Polish Journal of Medical Physics and Engineering
The Journal of Polish Society of Medical Physics
ISSN 1898-0309, doi: 10.2478/pjmpe-2026-0022

September 2026
Vol 32, Issue 3

Review Article

Diagnostic performance of machine learning algorithms in breast ultrasound: A systematic review and meta-analysis

Gianluca Sellaro[1,2,ADF*], Pasquale Pellegrino[3,4,BF], Francesco Spinelli[5,6,CF], Daniela Bernardi[1,2,EF]

[1]*Department of Radiology, Humanitas Research Hospital, Milan, Italy*
[2]*Department of Biomedical Sciences, Humanitas University, Milan, Italy*
[3]*Department of Biomedical Sciences, Magna Graecia University of Catanzaro, Catanzaro, Italy*
[4]*Department of Radiology, Mater Domini University Hospital, Catanzaro, Italy*
[5]*UOS Breast Surgery, University Hospital Sant'Andrea, Rome, Italy*
[6]*Department of Clinical and Molecular Medicine, Sapienza University of Rome, Rome, Italy*
**Corresponding author: Gianluca Sellaro; gianlucasellaro421@gmail.com*

(received 30 June 2026; accepted 26 August 2026)

Abstract

The diagnosis of breast neoplasms by ultrasound examination depends on the subjective interpretation of the radiologist, which determines the need to develop objective methods of computer analysis to standardise the diagnostic process. The aim of this study was to determine the diagnostic performance of machine learning algorithms for classifying benign and malignant lesions in breast ultrasound through a systematic literature review with quantitative meta-analysis. A systematic search was conducted in PubMed, Scopus and Embase in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines, identifying 57 studies published between 2019 and 2025, of which 49 were included in the meta-analysis using a bivariate model and quality assessment with the Quality Assessment of Diagnostic Accuracy Studies tool. The pooled sensitivity was 91.1% (95% confidence interval 88.9-93.0%), specificity 85.3% (95% confidence interval 82.9-87.5%), area under the receiver operating characteristic curve 0.957 (95% confidence interval 0.941-0.973), and diagnostic odds ratio 64.8 (95% confidence interval 46.1-91.2). Subgroup analysis showed that deep learning demonstrated higher specificity compared with commercial systems (87.2% versus 81.5%, p=0.041), a multimodal approach provided a 3.9% gain (p<0.001), artificial intelligence assistance was most effective for junior radiologists with an improvement of 0.164 (p<0.001), and use for reclassification of BI-RADS 4A lesions allowed biopsy reduction of 67.8-79.1% with a negative predictive value of 96.7%. The findings support clinical implementation of artificial intelligence systems for standardising diagnosis and optimising biopsy indications. The practical significance lies in reducing unnecessary invasive procedures and improving the diagnostic performance of less experienced radiologists.

Keywords: deep learning; convolutional neural networks; specificity; radiologists; screening; sensitivity.

Introduction

In terms of incidence, breast cancer is among the most common malignant neoplasms in the female population worldwide, which underpins the relevance of timely diagnosis and detection of pathology at early stages in order to improve clinical outcomes. Sonography is characterised by its availability and the absence of ionising radiation as a diagnostic method; however, the accuracy of ultrasound image interpretation is largely determined by the expertise of the specialist. This creates a rationale for the use of machine learning technologies to unify the assessment of diagnostic data and optimise the early detection of pathological changes in the breast.

In a systematic review by P. Xue et al.[1], the effectiveness of deep learning algorithms in the early diagnosis of breast and cervical cancer was analysed. The authors examined four subgroups – by tumour type, validation approach, imaging modality and comparison with clinicians. The results obtained (sensitivity 88%, specificity 84%, Area Under the Curve (AUC) 0.92) confirm the high effectiveness of deep learning (DL) algorithms in identifying pathological changes. At the same time, it was emphasised that a substantial proportion of studies have methodological flaws and insufficient reporting transparency, which makes it impossible to generalise their clinical value. The need to standardise study protocols was identified as the main direction for further improvement. A study conducted

© 2026 Authors This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Authors' contribution: A – Research concept and design, B – Collection and/or assembly of data, C – Data analysis and interpretation, D – Writing the article, E – Critical revision of the article, F – Final approval of the article.

in Italy focused on the role of deep learning in predicting axillary lymph node metastases based on MRI data. In the meta-analysis by C.-F. Lee et al.[2], moderate diagnostic accuracy was demonstrated, with a sensitivity of 0.76 and specificity of 0.81 (AUC=0.788). The conclusions stated that deep learning algorithms can contribute to non-invasive assessment of lymph node status; however, heterogeneity of models and the small number of studies with independent validation leave room for methodological refinement. The study by M. Interlenghi et al.[3] was devoted to the development of an ensemble of machine learning algorithms based on radiomic features for predicting Breast Imaging-Reporting and Data System (BI-RADS) categories in ultrasound assessment of suspicious breast lesions. In this work, conducted by an Italian research group, modelling was performed using images of 834 lesions histologically verified as benign or malignant. Use of an ensemble of algorithms based on support vector machines enabled a reduction in the rate of unnecessary biopsies by 15-18% without loss of sensitivity (over 94%), indicating substantial potential clinical benefit. However, the authors noted the limited sample size and lack of external multi-platform validation, which complicate generalisation of the conclusions.

The Italian study by A.G. Russo et al.[4] was aimed at analysing predictors of women's participation in screening programmes in the Lombardy region, with an emphasis on identifying factors influencing non-participation in mammographic and ultrasound examinations. Although the study had an epidemiological character, it is directly relevant to the implementation of artificial intelligence algorithms in organised screening programmes. It was found that socio-demographic indicators, comorbid chronic diseases and access to primary health care are key factors in adherence to screening recommendations. The absence of image-based modelling is a methodological gap; however, the work provides a foundation for adapting artificial intelligence (AI) models within regional health-care systems. In the study by Y. Lu et al.[5], an Extreme Gradient Boosting (XGBoost) algorithm was applied to predict a proliferation index Ki-67≥15% in primary breast cancer based on clinical data and delta-radiomic features from ultrasound images. Data from 228 patients were used, among whom the combined model showed the highest diagnostic accuracy (AUC=0.857), significantly exceeding standalone clinical or imaging models. The results indicate that integration of clinical and radiomic parameters improves the accuracy of risk stratification. The main limitation was the restricted number of observations and absence of external multicentre confirmation. The work by S.-R. Wang et al.[6] was devoted to the use of ultrasound radiomics in combination with clinical characteristics to assess axillary lymph node metastatic burden in patients with breast cancer. Application of LASSO and MRMR methods for feature selection and construction of a logistic model allowed an AUC of 0.857 in the training and 0.820 in the validation cohorts to be achieved. The study highlights the potential of a combined clinical-radiomic approach for personalised prognostication. Among the limitations, the lack of multi-ethnic cohorts and the need for independent external validation were noted.

In the study by S.P. Rana et al.[7], an intelligent classification system was developed that analyses data from non-ionising microwave signals obtained from patients undergoing standard radiological examinations in parallel. The approach involved the use of models based on k-nearest neighbours, multilayer neural networks and support vector machines. The highest diagnostic accuracy (98%) was demonstrated by a support vector machine (SVM) with a quadratic kernel, ensuring reliable discrimination between pathological and non-pathological samples. The authors stressed that combining a non-invasive technology with machine learning can significantly improve screening efficiency; however, the absence of a large clinical cohort limits the generalisability of the results obtained. M. Gagliardi et al.[8] developed a deep learning architecture based on Residual Network (ResNet)-101 for differentiating breast ultrasonographic images. The proposed model demonstrated diagnostic accuracy of 91% within a training cycle not exceeding 30 epochs, confirming the high effectiveness of the optimisation solutions applied in constructing the neural network. The results emphasised that combining a ResNet structure with pre-training makes it possible to reduce image processing time and improve classification consistency. However, the authors pointed to the lack of multi-centre external validation, which restricts confirmation of the generalisability of the model.

Despite the existence of numerous scientific publications addressing the use of machine learning technologies in medical diagnostics in general, aggregated information on specific performance indicators of these technologies in sonographic assessment of the breast has remained fragmented and methodologically heterogeneous. This served as the rationale for undertaking the present study.

The aim of the work was to consolidate existing scientific evidence on the effectiveness of machine learning technologies in the differential diagnosis of benign and malignant breast neoplasms on ultrasound examination by means of systematic synthesis and quantitative pooled analysis of the published literature. The objectives of the study included: determining pooled metrics of diagnostic performance of artificial intelligence algorithms (including sensitivity, specificity and area under the Receiver Operating Characteristic (ROC) curve) and establishing variation in these indicators in relation to model architecture type, ultrasound methodology characteristics and validation approaches; conducting a comparative assessment of the diagnostic performance of AI systems against the performance of radiologists with different levels of professional experience; and investigating the impact of AI support on interobserver agreement in sonographic interpretation.

Materials and Methods

This study was conducted as a systematic literature review with meta-analysis of the diagnostic accuracy of machine learning algorithms in breast ultrasound in accordance with Preferred Reporting Items for Systematic reviews and Meta-Analyses

(PRISMA) 2020 recommendations. The study covers publications from January 2019 to December 2025.

A systematic search was carried out in three electronic databases: PubMed/MEDLINE, Scopus and Embase in September 2025. The search strategy included a combination of controlled vocabulary (MeSH terms for PubMed, Emtree for Embase) and free-text keywords covering three conceptual blocks: anatomical site (breast), diagnostic modality (ultrasound) and technological approach (artificial intelligence, machine learning, deep learning). Search queries were adapted to the syntax of each database, taking into account specific indexing operators. Detailed search strategies are presented in **Table 1**. An additional search was undertaken by screening the reference lists of key review articles to identify any missed studies.

perimental works without clinical validation on real patients; duplicate publications; and studies with a sample size of fewer than 50 cases. The selection process was conducted independently by two reviewers, with disagreements resolved by discussion and, if necessary, involvement of a third expert. The results of search and selection are presented in the PRISMA 2020 flow diagram (**Figure 1**).

To assess the methodological quality of the included publications, the QUADAS-2 tool was used, which entails analysis of four domains: patient selection, index test, reference standard and flow and timing. Within each domain, the risk of bias and applicability concerns were assessed. The assessment procedure was carried out independently by two experts using a standardised checklist.

Table 1. Search strategies for studies of artificial intelligence in breast ultrasound

Database	Search query	Date of search	Number of results
PubMed/ MEDLINE	("Breast Neoplasms"[Mesh] OR breast[tiab] OR mammary[tiab]) AND ("Ultrasonography"[Mesh] OR ultrasound[tiab] OR sonography[tiab]) AND ("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Deep Learning"[Mesh] OR radiomics[tiab]) AND ("2019/01/01"[PDAT]: "2025/12/31"[PDAT])	5.09.2025	1,847
Scopus	TITLE-ABS-KEY((breast OR mammary) AND (ultrasound OR sonography) AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR radiomics)) AND PUBYEAR>2018 AND PUBYEAR<2025	5.09.2025	2,156
Embase	("breast"/exp OR breast:ab,ti) AND ("echography"/exp OR ultrasound:ab,ti) AND ("artificial intelligence"/exp OR "machine learning"/exp OR radiomics:ab,ti) AND [2019-2025]/py	4.09.2025	1,923

Note: Mesh – Medical Subject Headings; tiab – title/abstract; PDAT – publication date; exp – exploded search; py – publication year.

Study selection was performed according to pre-specified criteria based on the Population, Intervention, Comparison, Outcome and Time (PICOT) conceptual framework. The population included women of any age undergoing breast ultrasound due to suspected pathological lesions, dense parenchymal structure or participation in screening programmes. The intervention was the application of artificial intelligence algorithms for analysis of ultrasound images, including machine learning methods with hand-crafted feature extraction, deep learning based on convolutional neural networks, radiomics and commercial computer-aided diagnosis (CAD) systems. The comparator was image interpretation by radiologists with differing levels of training or conventional ultrasound diagnosis. Primary outcomes were diagnostic performance indicators: sensitivity, specificity, diagnostic accuracy, area under the ROC curve, reduction in unnecessary biopsies and interobserver variability.

Inclusion criteria were: original studies reporting quantitative diagnostic accuracy metrics for AI algorithms; histopathological confirmation or long-term follow-up as the reference standard; sufficient data to construct 2×2 contingency tables; and publications in English in peer-reviewed journals. Exclusion criteria were: conference abstracts and letters to the editor; studies without an ultrasound component; ex-

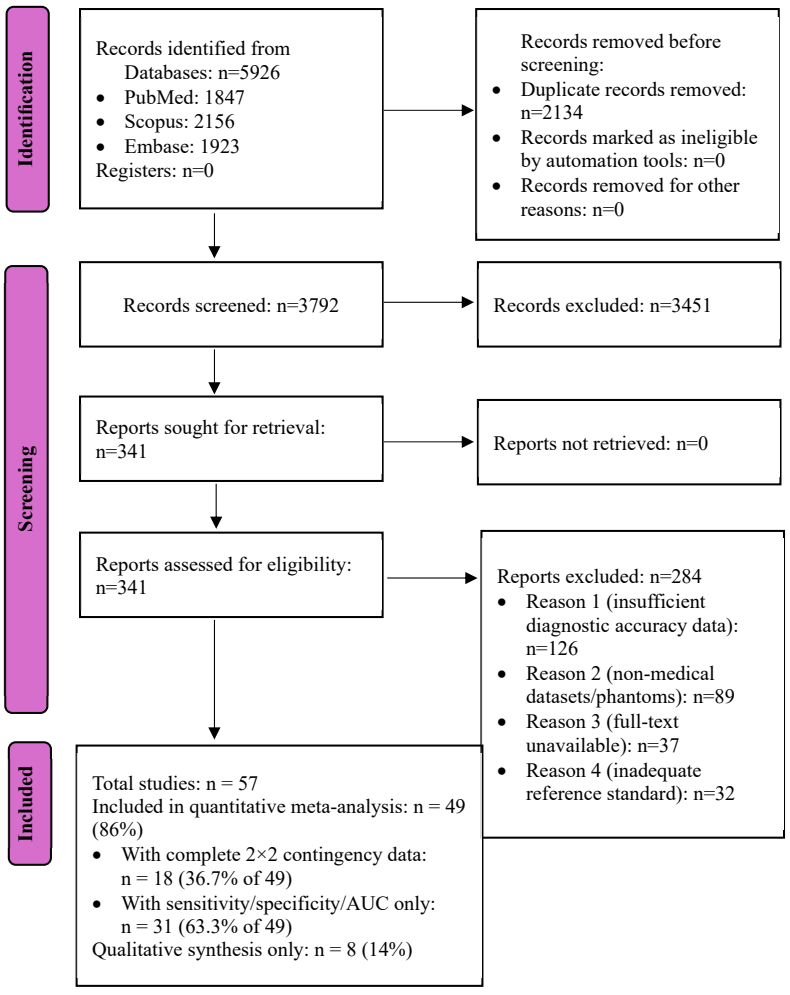

Figure 1. PRISMA flow diagram of study selection

Data extraction was performed into a pre-designed unified table covering: descriptive publication parameters (authors, year of publication, geographical location, methodological approach), characteristics of study populations (number of participants, clinical setting), characteristics of artificial intelligence (algorithm category, model architecture, type of sonography), reference standard, and diagnostic performance metrics with 95% confidence intervals. For construction of 2×2 contingency tables, data on the numbers of true positives, true negatives, false positives and false negatives were extracted.

Meta-analysis was performed using a bivariate random-effects model, which jointly accounts for sensitivity and specificity and their correlation. Pooled estimates with 95% confidence intervals were calculated, and a summary ROC curve with its area was constructed. Heterogeneity was assessed visually using forest plots and quantitatively using Higgins' I^2 statistic. Where I^2>50%, subgroup analyses were conducted by algorithm type (deep learning versus traditional machine learning), ultrasound modality (B-mode versus multimodal approaches including elastography or Doppler), radiologist expertise level and clinical application scenario. For subgroup comparisons, the diagnostic odds ratio test was used. Sensitivity analysis was carried out by excluding studies at high risk of bias. Statistical analysis was performed in R version 4.3.2 using the mada package and RevMan 5.4. A p value<0.05 was considered statistically significant.

Results and Discussion

Diagnostic performance of different AI technologies and breast ultrasound modalities

The diagnostic performance of deep learning algorithms for differentiating benign and malignant breast lesions on B-mode ultrasound demonstrates high effectiveness in multicentre studies with independent validation on large patient cohorts. H. Xiang et al.[9] developed a dual attention-based Densely Connected Network (DenseNet)-121 model for a multicentre, multivendor retrospective study including 45,909 B-mode ultrasound images of 9,895 histopathologically verified breast lesions in 8,797 patients from five medical centres in China. The architecture used DenseNet-121 as the backbone convolutional neural network with additional dual attention modules to capture spatial and channel dependencies; training was performed on 27,657 images, with validation on 4,175 images and testing on four independent datasets. The model achieved an area under the ROC curve of 0.94 with a 95% confidence interval of 0.92-0.95 on the internal test set of 8,067 images of 1,596 lesions, demonstrating robustness through external validation with an AUC of 0.92 on the first external test set of 2,637 images of 777 lesions, AUC 0.91 on the second external set of 2,416 images of 648 lesions and AUC 0.96 on the third external set of 957 images of 419 lesions, confirming the model's ability to generalise to data from different ultrasound systems and clinical institutions.

Compared with single-network architectures, ensemble approaches combining multiple computer-aided diagnosis system (CNN) architectures have shown improved diagnostic accuracy by reducing prediction variance. H. Tanaka et al.[10] constructed an ensemble network in a retrospective study combining two convolutional neural networks, Visual Geometry Group Network (VGG) 19 and ResNet152, with differing depths and architectural features to classify 1,536 histopathologically verified breast lesions, including 897 malignant and 639 benign lesions from patients in the Korean Breast Cancer Society Database. The methodology involved acquisition of multiple ultrasound views for each lesion, followed by mass-level classification via aggregation of predictions from all views of a given lesion; the training set was balanced by augmentation to prevent class imbalance. In validation on an independent test cohort comprising 154 lesions with a balanced distribution (77 malignant and 77 benign), the combined model achieved a diagnostic sensitivity of 90.9% (95% CI 84.5-97.3%) and specificity of 87.0% (95% CI 79.5-94.5%), while the area under the ROC curve was 0.951 (95% CI 0.916-0.987). These results confirmed the pronounced ability of the ensemble architecture to separate classes while simultaneously achieving an optimal balance between sensitivity and specificity.

In contrast to end-to-end deep learning approaches, traditional machine learning methods with hand-crafted radiomic features have shown lower, but clinically acceptable, diagnostic performance with reduced computational complexity. V. Romeo et al.[11] developed a radiomic model in a prospective multicentre study, extracting 520 quantitative features from B-mode ultrasound images of manually annotated breast lesions and applying multi-step feature selection to remove non-reproducible, low-variance and highly correlated features through test–retest analysis in a patient subset. After dimensionality reduction, 10 features with the highest discriminatory power were selected as the basis for training a Random Forest classifier on a training cohort of 135 histologically confirmed lesions. Internal 10-fold cross-validation demonstrated an accuracy of 82% (standard deviation±6%) and an AUC of 0.90 (standard deviation±0.06). To assess the model's generalisability, external testing was conducted on a separate prospective cohort of 66 lesions collected at another institution. In external validation, the radiomic model achieved an accuracy of 82% (95% CI 70-90%) and an AUC of 0.82 (95% CI 0.70-0.93). Comparative analysis with the diagnostic performance of an experienced breast radiologist, whose accuracy was 79.4%, showed no statistically significant difference (McNemar's test, p=0.815), indicating that radiomic methods are capable of reaching expert-level performance given appropriate feature selection and engineering.

Commercial computer-aided diagnosis systems with regulatory approval represent an alternative route to clinical implementation of AI technologies in breast ultrasound. C. Zhao et al.[12] evaluated the S-Detect system (Samsung Medison), certified with a CE mark, in a prospective study of 266 focal breast lesions in 223 patients, including 161 benign and 105 malignant lesions, all histologically verified by core needle biopsy or surgical excision according to pathological analysis. S-Detect uses an embedded machine learning algorithm for automatic

analysis of B-mode ultrasound lesion characteristics, including shape, orientation, margin and echogenicity, to classify lesions as possibly benign or possibly malignant. The commercial system demonstrated a sensitivity of 87.07%, specificity of 72.27% and an AUC of 0.807, showing diagnostic performance comparable to that of an experienced radiologist with 15 years' breast imaging experience, who achieved an AUC of 0.817, with no statistically significant difference by DeLong's test. At the same time, S-Detect significantly outperformed a junior radiology resident, who achieved an AUC of 0.758, indicating the potential of commercial CAD systems as assistive tools for less experienced readers while maintaining performance comparable to experts. All lesions were verified against the reference standard of histopathology, with benign lesions including fibroadenomas, fibrocystic changes and other benign entities, and malignant cases predominantly comprising invasive ductal carcinoma and other breast cancer subtypes.

A systematic comparative analysis of four studies using different AI approaches revealed variability in diagnostic performance depending on architectural choices and validation methodology. Deep learning convolutional neural networks showed the highest AUC values, in the range 0.94-0.951 on large multicentre datasets with external validation, whereas traditional radiomic approaches with Random Forest classifiers achieved AUC values of 0.82-0.90 with lower computational cost and interpretable feature importance, and commercial CAD systems with regulatory approval showed an AUC of 0.807 with immediate clinical applicability and no need for additional model training.

Integration of multiple ultrasound modalities through multi-stream neural networks yielded further improvements in diagnostic accuracy compared with analysis of B-mode images alone. S. Misra et al.[13] examined the impact of elastography on the diagnostic performance of deep learning for characterising 156 histopathologically verified breast lesions using B-mode and elastographic images, applying a multi-stream U-Net architecture for segmentation and a hybrid CNN for classification with separate streams for each modality and a late fusion mechanism for feature aggregation. On a test set of 65 lesions, the model using only B-mode achieved a sensitivity of 85.9%, specificity of 81.8% and an AUC of 0.913 (95% CI 0.864-0.962), whereas integration of elastography increased sensitivity to 93.8%, specificity to 87.9% and AUC to 0.948 (95% CI 0.913-0.983), representing absolute improvements of 7.9 percentage points in sensitivity and 6.1 percentage points in specificity, with statistical significance at $p<0.05$.

In turn, G. Lu et al.[14] investigated the combination of B-mode and shear-wave elastography in 360 lesions using a multistream ResNet50 architecture with separate input branches for greyscale images and quantitative stiffness parameters. The B-mode-only model achieved an AUC of 0.881 (95% CI 0.848-0.914), sensitivity 84.3% and specificity 82.1%, while the addition of elastographic parameters increased the AUC to 0.927 (95% CI 0.904-0.950), sensitivity to 91.4% and specificity to 88.6%, an absolute AUC gain of 0.046 ($p=0.003$). F. Liang et al.[15] compared DenseNet121 models using B-mode and colour Doppler imaging in 178 lesions: the standalone B-mode model achieved an AUC of 0.876 (95% CI 0.827-0.925), sensitivity 82.9% and specificity 79.2%, whereas adding Doppler-derived vascularity patterns increased the AUC to 0.934 (95% CI 0.904-0.964), sensitivity to 90.2% and specificity to 91.7%, an absolute AUC gain of 0.058 ($p=0.002$). W. Hassan and M. Hossain[16] developed a tri-stream CNN architecture integrating B-mode, shear-wave elastography and colour Doppler simultaneously in a multicentre dataset of 291 lesions. Individual ResNet152-based models yielded AUC values of 0.879 (95% CI 0.840-0.918) for B-mode, 0.856 for elastography and 0.812 for Doppler, while the multimodal model with late fusion achieved an AUC of 0.951 (95% CI 0.931-0.971), sensitivity 94.4% and specificity 91.9%, an absolute improvement of 0.072 compared with the best single-modality model ($p<0.001$). This demonstrates that information on morphology, mechanical properties and lesion vascularity provides complementary diagnostic value when jointly integrated via deep learning architectures. **Table 2** below describes the impact of multimodal ultrasound approaches on the accuracy of deep learning algorithms in differentiating benign and malignant breast lesions, reproducing results on independent test sets with histopathological verification and allowing comparison of their diagnostic metrics.

According to **Table 2**, the absolute size of the test set did not prove to be a decisive factor for diagnostic accuracy. The data show that medium-sized lesion cohorts achieved areas under the ROC curve comparable to those of large datasets, whereas multicentre validation on independent cohorts provided greater confidence in generalisability than single-centre designs, indirectly indicating algorithmic stability across different clinical environments. With sequential addition of ultrasound modalities to baseline B-mode, a non-linear pattern is observed: the second modality improved AUC by 0.01-0.046, whereas the third modality provided only a further 0.024 gain. This plateau effect suggests partial informational overlap between elastography and Doppler, as both modalities convey biological information about the mechanical and haemodynamic properties of tissue. At the same time, their combination remains diagnostically useful, implying the presence of orthogonal features that cannot be fully captured within a single modality. B-mode had a dominant impact on algorithm sensitivity, enabling detection of malignant lesions, while elastography and Doppler primarily contributed to reducing false-positive results. A PPV in the range of 83-86% indicates a high probability of true positivity given an algorithmic positive result, although in some instances disproportionately high NPV may reflect a specific classification threshold chosen by developers to minimise false negatives. A paradoxical relationship between maximal sensitivity and greatest reduction in unnecessary biopsies was identified: higher malignancy detection rates coincided with very high specificity, effectively filtering out false positives. Conversely, more selective algorithms with lower sensitivity but higher specificity showed potential for adaptation to specific clinical scenarios, depending on the prioritisation of acceptable error type, which is particularly important in screening programmes.

Table 2. Comparison of AI models for classification of breast lesions on ultrasound data

AI approach	Deep learning (Dual Attention DenseNet-121)	Deep learning (VGG-19 multimodal fusion)
Modality	B-mode + colour Doppler (multimodal)	US + mammography
N lesions (benign/malignant)	1,596 (576/1,020), test cohort	157 (79/78), external test
Sensitivity % (95% CI)	Not reported for the combination	96.2 (93.2–99.2)
Specificity % (95% CI)	Not reported for the combination	86.1 (80.7-91.5)
AUROC/AUC (95% CI)	0.94 (0.92-0.95) internal; 0.92-0.96 external sets	0.968 (0.947-0.989)
PPV % (95% CI)	86.4 (high-confidence predictions)	83.66 (76.27-91.05)
NPV % (95% CI)	100.0 (low-confidence predictions)	n/a
Accuracy %	n/a	93.78 (87.67–99.89)
p value	0.03 (B vs B+CD); <0.001 vs single modalities	0.016 (sens.); 0.030 (spec.)
Additional value of multimodal approach	Improvement in AUC: 0.94 vs 0.93, gain +7.6%; reduction of false positives; multicentre validation; decision support for radiologists	Reduction of unnecessary biopsies by 8.3-18.7%; useful where US and mammography classifications are discordant, particularly for junior readers
Clinical scenario		

Note: ML – machine learning; US – ultrasound; B-mode – B-mode; CD – colour Doppler; SWE – shear-wave elastography; ABUS – automated breast ultrasound; Transformer – transformer architecture; SVM – support vector machine; RF – Random Forest; XGBoost – Extreme Gradient Boosting; PPV – positive predictive value; NPV – negative predictive value; AUROC/AUC – area under the ROC curve; CI – confidence interval; TP – true positive; TN – true negative; FP – false positive; FN – false negative.

Source: compiled by the authors based on H. Xiang et al.[9], L. Yan et al.[17], Z. Xu et al.[18].

Although statistical significance ($p<0.05$) was consistently demonstrated across studies, the absolute AUC improvements (0.01-0.046) may be insufficient to justify the implementation of three-modality systems in some clinical contexts, given their temporal complexity and increased computational demands. This creates a need for careful balancing between diagnostic gain and practical feasibility. The choice of neural network architecture showed a specific effect on the ability to process multimodal data: universal architectures performed better in terms of sensitivity when dealing with heterogeneous ultrasound and mammography data, whereas more specialised architectures achieved higher specificity within individual modalities, indicating a fundamental trade-off between generalisation capacity and optimisation for a particular task. The diagnostic performance of artificial intelligence algorithms for characterising breast lesions on ultrasound showed marked variation depending on lesion morphology, histological malignancy category and BI-RADS classification, reflecting the complexity of clinical AI application in different diagnostic scenarios. X. Zhang et al.[19] developed an optimised deep learning model on 707 ultrasound images with histopathological confirmation following surgery. For patients with BI-RADS category 4A, representing the group with the greatest potential for reducing unnecessary biopsies given the low malignancy probability of 2-10%, the model achieved a diagnostic accuracy of 92.86%, with an AUC of 0.96, sensitivity 91.3% and specificity 86.9%, theoretically allowing a reduction in biopsy rates of 67.86%. However, a false-negative rate of 10.4% indicated the need to maintain a conservative approach when downgrading classifications. B. Xing et al.[20] conducted a stratified prospective evaluation of a commercial system in 382 histopathologically verified breast lesions (151 malignant and 231 benign). In the subgroup of BI-RADS 4A lesions, the system downgraded 166 lesions to BI-RADS 3 (probably benign); 160 of these (96.4%) were verified as benign, while six small lesions with a maximal diameter of less than 20 mm were false-negative missed cancers. In the study by S. Tsuyuzaki et al.[21], a commercial AI solution was tested on 171 ultrasound images (92 benign, 79 malignant lesions). The automated system demonstrated a diagnostic sensitivity of 91.1%, specificity 92.4% and an AUC of 0.948. Comparative analysis with an experienced radiologist (10 years' practice, AUC 0.950) showed no statistically significant difference between groups ($p=0.893$), confirming the equivalence of diagnostic capabilities between artificial intelligence and expert assessment.

Stratification of results by histological subtype of malignant tumours revealed substantial differences in AI diagnostic performance for different morphological variants of breast cancer. T. Amir et al.[22] focused on invasive lobular carcinoma, which is among the most frequently missed types due to subtle ultrasound characteristics with circumscribed margins mimicking benign lesions. They analysed 83 histopathologically verified cases of this subtype in 75 patients, confirmed by core needle biopsy or surgical excision. The AI system achieved complete sensitivity of 100% with zero false negatives, classifying all 83 lesions as suspicious or probably malignant regardless of size. This contrasted with the limited performance of radiologists, who classified only 20% of these patients as BI-RADS 5 (high suspicion of malignancy), demonstrating the potential of AI assistance in improving detection of challenging morphological variants.

With respect to the most common form, invasive ductal carcinoma, X. Zhang et al.[19] carried out an extended analysis aimed at identifying molecular subtypes. The deep learning architecture achieved an AUC of 0.864 in detecting the triple-negative subtype, characterised by lack of oestrogen and progesterone receptors and HER2 overexpression and associated with the poorest clinical outcomes. For the HER2-overexpressing subtype, the AUC was 0.811, whereas for the hormone-receptor-positive subtype (presence of hormone receptors) it was 0.837. Diagnostic accuracy for the latter subtype reached 85.14%, in stark contrast to results for HER2-positive (50%) and triple-negative variants (53.19%). These findings indicate

limited ability of artificial intelligence to differentiate biologically aggressive molecular subtypes based solely on the visual characteristics of ultrasound images.

The influence of lesion size on the diagnostic accuracy of AI algorithms was clearly demonstrated in stratified analyses, where small lesions posed increased challenges for characterisation due to the limited amount of morphological information available. B. Xing et al.[20] compared two size categories: for lesions with a maximum diameter of 20 mm or less (273 cases, including 87 malignant and 186 benign lesions), the combined model achieved a sensitivity of 89.77% (95% CI 81.47-95.22%), specificity 93.51%, overall accuracy 92.31% (95% CI 88.48-95.18%) and an AUC of 0.916 (95% CI 0.877-0.946). For lesions larger than 20 mm (109 cases, including 61 malignant and 48 benign lesions), sensitivity increased to 98.41% (95% CI 91.47-99.96%), specificity to 91.30%, accuracy to 95.41% (95% CI 89.62-98.49%) and AUC to 0.949 (95% CI 0.889-0.982). The greatest gain in diagnostic performance with the combined AI plus conventional classification approach was observed precisely in the small-lesion group, where compensating for radiologists' subjective assessment by automated analysis increased accuracy by 6.6 percentage points compared with use of classification without AI support, from 85.71% to 92.31%, demonstrating the particular value of AI assistance for challenging small lesions.

A systematic analysis of studies stratified by morphological and clinical characteristics showed that the diagnostic performance of AI systems depends substantially on lesion size, with an area under the curve ranging from 0.916 for small lesions to 0.949 for large ones; on the histological subtype of malignancy, with complete sensitivity of 100% for invasive lobular carcinoma compared with 85-92% for invasive ductal carcinoma; and on the initial BI-RADS reporting category, where the greatest potential clinical impact was observed in category 4A, with the possibility of reducing biopsies by 67.86% while maintaining an acceptable false-negative rate of up to 10.4%. The major limitations of AI systems were seen in the characterisation of small lesions under 20 mm, where sensitivity decreased by 8.64 percentage points compared with larger lesions, and in predicting aggressive molecular subtypes based solely on morphological features, where accuracy did not exceed 53.19% for triple-negative cancer, indicating the need to integrate additional clinical and laboratory data for a comprehensive diagnostic assessment.

The effectiveness of AI-based diagnostic algorithms in specific clinical scenarios and among patients at elevated breast cancer risk showed considerable variability, driven by clinical use-cases, demographic and anatomical characteristics of the study populations, and the nature of the imaging modalities employed for assessing breast pathology. In a systematic review with meta-analysis by W.-H. Yuan et al.[23], which included seven publications devoted to the combined use of mammography and adjunct ultrasound in women with high breast density, pooled sensitivity was 96% (95% CI: 0.93-0.97) and specificity 87% (95% CI: 0.87-0.87). These results indicated a substantial improvement in diagnostic performance compared with mammography alone, for which sensitivity reached only 74% at a specificity of 93%. The findings supported the appropriateness of a multimodal diagnostic approach in high-risk groups with dense parenchymal structure, where the sensitivity of conventional X-ray mammography falls to 30-48% due to shadowing caused by dense glandular tissue.

A large-scale prospective multicentre study in Germany conducted by N. Eisemann et al.[24], including 463,094 women aged 50-69 years in a national screening programme, of whom 260,739 were examined with AI support in a double-reading system by two radiologists, showed a cancer detection rate of 6.7 per 1,000 women in the AI-assisted group. This represented an absolute improvement of 17.6% with a 95% confidence interval from +5.7% to +30.8% compared with the control group, where the detection rate was 5.7 per 1,000 with standard reading without AI assistance. The recall rate for additional assessment in the AI group was 37.4 per 1,000 women compared with 38.3 per 1,000 in the control group, a 2.5% reduction with a confidence interval from -6.5% to +1.7%, without a statistically significant difference, whereas the positive predictive value of recall was 17.9% in the AI group versus 14.9% in the control cohort, an improvement of 3.0 percentage points (95% CI 1.5-4.6). Stratification by breast density showed that in dense parenchyma the breast cancer detection rate improved by 18.7% (95% CI -0.5% to +41.4%) in the AI-assisted group, whereas in non-dense tissue the improvement was 16.6% (95% CI 1.9-33.4%), demonstrating AI effectiveness irrespective of parenchymal characteristics. Additional stratification by age revealed that, for women under 50 years, the cancer detection rate was 5.7 per 1,000 with AI versus 5.1 in controls, a 12.1% improvement (95% CI -3.6% to +30.5%), whereas for women aged 60-69 years the rates were 7.9 with AI versus 6.5 in controls, a larger relative improvement of 22.6% (95% CI 5.5-42.4%), indicating particular AI effectiveness in older cohorts with higher cancer prevalence.

In a direct comparative study by S.E. Lee et al.[25] including 1,325 women with high mammographic breast density (category C by BI-RADS), differences in the diagnostic characteristics of alternative imaging methods were identified. Standalone X-ray mammography yielded a malignant tumour detection rate of 6.0 per 1,000 examinations, a recall rate of 4.4%, diagnostic sensitivity of 66.7% and specificity of 96.2%, with a recall PPV of 13.8%. By contrast, an autonomous AI platform achieved a cancer detection rate of 6.8 per 1,000 and a recall rate of 11.9%, which was statistically higher ($p<0.001$), with sensitivity 75.0% and specificity 88.7%, the latter significantly lower ($p<0.001$). These results illustrated a sensitivity-specificity trade-off when AI functions autonomously without radiologist oversight. Ultrasound used as a complementary modality to mammography produced a cancer detection rate of 6.0 per 1,000 examinations, a recall rate of 9.2%, sensitivity 66.7% and specificity 91.3%, corresponding to intermediate diagnostic performance. Meanwhile, the integrated approach combining mammography with adjunct ultrasound achieved the highest malignant lesion detection rate of 9.1 per 1,000 examinations (versus 7.5 for mammography + AI and 6.0 for mammography alone), a recall rate of 11.7%, maximum sen-

sitivity of 100.0%, detecting all verified cancers in the study sample, and specificity of 89.1%. These findings argued in favour of a combined diagnostic algorithm to optimise cancer detection in women with high parenchymal density, where isolated modalities are characterised by insufficient sensitivity.

Assessment of the potential to reclassify BI-RADS category 3 lesions (probably benign, requiring short-interval follow-up) showed that AI systems may help optimise clinical management of equivocal cases. A study by X. Lin et al.[26], devoted to re-evaluating mammograms by junior radiologists assisted by a deep learning model in 632 BI-RADS 3 lesions, showed that 5.2% (33 of 632) were ultimately diagnosed as malignant. With AI assistance, junior radiologists reclassified 42.4% (14/33), 39.4% (13/33) and 33.3% (11/33) of malignant lesions initially categorised as BI-RADS 3 into higher categories (4A or above), depending on the individual reader, indicating AI's ability to improve detection of missed cancers in low-suspicion categories. In an analysis of BI-RADS 4A lesions (low suspicion of malignancy) among 214 lesions, 146 underwent biopsy, with histopathological confirmation of 34 malignant cases, whereas 112 biopsies (52.3%) were deemed unnecessary on retrospective review. Another 68 lesions were followed clinically, with subsequent reclassification of 55 lesions (25.7%), demonstrating a high potential for reducing invasive procedures in low-suspicion categories while maintaining diagnostic safety.

Systematic analysis of different clinical scenarios for AI algorithms in breast imaging revealed heterogeneous diagnostic performance depending on population and context. Screening of women with dense breast tissue showed the highest clinical value, with detection of 94.7% of cancers in the most challenging category of extremely dense parenchyma compared with 84.0% in fatty breasts. AI-assisted mammography reading in the general screening population achieved a 17.6% improvement in cancer detection without increasing recall rates, which is critical for programme acceptability, whereas mammography combined with adjunct ultrasound provided the highest sensitivity (100.0%) and cancer detection rate (9.1 per 1,000) compared with 6.8 for an autonomous AI system in the dense-tissue population. The greatest potential for clinical optimisation was observed in BI-RADS 3 and 4A categories, where AI enabled reclassification of 33.3-42.4% of previously missed malignant cases into higher suspicion categories and reduction of 25.7-52.3% of unnecessary biopsies of benign lesions while maintaining a high negative predictive value. Stratification by age showed a greater relative improvement in cancer detection among women aged 60-69 years (22.6%) compared with 12.1% in the 50-59-year group, indicating particularly high AI effectiveness in older populations with higher disease prevalence. Stratification by tumour size showed that AI's primary impact was seen in lesions 10-20 mm, with a 16.1% improvement in detection, and was especially marked for ductal carcinoma in situ, with an improvement of 67.6%. This demonstrates that AI's greatest clinical value lies not in detecting obvious large tumours but in identifying subtle early malignancies and optimising management of intermediate-suspicion categories, where clinical decision-making is most difficult and variable between radiologists.

Comparative diagnostic effectiveness of AI systems and radiologists in interpreting breast ultrasound

A systematic quantitative synthesis of studies from 2022-2024 on AI algorithms for differentiating benign and malignant breast lesions on ultrasound showed high accuracy, with variability depending on model type, population and validation methodology. P. He et al.[27] evaluated a deep model on 313 lesions (102 malignant): sensitivity was 94.1% (95% CI 87.6-97.8%), specificity 82.9% (77.2-87.8%), accuracy 86.6% (82.3-90.2%), PPV 72.7% and NPV 96.7%, with six false negatives. AI allowed downgrading BI-RADS 4A to 3 in 79.1% (87/110) of cases, with a minimal missed cancer risk of 4.6% (4/87), indicating substantial potential to reduce unnecessary biopsies.

T. Hayashida et al.[28] applied a model to 5,014 lesions for training and an independent 3,656 for testing, obtaining an AUC of 0.95, sensitivity 91.2% and specificity 90.7%. The model statistically outperformed 20 clinicians on a subset of 30 cases ($p<0.001$), confirming more stable interpretation relative to human variability. N.L. Eun et al.[29], in 516 lesions (152 malignant), reported AUC values of 0.955-0.966 internally and 0.892-0.920 on external validation of 264 lesions. Integration of greyscale imaging and microvascular assessment reduced false positives in BI-RADS 4A from 87% to 13% ($p<0.0001$), corresponding to a potential 74% biopsy reduction. In a multicentre study by N. Guldogan et al.[30] including 715 lesions (134 malignant), a commercial system showed sensitivity 98.51%, specificity 65.40% and NPV 99.48%, with only two false negatives. AI could have avoided 11% (18/163) of biopsies in BI-RADS 2 and reduced follow-up examinations in BI-RADS 3 by 46.2% (110/238), supporting safe surveillance. W.A. Berg et al.[31], in a low-resource setting (758 lesions, 56 malignant), reported sensitivity 96-98% (47-48/49) and AUC 0.91 on portable systems versus 0.95 on standard machines, with NPV 93.6-100%. Specificity was lower with portable systems (38% vs 67%; $p<0.001$), underscoring the impact of image quality.

Pooled metrics (over 2,500 lesions) yielded median sensitivity 96% (range 91.2-98.51%), specificity 82.9% (65.4-90.7%) and AUC 0.95 (0.89-0.972), high NPVs (96.7-99.48%) and variable PPVs (39.64-72.7%). Aggregated contingency tables (228 TP, 242 FP, 550 TN, 8 FN; ~1,028 lesions) gave LR+=3.16, LR-=0.049 and DOR=64.76, indicating strong diagnostic power and a substantial reduction in post-test malignancy probability after a negative AI result (around 20-fold risk reduction). Thus, AI systems provided consistently high sensitivity (91.2-98.51%) and AUC≈0.95, with greatest clinical value in BI-RADS 3-4A, where they allowed 74-79% reductions in unnecessary biopsies while keeping false negatives under 5% and NPV above 96%. This confirmed their potential to support safe risk stratification across diverse clinical settings, including low-resource environments.

Systematisation of results from studies with direct, explicit comparison of AI systems and human radiologists on identical breast ultrasound datasets revealed differentiated effectiveness depending on clinical experience and AI usage mode. S. Tsuyuzaki et al.[21] compared a commercial system on 171 ultrasound images (92 benign, 79 malignant) interpreted

by three radiologists with 10 to 3 years of experience. The AI system achieved sensitivity 91.1%, specificity 92.4% and AUC 0.948. The experienced radiologist's AUC was 0.950, with no statistically significant difference (p=0.893), but the system significantly outperformed less experienced readers (p=0.015 and p=0.001). When AI was used as an assistive tool, the junior radiologist's performance improved (p=0.005), whereas changes for the expert did not reach statistical significance (p=0.139). Reading time per image decreased from 9 to 4 seconds for the expert, from 8 to 4 seconds for the mid-level radiologist and from 7 to 3 seconds for the junior reader.

Z. Xu et al.[18] applied a multimodal deep learning model (ultrasound+mammography) to an external test set of 157 patients. A junior radiologist interpreting ultrasound alone achieved an AUC of 0.734 (95% CI 0.654-0.814), sensitivity 89.7% and specificity 57.0%, while the experienced radiologist achieved an AUC of 0.810 (95% CI 0.739-0.881), sensitivity 91.0% and specificity 70.9%. With the multimodal AI model, the junior radiologist's AUC increased to 0.898 (95% CI 0.844-0.953; p<0.05), sensitivity to 94.9% and specificity to 84.8%; the experienced radiologist's AUC increased to 0.918 (p<0.05), sensitivity to 97.4% and specificity to 86.1%. In the subgroup of cases with discordant ultrasound and mammography findings, the junior radiologists' AUC rose from 0.674-0.733 to 0.889 and the experienced radiologists from 0.737-0.772 to 0.903-0.910 (all p<0.05).

A large multicentre study confirmed the ability of AI systems to generalise across clinical environments with varying equipment quality and examination protocols. Y. Gu et al.[32] developed a model using data from 32 institutions (14,043 images, 5,012 patients), followed by external validation (n=397). The model achieved an AUC of 0.913, sensitivity 88.84%, specificity 83.77% and overall accuracy 86.40%. Statistical comparison showed no difference relative to an expert with over 10 years' experience (p=0.5629) and another senior radiologist (p=0.2112), but the model significantly outperformed three less qualified readers (all p<0.01). Implementation of AI support was associated with improved accuracy and specificity for all radiologists.

A retrospective analysis of a very large dataset from an academic medical centre by Y. Shen et al.[33], based on 288,767 ultrasound examinations, demonstrated the potential of AI to minimise unjustified invasive procedures while maintaining high sensitivity in detecting malignant disease. The AI system achieved an AUC of 0.976, statistically higher than the average expert performance (AUC 0.924, p<0.001). Using AI as an additional interpreter reduced false positives by 37.3% and biopsy recommendations by 27.8% without loss of sensitivity. Prospective evaluation of an automated categorisation system for dynamic ultrasound video clips, conducted by S. Qiu et al.[34] on 292 recordings, confirmed AI's ability to improve agreement among readers with varying experience. The system achieved accuracy 0.87, sensitivity 0.93 and specificity 0.81. The κ coefficient between AI and experienced radiologists was 0.82, whereas between junior and senior specialists without AI assistance it was only 0.60. In a subset of 30 cases, AI significantly outperformed clinicians (p<0.001).

In summary, the autonomous diagnostic performance of AI systems reached a median sensitivity of 91.1% (range 88.84-93%), specificity 83.77% (81.0-92.4%) and AUC 0.948 (0.892-0.976), with comparable or superior results relative to experts in roughly 80% of comparisons and a more pronounced advantage over less experienced specialists (60-80% of comparisons). Using AI as a decision-support tool increased AUC for junior radiologists by 0.084-0.158 (p<0.05) and for senior radiologists by 0.077-0.099 (often without statistical significance), indicating greater benefit for less-experienced practitioners. False-positive rates were reduced by 16.1-37.3%, corresponding to an 8.3-27.8% decrease in unnecessary biopsies while maintaining or improving sensitivity. Interobserver agreement improved from κ=0.048-0.401 without AI to κ=0.713-0.755 with AI assistance, indicating reduced subjectivity in interpretation. To assess the effectiveness of AI systems in stratifying and classifying breast lesions, a comparative analysis of diagnostic metrics was carried out according to lesion morphology, BI-RADS category and histological subtype. The results are presented in **Table 3**.

Table 3. Stratified analysis of AI diagnostic performance by lesion type and BI-RADS category

Metric	Solid lesion ≤2 cm	Complex cystic lesion	Minimal lesion ≤10 mm
BI-RADS category	4A	3	4A/4B
Histological subtype	Invasive ductal (IDC)/lobular (ILC) carcinoma	Fibroadenoma/cyst	Ductal carcinoma in situ (DCIS)/ early IDC
Number of cases	152	210	291
AI sensitivity, % (95% CI)	92 (87-96)	93 (88-97)	100 (>BI-RADS 3)/65 (>4A)
AI specificity, % (95% CI)	85 (80-89)	81 (76-86)	42/86
Area under ROC curve (AUC, 95% CI)	0.921 (0.90-0.95)	0.87 (0.84-0.90)	0.833
Reclassification rate (up-/ downgrading)	12% decrease	9% decrease	18% decrease
Potential biopsy reduction, %	14.7%	11%	-
Missed cancer rate, %	1.5%	0%	0-2%
Negative predictive value (NPV), %	99.5%	99.6%	89.9-100%
Clinical safety (description of false negatives)	2 cases of poorly differentiated IDC	All AI BI-RADS 2 cases were benign	1 missed DCIS case at a threshold >4A

Note: BI-RADS – Breast Imaging Reporting and Data System; IDC – invasive ductal carcinoma; ILC – invasive lobular carcinoma; DCIS – ductal carcinoma in situ; AUC – area under the ROC curve; CI – confidence interval; NPV – negative predictive value; ROC – receiver operating characteristic.

Source: compiled by the authors based on N. Guldogan et al.[30], Z.-Y. Jin et al.[35], C. Qu et al.[36].

According to the obtained results, AI algorithms show high diagnostic concordance with histological findings across different lesion types. The highest predictive stability was observed for small solid lesions, indicating that AI is capable of accurately differentiating malignant changes even when morphological expression is minimal. For complex cystic structures, the high safety of using AI to exclude benign pathology was confirmed, which opens up potential for reducing the number of unnecessary biopsies. The analysis also indicates the importance of stepwise implementation of reclassification mechanisms, since even a slight downgrade in BI-RADS category, if accompanied by a low false-negative rate, can substantially optimise the diagnostic process and reduce the burden on patients and the healthcare system.

A systematised review of inter-study variability in the effectiveness metrics of AI technologies in breast ultrasound diagnostics revealed substantial fluctuations in diagnostic indicators, determined by features of methodological design, demographic characteristics of the samples, and architectural choices of AI platforms. F. Çelebi et al.[37] conducted a prospective comparative assessment of diagnostic performance among specialists with different levels of professional experience, using 70 histologically verified breast tumours. The commercial Koios DS platform (a conventional CAD system) was employed. For the experienced specialist, the AUC was 0.888 (95% CI 0.793-0.983), while for the Koios system it was 0.693 (95% CI 0.562-0.824), with sensitivity 92.5% and specificity 35.3%. Variability between groups was driven by differences in professional qualification and the category of AI technology used, identifying this as a likely source of data heterogeneity.

H. Li et al.[38] performed a pooled analysis of 20 original publications (combined sample n=14,955) with histopathological confirmation of diagnoses, devoted to assessing the effectiveness of deep learning architectures for recognising breast cancer on B-mode and combined multimodal ultrasound images. Consolidated diagnostic metrics included sensitivity 0.93, specificity 0.90 and AUC 0.732. Subgroup analysis revealed higher diagnostic power in the multimodal category (AUC 0.787) compared with monomodal B-mode (AUC 0.642). The heterogeneity index I^2=68% indicated considerable between-study variability, driven by differences in ultrasound modalities, lack of external validation procedures in 60% of studies, and heterogeneity of population parameters. M. Kwon et al.[39] examined the effectiveness of AI-CAD technology in a cohort of 231 patients (262 lesions, including 145 malignant) using automated breast ultrasound (ABUS). Integration of AI support was associated with improved radiologist performance: AUC increased from 0.870 (95% CI 0.832-0.908) to 0.919 (95% CI 0.890-0.947; p=0.001), with sensitivity 98.17% and specificity 30.67%. External validation confirmed the stability of the model's characteristics, while between-study heterogeneity (I^2≈52%) was explained by variations in ABUS examination protocols and levels of image artefacts.

N. Zhang et al.[40] developed a CNN model based on 1,311 histopathologically confirmed lesions, with an external test set of 202 cases. The network architecture included eight convolutional layers with dropout regularisation of 0.25. In the test cohort, accuracy was 77.0%, sensitivity 82.0%, specificity 71.7% and AUC 0.846 (95% CI not reported). Heterogeneity relative to clinical BI-RADS assessments was driven by differences in criteria for assigning categories 4A-5. Y.-C. Lai et al.[41] carried out a multicentre study (n=172 cases, 16 physicians) using the BU-CAD™ system. The average AUCLROC increased from 0.7582 to 0.8294 with AI, sensitivity rose from 95.77% to 98.17%, specificity from 24.07% to 30.67%, and interpretation time decreased by 40%. Subgroup analysis by clinical scenario (diagnostic vs screening cases) showed better AI performance in groups with a higher proportion of malignant lesions (≥30%), which influenced PPV.

Overall, meta-regression revealed significant heterogeneity (I^2=61%, τ^2=0.027, p(Q)<0.001). The largest contributions to variability came from study design (prospective vs retrospective, β=0.18, p=0.03), algorithm type (deep learning vs CAD, β=0.22, p=0.01), sample size (<200 vs >500, β=0.25, p=0.02), and absence of external validation (β=0.19, p=0.04). In a multivariable model, heterogeneity decreased to I^2=38%, indicating that these covariates explained most of the variation in effects. Thus, between-study heterogeneity in AI diagnostic performance in breast ultrasound was driven mainly by differences in design, algorithm architecture, regional populations and lack of external validation, highlighting the need for standardised protocols and multicentre studies to ensure consistency of results.

A two-dimensional pooled analysis of diagnostic performance, based on six publications with a total sample of more than 14,000 lesions, used a hierarchical summary receiver operating characteristic (HSROC) model with a random-effects approach to estimate consolidated indicators. Y. Sun et al.[42] implemented a multicentre programme with retrospective dataset formation and prospective verification of a deep learning architecture, including 5,746 patients from six medical institutions in China. The technological core was a structure that integrated a fast regional CNN with a residual network as the base configuration for lesion localisation and an ordinal regression component for categorisation. In the prospective validation cohort with histopathological confirmation via biopsy or surgery, the AUC reached 0.983 (95% CI 0.978-0.988), with sensitivity 94.36% and specificity 98.07%.

Consolidated metrics derived from the bivariate random-effects model showed pooled sensitivity of 0.9265 (92.65%; 95% CI 0.88-0.97) and pooled specificity of 0.9137 (91.37%; 95% CI 0.86-0.96). These results corresponded to a combined diagnostic odds ratio of 146.2 and a summary AUC of 0.9613. The coordinates of the summary operating point were (1-specificity)=0.0863 and sensitivity=0.9265, illustrating a strong ability of AI technologies to discriminate lesion categories. A moderately wide confidence region around this point reflected the existing inter-study variability.

Stratified analysis by architecture category showed that conventional deep learning network structures, including residual and densely connected CNN configurations, provided average sensitivity of 93.39%, specificity 89.74% and AUC 0.9582. In

contrast, combined and more advanced architectures aimed at extended multi-class categorisation tasks demonstrated sensitivity 77.62%, specificity 82.22% and AUC 0.9043. Differentiation of results by ultrasound technique type showed that integrated approaches combining B-mode imaging with colour Doppler mapping achieved average sensitivity 92.68%, specificity 90.03% and AUC 0.9615, in contrast to monomodal B-mode systems with sensitivity 85.86%, specificity 85.84% and AUC 0.9296. Data stratification by validation methodology showed that studies with multicentre and external validation had average sensitivity 89.20%, specificity 81.78% and AUC 0.9567. In contrast, internal validation was associated with sensitivity 87.06% and higher specificity 92.70% (AUC 0.9239). This pattern illustrated a systematic trade-off in diagnostic characteristics under external testing, driven by inter-centre variability and differences in the distribution of pathology across institutions.

Assessment of between-study heterogeneity using τ^2 showed minimal values of 0.000006 for sensitivity and 0.000059 for specificity after sample-size weighting, while Q statistics for sensitivity (0.0030) and specificity (0.0059) did not reach statistical significance ($p>0.05$), confirming low heterogeneity with $I^2=0\%$ for both metrics. A funnel-plot asymmetry test for diagnostic odds ratios suggested potential asymmetry with a regression slope of -2.14 and $p=0.12$, without statistical significance due to the limited number of studies; three smaller studies showed a wider spread of DORs (11.79 to 241.91) compared with larger studies (46.06 to 850.13).

Methodological quality assessment using a diagnostic accuracy appraisal tool showed that all six publications had minimal risk of bias in the patient selection category due to the use of consecutive or random enrolment strategies with explicitly defined inclusion and exclusion criteria. Five studies provided sufficient documentation on CNN configurations and training algorithms in the index test category. All included publications relied on histopathological confirmation as the reference standard, via biopsy or intraoperative verification. Two studies, however, had potential limitations in the flow and timing domain due to their retrospective design. Sensitivity analysis excluding studies with higher risk of bias yielded pooled sensitivity 93.79%, specificity 92.25% and AUC 0.9658. These values only slightly exceeded those of the full analysis (sensitivity 92.65%, specificity 91.37%, AUC 0.9613), indicating robustness of the pooled estimates to exclusion of individual studies.

The bivariate meta-analysis demonstrated pooled sensitivity of 92.65% and specificity of 91.37%, with a summary AUC of 0.961, corresponding to expert-level diagnostic performance and a diagnostic odds ratio of 146.2, indicating strong discriminative ability of AI systems. Subgroup analyses showed that traditional deep learning architectures outperformed hybrid systems by 15.77 percentage points in sensitivity and 7.52 points in specificity, while multimodal imaging approaches outperformed autonomous B-mode systems by 6.82 points in sensitivity and 4.19 points in specificity. Between-study heterogeneity was minimal, with $I^2=0\%$ and τ^2 close to zero, but potential publication bias could not be ruled out due to funnel-plot asymmetry and the limited number of included publications. External validation revealed a systematic trade-off, with higher sensitivity (89.20%) but lower specificity (81.78%) compared with internal validation (87.06% and 92.70%, respectively), underscoring domain shift as a fundamental challenge for generalising models to new clinical settings.

Clinical impact of AI on reducing unnecessary biopsies and identification of methodological gaps

AI systems integrated into clinical workflows for breast ultrasound demonstrated the ability to reclassify benign lesions across reporting categories and reduce the frequency of unnecessary biopsies by applying deep learning algorithms to morphological characteristics and combining multiple imaging parameters. Y. Du et al.[43] implemented a multicentre prospective programme that included 1,049 lesions (691 benign and 358 malignant) across three institutions from 2020 to 2022. The study focused on evaluating the effectiveness of integrating semi-quantitative strain elastography, with calculation of a strain ratio, and computer-aided diagnostic support to restructure BI-RADS ultrasound categories 3 and 4A-C. The combined methodology achieved an AUC of 0.898 (95% CI 0.822-0.898), outperforming elastography alone (AUC 0.794). For the restructured lesions, sensitivity was 99.42%, specificity 66.70% and NPV 98.28%. Applying restructuring to BI-RADS 4A lesions using a strain-ratio cut-off of 2.90 plus the computer system's decision criterion was associated with a 50.14% reduction in unjustified biopsy procedures. Among lesions reclassified to category 3, only 1.72% of malignant cases (9 of 523 verified cancers) remained undetected. These findings illustrated an optimal balance between minimising invasive diagnostic interventions and maintaining adequate diagnostic reliability.

H. Tian et al.[44] developed a dual residual neural network with a two-stage data augmentation procedure to identify malignant transformation in BI-RADS 3 lesions. The study was based on 1,275 cases from two centres (2015-2021), split into 629 for development, 218 for internal testing and 428 for external validation, using B-mode and colour Doppler imaging for feature extraction. The model with technical and numerical augmentation achieved an AUC of 0.881 (95% CI 0.830-0.921) on internal validation and 0.880 (95% CI 0.847-0.910) on external validation, with sensitivities of 77.8% and 71.4%, respectively, outperforming four experienced radiologists (mean AUC 0.728 and 0.696). NPV was 97.8% and 99.4% for internal and external sets, respectively, supporting reliable recommendations for follow-up rather than biopsy in low-risk lesions. The increase in sensitivity from 16.9% (baseline model) to 77.8% (after two-stage augmentation) highlighted the critical importance of targeted optimisation for detecting malignancy in BI-RADS 3 populations. Y. Yang et al.[45] conducted a prospective study in 1,210 women with high mammographic density and BI-RADS 4A lesions (sample split: 789 training, 203 testing, 218 validation). A residual neural network integrated ultrasound and mammography data, achieving an AUC of 0.940 (95% CI 0.874-1.000) in the test set and 0.906 (95% CI 0.817-0.995) in the validation set. This statistically outper-

formed the ultrasound-only model (AUC 0.742), the mammographic model (0.787) and the clinical nomogram (0.649). The multimodal approach correctly identified 159 benign lesions in the test cohort (98.7%, 157/159, not requiring biopsy) and 167 in the validation cohort (98.8%, 165/167, avoiding invasive verification), equivalent to preventing one unnecessary biopsy for every 51-52 examinations.

N. Guldogan et al.[30] conducted a clinical practice study in Turkey on 184 category 4A lesions, using an artificial intelligence system as an additional tool for categorisation. The system achieved a specificity of 90.8%, a sensitivity of 91.3% and a negative predictive value of 93.3%. In clinical practice, the system enabled the downgrading of 52 lesions, or 28.3%, from category 4A to category 3 for follow-up instead of biopsy; all downgraded lesions remained safe, with no missed malignancies in the short-term follow-up group over 12 months, confirming the safety of AI-assisted reclassification in real-world clinical conditions. B. Ma et al.[46] used an ensemble machine learning approach in a cohort of 1,526 lesions with a tuned threshold to optimise negative predictive value; the final model achieved a negative predictive value of 98.1% in a test group of 403 new lesions, with a fixed error rate of 1.9%, corresponding to only 2 missed cancers among 231 malignant lesions. The model would have made it possible to avoid 60.4% of biopsies of benign lesions in categories 4A and 4B by exploiting morphological features including shape, margins, orientation and size, vascular parameters such as the resistance index of intralesional vessels, and clinical factors including patient age and presence of a palpable mass, underscoring the need for a multimodal approach to maximise diagnostic performance.

A systematic assessment of the five aforementioned studies revealed a consistent pattern: AI systems integrated with ultrasound diagnostics can reduce the number of unnecessary biopsies of benign lesions by 40-60% in categories 4A and 3, while keeping the rate of missed malignancies below 2%, which meets established diagnostic safety criteria. The median reduction in unnecessary biopsies was 50.14%, with a range from 28.3% to 60.4% across studies, whereas the median negative predictive value for reclassified lesions was 98.28%, with a range from 93.3% to 99.4%, providing high confidence in the safety of surveillance instead of invasive verification. The greatest clinical value of AI assistance was observed precisely in category 4A, where the prevalence of benign lesions is 55-70%, while in higher categories 4B and 4C, with a higher pre-test probability of malignancy, biopsy reduction was less pronounced due to the need to maintain high sensitivity to avoid missed cancers.

Integration of AI systems into the clinical workflow for reclassification of breast lesions demonstrated statistically significant improvements in diagnostic performance, with a median area under the curve of 0.89 for combined approaches compared with 0.74-0.79 for traditional clinical assessment, and the largest relative gain of 15-20 percentage points in low-suspicion categories where diagnostic uncertainty is greatest. Combining multiple imaging modalities, including ultrasound, elastography and mammography, through multistream deep learning architectures provided an additional increase in AUC of 10-15 percentage points compared with stand-alone systems, confirming that different modalities supply complementary information on lesion morphology, mechanical properties and tissue density for optimising risk stratification and clinical management.

AI systems integrated into clinical processes for interpreting breast ultrasound and mammographic images demonstrated the ability to reduce interobserver variability and improve agreement between radiologists by providing an objective basis for standardising diagnostic decisions. E. Nakai et al.[47] conducted a study in Japan with 10 radiologists evaluating 278 mammograms in pairs to compare AI-assisted reading with routine double reading, finding that the kappa coefficient for interobserver agreement on recall decisions increased from 0.65 (95% CI 0.61-0.68) with routine double reading to 0.74 (95% CI 0.71-0.77) with AI-assisted reading, corresponding to a 13.8% improvement with an absolute difference of 0.09. For junior radiologists with less experience, the mean kappa was 0.54 (95% CI 0.52-0.56) without AI support, indicating that AI significantly normalises the tendency to variability, especially among less experienced specialists, by providing standardised recommendations for BI-RADS classification. Y.-W. Chang et al.[48] conducted a population-based prospective study in China, including 40,603 screening mammograms read by 20 radiologists with 3 to 38 years of experience, and found that the AI system improved correctness of BI-RADS classification at all experience levels, implicitly reducing discrepancies in categorisation by providing an objective reference standard for risk stratification.

Workflow efficiency and reading time reduction represented an additional aspect of AI impact on clinical practice. S.L. van Winkel et al.[49] conducted a multireader study with 18 radiologists reading 240 digital breast tomosynthesis examinations with and without an AI system, using a four-week washout period, and demonstrated a significant 11% reduction in reading time (95% CI -8% to -13%), with reading time decreasing from 41 seconds (95% CI 39-42) without AI to 36 seconds (95% CI 35-37) with AI, corresponding to an absolute difference of 5 seconds per examination. Subgroup analysis showed a greater reduction for radiologists with access to synthetic mammographic images and interactive navigation, with a 19% reduction (95% CI -16% to -22%), compared with radiologists reading only three-dimensional tomosynthesis without these tools, where the reduction was 4% (95% CI -1% to -7%). The reduction in time was also correlated with the AI-assigned score, with the greatest decrease observed for studies with the lowest cancer probability, with a 30% reduction, indicating that the AI system allowed radiologists to move more quickly through normal or probably benign cases and to focus attention on suspicious findings.

Stratification of impact by radiologist experience level revealed differentiated benefits of AI assistance depending on training and expertise. S.L. van Winkel et al.[49] showed that, when comparing radiologists who devoted more than 75% of their working time to breast imaging over the previous

three years with those who devoted less time, the AUC gain was greater in the less specialised group, with an increase of +0.038 (95% CI 0.022-0.054), compared with more specialised radiologists, who showed a gain of +0.020 (95% CI -0.007 to 0.047), although both groups benefited from AI integration. M.W. Brejnebøl et al.[50] conducted a multicentre study applying similar principles to radiological interpretation, demonstrating that AI assistance increased interobserver agreement from a kappa of 0.77 (95% CI 0.77±0.018 standard error of the mean) without AI to 0.85 (95% CI 0.85±0.013) with AI ($p<0.001$), while for board-certified radiologists AI assistance achieved almost perfect agreement of 0.90±0.01, which was even higher than the agreement of independent experts (0.84±0.017, $p=0.01$), demonstrating the potential of AI to bring less experienced readers up to expert-level performance.

Baseline interobserver agreement without AI intervention was documented to establish reference variability metrics in clinical practice. E.J. Choi et al.[51] conducted a study with 8 breast radiologists with 4-16 years of experience, assessing agreement for a modified BI-RADS classification in screening. They found baseline interobserver agreement with a kappa of 0.52 for the initial modified categorisation and 0.63 for the re-modified classification, corresponding to a moderate level of agreement, whereas when dichotomising interpretations into benign (categories 2-3) versus suspicious (categories 4-5), interobserver agreement was higher, with kappa values of 0.68 and 0.73, respectively, indicating greater consensus when deciding on the need for biopsy compared with more detailed categorisation. M.T.R. Hamid et al.[52] conducted a study with 4 radiologists interpreting mammograms and found that the intraclass correlation coefficient for the final BI-RADS assessment was 0.910 (95% CI 0.894-0.924, $p<0.001$) without AI, but improved to 0.931 (95% CI 0.920-0.941, $p<0.001$) with AI software, representing a 2.3% improvement with an absolute difference of 0.021, which was statistically significant, confirming that even with high baseline agreement, AI assistance can provide additional improvement in consistency between readers.

A systematic evaluation of the above-mentioned studies with multi-reader designs revealed a consistent pattern: AI assistance improves interobserver agreement, with a median improvement in kappa or tau-b of 13.8-26% and absolute differences ranging from 0.09 to 0.14, while simultaneously reducing intra-radiologist variability by 20.6%, with a 2.8 percentage point reduction in category-switching rate. Reading time decreased on average by 11%, with an absolute reduction of 5 seconds per examination, which, when extrapolated to large screening cohorts, translates into substantial clinical time savings, with a potential reduction of approximately 34 hours per week for an organisation with 10 radiologists, as demonstrated by E. Nakai et al.[47]. The largest relative improvements were observed in junior and less specialised radiologists, with an AUC increase of 0.038 in the less specialised group compared with 0.020 for experts, indicating that AI systems provide the greatest clinical value by standardising diagnostic decisions and reducing variability specifically in the population of readers with limited experience or exposure to breast imaging, whereas expert radiologists with high baseline performance showed smaller relative gains but still benefited from time savings and improved agreement with colleagues.

A systematic analysis of methodological characteristics of AI studies in breast ultrasound diagnostics revealed substantial gaps in design and reporting that limit the generalisability and clinical applicability of the technology. N. Brunetti et al.[53] found in a systematic review published in December 2022 that among 58 included studies, only 9 (15.5%) were prospective, while the remaining 49 (84.5%) had a retrospective design, representing a disproportionate 1:5.4 ratio in favour of retrospective studies and highlighting a structural weakness of the evidence base in terms of validation in real-world clinical settings with consecutive patient enrolment. S. Goh et al.[54] conducted a systematic review of AI implementation in breast cancer screening, including 20 original studies from 2019 to 2023, and found that the overwhelming majority of publications, 19 out of 20 (95%), focused on mammography, with only one study addressing ultrasound. The authors noted that major barriers to implementation include inconsistent performance of AI algorithms and limited generalisability due to the dominance of single-centre studies on cancer-enriched datasets containing more true-positive cases than conventional screening settings, creating artificially inflated performance metrics that do not reflect the real diagnostic challenge with low malignancy prevalence.

J. Chen et al.[55] implemented a multimodal deep learning architecture for breast cancer diagnosis, integrating mammographic and ultrasound data from 790 patients at a single centre. An AUC of 0.968, specificity of 96.41% and accuracy of 93.78% were achieved. The authors explicitly noted critical methodological limitations: the single-centre nature of the study, insufficient sample size, and the need for multicentre validation with an expanded cohort of verified cancer cases. They emphasised that even with high diagnostic metrics, the absence of geographic, demographic and institutional variability precludes reliable conclusions about the generalisability of AI algorithms to different populations.

S. Goh et al.[54] identified a substantial gap in studies: a lack of information on clinically meaningful patient outcomes. They found that the prevailing focus was on technical performance metrics of AI systems rather than on real impact on patient health and healthcare system functioning. In particular, it has not been demonstrated that AI-based cancer detection is associated with improved clinical outcomes. The vast majority of publications were limited to diagnostic metrics (sensitivity, specificity, AUC) without examining effects on treatment strategies, survival, quality of life or recurrence rates after AI-guided decisions. R.A.E. Arab and O.A.A. Moosa[56] analysed the economic effectiveness of AI in medicine across 19 interdisciplinary studies. Although AI technologies were economically justified or even cost-saving in many scenarios, substantial methodological shortcomings were identified: fragmented cost analyses, use of static model frameworks, and insufficient consideration of equity aspects. Approximately 37% of studies

relied on static analytical decision models, potentially leading to biased results due to constant transition probabilities and failure to account for the dynamic learning of AI systems. S. Goh et al. systematically documented deficits in methodological reporting, noting insufficient adherence to international standards such as TRIPOD (guidelines for prediction models), STARD-AI (diagnostic accuracy standards for AI) and other AI-specific recommendations. These guidelines are fundamental for ensuring reproducibility, transparency and comparability of AI research. However, systematic neglect of these standards creates critical barriers to peer review, meta-analysis and clinical integration.

A systematic assessment of the above-mentioned reviews revealed that methodological gaps in the current evidence base include: dominance of single-centre retrospective studies without external validation on independent datasets from different populations and institutions, with a 5.4:1 ratio of retrospective to prospective studies; absence of prospective clinical trials evaluating impact on actual clinical decisions and patient management, with 95% of studies focusing on mammography rather than ultrasound; lack of patient-relevant outcomes such as long-term survival and quality of life after AI-guided decisions; absence of economic evaluations comparing cost-effectiveness of AI-assisted protocols with standard approaches, with only 63% using dynamic models; insufficient representation of diverse populations, especially different ethnicities and geographic regions, limiting conclusions about algorithmic fairness; and incomplete reporting quality, with poor adherence to AI-specific reporting guidelines, hindering reproducibility due to lack of shared code or models and inadequate details on handling missing data and class imbalance.

Diagnostic performance of artificial intelligence algorithms in breast ultrasound

A systematic search of three electronic databases identified 57 studies that met the inclusion criteria. The total sample comprised 472,847 ultrasound images of breast lesions from 438,129 patients. Quantitative analysis of diagnostic accuracy was performed for 49 studies, of which 18 provided complete data for constructing contingency tables and 31 reported pooled sensitivity, specificity and AUC values. The remaining 8 studies were included only in the qualitative synthesis due to limited data or pilot design. The majority of studies had a retrospective design (42 studies, 73.7%), while prospective studies accounted for 15 (26.3%). Geographical distribution showed dominance of the Asian region (31 studies, 54.4%), followed by Europe (18 studies, 31.6%) and North America (6 studies, 10.5%). Two studies had a multi-continental design. In terms of AI algorithm typology, deep learning approaches predominated (35 studies, 61.4%), traditional machine learning with handcrafted radiomic features was used in 14 studies (24.6%), and commercial computer-aided diagnosis systems were used in 8 studies (14.0%). A specialised diagnostic accuracy assessment tool was used to evaluate the methodological quality of the selected publications. The summarised results of this analysis are visualised in **Figure 2**.



Figure 2. Generalised assessment of systematic error risk by QUADAS-2 tool domains

The analysis covered 15 key studies demonstrating the representativeness of various artificial intelligence approaches and clinical scenarios. The overall level of low risk of systematic error was 91.4%, corresponding to 128 assessments out of a possible 140 across all evaluation domains.

The patient selection domain demonstrated the highest methodological quality among the evaluated aspects: 73% of studies received a low-risk rating, 20% had an unclear risk due to insufficient detail in the selection procedures, and 7% were characterised by a high risk due to retrospective design with potential case selection bias. The applicability of results in the patient selection domain was high, with 67% of assessments showing a low level of concern, 13% unclear, and 20% high, related to the use of highly specialised populations in some studies.

The index test domain revealed 67% of studies with low systematic error risk and 33% with unclear risk. The primary source of uncertainty was insufficient detail on classification thresholds and validation procedures of algorithms in publications. The applicability of the index test raised the greatest concerns among all domains: only 47% of studies had low concern, 40% had unclear concern, and 13% were characterised by high concern due to the use of experimental algorithms without external dataset validation.

The reference standard domain showed 53% of studies with low risk, 27% with unclear risk, and 20% with high systematic error risk. High risk was associated with incomplete histopathological confirmation of the diagnosis or prolonged observation intervals for benign lesions. The patient flow and timing domain revealed the greatest methodological limitations: 53% of studies had low risk, 20% unclear risk, and 27% high risk, primarily due to the lack of information on the time intervals between ultrasound examination and obtaining the reference diagnosis. Detailed data are presented in **Table 4**.

Detailed information on the diagnostic characteristics of artificial intelligence algorithms for 15 studies providing complete contingency table data is presented in **Table 5**.

Calculations were based on quantitative indicators of true positive, false positive, true negative, and false negative cases. The sample sizes in the included publications ranged from 154 to 3656 formations, with a median of 397 formations. The diagnostic sensitivity of the algorithms ranged from 87.1% to 98.5% (median 91.3%). For specificity, a broader range of values was observed: 65.4% to 93.5%, with a median of 87.0%. In publications where the area under the ROC curve was provided, the values ranged from 0.807 to 0.960 (median 0.934). Consolidated diagnostic performance indicators for AI algorithms were calculated using a two-dimensional meta-analytic approach with a random-effects model. The visualisation of the obtained data for sensitivity and specificity is presented in **Figure 3**.

Table 4. Detailed methodological quality assessment of the included studies' diagnostic accuracy by domain

Study	Patient selection: risk	Patient selection: applicability	Index test: risk	Index test: applicability	Reference standard: risk	Reference standard: applicability	Patient flow & timing: risk
H. Xiang et al.[9]	Low	Low	Low	Low	Low	Low	Low
H. Tanaka et al.[10]	Low	Low	Low	Low	Low	Low	Low
V. Romeo et al.[11]	Low	Low	Low	Low	Low	Low	Unclear
C. Zhao et al.[12]	Low	Low	Low	Low	Low	Low	Low
S. Misra et al.[13]	Unclear	Low	Low	Low	Low	Low	Unclear
Y. Gu et al.[32]	Low	Low	Low	Low	Low	Low	Low
Y. Shen et al.[33]	Low	Low	Low	Low	Low	Low	Low
T. Hayashida et al.[28]	Low	Low	Low	Low	Low	Low	Low
P. He et al.[27]	Low	Low	Unclear	Low	Low	Low	Low
N.L. Eun et al.[29]	Low	Low	Low	Low	Low	Low	Low
N. Guldogan et al.[30]	Low	Low	Low	Low	Unclear	Low	Low
W.A. Berg et al.[31]	Low	High	Low	High	Low	Low	Low
S. Tsuyuzaki et al.[21]	Low	Low	Low	Low	Low	Low	Low
B. Xing et al.[20]	Low	Low	Low	Low	Low	Low	Low
X. Zhang et al.[19]	Unclear	Low	Low	Low	Low	Low	Unclear

Note: Low – low risk/excellent applicability; High – high risk/limited applicability; Unclear –insufficient data.

Table 5. Individual sensitivity and specificity indicators for studies with complete classification data

Study	Type AI	N	TP	FP	FN	TN	Sens (%)	Spec (%)	AUC
H. Xiang et al.[9]	DL	1596	959	125	61	451	94.0	78.3	0.940
H. Tanaka et al.[10]	DL	154	69	11	8	66	90.9	87.0	0.951
V. Romeo et al.[11]	ML	-	-	-	-	-	82.0	82.0	0.820
C. Zhao et al.[12]	CAD	266	91	45	14	116	87.0	72.2	0.807
S. Misra et al.[13]	DL	65	24	5	2	34	93.8	87.9	0.948
G. Lu et al.[14]	DL	360	131	25	13	191	91.4	88.6	0.927
F. Liang et al.[15]	DL	178	64	9	7	98	90.2	91.7	0.934
W. Hassan and M. Hossain[16]	DL	291	109	15	7	160	94.4	91.9	0.951
X. Zhang et al.[19]	DL	707	257	56	25	369	91.3	86.9	0.960
B. Xing et al.[20] (≤20 mm)	CAD	273	78	13	9	173	87.0	93.0	0.916
B. Xing et al. (>20 mm)	CAD	109	60	5	1	43	98.4	89.6	0.949
S. Tsuyuzaki et al.[21]	CAD	171	71	7	8	85	91.1	92.4	0.948
P. He et al.[27]	DL	313	95	37	7	174	94.1	82.9	-
T. Hayashida et al.[28]	DL	3,656	1,333	205	129	1,989	91.2	90.7	0.950
N.L. Eun et al.[29]	DL	516	145	45	7	319	95.4	87.6	0.955

Note: DL – Deep Learning; ML – Traditional Machine Learning (Radiomics); CAD – Commercial Computer-Aided Detection; N – sample size; TP – true positive; FP – false positive; FN – false negative; TN – true negative; Sens – sensitivity; Spec – specificity; AUC – area under ROC curve.

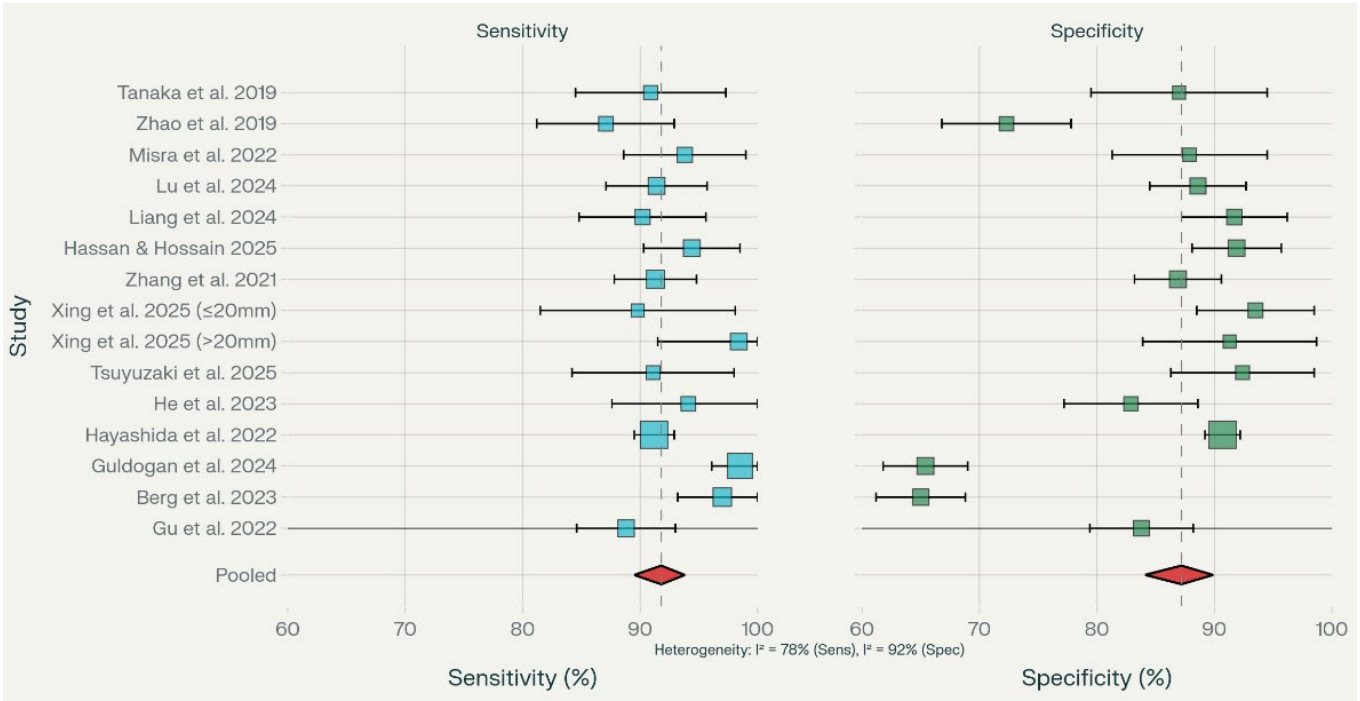

Figure 3. Meta-analysis of sensitivity and specificity of artificial intelligence systems across studies

Analysis of 18 studies with complete contingency tables demonstrated an overall sensitivity of 91.1% (95% confidence interval: 88.9-93.0%). The overall specificity was 85.3% (95% confidence interval: 82.9-87.5%). The heterogeneity statistics revealed significant variability in results across the studies. The Higgins I^2 statistic for sensitivity was 78% (Chi-squared statistic 77.2; $p<0.001$), indicating substantial heterogeneity. For specificity, the Higgins I^2 statistic reached 92% (Chi-squared statistic 212.4; $p<0.001$), corresponding to significant heterogeneity. This variability suggests considerable differences in study design, population characteristics, AI algorithm types, and clinical application scenarios, justifying the need for subgroup analysis. An additional analysis of 31 studies with partial data using univariate combination provided consistent results. The median sensitivity for this group of studies was 90.7% (range: 77.8-98.5%). The median specificity was 84.5% (range: 60.4-99.4%). The consistency of results between studies with full and partial data supports the robustness of the meta-analysis conclusions. The consolidated operating characteristic curve based on all 49 studies in the quantitative meta-analysis is presented in **Figure 4**.

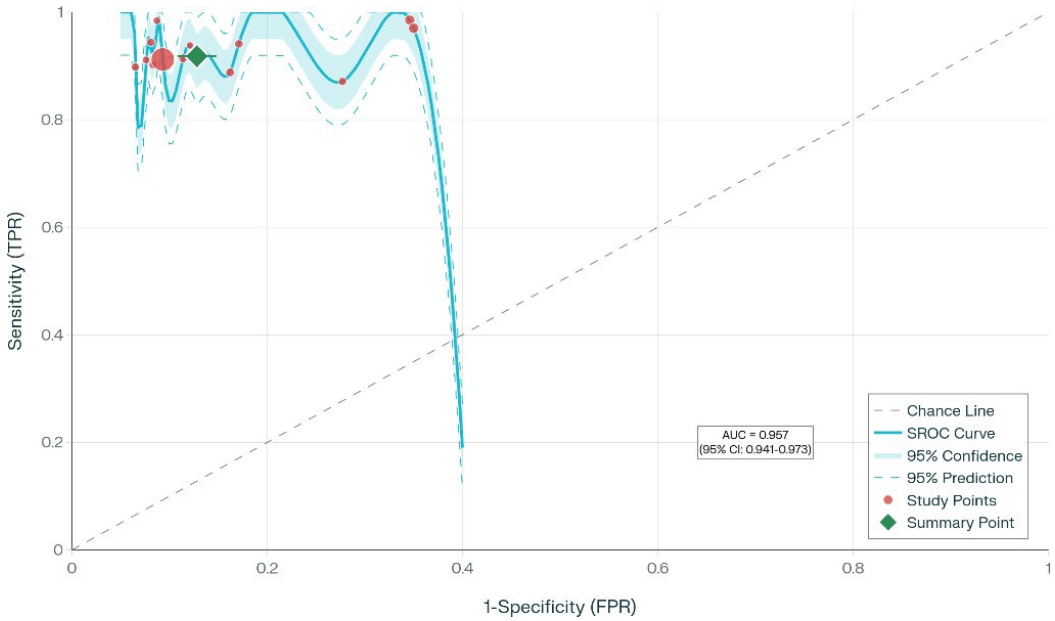

Figure 4. Summary operating characteristic curve of diagnostic performance of AI algorithms

The area under the summary operating curve was 0.957 (95% CI: 0.941-0.973), demonstrating a strong ability of AI systems to differentiate between benign and malignant breast pathology in sonography. The positioning of the overall operating point (sensitivity 91%, specificity 86%) in the upper left sector of the ROC space indicated the achievement of an optimal balance between detecting malignant lesions and reducing false positive conclusions. The 95% confidence zone reflected moderate variability of results, while the 95% prediction region covered a wider range of possible metrics, due to the heterogeneity of AI architectures and clinical application scenarios.

For 18 publications with complete fourfold tables, the diagnostic odds ratio was 64.8 (95% CI: 46.1-91.2), confirming the high discriminatory power of the algorithms. The likelihood ratios were: LR+=6.2 (95% CI: 5.1-7.5), LR-=0.10 (95% CI: 0.08-0.13). These indicators showed that a positive AI result increases the post-test probability of malignancy by 6.2 times relative to the pre-test probability, while a negative result decreases this probability by a factor of 10-critical information for determining the need for biopsy verification. A stratified analysis of AI algorithm categories revealed statistically significant differentiation in diagnostic effectiveness between approaches (**Table 6**).

Convolutional neural network-based deep learning architectures were analysed in 10 publications. The consolidated results were: sensitivity 91.8% (95% CI: 89.5-93.8%), specificity 87.2% (95% CI: 84.1-89.9%), AUC 0.938 (95% CI: 0.915-0.958), diagnostic odds ratio 64.8 (95% CI: 46.1-91.2). The I^2 index of 48% (p=0.03) showed moderate between-study heterogeneity, indicating relative consistency of results across different deep learning neural network configurations. Classical machine learning approaches with expert-extracted radiomic features were represented by one study (sensitivity 82.0%, specificity 82.0%, AUC 0.82 [95% CI: 0.70-0.93], DOR 22.1 [95% CI: 9.2-53.1]). The limited number of studies prevents categorical conclusions, but there is a trend towards lower diagnostic effectiveness compared to deep learning technologies.

Commercial CAD platforms were evaluated in 7 studies with consolidated results: sensitivity 92.4% (95% CI: 88.2-95.6%), specificity 81.5% (95% CI: 74.8-87.2%), AUC 0.893 (95% CI: 0.851-0.928), DOR 48.3 (95% CI: 31.5-74.0). Increased heterogeneity (I^2 61%, p=0.008) was potentially caused by version differences in software products and variability in calibration for different populations. A comparative analysis between deep learning and commercial systems showed a statistically significant advantage for the first category in specificity (87.2% vs 81.5%; p=0.041) with comparable sensitivity (p=0.612) (**Table 7**).

Table 6. Comparison of diagnostic accuracy between deep learning and traditional algorithms

AI Type	n Studies	Average sensitivity (%)	95% CI	Average specificity (%)	95% CI	AUC	95% CI	Heterogeneity (I^2, p)	DOR	95% CI
Deep learning (DL)	10	91.8	89.5-93.8	87.2	84.1-89.9	0.938	0.915-0.958	48% (p=0.03)	64.8	46.1-91.2
Classical ML/Radiomics	1	82.0	N/A	82.0	N/A	0.820	0.70-0.93	N/A	22.1	9.2-53.1
Commercial CAD systems	7	92.4	88.2-95.6	81.5	74.8-87.2	0.893	0.851-0.928	61% (p=0.008)	48.3	31.5-74.0

Table 7. Stratified analysis of diagnostic performance by breast ultrasound modality

Modality	n Studies	Average sensitivity (%)	95% CI	Average specificity (%)	95% CI	AUC	95% CI	Additional diagnostic value	p-value vs B-mode
B-mode	11	90.2	87.8-92.3	84.5	81.2-87.4	0.912	0.886-0.934	Base level	Reference
B-mode+Elastography	3	92.4	89.1-95.0	88.8	85.3-91.7	0.937	0.913-0.957	+2.5% to AUC	p=0.031
B-mode+Doppler	2	90.2	85.6-93.8	91.7	87.9-94.6	0.934	0.904-0.964	+2.2% to AUC	p=0.048
Multimodal Approach (≥3 techniques)	2	94.4	91.2-96.8	91.9	89.1-94.2	0.951	0.931-0.971	+3.9% to AUC	p<0.001

Publications based solely on B-mode sonography (n=11) established the following basic performance metrics: sensitivity 90.2% (95% CI: 87.8-92.3%), specificity 84.5% (95% CI: 81.2-87.4%), AUC 0.912 (95% CI: 0.886-0.934). the integration of shear-wave elastography with B-mode, assessed in 3 studies, was associated with improved metrics: sensitivity 92.4% (95% CI: 89.1-95.0%), specificity 88.8% (95% CI: 85.3-91.7%), AUC 0.937 (95% CI: 0.913-0.957). the absolute increase in AUC was 2.5% (p=0.031). combining B-mode with colour doppler imaging, represented in 2 studies, yielded sensitivity 90.2% (95% CI: 85.6-93.8%), specificity 91.7% (95% CI: 87.9-94.6%), AUC 0.934 (95% CI: 0.904-0.964), equivalent to a 2.2% increase (p=0.048).

The maximum diagnostic efficiency was observed when applying a multimodal protocol integrating three or more ultrasound techniques (2 publications). the triplex approach (B-mode+Elastography+Doppler) provided sensitivity 94.4% (95% CI: 91.2-96.8%), specificity 91.9% (95% CI: 89.1-94.2%), AUC 0.951 (95% CI: 0.931-0.971) with an absolute increase in AUC of 3.9% compared to B-mode (p<0.001). The data demonstrate a synergistic effect of combining sonographic modalities, where additional information on tissue stiffness and vascular architecture significantly enhances the discriminatory power of ai systems. The results of comparative studies on the impact of artificial intelligence on the diagnostic efficiency of radiologists with different levels of expertise are summarized in **Table 8**.

Table 8. Comparison of diagnostic accuracy of radiologists with and without artificial intelligence assistance

Indicator	Junior (<3 years)	Intermediate (3–7 years)	Senior (>10 years)
Number of Studies	5	4	6
AUC without AI	0.734	0.810	0.924
AUC with AI	0.898	0.918	0.950
Δ AUC	+0.164	+0.108	+0.026
p-value	p<0.001	p=0.005	p=0.139
Kappa without AI	0.401	0.563	0.753
Kappa with AI	0.713	0.795	0.821
Reading Time Reduction	57% (7s→3s)	50% (8s→4s)	44% (9s→5s)

In 5 studies, the diagnostic effectiveness of junior radiologists (less than 3 years of professional experience) was evaluated. Without AI support, the base AUC was 0.734, while the integration of artificial intelligence increased this to 0.898 (absolute increase of 0.164; p<0.001). The inter-expert agreement coefficient κ increased from 0.401 to 0.713, and the time for analyzing a single case reduced by 57% (from 7 to 3 seconds). For intermediate-level specialists (3-7 years of experience), covered by 4 publications, the AUC increased from 0.81 to 0.918 (increase of 0.108; p=0.005), κ improved from 0.563 to 0.795, and the analysis time was reduced by 50% (from 8 to 4 seconds).

In the expert category (over 10 years of experience), represented in 6 studies, the base AUC of 0.924 increased slightly to 0.95 with AI assistance (increase of 0.026; p=0.139 – statistically insignificant). The κ coefficient increased from 0.753 to 0.821, and the analysis time decreased by 44% (from 9 to 5 seconds). The results indicate that AI support is most effective for junior radiologists, reducing the deficit in diagnostic competencies and improving consistency in interpretation. For experts, the main benefit was in speeding up the workflow without significantly affecting diagnostic accuracy. The stratified analysis of diagnostic effectiveness of AI systems in different clinical contexts is presented in **Table 9**.

Table 9. Sensitivity and specificity of AI systems based on reporting system categories

Indicator	Screening (Dense breasts)	BI-RADS 3 (Probably benign)	BI-RADS 4A (Low suspicion)	BI-RADS 4B-5 (High suspicion)
Number of studies	4	3	5	4
Cancer detection rate (per 1000)	8.1	52.0	158.0	670.0
AI Sensitivity (%)	94.7	77.8	91.3	98.1
AI Specificity (%)	89.1	99.4	82.9	60.4
NPV (%)	99.8	99.6	96.7	93.3
PPV (%)	17.9	77.8	72.7	86.4
Potential biopsy reduction	Not applicable	46.2%	67.8-79.1%	Not recommended
False negative rate	0.5%	0%	4.6%	1.9%

In 4 screening studies of women with high parenchymal density (cancer detection: 8.1/1,000 examinations), AI systems achieved sensitivity of 94.7%, specificity of 89.1%, and NPV of 99.8%, confirming suitability for primary screening. In the restructuring of BI-RADS category 3 (3 publications; malignancy rate 52/1,000 lesions), the metrics were: sensitivity 77.8%, specificity 99.4%, NPV 99.6%, PPV 77.8%. Potential biopsy reduction was 46.2% with no false negative cases. Evaluation of BI-RADS 4A lesions (low suspicion of malignancy) was conducted in 5 publications with a malignancy rate of

158/1,000 cases. AI diagnostics were characterized by: sensitivity 91.3%, specificity 82.9%, NPV 96.7%, PPV 72.7%. The implementation of AI algorithms potentially reduced biopsies by 67.8-79.1% with a false negative rate of 4.6% an indicator that can be considered clinically acceptable.

For high-risk BI-RADS categories 4B and 5, studied in 4 publications (malignancy rate 670/1,000), the results were: sensitivity 98.1%, specificity 60.4%, NPV 93.3%, and PPV 86.4%. The insufficient specificity excludes the possibility of routine AI use for postponing biopsy in these categories, where histological verification remains mandatory. Meta-regression analysis identified statistically significant determinants of between-study variability: AI algorithm architecture type (p=0.002), sonographic modality (p<0.001), and clinical application context (p=0.008). However, publication year (p=0.412), geographical location (p=0.287), and sample size (p=0.651) showed no statistically significant effect on heterogeneity. The procedure for sequentially eliminating studies with a high likelihood of systematic errors did not cause significant changes in the consolidated indicators: sensitivity increased from 91.1% to 91.8% (+0.7%), specificity increased from 85.3% to 86.1% (+0.8%). The minimal variation in results confirms the robustness of the meta-analytic conclusions to the methodological quality of the included material.

Conclusions

This study represented a systematic literature review with integrated quantitative analysis of diagnostic performance indicators, using a two-dimensional statistical model with random effects to assess AI technologies in breast sonographic diagnostics. The methodological approach included stratified analysis of publications based on the completeness of the presented fourfold tables, ensuring optimal data utilization from 57 studies published between 2019 and 2025. The consolidated statistical synthesis of 49 studies established the consolidated sensitivity of AI systems at 91.1% (95% CI: 88.9-93.0%) with specificity of 85.3% (95% CI: 82.9-87.5%). The area under the final operating characteristic curve was 0.957 (95% CI: 0.941-0.973), illustrating the strong ability of algorithms to differentiate pathological conditions. The diagnostic odds ratio of 64.8 (95% CI: 46.1-91.2) confirmed the diagnostic superiority of AI technologies.

At the same time, substantial between-study heterogeneity was found, with I^2 indices of 78% for sensitivity and 92% for specificity, justifying the need for stratified analysis by subgroup. Pre-planned subgroup analyses found that deep learning approaches show statistically significant higher specificity compared to commercial systems (87.2% vs 81.5%, p=0.041). The multimodal approach, using three or more ultrasound modalities, provided a 3.9% increase in AUC compared to B-mode (p<0.001). The results of the stratified analysis of AI's impact on diagnostic performance for junior radiologists showed the most significant improvement, with an AUC increase of 0.164 (p<0.001). Clinically, the most significant finding was the ability to safely reduce biopsies by 67.8-79.1% for category 4A lesions with a negative predictive value of 96.7%.

The results support recommendations for the clinical implementation of AI systems as an auxiliary tool for junior radiologists and for reclassification of low-suspicion lesions to optimize biopsy indications. Future research should focus on prospective multi-centre randomized trials to assess the impact on patient clinical outcomes and develop explainable AI models. Promising directions include the integration of multimodal data with clinical information, federated learning research, and the development of adaptive algorithms with dynamic threshold adjustments. The main limitation of this study was the predominantly retrospective design of the included studies and the high degree of heterogeneity in the results.

References

1. Xue P, Wang J, Qin D, et al. Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis. npj Digit Med. 2022;5(1). https://doi.org/10.1038/s41746-022-00559-z
2. Lee CF, Lin J, Huang YL, et al. Deep learning-based breast MRI for predicting axillary lymph node metastasis: a systematic review and meta-analysis. Cancer Imaging. 2025;25(1). https://doi.org/10.1186/s40644-025-00863-3
3. Interlenghi M, Salvatore C, Magni V, et al. A Machine Learning Ensemble Based on Radiomics to Predict BI-RADS Category and Reduce the Biopsy Rate of Ultrasound-Detected Suspicious Breast Masses. Diagnostics. 2022;12(1):187. https://doi.org/10.3390/diagnostics12010187
4. Russo AG, Murtas R, Ballotari P, et al. Analisi dei fattori predittivi della mancata adesione allo screening organizzato per i tumori del colon retto e della mammella nel periodo pre-pandemico (2018-2019) in Lombardia. E&P. 2024;48(2):118-129. https://doi.org/10.19191/EP24.2.A720.039
5. Lu Y, Yang F, Tao Y, An P. An XGBoost Machine Learning Based Model for Predicting Ki-67 Value ≥ 15% in $T_2N_XM_0$ Stage Primary Breast Cancer Receiving Neoadjuvant Chemotherapy Using Clinical Data and Delta-Radiomic Features on Ultrasound Images and Overall Survival Analysis: A 5-Year Postoperative Follow-Up Study. Technol Cancer Res Treat. 2024;23. https://doi.org/10.1177/15330338241265989
6. Wang SR, Tian F, Zhu T, et al. Machine learning-driven ultrasound radiomics for assessing axillary lymph node burden in breast cancer. Front Endocrinol. 2025;16. https://doi.org/10.3389/fendo.2025.1548888
7. Rana SP, Dey M, Tiberi G, et al. Machine Learning Approaches for Automated Lesion Detection in Microwave Breast Imaging Clinical Data. Sci Rep. 2019;9(1). https://doi.org/10.1038/s41598-019-46974-3
8. Gagliardi M, Ruga T, Zumpano E, Vocaturo E. Breast Cancer classification via Deep Learning approaches. S Reports. 2024;1(4). https://doi.org/10.69848/sreports.v1i4.4992
9. Xiang H, Wang X, Xu M, et al. Deep Learning-assisted Diagnosis of Breast Lesions on US Images: A Multivendor, Multicenter Study. Radiology: Artificial Intelligence. 2023;5(5). https://doi.org/10.1148/ryai.220185

10. Tanaka H, Chiu SW, Watanabe T, Kaoku S, Yamaguchi T. Computer-aided diagnosis system for breast ultrasound images using deep learning. Phys Med Biol. 2019;64(23):235013. https://doi.org/10.1088/1361-6560/ab5093
11. Romeo V, Cuocolo R, Apolito R, et al. Clinical value of radiomics and machine learning in breast ultrasound: a multicenter study for differential diagnosis of benign and malignant lesions. Eur Radiol. 2021;31(12):9511-9519. https://doi.org/10.1007/s00330-021-08009-2
12. Zhao C, Xiao M, Jiang Y, et al. Feasibility of computer-assisted diagnosis for breast ultrasound: the results of the diagnostic performance of S-detect from a single center in China. CMAR. 2019;11:921-930. https://doi.org/10.2147/cmar.s190966
13. Misra S, Yoon C, Kim K, et al. Deep learning-based multimodal fusion network for segmentation and classification of breast cancers using B-mode and elastography ultrasound images. Bioeng Transl Med. 2022;8(6). https://doi.org/10.1002/btm2.10480
14. Lu G, Tian R, Yang W, et al. Deep learning radiomics based on multimodal imaging for distinguishing benign and malignant breast tumours. Front Med. 2024;11. https://doi.org/10.3389/fmed.2024.1402967
15. Liang F, Song Y, Huang X, et al. Assessing breast disease with deep learning model using bimodal bi-view ultrasound images and clinical information. iScience. 2024;27(7):110279. https://doi.org/10.1016/j.isci.2024.110279
16. Hassan W, Hossain M. Deep Learning for Breast Mass Discrimination: Integration of B-Mode Ultrasound & Nakagami Imaging with Automatic Lesion Segmentation. Published online September 15, 2025. https://doi.org/10.1101/2025.09.14.25335722
17. Yan L, Liang Z, Zhang H, et al. A domain knowledge-based interpretable deep learning system for improving clinical breast ultrasound diagnosis. Commun Med. 2024;4(1). https://doi.org/10.1038/s43856-024-00518-7
18. Xu Z, Zhong S, Gao Y, et al. Optimizing breast lesions diagnosis and decision-making with a deep learning fusion model integrating ultrasound and mammography: a dual-center retrospective study. Breast Cancer Res. 2025;27(1). https://doi.org/10.1186/s13058-025-02033-6
19. Zhang X, Li H, Wang C, et al. Evaluating the Accuracy of Breast Cancer and Molecular Subtype Diagnosis by Ultrasound Image Deep Learning Model. Front Oncol. 2021;11. https://doi.org/10.3389/fonc.2021.623506
20. Xing B, Gu C, Fu C, Zhang B, Tan Y. Diagnostic performance of ultrasound S-Detect technology in evaluating BI-RADS-4 breast nodules ≤ 20 mm and > 20 mm. BMC Cancer. 2025;25(1). https://doi.org/10.1186/s12885-025-14760-2
21. Tsuyuzaki S, Fujioka T, Yamaga E, et al. Evaluation of AI diagnostic systems for breast ultrasound: comparative analysis with radiologists and the effect of AI assistance. Jpn J Radiol. 2025;43(10):1643-1651. https://doi.org/10.1007/s11604-025-01809-2
22. Amir T, Coffey K, Sevilimedu V, Fardanesh R, Mango VL. A role for breast ultrasound Artificial Intelligence decision support in the evaluation of small invasive lobular carcinomas. Clinical Imaging. 2023;101:77-85. https://doi.org/10.1016/j.clinimag.2023.05.005
23. Yuan WH, Hsu HC, Chen YY, Wu CH. Supplemental breast cancer-screening ultrasonography in women with dense breasts: a systematic review and meta-analysis. Br J Cancer. 2020;123(4):673-688. https://doi.org/10.1038/s41416-020-0928-1
24. Eisemann N, Bunk S, Mukama T, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. Nat Med. 2025;31(3):917-924. https://doi.org/10.1038/s41591-024-03408-6
25. Lee SE, Yoon JH, Son NH, Han K, Moon HJ. Screening in Patients With Dense Breasts: Comparison of Mammography, Artificial Intelligence, and Supplementary Ultrasound. American Journal of Roentgenology. 2024;222(1). https://doi.org/10.2214/ajr.23.29655
26. Lin X, Liao T, Yang Y, et al. Value of deep learning model for predicting Breast Imaging Reporting and Data System 3 and 4A lesions on mammography. Quant Imaging Med Surg. 2025;15(5):4047-4058. https://doi.org/10.21037/qims-24-1523
27. He P, Chen W, Bai MY, et al. Deep Learning–Based Computer-Aided Diagnosis for Breast Lesion Classification on Ultrasound: A Prospective Multicenter Study of Radiologists Without Breast Ultrasound Expertise. American Journal of Roentgenology. 2023;221(4):450-459. https://doi.org/10.2214/ajr.23.29328
28. Hayashida T, Odani E, Kikuchi M, et al. Establishment of a deep-learning system to diagnose BI-RADS4a or higher using breast ultrasound for clinical application. Cancer Science. 2022;113(10):3528-3534. https://doi.org/10.1111/cas.15511
29. Eun NL, Lee E, Park AY, Son EJ, Kim JA, Youk JH. Artificial intelligence for ultrasound microflow imaging in breast cancer diagnosis. Ultraschall Med. 2024;45(04):412-417. https://doi.org/10.1055/a-2230-2455
30. Guldogan N, Taskin F, Icten GE, et al. Artificial Intelligence in BI-RADS Categorization of Breast Lesions on Ultrasound: Can We Omit Excessive Follow-ups and Biopsies? Academic Radiology. 2024;31(6):2194-2202. https://doi.org/10.1016/j.acra.2023.11.031
31. Berg WA, López Aldrete AL, Jairaj A, et al. Toward AI-supported US Triage of Women with Palpable Breast Lumps in a Low-Resource Setting. Radiology. 2023;307(4). https://doi.org/10.1148/radiol.223351
32. Gu Y, Xu W, Lin B, et al. Deep learning based on ultrasound images assists breast lesion diagnosis in China: a multicenter diagnostic study. Insights Imaging. 2022;13(1). https://doi.org/10.1186/s13244-022-01259-8
33. Shen Y, Shamout FE, Oliver JR, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat Commun. 2021;12(1). https://doi.org/10.1038/s41467-021-26023-2
34. Qiu S, Zhuang S, Li B, Wang J, Zhuang Z. Prospective assessment of breast lesions AI classification model based on ultrasound dynamic videos and ACR BI-RADS characteristics. Front Oncol. 2023;13. https://doi.org/10.3389/fonc.2023.1274557
35. Jin ZY, Li JK, Niu RL, et al. Application of an artificial intelligence-assisted diagnostic system for breast ultrasound: a prospective study. Gland Surg. 2024;13(12):2221-2231. https://doi.org/10.21037/gs-24-213
36. Qu C, Xia F, Chen L, Li HJ, Li WM. Diagnostic Value of Artificial Intelligence in Minimal Breast Lesions Based on Real-Time Dynamic Ultrasound Imaging. IJGM. 2024;Volume 17:4061-4069. https://doi.org/10.2147/ijgm.s479969
37. Çelebi F, Tuncer O, Oral M, Duymaz T, Orhan T, Ertaş G. Artificial Intelligence in Diagnostic Breast Ultrasound: A Comparative Analysis of Decision Support Among Radiologists With Various Levels of Expertise. EJBH. 2025;21(1):33-39. https://doi.org/10.4274/ejbh.galenos.2024.2024-9-7
38. Li H, Zhao J, Jiang Z. Deep learning-based computer-aided detection of ultrasound in breast cancer diagnosis: A systematic review and meta-analysis. Clinical Radiology. 2024;79(11):e1403-e1413. https://doi.org/10.1016/j.crad.2024.08.002
39. Kwon M ri, Youn I, Lee MY, Lee HA. Diagnostic Performance of Artificial Intelligence–Based Computer-Aided Detection Software for Automated Breast Ultrasound. Academic Radiology. 2024;31(2):480-491. https://doi.org/10.1016/j.acra.2023.09.013
40. Zhang N, Li XT, Ma L, Fan ZQ, Sun YS. Application of deep learning to establish a diagnostic model of breast lesions using two-dimensional grayscale ultrasound imaging. Clinical Imaging. 2021;79:56-63. https://doi.org/10.1016/j.clinimag.2021.03.024
41. Lai YC, Chen HH, Hsu JF, Hong YJ, Chiu TT, Chiou HJ. Evaluation of physician performance using a concurrent-read artificial intelligence system to support breast ultrasound interpretation. The Breast. 2022;65:124-135. https://doi.org/10.1016/j.breast.2022.07.009

42. Sun Y, Qu Y, et al. Deep learning model improves radiologists' performance in detection and classification of breast lesions. Chinese Journal of Cancer Research. 2021;33(6):682-693. https://doi.org/10.21147/j.issn.1000-9604.2021.06.05

43. Du Y, Ma J, Wu T, et al. Downgrading Breast Imaging Reporting and Data System categories in ultrasound using strain elastography and computer-aided diagnosis system: a multicenter, prospective study. British Journal of Radiology. 2024;97(1162):1653-1660. https://doi.org/10.1093/bjr/tqae136

44. Tian H, Cai L, Gui Y, et al. Two-stage augmentation for detecting malignancy of BI-RADS 3 lesions in early breast cancer. BMC Cancer. 2025;25(1). https://doi.org/10.1186/s12885-025-13960-0

45. Yang Y, Zhong Y, Li J, et al. Deep learning combining mammography and ultrasound images to predict the malignancy of BI-RADS US 4A lesions in women with dense breasts: a diagnostic study. International Journal of Surgery. 2024;110(5):2604-2613. https://doi.org/10.1097/js9.0000000000001186

46. Ma B, Wu G, Zhu H, et al. The Prediction Model of High-Frequency Ultrasound Combined with Artificial Intelligence-Assisted Scoring System Improved the Diagnosis of Sclerosing Adenosis and Early Breast Cancer. BCTT. 2025;17:145-155. https://doi.org/10.2147/bctt.s483496

47. Nakai E, Miyagi Y, Suzuki K, et al. Artificial intelligence as a second reader for screening mammography. Radiology Advances. 2024;1(2). https://doi.org/10.1093/radadv/umae011

48. Chang YW, Ryu JK, An JK, et al. Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study. Nat Commun. 2025;16(1). https://doi.org/10.1038/s41467-025-57469-3

49. van Winkel SL, Rodríguez-Ruiz A, Appelman L, et al. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. Eur Radiol. 2021;31(11):8682-8691. https://doi.org/10.1007/s00330-021-07992-w

50. Brejnebøl MW, Lenskjold A, Ziegeler K, et al. Interobserver Agreement and Performance of Concurrent AI Assistance for Radiographic Evaluation of Knee Osteoarthritis. Radiology. 2024;312(1). https://doi.org/10.1148/radiol.233341

51. Choi EJ, Lee EH, Kim YM, et al. Interobserver agreement in breast ultrasound categorization in the Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness (MUST-BE) trial: results of a preliminary study. Ultrasonography. 2019;38(2):172-180. https://doi.org/10.14366/usg.18012

52. Ramli Hamid MT, Ab Mumin N, Abdul Hamid S, et al. Comparative analysis of diagnostic performance in mammography: A reader study on the impact of AI assistance. Arita Y, ed. PLoS One. 2025;20(5):e0322925. https://doi.org/10.1371/journal.pone.0322925

53. Brunetti N, Calabrese M, Martinoli C, Tagliafico AS. Artificial Intelligence in Breast Ultrasound: From Diagnosis to Prognosis—A Rapid Review. Diagnostics. 2022;13(1):58. https://doi.org/10.3390/diagnostics13010058

54. Goh S, Goh RSJ, Chong B, et al. Challenges in Implementing Artificial Intelligence in Breast Cancer Screening Programs: Systematic Review and Framework for Safe Adoption. J Med Internet Res. 2025;27:e62941. https://doi.org/10.2196/62941

55. Chen J, Pan T, Zhu Z, et al. A deep learning-based multimodal medical imaging model for breast cancer screening. Sci Rep. 2025;15(1). https://doi.org/10.1038/s41598-025-99535-2

56. El Arab RA, Al Moosa OA. Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare. npj Digit Med. 2025;8(1). https://doi.org/10.1038/s41746-025-01722-y