast lesion classification, and cancer subtype differentiation (n=16).

Author	Country	Year of publication	Study objective	Sample characteristics: images and/or participants	AI used	Comparator	Results	Other results
O'Connell *et al.*[36]	United States and Italy	2021	To evaluate the performance of the S-Detect AI program for the classification of suspicious breast lesions in ultrasound images compared to a group of radiologists	299 breast ultrasound images from 299 women (average age: 52.3 years) with at least one suspicious lesion and for whom a final diagnosis was independently determined, obtained between 2018 and 2019 at the University of Rochester (n=150 [95 benign and 55 malignant]) and the University Hospital of Palermo (n=149 [54 benign and 95 malignant])	The S-Detec for Breast program based on a CNN trained with more than 10 000 breast scans compared to gold standard biopsy assessments to classify breast lesions as possibly benign and possibly malignant	Human diagnosis (5 radiologists with less than 5 years of experience and 5 radiologists with more than 10 years of experience)	- Performance of the S-Detect program: sensitivity: 0.810, specificity: 0.827, accuracy: 0.818. - Performance of radiologists: sensitivity: 0.703, specificity: 0.755, accuracy: 0.731	-
Busaleh *et al.*[38]	Saudi Arabia	2021	To evaluate the efficacy of a computer-aided system based on a ResNet-50 model (and density-specific modifications) and fusion techniques in the classification of breast masses (benign and malignant)	CBIS-DDSM dataset: digitized film images of 753 calcifications and 891 masses converted to DICOM format with updated annotations of mass regions on mediolateral oblique and bilateral craniocaudal views. INbreast dataset: 410 full-field digital mammographic images in DICOM format with mediolateral oblique and bilateral craniocaudal views	AI system based on the ResNet-50 model and density-specific modifications (DIResNet-50, DIIResNet-50, DIIIResNet-50, DIVResNet-50), as well as fusion techniques (backbone model and SVM with the polynomial kernel as a fusion method).	Human diagnosis	- AI system performance with the CBIS-DDSM dataset: DIRresNet-50: sensitivity: 98.48%, specificity: 92.31%; DIIRresNet-50: sensitivity: 96.97%, specificity: 92.31%; DIIRresNet-50: sensitivity: 97.73%, specificity: 91.21%; DIVRresNet-50: sensitivity: 91.67%, specificity: 96.70%; RresNet-50: sensitivity: 99.24%, specificity: 87.36%. - AI system performance with the INbreast dataset: DIRresNet-50: sensitivity: 100%, specificity: 97.5%; DIIRresNet-50: sensitivity: 100%, specificity: 100%; DIIIRresNet-50: sensitivity: 97.14%, specificity: 97.5%; DIVRresNet-50: sensitivity: 100%, specificity: 97.14%; RresNet-50: sensitivity: 98.57%, specificity: 97.5%.	-
Cai *et al.*[39]	United Kingdom	2021	To evaluate the performance of an optimized CNN for breast cancers diagnosed in mammogram images	Breast mammogram images from the Image xAnalysis Society Digital Mammogram Database (322 digital mammography images measuring 1 024×1 024 taken from the UK National Breast Screening Program)	Advanced Thermal Exchange Optimizer AI algorithm, an optimized version of CNN and a new improved metaheuristic	Human diagnosis	AI algorithm performance: sensitivity: 96.89%, specificity: 67.7%, accuracy: 93.79%.	-
Shen *et al.*[40]	United States	2019	To evaluate the efficacy of different deep learning algorithms for detecting breast cancer in screening mammograms	CBIS-DDSM dataset: 2 478 mammography images from 1 249 women. INbreast dataset: 410 mammography images from 115 patients	Deep learning algorithms, including ResNet50 and VGG16	Human diagnosis	- Performance of the ensemble model (based on the average of the 4 best models) on the CBIS-DDSM dataset: AUC: 0.91, sensitivity: 86.1%, specificity: 80.1% - Ensemble model performance (based on the average of the 4 best models) on the INbreast dataset: AUC: 0.98, sensitivity: 86.7%, specificity: 96.1%	-

Table 3. Key features and results of the included diagnostic test studies comparing the performance of artificial intelligence algorithms to human diagnosis in the detection and/or classification of breast cancer, breast lesion classification, and cancer subtype differentiation (n=16). (Continued)

Author	Country	Year of publication	Study objective	Sample characteristics: images and/or participants	AI used	Comparator	Results	Other results
Suh *et al.*[41]	South Korea	2020	To evaluate the performance of two CNNs (DenseNet-169 and EfficientNet-B5) in the detection and classification of breast cancer in digital mammography images of patients with different grades of breast density	3 002 fused digital mammograms (537 malignant and 2 465 non-malignant images) from 1 501 patients (average age: 48.9 years, 532 with malignant lesions) performed between February 2007 and May 2015	Deep learning model developed from two CNNs: DenseNet-169 and EfficientNet-B5 for breast cancer detection in mammograms based on breast density	Human diagnosis	Overall performance for breast cancer detection: - DenseNet-169: AUC: 0.952, sensitivity: 87.0%, specificity: 88.4%, accuracy: 88.1%, PPV: 62.1%, and NPV: 96.9%. - EfficientNet-B5: AUC: 0.954, sensitivity: 88.3%, specificity: 87.9%, accuracy: 87.9%, PPV: 62.1%, and NPV: 97.2%.	Model performance in detecting malignancy decreased as breast density increased: - Density A: DenseNet-169: AUC: 0.984, sensitivity: 100%, specificity: 92.9%, accuracy: 95.0%, PPV: 85.7%, and NPV: 100%; EfficientNet-B5: AUC: 0.988, sensitivity: 100%, specificity: 95.3%, accuracy: 96.7%, PPV: 90.5%, and NPV: 100%. - Density B: DenseNet-169: AUC: 0.962, sensitivity: 97.0%, specificity: 96.1%, accuracy: 96.2%, PPV: 85.3%, and NPV: 99.3%; EfficientNet-B5: AUC: 0.990, sensitivity: 97.0%, specificity: 94.8%, accuracy: 95.2%. PPV: 81.0% and NPV: 99.3%. - Density C: DenseNet-169: AUC: 0.950, sensitivity: 87.7%, specificity: 86.2%, accuracy: 86.4%, PPV: 58.8%, and NPV: 97.0%; EfficientNet-B5: AUC: 0.940, sensitivity: 84.0%, specificity: 81.5%, accuracy: 81.9%, PPV: 49.6%, and NPV: 96.0%. - Density D: DenseNet-169: AUC: 0.902, sensitivity: 83.3%, specificity: 84.6%, accuracy: 84.3%, PPV: 51.0%, and NPV: 96.5%; EfficientNet-B5: AUC: 0.925, sensitivity: 86.7%, specificity: 85.8%, accuracy: 85.9%, PPV: 58.4%, and NPV: 97.1%.
Guo *et al.*[42]	China	2022	To develop and evaluate a fully automated segmentation method based on a CNN and an SVM to discriminate TN breast cancer	Breast MRI data from 272 patients (165 cases of other molecular types of breast cancer and 107 cases of TN breast cancer)	AI algorithm for breast tumor segmentation based on a CNN and an SVM	Human diagnosis	Performance of the AI algorithm for TN breast tumor segmentation: sensitivity: 0.92, PPV: 0.95.	Dice Similarity Coefficient: 0.93
Rodríguez-Ruiz *et al.*[43]	-	2019	To evaluate breast cancer detection performance of Transpara AI system and screening mammogram interpretation by radiologists versus conventional interpretation (i.e., without AI assistance).	Screening digital mammographic examinations from 240 women (average age: 61 years, 100 with cancer, 40 with false positive results, and 100 with normal results) performed between 2013 and 2017	Transpara AI System (version 1.3.0). This system uses a DCNN and feature classifiers, and image analysis algorithms for automated breast cancer detection in mammograms and breast tomosynthesis	Human diagnosis (14 radiologists)	- Performance of radiologists with AI assistance: AUC: 0.89, sensitivity: 86%, specificity: 79% - Performance of radiologists without AI assistance: AUC: 0.87, sensitivity: 83%, specificity: 77%.	Reading time per case: 146 seconds without assistance and 149 seconds with AI support
Pang *et al.*[44]	-	2021	To evaluate the efficacy of a radiomics model based on a semi-supervised GAN and a CNN in the classification of benign and malignant breast masses in ultrasounds	1 447 breast ultrasound images (767 of benign masses and 680 of malignant masses) from 357 patients (average age: 51.6 years) obtained between January 2018 and January 2019	Semi-supervised GAN model (TripleGAN) and a CNN to synthesize breast masses and subsequent classification, respectively	Human diagnosis (2 experienced radiologists)	Radiomics model performance: sensitivity: 87.94%, specificity: 85.86%, accuracy: 90.41%.	-
Ye *et al.*[45]	China	2021	To develop and evaluate the accuracy of a DCNN (ResNet50) in classifying benign and malignant lesions in breast ultrasound images	1 844 ultrasound images (910 with benign lesions and 934 with malignant lesions [110 TN and 824 NTN]) from 1 446 patients obtained between February 2018 and March 2019: - Training cohort: 1 618 images from 1 261 patients (benign cohort: 820 images from 598 women with an average age of 39.65 years and malignant cohort: 798 images from 3 men with an average age of 48.30 years and 660 women with an average age of 52.80 years). - Test cohort: 226 images from 185 patients (90 benign images from 79 women with an average age of 40.54 years and 136 malignant images from 106 women with an average age of 51.02 years).	Resnet50, a fine-tuned DCNN for the diagnosis of breast masses	Human diagnosis (3 radiologists)	- Performance in discriminating between benign lesions and TN breast cancer: AUC: 0.9789, sensitivity: 92.50%, specificity: 95.56%. - Performance in discriminating between benign lesions and NTN breast cancer: AUC: 0.9689, sensitivity: 90.62%, specificity: 95.56%	Performance in discriminating between TN and NTN breast cancer: AUC: 0.9000, sensitivity: 87.50%, specificity: 90.00%

Table 3. Key features and results of the included diagnostic test studies comparing the performance of artificial intelligence algorithms to human diagnosis in the detection and/or classification of breast cancer, breast lesion classification, and cancer subtype differentiation (n=16). (Continued)

Author	Country	Year of publication	Study objective	Sample characteristics: images and/or participants	AI used	Comparator	Results	Other results
Ferre *et al.*[46]	Canada	2023	To evaluate the diagnostic performance of machine learning (ML) classification in differentiating breast cancer subtypes (TN vs. NTN and HER2+ vs. HER2-) using radiomic features extracted from grayscale ultrasound b-mode images	88 women who underwent diagnostic breast ultrasounds between June 1, 2011, and July 31, 2019, with confirmation of invasive malignancy in pathology and receptor status determined by immunohistochemistry available: - TN breast cancer cohort: 22 patients (average age: 52 years). - NTN breast cancer cohort: 66 patients (average age: 59 years). - HER2+ cohort: 21 patients (average age: 51 years). - HER2- cohort: 45 patients (average age: 62 years).	Supervised ML classifiers: logistic regression, k-nearest neighbor, and Naïve Bayes	Human diagnosis (1 board-certified breast radiologist and 1 breast imaging fellow)	- Performance of the logistic regression classifier: * TN vs. NTN: AUC: 0.824, sensitivity: 81.8%, specificity: 74.2%. * HER2+ vs HER2-: AUC: 0.778, sensitivity: 71.4%, specificity: 71.6%. - Performance of the k-nearest neighbor classifier: * TN vs. NTN: AUC: 0.739, sensitivity: 85.7%, specificity: 65.0%. * HER2+ vs HER2-: AUC: 0.679, sensitivity: 83.3%, specificity: 52.4%. - Performance of the Naïve Bayes classifier * TN vs. NTN: AUC: 0.807, sensitivity: 71.4%, specificity: 90.0%. * HER2+ vs HER2-: AUC: 0.535, sensitivity: 50.0%, specificity: 57.1%.	-
Feng *et al.*[47]	China	2020	To evaluate the accuracy of a knowledge-driven feature learning and integration (KFLI) framework for classifying benign and malignant breast lesions based on MRI	MRI images of 100 female patients at high risk of breast cancer (68 with malignant lesions)	KFLI, an ensemble of deep networks and domain knowledge that enables the extraction of features from multi-sequence MRI images	Human diagnosis (2 experienced breast radiologists)	Performance of KFLI in classifying benign and malignant breast lesions: sensitivity: 84.6%, specificity: 85.7%, accuracy: 85.0%.	-
Ciritsis *et al.*[48]	Switzerland	2019	To evaluate the accuracy of a DCNN in classifying breast lesions in ultrasound images based on the ACR BI-RADS catalog and compare its performance to human readers	1 019 breast ultrasound images from 582 patients (average age: 56.3 years) obtained between 2012 and 2016: - Internal test dataset: 101 images. - External test dataset: 43 images.	DCNN with sliding window approach for fully automatic classification of breast lesions in ultrasound	Human diagnosis (2 radiologists with more than 5 years of experience)	- DCNN performance in the internal test dataset: AUC: 83.8, sensitivity: 76.00%, specificity: 92.11%. - DCNN performance in the external test dataset: AUC: 96.7, sensitivity: 89.47%, specificity: 100%.	-
Koch *et al.*[35]	Norway	2023	To evaluate the diagnostic performance of the Transpara AI system in detecting breast cancer in women with high breast density (VDG4) and compare it with an independent double reading by radiologists	14 900 digital mammograms (1 254 cases of breast cancer, 12 642 negative controls, and 1 004 previous examinations of patients diagnosed with breast cancer) of women (average age: 58 years) screened through the BreastScreen Norway program between 2010 and 2018	Transpara AI system (version 1.7.0) that uses CNN to analyze mammograms and has been trained with mammograms from different screening programs and providers	Human diagnosis (13 breast radiologists with an average annual reading volume of around 7 000 readings and between 1 and 22 years of experience in interpreting screening exams)	Performance in breast cancer detection in women with the highest breast density (VDG4): - Transpara: sensitivity: 80.9%. - Independent double reading by radiologists: sensitivity: 62.8%.	-
Li *et al.*[49]	China	2023	To evaluate the efficacy of CNN models combined with radiomics and clinical data for breast lesion classification (sclerosing adenosis vs. breast cancer)	197 patients (average age: 61.25 years, 100 with breast cancer and 97 with sclerosing adenosis)	CNN models (VGG16, Resnet18, Resnet50, and Densenet121) combined with radiomics and clinical data	Human diagnosis	Performance of the different CNNs: - VGG16+radiomics+clinical data: AUC: 0.858, sensitivity: 84.26%, specificity: 83.17%, accuracy: 83.65, PPV: 84.03%, NPV: 83.25%. - Renet18+radiomics+clinical data: AUC: 0.864, sensitivity: 84.97%, specificity: 84.78%, accuracy: 84.82%, PPV: 84.79%, NPV: 84.65%. - Resnet50+radiomics+clinical data: AUC: 0.873, sensitivity: 86.99, specificity: 85.40%, accuracy: 85.88%, PPV: 85.42%, NPV: 86.43%. - Densenet121+radiomics+clinical data: AUC: 0.915, sensitivity: 87.60%, specificity: 86.20%, accuracy: 86.80%, PPV: 87.42%, NPV: 86.01 **(model with the best performance).**	Performance of radiologists without AI assistance: AUC: 0.716, sensitivity: 100%, specificity: 43.30%, accuracy: 72.08%, PPV: 64.52%, NPV: 100%.
Gu *et al.*[34]	China	2022	To develop a deep learning model and evaluate its diagnostic performance in the classification of benign and malignant breast tumors in ultrasound images	14 043 ultrasound images of 5 012 women with breast lesions taken between December 2018 and December 2020: - Training set: 4 149 patients (average age: 43.67 years). - Internal test set: 466 patients (average age: 43.21 years). - External test set: 397 patients (average age: 44.91 years).	VGG-19 DCNN.	Human diagnosis (5 radiologists)	- Model performance in the internal test set: AUC: 0.908, sensitivity: 83.23%, specificity: 83.61, accuracy: 83.48%, PPV: 72.83%, NPV: 90.43%. - Model performance in the external test set: AUC: 0.913, sensitivity: 88.84%, specificity: 83.77%, accuracy: 86.40%, PPV: 85.51%, NPV: 87.43%.	-

Table 3. Key features and results of the included diagnostic test studies comparing the performance of artificial intelligence algorithms to human diagnosis in the detection and/or classification of breast cancer, breast lesion classification, and cancer subtype differentiation (n=16). (Continued)

Author	Country	Year of publication	Study objective	Sample characteristics: images and/or participants	AI used	Comparator	Results	Other results
Park *et al.*[37]	South Korea	2019	To evaluate the added value of computer-aided diagnosis (CAD system S-Detect) in improving the diagnostic accuracy of radiologists with different levels of experience in classifying breast masses in ultrasound images	100 breast masses (41 malignant and 59 benign) in 91 women (average age: 46.5 years) scheduled for breast ultrasound or ultrasound-guided biopsy between May and June 2015	S-Detect (Samsung Medison) is a CAD program that provides analyses based on morphologic features, using a novel feature extraction technique and an SVM classifier	Human diagnosis (3 less experienced radiologists [first-year fellows in breast imaging] and 2 experienced radiologists [8 and 10 years of experience in breast imaging])	- Performance of the 5 radiologists without the assistance of S-Detect: AUC: 0.623-0.889, sensitivity: 97.6-90.2%, specificity: 27.1-54.2%, accuracy: 43.0-70.0%, PPV: 38.6-58.5%, NPV: 53.3-91.4%. - Performance of the five radiologists with the assistance of the S-Detect program: AUC: 0.823-0.907, sensitivity: 65.9-92.7%, specificity: 23.7-66.1%, accuracy: 54.0-76.0%, PPV: 47.1-64.9%, NPV: 86.7-94.7%.	Compared with experienced radiologists, a significant improvement in NPV (86.7–94.7% vs. 53.3–76.2%) and AUC (0.823–0.839 vs. 0.623–0.759) was observed in the less experienced radiologists when using the CAD program. In contrast, compared to less experienced radiologists, specificity and PPV improved significantly in experienced radiologists with the assistance of S-Detect (55.6% and 58.5% vs. 64.9% and 64.9%).

AI: artificial intelligence; CNN: convolutional neural network; DCNN: deep convolutional neural network; AUC: area under curve; PPV: positive predictive value; NPV: negative predictive value; MRI: magnetic resonance imaging; TN: triple negative; NTN: not triple negative. VGG: visual geometry group; SVM: support vector machine; CBIS-DDSM: curated breast imaging subset of digital database for screening mammography; CAD: computer-aided detection/diagnosis; GAN: generative adversarial network; HER2: human epidermal growth factor receptor 2; HUSS: handheld ultrasound. ROC: receiver operating characteristic.

Assessment of risk of bias in cross-sectional studies

Sample representativeness: all the studies had representative samples, so each one received one point for this criterion.[30-33]
Sample size: all four studies had an adequate number of participants for the analyses performed, so they each obtained one point for this criterion.[30-33]
Excluded subjects: none of the studies provided sufficient information to obtain a score regarding the description of excluded subjects.
Adjustment for confounding factors: three studies adjusted diagnostic accuracy measures adequately for confounding factors,[30,31,33] thus receiving one point each, while the fourth study performed a more comprehensive adjustment,[31] covering a wider range of potential confounding variables and obtaining two points for this criterion.
Adjustment for other factors: only one of the studies obtained a point for adjusting for other factors that could affect the result.[31]
Evaluation of results: all studies adequately and reliably described how they measured outcomes, so each of them received one point.[30-33]
Statistical test: three studies used statistical tests appropriate for analysis,[30,32,33] scoring one point each, while one did not include these tests[31] and was therefore not awarded a score.
Total score: out of a maximum score of eight, the studies by Shia *et al.*[30] Trnag *et al.*[32] and Yoon *et al.*[33] obtained five points, while the study by Shia *et al.*[31] obtained six points. This suggests that, although all studies have a relatively low risk of bias, there are variations in methodological quality and in the degree of adjustment for potentially confounding factors.

Table 4. Assessment of risk of bias in cross-sectional studies using the Newcastle-Ottawa scale.

Criterion	Shia ***et al.***[30] 2021	Trang ***et al.***[32] 2023	Shia & Chen[31], 2021	Yoo ***et al.***[33] 2023
Sample representativeness	1	1	1	1
Sample size	1	1	1	1
Excluded subjects	0	0	0	0
Adjustment for confounding factors	1	1	2	1
Adjustment for other factors	0	0	1	0
Outcome assessment	1	1	1	1
Statistical test	1	1	0	1
Total	5	5	6	5

Assessment of risk of bias in diagnostic test studies

Patient selection: five studies had a high risk of bias,[32,34,35,40,46] which suggests that patient selection in these studies may have introduced bias into the results; five studies had an unclear risk of bias,[38,39,41,44,47] which points to a lack of clarity or insufficient information about patient selection; and six studies showed a low risk of bias,[35,40,43,45,46,48] reflecting an appropriate and well-defined selection of patients.

Index test: six studies had a high risk of bias in the application or interpretation of the diagnostic test,[34,38-40,44,48] revealing potential problems in the performance or interpretation of the index test that could affect the validity of the results; three studies had an unclear risk of bias due to a lack of details or unclear methodology regarding the index test;[41,42,46] and six studies showed a low risk of bias.[35-37,43,45,47]

Reference standard: one study presented a high risk of bias,[34] which could call into question the accuracy of the reference standard used; eight studies had an unclear risk of bias,[36,38-40,43,44,46,48] demonstrating that the information provided on the reference standard was insufficient or unclear; and seven studies obtained a low risk of bias assessment,[35,37,41,42,45,47,49] which indicates that the reference standard used was appropriate and correctly applied.

Flow and timing: five studies had an unclear risk of bias[38-40,44,48] and 11 studies had a low risk of bias.[34-37,41-43,45-47,49] These results demonstrate that there were no significant problems in patient flow across the study or in the interval between the index test and the reference standard in most studies.

Table 5. Risk of bias in diagnostic test studies according to the QUADAS-2 tool.

Item	Gu *et al.*[34] 2022	Koch *et al.*[35] 2023	O'Connell *et al.*[36] 2021	Park *et al.*[37] 2019	Busaleh *et al.*[38] 2021	Cai *et al.*[39] 2021	Shen *et al.*[40] 2019	Suh *et al.*[41] 2020	Guo *et al.*[42] 2022	Rodríguez-Ruiz *et al.*[43] 2019	Pang *et al.*[44] 2021	Ye *et al.*[45] 2021	Ferre *et al.*[46] 2023	Feng *et al.*[47] 2020	Ciritsi *et al.*[48] 2019	Li *et al.*[49] 2023
Patient selection	High	Low	High	High	Unclear	Unclear	Low	Unclear	High	Low	Unclear	Low	Low	Unclear	Low	High
Index test	High	Low	Low	Low	High	High	High	Unclear	Unclear	Low	High	Low	Unclear	Low	High	Low
Reference standard	High	Low	Unclear	Low	Unclear	Unclear	Unclear	Low	Low	Unclear	Unclear	Low	Unclear	Low	Unclear	Low
Flow and timing	Low	Low	Low	Low	Unclear	Unclear	Unclear	Low	Low	Low	Unclear	Low	Low	Low	Unclear	Low

Risk of bias assessment in cohort studies

Selection bias: 8 studies attained a score ≥2 (5 studies had 2 points;[18,24,25,27,29] 1 study had 3 points;[19] and 2 studies had 4 points),[22,26] suggesting an adequate cohort selection, as well as a clear definition of exposure and outcome indicators. In contrast, 4 studies obtained a score of 1,[23,28,20,21] indicating that aspects related to subject selection or exposure definition could be improved.

Comparability bias: 5 studies met both comparability criteria and obtained the maximum score of 2,[18,19,22,25,28] which implies that these studies were adequately adjusted for the main confounding factors. Concerning the remaining studies, 6 obtained a single point[20,21,24,26,27,29] and 1 did not obtain any points,[23] suggesting a potential risk of bias due to the lack of adjustment for confounding factors.

Outcome bias: 9 studies achieved a score ≥2 (5 studies scored 2 points,[20,21,25,27,28] and 4 studies scored 3 points),[19,22,24,26] indicating that the outcome was adequately defined and that the follow-up period was long enough to observe the events of interest. Two studies

obtained a score of 1[18,29] and another obtained a score of 0,[23] suggesting that improvements could be made in outcome assessment or subject follow-up.
Total score: 1 study obtained the maximum total score of 9,[22] indicating a low risk of bias in all areas assessed. Two studies obtained a total score of 8,[19,26] which also suggests a low risk of bias. Five studies achieved a total score of 5 or 6,[18,24,25,27,28] reflecting a moderately low risk of bias. Two studies had total scores of 4,[20,29] suggesting a moderate risk of bias. Finally, only one study had the lowest score, with a total of 1,[23] demonstrating a high risk of bias in the domains evaluated.

Table 6. Risk of bias in cohort studies.

Domain	Ma *et al.*[18] 2022	Sun *et al.*[19] 2021	Gu *et al.*[20] 2023	Schönenberger *et al.*[21] 2021	Leibig *et al.*[22] 2022	Lee *et al.*[23] 2022	Sasaki *et al.*[24] 2020	Elhakim *et al.*[25] 2023	Marinovich *et al.*[26] 2023	Raafat *et al.*[27] 2022	Wang *et al.*[28] 2022	Jiang *et al.*[29] 2021
Selection	2	3	1	1	4	1	2	2	4	2	1	2
Comparability	2	2	1	1	2	0	1	2	1	1	2	1
Result	1	3	2	2	3	0	3	2	3	2	2	1
Total	5	8	4	4	9	1	6	6	8	5	5	4

Source: Own elaboration.

Sensitivity and specificity results for AI in breast cancer detection

The meta-analysis included only 26 studies focusing on the capacity of AI to detect breast cancer in imaging tests (mammography, ultrasound, and MRI) versus human diagnosis (confirmed histopathologically or in patients followed up for 12 months after diagnosis) through image interpretation by a radiologist without AI assistance. The pooled sensitivity and specificity were 87% (95%CI: 83-90) and 89% (95%CI: 84-92), respectively.

Individual sensitivity varied considerably between studies, with a minimum of 40% and a maximum of 100%, both observed in Schönenberger *et al.*[21] In turn, specificity was generally above 70%, although the studies by Jiang *et al.*,[29] Park *et al.*,[37] and Wang *et al.*[46] found values between 53% and 67%.

Some studies showed wide confidence intervals (CI), which indicates small sample sizes or methodological variability. For example, the studies by Schönenberger *et al.*[21] (2021), Ferre *et al.*[46] (2023 and 2023a), and Yoon *et al.*[33] repoted sensitivity 95%CI of 0.27-0.54, 0.60-0.95, 0.48-0.89, and 0.63-0.94, respectively, while the studies by Ferre *et al.*[46] (2023 and 2023^a), Wang *et al.*,[28] Jiang *et al.*,[29] and Park *et al.*[37] reported specificity 95%CI of 0.62-0.84, 0.56-0.84, 0.54-0.78, 0.39-0.66, and 0.53-0.78, respectively. Although Schönenberger *et al.*21 report BI-RADS classification data (accuracy in BI-RADS 4 and 5), this study was only included once in the detection meta-analysis despite being listed three times in Figure 2. The diamond summary of this figure illustrates the solid balance achieved between sensitivity and specificity by AI algorithms.

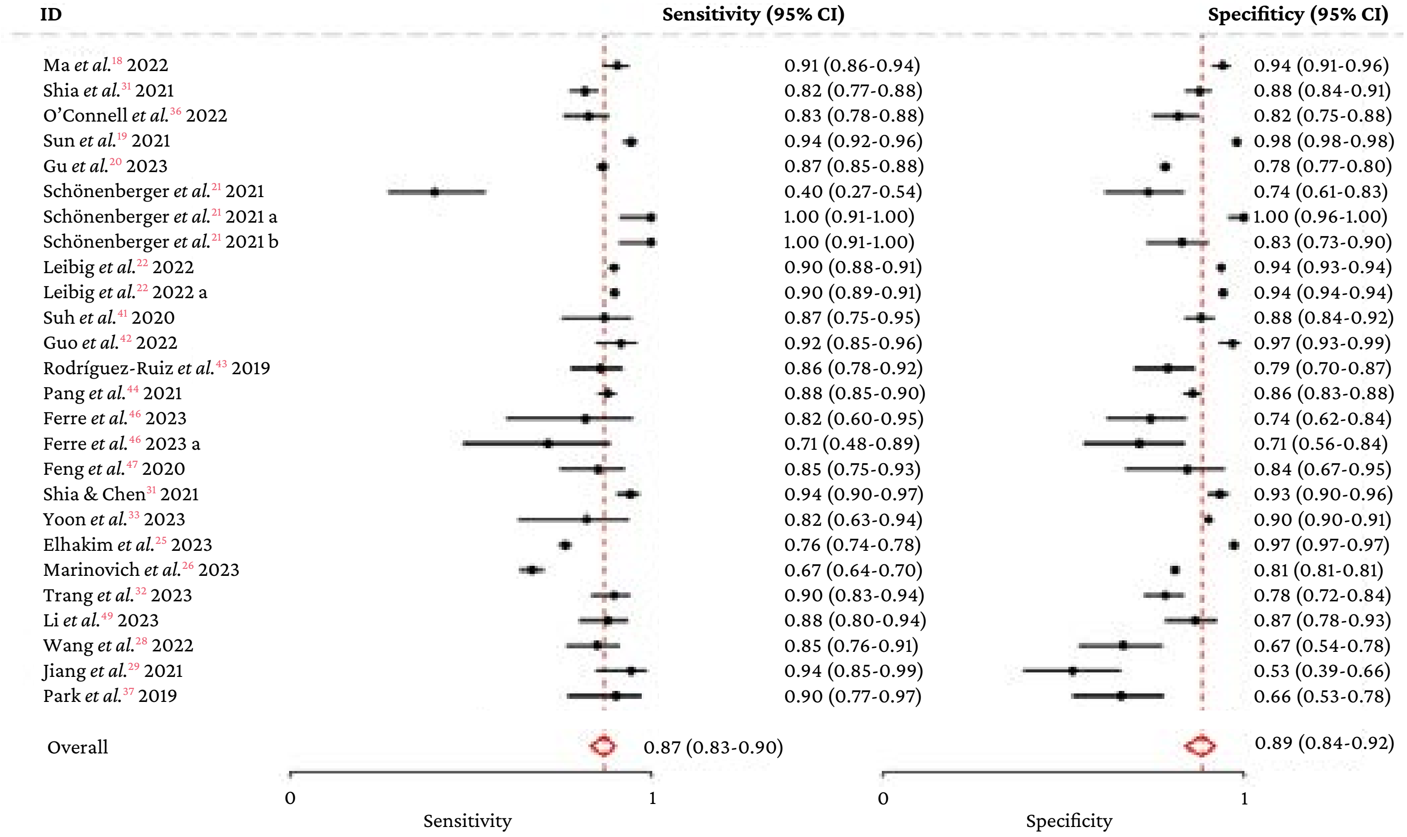

Figure 2. Cumulative sensitivity and specificity of artificial intelligence in breast cancer detection: a meta-analysis.
CI: confidence interval.

SROC curve results for AI in breast cancer detection

The SROC curve synthesizes the sensitivity and specificity data from the studies included in the meta-analysis to evaluate the overall performance of AI algorithms in breast cancer detection. Figure 3 shows individual points representing the data reported in the studies and diamonds representing the summary point, which provides an estimate of the average performance of AI in these studies.

The area bounded by the dashed-dotted line represents the prediction region and indicates where the true results of future studies are expected to be found based on the current variability of the data reported in the included studies. The area bounded by the dashed line shows the confidence region, which provides an estimate of the uncertainty surrounding the true summary point. Finally, the solid line represents the SROC curve, which indicates the overall relationship between sensitivity and specificity across the different diagnostic cut-off points used in the included studies. The proximity of the curve to the upper left point of the graph (sensitivity 1, specificity 1) denotes better diagnostic performance.

In Figure 3, the SROC curve seems to imply that AI has a high diagnostic performance, although there is some variability between studies. Furthermore, the summary point at the top of the curve suggests that AI tends to have a higher sensitivity at the expense of slightly lower specificity.

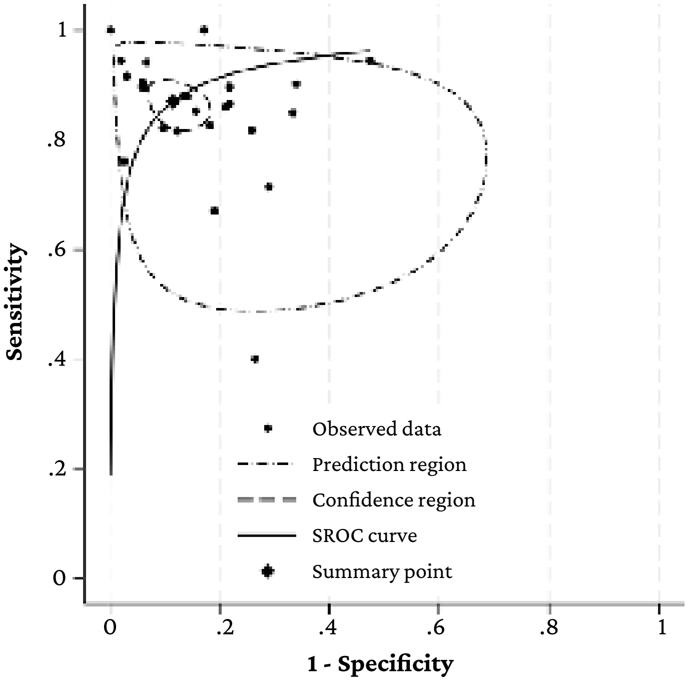

Figure 3. SROC curve of artificial intelligence performance in breast cancer detection in studies included in the meta-analysis.
SROC: summary receiver operating characteristic.

Discussion

This meta-analysis summarizes the evidence on the capacity of AI to detect breast cancer and demonstrates that it has significant sensitivity and specificity, thereby reflecting its potential as a complementary diagnostic tool.

According to the results of the meta-analysis, the pooled sensitivity and specificity of AI for breast cancer detection in imaging tests (mammograms, ultrasound, MRI) were 87% and 89%, respectively, demonstrating that it has robust diagnostic performance compared to standard evaluations. Furthermore, these values are similar to the pooled measures of sensitivity and specificity reported in previous systematic reviews and meta-analyses (searches up to 2022) on the performance of machine learning algorithms in the detection of breast cancer in mammograms (75.4-91.4% and 85.7-91.6%).[50-52]

Considering that our meta-analysis included seven studies published in 2023 (searches conducted up to May 2024), the aforementioned similarity implies that new evidence on the diagnostic performance of AI algorithms continues to demonstrate that these technologies have a high diagnostic performance in the detection of breast cancer in imaging tests, mainly mammograms in screening programs. Finally, although it was not possible to obtain a pooled AUC measure due to the heterogeneity between studies, the AUCs observed in most individual studies (≥0.90) are in line with the pooled AUCs reported in two of these meta-analyses (0.89[50] and 0.945 [0.974 for CNN, 0.881 for ANN, and 0.914 for SVM][51]).

Meanwhile, regarding the classification of breast lesions (benign vs. malignant) and tumor subtypes, even though sensitivity and specificity data were obtained for all studies, it was not possible to calculate aggregate metrics due to methodological and population heterogeneity. The variations observed in these three AI uses (detection, lesion classification, and subtype differentiation) can be attributed to both differences in the algorithms employed and the characteristics of the study populations.

Our findings are consistent with previous studies that report the high accuracy of AI in interpreting mammographic images. For example, according to Kim *et al.*,[53] AI has been shown to significantly improve diagnostic accuracy compared to radiologists, especially when used as a support tool. However, the variability observed in sensitivity suggests that some AI algorithms may require further adjustments to match or exceed the specificity of human diagnosis, a challenge that has been identified in prior systematic reviews.[54] Moreover, although some researchers have pointed out the risk of overdiagnosis associated with the use of AI,[8] the evidence suggests that, with appropriate calibrations of decision thresholds to define an "abnormal result," AI can optimize the balance between sensitivity and specificity, thus minimizing the number of false positives and potentially reducing the risk of overdiagnosis.

Our study has some limitations that should be considered when interpreting the results. First, the heterogeneity of the designs and population characteristics in the included studies, which may result in considerable variability in sensitivity and specificity estimates. Second, the decision thresholds used by the different algorithms are not standardized, complicating direct comparison of their results. Third, there is a probable publication and language bias, given that international databases tend to favor studies with positive findings. Fourth, most of the studies analyzed are retrospective and lack external validation in independent cohorts, which reduces certainty about the applicability of the models in clinical contexts other than those in which they were developed.

However, this systematic review also has several strengths. First, an exhaustive and updated search was conducted in Medline, Embase, and LILACS, which was complemented by a manual search of references. This ensured the inclusion of the most recent publications on the use of AI algorithms for the interpretation of breast imaging studies (mammograms, ultrasound, and MRI). Second, the protocol was registered in PROSPERO and the study followed the PRISMA guidelines, which adds transparency and reproducibility to the process. Third, the risk of bias assessment was performed using the QUADAS-2 tool, which is specifically designed for diagnostic accuracy studies. Lastly, the results were grouped based on clearly defined clinical scenarios (population screening, lesion classification, and subtype differentiation), making the interpretation of AI's usefulness in each context easier.

The results of this systematic review and meta-analysis underscore the need for prospective studies evaluating the integration of AI into clinical workflows and its impact on long-term health outcomes. Moreover, further studies should focus on standardizing AI protocols and evaluating their cost-effectiveness to facilitate their adoption in various clinical settings. The customization of AI algorithms according to demographic and clinical characteristics may also be a fruitful area for further research.

Conclusion

AI demonstrated a high diagnostic performance in screening and classification of breast lesions, outperforming humans in several studies. Therefore, AI has the potential to become a complementary tool for improving the efficacy and quality of screening programs, although prospective studies are needed to evaluate its clinical implementation, standardize protocols, and determine its impact on long-term health outcomes.

Therefore, AI systems could be used as a second opinion to increase diagnostic accuracy and as a triage tool to prioritize cases requiring urgent care, hence optimizing healthcare resources.

Conflicts of interest

None stated by the authors.

Funding

None stated by the authors.

Acknowledgements

None stated by the authors.

References:

1. Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, *et al.* Global Cancer Observatory: Cancer Today [Internet]. Lyon: International Agency for Research on Cancer; 2022 [cited 2025 Jul 2]. Available from: https://gco.iarc.who.int/media/globocan/factsheets/populations/900-world-fact-sheet.pdf.
2. Harbeck N, Gnant M. Breast cancer. The Lancet. 2017;389(10074):1134-50. doi: 10.1016/S0140-6736(16)31891-8. PMID: 27865536.
3. Chen Y, Lu J, Li J, Liao J, Huang X, Zhang B. Evaluation of diagnostic efficacy of multimode ultrasound in BI-RADS 4 breast neoplasms and establishment of a predictive model. Front Oncol. 2022;12:1053280. doi: 10.3389/fonc.2022.1053280. PMID: 36505867; PMCID: PMC9730703.
4. Akselrod-Ballin A, Chorev M, Shoshan Y, Spiro A, Hazan A, Melamed R, *et al.* Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms. Radiology. 2019;292(2):331-42. doi: 10.1148/radiol.2019182622. PMID: 31210611.
5. World Health Organization (WHO). Breast cancer [Internet]. Geneva: WHO; 2024 [cited 2024 Sep 3]. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
6. Manson EN, Achel DG. Fighting breast cancer in low-and-middle-income countries - What must we do to get every woman screened on regular basis? Scientific Africa. 2023;21:e01848. doi: 10.1016/j.sciaf.2023.e01848.
7. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, *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. PMID: 31894144.
8. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, *et al.* Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019;111(9):916-22. doi: 10.1093/jnci/djy222. PMID: 30834436; PMCID: PMC6748773.
9. Verburg E, van Gils CH, Van Der Velden BHM, Bakker MF, Pijnappel RM, Veldhuis WB, *et al.* Validation of Combined Deep Learning Triaging and Computer-Aided Diagnosis in 2901 Breast MRI Examinations from the Second Screening Round of the Dense Tissue and Early Breast Neoplasm Screening Trial. Invest Radiol. 2023;58(4):293-8. doi: 10.1097/RLI.0000000000000934. PMID: 36256783; PMCID: PMC9997620.
10. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson AN, Miglioretti DL, *et al.* Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med. 2015;175(11):1828-37. doi: 10.1001/jamainternmed.2015.5231. PMID: 26414882; PMCID: PMC4836172.
11. Watanabe AT, Lim V, Vu HX, Chim R, Weise E, Liu J, *et al.* Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography. J Digit Imaging. 2019;32(4):625-37. doi: 10.1007/s10278-019-00192-5. PMID: 31011956; PMCID: PMC6646649.
12. Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, *et al.* Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. JAMA Netw Open. 2020;3(3):e200265. doi: 10.1001/jamanetworkopen.2020.0265. Erratum in: JAMA Netw Open. 2020;3(3):e204429. doi: 10.1001/jamanetworkopen.2020.4429. PMID: 32119094; PMCID: PMC7052735.
13. Hsu W, Hippe DS, Nakhaei N, Wang PC, Zhu B, Siu N, *et al.* External Validation of an Ensemble Model for Automated Mammography Interpretation by Artificial Intelligence. JAMA Netw Open.2022;5(11):e2242343. doi: 10.1001/jamanetworkopen.2022.42343. PMID: 36409497; PMCID: PMC9679879.
14. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700. doi: 10.1136/bmj.b2700. PMID: 19622552; PMCID: PMC2714672.

15. Centro Cochrane Iberoamericano, traductores. Manual Cochrane de Revisiones Sistemáticas de Intervenciones, versión 5.1.0 [Internet]. Barcelona: Centro Cochrane Iberoamericano; 2012 [cited 2024 Apr 29]. Available from: https://training.cochrane.org/es/manual-cochrane-de-revisiones-sistem%C3%A1ticas-de-intervenciones.
16. Hillen MA, Medendorp NM, Daams JG, Smets EMA. Patient-Driven Second Opinions in Oncology: A Systematic Review. Oncologist. 2017;22(10):1197-211. doi: 10.1634/theoncologist.2016-0429. PMID: 28606972; PMCID: PMC5634767.
17. Blanchard L, Ray S, Law C, Vega-Salas MJ, Bidonde J, Bridge G, *et al.* The effectiveness, cost-effectiveness and policy processes of regulatory, voluntary and partnership policies to improve food environments: an evidence synthesis. Southampton (UK): National Institute for Health and Care Research; 2024 Sep. (Public Health Research, No. 12.08.) Available from: https://www.ncbi.nlm.nih.gov/books/NBK607533/. doi: 10.3310/JYWP4049
18. Ma M, Liu R, Wen C, Xu W, Xu Z, Wang S, *et al.* Predicting the molecular subtype of breast cancer and identifying interpretable imaging features using machine learning algorithms. Eur Radiol. 2022;32(3):1652-62. doi: 10.1007/s00330-021-08271-4. PMID: 34647174.
19. Sun Y, Qu Y, Wang D, Li Y, Ye L, Du J, *et al.* Deep learning model improves radiologists' performance in detection and classification of breast lesions. Chin J Cancer Res. 2021;33(6):682-93. doi: 10.21147/j.issn.1000-9604.2021.06.05. PMID: 35125812; PMCID: PMC8742176.
20. Gu Y, Xu W, Liu T, An X, Tian J, Ran H, *et al.* Ultrasound-based deep learning in the establishment of a breast lesion risk stratification system: a multicenter study. Eur Radiol. 2023;33(4):2954-64. doi: 10.1007/s00330-022-09263-8. PMID: 36418619.
21. Schönenberger C, Hejduk P, Ciritsis A, Marcon M, Rossi C, Boss A. Classification of Mammographic Breast Microcalcifications Using a Deep Convolutional Neural Network: A BI-RADS-Based Approach. Invest Radiol. 2021;56(4):224-31. doi: 10.1097/RLI.0000000000000729. PMID: 33038095.
22. 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):e507-19. doi: 10.1016/S2589-7500(22)00070-X. PMID: 35750400; PMCID: PMC9839981.
23. Lee SE, Son NH, Kim MH, Kim EK. Mammographic Density Assessment by Artificial Intelligence-Based Computer-Assisted Diagnosis: A Comparison with Automated Volumetric Assessment. J Digit Imaging. 2022;35(2):173-9. doi: 10.1007/s10278-021-00555-x. PMID: 35015180; PMCID: PMC8921363.
24. Sasaki M, Tozaki M, Rodríguez-Ruiz A, Yotsumoto D, Ichiki Y, Terawaki A, *et al.* Artificial intelligence for breast cancer detection in mammography: experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women. Breast Cancer. 2020;27(4):642-51. doi: 10.1007/s12282-020-01061-8. PMID: 32052311.
25. Elhakim MT, Stougaard SW, Graumann O, Nielsen M, Lång K, Gerke O, *et al.* Breast cancer detection accuracy of AI in an entire screening population: a retrospective, multicentre study. Cancer Imaging. 2023;23(1):127. doi: 10.1186/s40644-023-00643-x. PMID: 38124111; PMCID: PMC10731688.
26. Marinovich ML, Wylie E, Lotter W, Lund H, Waddell A, Madeley C, *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. PMID: 36863255; PMCID: PMC9996220.
27. Raafat M, Mansour S, Kamel R, Ali HW, Shibel PE, Marey A, *et al.* Does artificial intelligence aid in the detection of different types of breast cancer? Egypt J Radiol Nucl Med. 2022;53(1):182. doi: 10.1186/s43055-022-00868-z.
28. Wang Q, Chen H, Luo G, Li B, Shang H, Shao H, *et al.* Performance of novel deep learning network with the incorporation of the automatic segmentation network for diagnosis of breast cancer in automated breast ultrasound. Eur Radiol. 2022;32(10):7163-72. doi: 10.1007/s00330-022-08836-x. PMID: 35488916.
29. Jiang Y, Edwards AV, Newstead GM. Artificial intelligence applied to breast MRI for improved diagnosis. Radiology. 2021;298(1):38-46. doi: 10.1148/radiol.2020200292. PMID: 33078996.
30. Shia WC, Lin LS, Chen DR. Classification of malignant tumours in breast ultrasound using unsupervised machine learning approaches. Sci Rep. 2021;11(1):1418. doi: 10.1038/s41598-021-81008-x. PMID: 33446841; PMCID: PMC7809485.
31. Shia WC, Chen DR. Classification of malignant tumors in breast ultrasound using a pretrained deep residual network model and support vector machine. Comput Med Imaging Graph. 2021;87:101829. doi: 10.1016/j.compmedimag.2020.101829. PMID: 33302247.
32. Trang NTH, Long KQ, An PL, Dang TN. Development of an Artificial Intelligence-Based Breast Cancer Detection Model by Combining Mammograms and Medical Health Records. Diagn (Basel). 2023;13(3):346. doi: 10.3390/diagnostics13030346. PMID: 36766450; PMCID: PMC9913958.
33. Yoon JH, Han K, Suh HJ, Youk JH, Lee SE, Kim EK. Artificial intelligence-based computer-assisted detection/diagnosis (AI-CAD) for screening mammography: Outcomes of AI-CAD in the mammographic interpretation workflow. Eur J Radiol Open. 2023;11:100509. doi: 10.1016/j.ejro.2023.100509. PMID: 37484980; PMCID: PMC10362167.
34. Gu Y, Xu W, Lin B, An X, Tian J, Ran H, *et al.* Deep learning based on ultrasound images assists breast lesion diagnosis in China: a multicenter diagnostic study. Insights Imaging. 2022;13(1):124. doi: 10.1186/s13244-022-01259-8. PMID: 35900608; PMCID: PMC9334487.

35. Koch HW, Larsen M, Bartsch H, Kurz KD, Hofvind S. Artificial intelligence in BreastScreen Norway: a retrospective analysis of a cancer-enriched sample including 1254 breast cancer cases. Eur Radiol. 2023;33(5):3735-43. doi: 10.1007/s00330-023-09461-y. PMID: 36917260; PMCID: PMC10121532.
36. O'Connell AM, Bartolotta TV, Orlando A, Jung SH, Baek J, Parker KJ. Diagnostic Performance of an Artificial Intelligence System in Breast Ultrasound. J Ultrasound Med. 2022;41(1):97-105. doi: 10.1002/jum.15684. PMID: 33665833.
37. Park HJ, Kim SM, La Yun B, Jang M, Kim B, Jang JY, *et al.* A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: Added value for the inexperienced breast radiologist. Medicine (Baltimore). 2019;98(3):e14146. doi: 10.1097/MD.0000000000014146. PMID: 30653149; PMCID: PMC6370030.
38. Busaleh M, Hussain M, Aboalsamh HA, Amin FE. Breast Mass Classification Using Diverse Contextual Information and Convolutional Neural Network. Biosensors (Basel). 2021;11(11):419. doi: 10.3390/bios11110419. PMID: 34821634; PMCID: PMC8615673.
39. Cai X, Li X, Razmjooy N, Ghadimi N. Breast Cancer Diagnosis by Convolutional Neural Network and Advanced Thermal Exchange Optimization Algorithm. Comput Math Methods Med. 2021;2021:5595180. doi: 10.1155/2021/5595180. PMID: 34790252; PMCID: PMC8592754.
40. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Sci Rep. 2019;9(1):12495. doi: 10.1038/s41598-019-48995-4. PMID: 31467326; PMCID: PMC6715802.
41. Suh YJ, Jung J, Cho BJ. Automated Breast Cancer Detection in Digital Mammograms of Various Densities via Deep Learning. J Pers Med. 2020;10(4). doi: 10.3390/jpm10040211. PMID: 33172076; PMCID: PMC7711783.
42. Guo YY, Huang YH, Wang Y, Huang J, Lai QQ, Li YZ. Breast MRI Tumor Automatic Segmentation and Triple-Negative Breast Cancer Discrimination Algorithm Based on Deep Learning. Comput Math Methods Med. 2022;2022:2541358. doi: 10.1155/2022/2541358. PMID: 36092784; PMCID: PMC9453096.
43. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, *et al.* Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology. 2019;290(2):305-14. doi: 10.1148/radiol.2018181371. PMID: 30457482.
44. Pang T, Wong JHD, Ng WL, Chan CS. Semi-supervised GAN-based Radiomics Model for Data Augmentation in Breast Ultrasound Mass Classification. Comput Methods Programs Biomed. 2021;203:106018. doi: 10.1016/j.cmpb.2021.106018. PMID: 33714900.
45. Ye H, Hang J, Zhang M, Chen X, Ye X, Chen J, *et al.* Automatic identification of triple negative breast cancer in ultrasonography using a deep convolutional neural network. Sci Rep. 2021;11(1):20474. doi: 10.1038/s41598-021-00018-x. PMID: 34650065; PMCID: PMC8517009.
46. Ferre R, Elst J, Senthilnathan S, Lagree A, Tabbarah S, Lu FI, *et al.* Machine learning analysis of breast ultrasound to classify triple negative and HER2+ breast cancer subtypes. Breast Dis. 2023;42(1):59-66. doi: 10.3233/BD-220018. PMID: 36911927.
47. Feng H, Cao J, Wang H, Xie Y, Yang D, Feng J, *et al.* A knowledge-driven feature learning and integration method for breast cancer diagnosis on multi-sequence MRI. Magn Reson Imaging. 2020;69:40-8. doi: 10.1016/j.mri.2020.03.001. PMID: 32173583.
48. Ciritsis A, Rossi C, Eberhard M, Marcon M, Becker AS, Boss A. Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making. Eur Radiol. 2019;29(10):5458-68. doi: 10.1007/s00330-019-06118-7. PMID: 30927100.
49. Li C, Zhang H, Chen J, Shao S, Li X, Yao M, *et al.* Deep learning radiomics of ultrasonography for differentiating sclerosing adenosis from breast cancer. Clin Hemorheol Microcirc. 2023;84(2):153-63. doi: 10.3233/CH-221608. PMID: 36373313.
50. Hickman SE, Woitek R, Le EPV, Im YR, Mouritsen Luxhøj C, *et al.* Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis. Radiology. 2022;302(1):88-104. doi: 10.1148/radiol.2021210391. PMID: 34665034; PMCID: PMC8717814.
51. Liu J, Lei J, Ou Y, Zhao Y, Tuo X, Zhang B, *et al.* Mammography diagnosis of breast cancer screening through machine learning: a systematic review and meta-analysis. Clin Exp Med. 2023;23(6):2341-2356. doi: 10.1007/s10238-022-00895-0. PMID: 36242643.
52. Yoon JH, Strand F, Baltzer PAT, Conant EF, Gilbert FJ, Lehman CD, *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. PMID: 37219445; PMCID: PMC10315526.
53. Kim HE, Kim HH, Han BK, Kim KH, Han K, Nam H, *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):e138-48. doi: 10.1016/S2589-7500(20)30003-0. PMID: 33334578.
54. Aparecida-Roela R, Vansuita-Valente G, Shimizu C, Mendoza-Lopez RV, de Melo-Tucunduva TC, Koike-Folgueira G, *et al.* Deep learning algorithm performance in mammography screening: A systematic review and meta-analysis. J Clin Oncol. 2021;39(Suppl 15):e13553. doi: 10.1200/JCO.2021.39.15_suppl.e1355.

Annexes

Annex 1. Search strategies

Medline—via PubMed:
(“Breast Neoplasms/diagnosis”[MeSH] OR “Breast Cancer/diagnosis”[tiab] OR “mammary neoplasms”[tiab] OR “breast cancer”[tiab] OR “breast carcinoma”[tiab] OR “breast tumor”[tiab] OR “mammary carcinoma”[tiab] OR “mammary cancer”[tiab])
AND
(“Artificial Intelligence”[MeSH] OR “Machine Learning”[MeSH] OR “Deep Learning”[MeSH] OR “Neural Networks, Computer”[MeSH] OR “AI”[tiab] OR “machine learning”[tiab] OR “deep learning”[tiab] OR “neural network”[tiab] OR “algorithms”[tiab])
AND
(“Radiology”[MeSH] OR “Diagnostic Imaging”[MeSH] OR “radiology”[tiab] OR “imaging”[tiab] OR “mammography”[tiab] OR “breast imaging”[tiab])

Embase:
(‘breast cancer’/exp OR ‘breast neoplasm’/exp OR ‘mammary neoplasm’/exp OR ‘breast carcinoma’/exp OR ‘breast tumor’/exp OR ‘mammary carcinoma’/exp OR ‘breast cancer’:ti,ab OR ‘breast neoplasm’:ti,ab OR ‘mammary neoplasm’:ti,ab OR ‘breast carcinoma’:ti,ab OR ‘mammary carcinoma’:ti,ab OR ‘breast tumor’:ti,ab)
AND
(‘artificial intelligence’/exp OR ‘machine learning’/exp OR ‘deep learning’/exp OR ‘neural network’/exp OR ‘ai’:ti,ab OR ‘machine learning’:ti,ab OR ‘deep learning’:ti,ab OR ‘neural network’:ti,ab OR ‘algorithm’:ti,ab)
AND
(‘radiology’/exp OR ‘diagnostic imaging’/exp OR ‘mammography’/exp OR ‘breast imaging’/exp OR ‘radiology’:ti,ab OR ‘imaging’:ti,ab OR ‘mammography’:ti,ab OR ‘breast imaging’:ti,ab)

LILACS:
(“Neoplasias de la Mama/diagnóstico”[DeCS] OR “Cáncer de Mama/diagnóstico”[-DeCS] OR “neoplasias de mama”[tiab] OR “cáncer de mama”[tiab] OR “carcinoma de mama”[tiab] OR “tumor de mama”[tiab] OR “carcinoma mamario”[tiab] OR “cáncer mamario”[tiab])
AND
(“Inteligencia Artificial”[DeCS] OR “Aprendizaje de Máquinas”[DeCS] OR “Aprendizaje Profundo”[DeCS] OR “Redes Neuronales, Computacionales”[DeCS] OR “AI”[tiab] OR “aprendizaje automático”[tiab] OR “aprendizaje profundo”[tiab] OR “redes neuronales”[tiab] OR “algoritmos”[tiab])
AND
(“Radiología”[DeCS] OR “Imagen por Diagnóstico”[DeCS] OR “radiología”[tiab] OR “imágenes”[tiab] OR “mamografía”[tiab] OR “imágenes mamarias”[tiab])