narra R

Systematic Review

Diagnostic accuracy of artificial intelligence-assisted early detection of breast cancer using ultrasound imaging: A systematic review

Wilia A. Utami[1*], Siti Wahyuni[1], Miska Zamharira[1] and Richard Y. Akele[2]

[1]Postgraduate Program of Public Health, Universitas Muhammadiyah Aceh, Banda Aceh, Indonesia; [2]Department of Biomedical Science, University of Brighton, Brighton, United Kingdom

*Corresponding author: aprilisautamiwiliai@gmail.com

Abstract

Artificial intelligence (AI) has emerged as a promising tool for improving the diagnostic accuracy of breast ultrasound and facilitating early breast cancer detection. This systematic review evaluated the current evidence on AI-based diagnostic models for breast ultrasound imaging. A systematic literature search was performed in PubMed, Scilit, and PMC following the PRISMA 2020 guidelines. A total of 1,296 records were identified, and after duplicate removal and eligibility screening, 18 studies published between 2021 and 2026 were included. Methodological quality was assessed using the QUADAS-3 tool. Owing to heterogeneity in AI models, study designs, and reported outcomes, a qualitative synthesis was conducted without quantitative meta-analysis. AI-assisted breast ultrasound demonstrated generally favourable diagnostic performance, although the reported parameters varied across studies. Sensitivity ranged from 71.4% to 100.0%, specificity from 71.6% to 96.2%, and AUC from 0.778 to 1.00. One EfficientNet-B7 model integrated with explainable AI achieved an AUC of 1.00, sensitivity of 99.5%, an F1-score of 98.9%, and accuracy of 99.14%, representing one of the highest-performing models identified. Other high-performing approaches included hybrid deep learning, Vision Transformers, interpretable ensemble transformers, convolutional neural networks, and machine learning models, which improved lesion classification, reduced false-positive findings, enhanced interpretability, and supported standardized reporting. However, most studies were retrospective, single-center investigations with limited external validation. AI-assisted breast ultrasound is a promising adjunctive tool for early breast cancer detection. Although current evidence supports its clinical potential, further prospective multicenter studies with standardized methodologies and external validation are needed before routine clinical implementation.

Keywords: Artificial intelligence, breast ultrasound, breast cancer, diagnostic accuracy, systematic review

Introduction

Breast cancer is the most frequently diagnosed malignancy among women worldwide and remains a major cause of cancer-related mortality. According to recent GLOBOCAN estimates and future projections, the global burden of breast cancer is expected to increase substantially by 2050, underscoring the critical importance of early detection and timely diagnosis to improve

Received: June 15, 2026 - Accepted: July 29, 2026 - Published online: July 31, 2026

patient outcomes and reduce mortality [1]. Breast ultrasound (BUS) is widely used because it is accessible, radiation-free, cost-effective, and particularly effective for evaluating dense breast tissue. However, its diagnostic performance is highly dependent on operator expertise, leading to inter-observer variability and unnecessary biopsies [2]. Therefore, improving the accuracy and consistency of BUS interpretation remains a major priority in breast imaging.

Recent advances in artificial intelligence (AI), including machine learning, deep learning, convolutional neural networks (CNNs), transformer-based models, and explainable AI (XAI), have significantly improved BUS image analysis. AI can automatically identify suspicious lesions, distinguish benign from malignant tumors, and support radiologists in clinical decision-making. Recent studies have reported high diagnostic performance, with improvements in sensitivity, specificity, and area under the receiver operating characteristic curve (AUC), while reducing false positive findings and enhancing diagnostic consistency [3]. Despite these advances, existing studies show substantial heterogeneity in patient populations, ultrasound systems, AI models, validation methods, and reference standards, limiting comparison across studies and clinical implementation. Therefore, this systematic review aims to synthesize the current evidence on the diagnostic accuracy of AI for early breast cancer detection using BUS, evaluate the performance of different AI models, and identify gaps to guide future research and clinical practice [2].

Methods

Study design

This study was designed as a systematic review to evaluate the diagnostic accuracy of AI for early breast cancer detection using BUS imaging. The review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement [4].

Search strategy

A comprehensive literature search was performed in three electronic databases: PubMed, PubMed Central (PMC), and Scilit, without restrictions on country of origin. The final literature search was conducted on June 30, 2026. The search strategy combined controlled vocabulary (Medical Subject Headings, MeSH) and free-text keywords related to breast cancer, BUS, artificial intelligence, and diagnostic accuracy. The Boolean search syntax used in PubMed was as follows: ("diagnostic accuracy" OR "diagnostic performance" OR "sensitivity and specificity" OR "ROC curve" OR "AUC" OR "early detection" OR screening) AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "computer aided diagnosis" OR "AI" OR "Computer aided design" OR “Intelligent System”[MeSHTerm]) OR ("Visual interpretation" OR Radiologist) AND ("breast cancer" OR "breast neoplasm" OR "breast tumor" OR "mammary carcinoma" OR "Breast Cancer Lymphedema") AND ("medical imaging" OR "breast ultrasound" OR "breast imaging" OR “diagnostic Imaging” [MeSHTerm]). Equivalent search strategies were adapted for PMC and Scilit. Reference lists of all eligible articles were also manually screened to identify additional relevant studies that might have been missed during the electronic database search [5].

Eligibility criteria

Eligibility criteria were established using the PICOS (Population, Index test, Comparator, Outcomes, Study design) framework. Population (P): women with suspected or confirmed breast lesions who underwent breast ultrasound examination. Index test (I): artificial intelligence algorithms applied to breast ultrasound images, including machine learning, deep learning, CNN, and computer aided design. Comparator (C): Immediate visual interpretation by a radiologist. Outcomes (O): Including accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), F1-score, reduction of false-positive findings, lesion classification performance, standardized reporting, model interpretability, and clinical utility. Study design (S): original observational studies, retrospective or prospective diagnostic studies, and AI model development or validation studies.

Studies were excluded if they (1) investigated imaging modalities other than breast ultrasound, such as mammography, MRI, CT, or histopathology; (2) were review articles, editorials, conference abstracts, letters, case reports, or book chapters; (3) involved animal experiments or phantom studies; (4) lacked sufficient diagnostic performance outcomes; or (5) were not published in English.

Screening and selection

All records identified from the database search were exported into EndNote X9 for duplicate removal. Subsequently, two independent reviewers (WAU and SW) screened article titles and abstracts according to the predefined eligibility criteria. Potentially relevant articles underwent full-text assessment to determine their final eligibility. Disagreements between reviewers were resolved through discussion and consensus. If consensus could not be achieved, a third reviewer was consulted (MZ). The study selection process was summarized using a PRISMA 2020 flow diagram.

Quality assessment

Methodological quality and risk of bias of the included studies were independently evaluated by two reviewers (WAU and SW) using the Quality Assessment of Diagnostic Accuracy Studies-3 (QUADAS-3) tool, which is the recommended instrument for diagnostic accuracy systematic reviews. QUADAS-3 evaluates four domains, including patient selection, index test, reference standard, and flow and timing. Each domain was classified as low, high, or unclear risk of bias. Applicability concerns were additionally assessed for patient selection, index test, and reference standard. Any disagreements between reviewers were resolved through consensus discussion [6].

Data extraction

A standardized and piloted data extraction form was developed before formal study selection. Two reviewers (WAU and SW) independently extracted data from each eligible study. The extracted information included the first author, publication year, country, study design, study setting, sample size, participant characteristics, breast ultrasound modality, dataset source, data partitioning strategy, artificial intelligence algorithm, model input features, diagnostic or classification task, reference standard, validation approach, comparator, and reported diagnostic performance.

Diagnostic outcomes included sensitivity, specificity, area under the receiver operating characteristic curve, accuracy, precision, recall, F1-score, positive predictive value, and negative predictive value, as available. Information on internal validation, cross-validation, independent testing, and external or multicentre validation was also recorded. When multiple models or datasets were evaluated within the same study, results were extracted separately for each relevant model, dataset, validation cohort, or classification task. Where several performance estimates were available, priority was given to results from external or independent validation cohorts; otherwise, results from the principal test set were used. For studies reporting binary and multiclass outcomes, each classification task was recorded separately. Diagnostic parameters were extracted as reported in the original publication without recalculation. Unreported information was recorded as not reported. Any discrepancies between the two reviewers were resolved through discussion until consensus was reached or, when necessary, through consultation with a third reviewer (MZ).

Covariates and subgroup analysis

Because of methodological heterogeneity among the included studies, subgroup analyses were planned according to several predefined characteristics, including: AI methodology (machine learning and deep learning), deep learning architecture (CNN, EfficientNet, Vision Transformer, Swin Transformer, ensemble learning), computer aided design, Explainable AI versus conventional AI, Internal versus external validation, Clinical application (lesion detection, lesion classification, subtype prediction, structured reporting), Publication year (2021-2023 vs 2024-2026). These subgroup analyses were intended to identify potential sources of variability in diagnostic performance across different AI approaches and study designs.

Data analysis

Given the substantial heterogeneity in AI architectures, ultrasound datasets, validation strategies, reference standards, and reported diagnostic performance metrics, a quantitative meta-analysis was not performed. Instead, findings were synthesized using a narrative synthesis in accordance with the Synthesis Without Meta-analysis (SWiM) reporting recommendations. Diagnostic performance outcomes including accuracy, sensitivity, specificity, and AUC were summarized descriptively and compared across studies. Study characteristics, AI models, clinical applications, and methodological quality were synthesized in tables and discussed qualitatively to identify emerging trends, strengths, limitations, and future research directions for AI-assisted breast ultrasound diagnosis. This narrative approach is recommended when substantial clinical and methodological heterogeneity precludes valid statistical pooling [5].

Results

Search results

The electronic database search identified 1,296 potentially relevant records, comprising 982 articles from PubMed, 117 articles from Scilit, and 197 articles from PMC. After removing duplicate records, 1,118 unique articles remained for title and abstract screening. A total of 1,062 records were excluded because they were unrelated to breast ultrasound imaging, did not evaluate artificial intelligence, involved other imaging modalities (such as mammography, MRI, and CT), were review articles, conference abstracts, editorials, or did not report diagnostic outcomes. Subsequently, 56 full-text articles were assessed for eligibility. Of these, 38 studies were excluded after full-text review due to the following reasons: use of non-ultrasound imaging modalities (n=11), not diagnostic study (n=2), insufficient diagnostic outcome data (n=8), review articles (n=6), duplicate datasets (n=5), conference papers without accessible full text (n=3), and non-English publications (n=3). Ultimately, 18 studies fulfilled all inclusion criteria and were included in the qualitative synthesis. Because of substantial methodological heterogeneity among the included studies, no quantitative meta-analysis was performed. A summary of the characteristics and outcomes of the included studies is presented in **Table 1**.

Characteristics of included studies

Most of the included studies employed retrospective diagnostic designs, while several were multicenter investigations or pilot studies. The majority focused on developing and validating deep learning-based computer-aided diagnostic systems for breast ultrasound imaging. Artificial intelligence approaches included CNNs, transfer learning, hybrid deep learning architectures, XAI, ensemble learning, transformer-based models, and large language models for structured reporting.

Among the included studies, the principal clinical applications comprised lesion detection, benign versus malignant lesion classification, prediction of molecular breast cancer subtypes, automated breast ultrasound interpretation, structured reporting, and prediction of tumor biological characteristics. Histopathological examination served as the reference standard in nearly all diagnostic studies. Most investigations demonstrated promising diagnostic performance, with consistently high sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC), suggesting that AI-assisted breast ultrasound has substantial potential to improve early breast cancer detection and reduce observer variability. Earlier studies primarily implemented conventional CNN-based algorithms, whereas more recent publications increasingly investigated transformer-based architectures, XAI, multimodal learning, and large language models to enhance both diagnostic performance and clinical interpretability.

Quality of the included studies

Methodological quality was assessed using the QUADAS-3 framework (**Table 2**). Overall, the included studies demonstrated satisfactory methodological quality with an acceptable risk of bias. Most studies showed a low risk of bias in the domains of patient selection, index test, and

reference standard, as histopathological examination was consistently used as the diagnostic reference standard for confirming breast cancer.

Table 1. Characteristics of and findings from included studies (n=18)

Author, year	Study design	Dataset	AI model	Validation approach	Diagnostic performance*
Kim *et al.*, 2021 [7]	Retrospective diagnostic study	Hospital breast ultrasound images	Three model (VGG16, ResNet34, and GoogLeNet)	Internal training/validation/ test split	The highest AUC=0.90; The highest Sens=91%
Shen *et al.*, 2021 [8]	Retrospective diagnostic study	Large multicenter breast ultrasound examinations	AI-assisted breast ultrasound interpretation	Independent external validation	Sens=95.8%; Spec=80. 1%; AUC=0.976 (95%CI: 0.972–0.980)
Ye *et al.*, 2021 [9]	Retrospective diagnostic study	Ultrasound images of triple-negative breast cancer	CNN	Training/testing cohort	Sens=85.4%; F1: 86.0%
Luo *et al.*, 2022 [10]	Pilot diagnostic study	ABUS dataset	3D CNN	Training, independent test, and control groups	Sens=92.7% per volume, 94.5 per lesion, and 96.5% per patient.
Ferre *et al.*, 2023 [11]	Retrospective study	BUSI molecular subtype dataset	Logistic Regression	Cross validation	Sens=81.8% (TN)/71.4% (HER2+); Spec: 74.2% (TN)/71.6% (HER2+); AUC: 0.824 (TN)/0.778 (HER2+)
Liu *et al.*, 2024 [12]	Comparative study	BUSI reports	GPT-4 based structured reporting for breast ultrasound	Comparative evaluation (GPT-4 vs Bing GPT-4)	Sens=95.9%
Ru *et al.*, 2024 [13]	Multicenter retrospective study	Multicenter ultrasound dataset	GNN and CNN	External multicenter validation	Sens=83.3%; Spec=92.0%; AUC=0.911; F1=0.833
Buzatto *et al.*, 2025 [14]	Retrospective cohort study	Clinical and ultrasound features	XGBoost	Internal validation	Sens=94.6%; Spec=82.9%; AUC=0.96; F1=0.918
Chen *et al.*, 2025 [15]	Retrospective diagnostic study	BUSI dataset	BIScreener interpretable deep learning framework	Independent test dataset	AUC=0.982
Liu *et al.*, 2025 [16]	Retrospective predictive study	BUSI with pathology	Radiomics-Clinical (R-C) model	Internal validation	Sens=88.4%; Spec=76.5%; AUC=0.869; F1=0.854
Baek *et al.*, 2022 [17]	Experimental study	Raw ultrasound parameter dataset	SVM	Cross validation	Sens=100.0%; Spec=90.3%; AUC=0.982
Hossain *et al.*, 2023 [18]	Retrospective image classification study	BUSI dataset	VGG16 transfer learning	Train/test split	Sens=96.0%; F1=96.0%
Zhang *et al.*, 2023 [19]	Retrospective image analysis study	BUSI dataset	REAF deep learning model	Cross validation	Sens=97.1%; Spec=96.2%; AUC=0.992; F1=96.7%
Latha *et al.*, 2024 [20]	Retrospective diagnostic study	BUSI dataset	EfficientNet-B7 with XAI	Independent testing	Sens=99.5%; AUC=1.00; F1=98.9%; Accuracy=99.14%
Al-Tam *et al.*, 2026 [21]	Retrospective diagnostic study	BUSI dataset	Interpretable ensemble transformer framework	External validation	AUC=97.10% (binary), 94.81% (three-class)
Azhar *et al.*, 2024 [22]	Comparative study	BUSI reports	SWE and AI-Powered Synthesis of Structured Multimodal Breast Ultrasound	Comparative validation	Sens: 84.7%; AUC=0.92; F1=85.6%

Güler, 2025 [23]	Retrospective classification study	BUSI dataset	DB-DETL-3	Cross-validation	Sens=99.60%; AUC=0.92; F1=99.6%
Prananda *et al.*, 2022 [24]	Retrospective image analysis study	BUSI dataset	CNN + LIME	Test split	Validation accuracy: 80%

*Diagnostic performance parameters differ across studies because of heterogeneity in the outcomes and metrics reported in the original publications.

ABUS: automated breast ultrasound; AI: artificial intelligence; AUC: area under the receiver operating characteristic curve; BUSI: Breast Ultrasound Images dataset; CAD: computer-aided diagnosis; CI: confidence interval; CNN: convolutional neural network; DB-DETL-3: model designation for a Dirichlet distribution-based deep ensemble transfer-learning framework; F1-score: harmonic mean of precision and recall; GAN: generative adversarial network; GNN: graph neural network; GPT-4: Generative Pre-trained Transformer 4; HER2+: human epidermal growth factor receptor 2-positive; LIME: Local Interpretable Model-Agnostic Explanations; NR: not reported; R-C: radiomics–clinical; REAF: region-of-interest extraction and adaptive fusion; ResNet34: 34-layer residual network; Sens: sensitivity; Spec: specificity; SVM: support vector machine; SWE: shear-wave elastography; TN: triple-negative; VGG16: 16-layer Visual Geometry Group network; XAI: explainable artificial intelligence; XGBoost, extreme gradient boosting. BIScreener, DB-DETL-3, GoogLeNet, and EfficientNet-B7 are model or architecture names rather than conventional abbreviations.

Table 2. QUADAS-3 quality assessment of included studies (n=18)

Author (year)	Patient selection	Index test (AI model)	Reference standard	Flow and timing	Applicability concerns	Overall risk of bias
Kim *et al.*, 2021 [7]	Low	Low	Low	Low	Low	Low
Shen *et al.*, 2021 [8]	Low	Low	Low	Low	Low	Low
Ye *et al.*, 2021 [9]	Low	Low	Low	Low	Low	Low
Luo *et al.*, 2022 [10]	Low	Low	Low	Unclear	Low	Low
Ferre *et al.*, 2023 [11]	Low	Low	Low	Low	Low	Low
Liu *et al.*, 2024 [12]	Low	Low	Low	Low	Low	Low
Ru *et al.*, 2024 [13]	Low	Low	Low	Low	Low	Low
Buzatto *et al.*, 2025 [14]	Low	Low	Low	Low	Low	Low
Chen *et al.*, 2025 [15]	Low	Low	Low	Low	Low	Low
Liu *et al.*, 2025 [16]	Low	Low	Low	Low	Low	Low
Baek *et al.*, 2022 [17]	Low	Low	Unclear	Unclear	Low	Moderate
Hossain *et al.*, 2023 [18]	Low	Low	Low	Unclear	Low	Moderate
Zhang *et al.*, 2023 [19]	Low	Low	Low	Low	Low	Low
Latha *et al.*, 2024 [20]	Low	Low	Low	Low	Low	Low
Al-Tam *et al.*, 2026 [21]	Low	Low	Low	Low	Low	Low
Azhar *et al.*, 2024 [22]	Low	Low	Low	Unclear	Low	Moderate
Güler, 2025 [23]	Low	Low	Low	Low	Low	Low
Prananda *et al.*, 2022 [24]	Low	Unclear	Low	Low	Moderate	Moderate

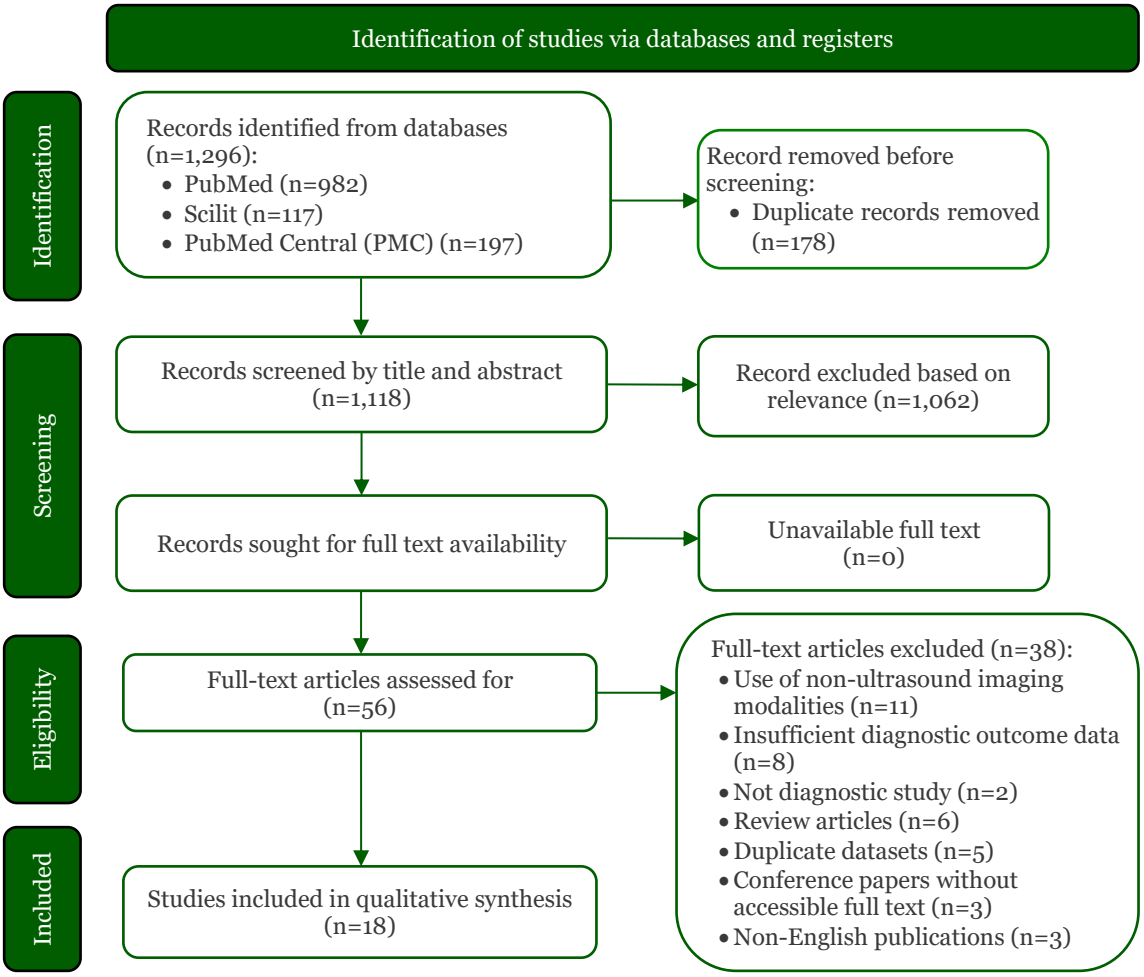

Figure 1. PRISMA 2020 flow diagram for the selection of studies on the diagnostic accuracy of artificial intelligence for early breast cancer detection using imaging data.

Furthermore, the majority of studies clearly described the AI algorithms and diagnostic procedures, thereby reducing concerns regarding the interpretation of the index test. Moderate concerns were primarily identified in the flow and timing domain. Most studies employed retrospective single-center designs, with limited information regarding patient recruitment, consecutive enrollment, or the interval between ultrasound examination and histopathological confirmation. In addition, external validation was unavailable in several studies. Collectively, 14 of the 18 included studies (77.8%) were judged to have a low overall risk of bias, whereas four studies (22.2%) were considered to have a moderate risk of bias, mainly because they relied on single-center datasets or lacked independent external validation cohorts. No study was classified as having a high overall risk of bias.

Diagnostic performance for breast lesion detection and classification

Across the 18 included studies, AI models generally demonstrated favourable performance for breast cancer detection and classification using ultrasound images, ultrasound-derived parameters, clinical and radiomic features, or structured ultrasound reports. Reporting was heterogeneous: 14 studies reported sensitivity, eight reported specificity, 12 reported AUC, nine reported F1-score, and two reported accuracies. Sensitivity ranged from 71.4% to 100.0%, specificity from 71.6% to 96.2%, AUC from 0.778 to 1.00, and F1-score from 0.833 to 0.996. These findings could not be directly compared because the studies differed in diagnostic tasks, datasets, model architectures, outcome reporting, and validation approaches.

Most studies reported high diagnostic performance. Nine studies reported sensitivities of at least 90%, including five with values exceeding 95% and three with values approaching or reaching 100%. Among studies reporting AUC, nine achieved values of at least 0.90, of which five reported AUCs of 0.96 or higher. Similarly, five studies reported F1-scores above 0.90. The

strongest results were generally observed in models developed for binary classification of benign and malignant lesions, with sensitivity ranging from 96.0% to 100.0%, specificity reaching 96.2%, AUC ranging from 0.982 to 1.00, and F1-scores ranging from 0.960 to 0.996.

Models integrating clinical, radiomic, or multimodal information showed favourable but generally more moderate performance. Across these studies, sensitivity ranged from 84.7% to 94.6%, specificity from 76.5% to 82.9%, AUC from 0.869 to 0.96, and F1-score from 0.854 to 0.918. This suggests that incorporating multiple data types may provide clinically relevant discrimination, although these models did not consistently achieve the near-perfect estimates reported by some image-only deep-learning models.

Molecular subtype classification

AI performance was less consistent when the diagnostic objective involved differentiating breast cancer molecular subtypes rather than distinguishing malignant from benign lesions. Ye *et al.* reported sensitivity of 85.4% and an F1-score of 86.0% for CNN-based detection of triple-negative breast cancer. Ferre *et al.* reported sensitivity of 81.8%, specificity of 74.2%, and an AUC of 0.824 for triple-negative breast cancer. Performance was lower for HER2-positive disease, with sensitivity of 71.4%, specificity of 71.6%, and an AUC of 0.778. These findings indicate that subtype prediction may represent a more complex diagnostic task than binary lesion classification. The lower and more variable estimates may reflect subtle and overlapping ultrasound characteristics between molecular subtypes, smaller subtype-specific samples, and differences in reference standards and predictor selection.

AI-assisted interpretation and reporting

Studies assessing AI-assisted interpretation or structured reporting also showed favourable results. Shen *et al.* evaluated AI-assisted breast ultrasound interpretation using independent external validation and reported sensitivity of 95.8%, specificity of 80.1%, and an AUC of 0.976 (95%CI: 0.972–0.980). Liu C. *et al.* reported sensitivity of 95.9% for GPT-4-based structured reporting compared with Bing GPT-4. However, specificity, AUC, and F1-score were not reported, limiting assessment of false-positive classifications and overall discrimination.

Performance according to validation approach

Performance estimates tended to be highest in studies using internal validation, cross-validation, or simple train–test splits. For example, internally or cross-validated models reported AUC values of 0.982–1.00 and sensitivities reaching 100.0%. Studies using public datasets, including BUSI, also frequently reported sensitivities and F1-scores above 96%. Nevertheless, favourable performance was not limited to internally validated studies. Shen *et al.* reported an externally validated AUC of 0.976, while Ru *et al.* reported sensitivity of 83.3%, specificity of 92.0%, an AUC of 0.911, and an F1-score of 0.833 using external multicentre validation. Chen *et al.* reported an AUC of 0.982 on an independent test dataset. Al-Tam *et al.* reported externally validated AUC values of 97.10% for binary classification and 94.81% for three-class classification. Despite these favourable findings, externally validated performance was generally more moderate and less uniformly near-perfect than performance observed in internally validated or public-dataset studies. Ru *et al.*, for example, reported lower sensitivity and F1-score than several internally validated models, although specificity remained high.

Binary versus multiclass classification

Where direct comparisons were available, binary classification appeared to outperform more complex classification tasks. Al-Tam *et al.* reported an AUC of 97.10% for binary classification compared with 94.81% for three-class classification. Similarly, the relatively lower performance reported for molecular subtype prediction supports the interpretation that AI models more readily distinguish broad benign and malignant categories than multiple pathological or molecular classes.

Certainty and limitations of the synthesis

The overall direction of evidence consistently favoured the diagnostic potential of AI, and no study reported clearly poor discriminatory performance. However, confidence in the magnitude

and generalisability of the reported estimates remains limited. Most studies were retrospective, and several used internal validation, cross-validation, or public datasets without independent clinical evaluation. Near-perfect estimates, including AUC values approaching 1.00 and sensitivities above 99%, may therefore reflect overfitting, limited dataset diversity, class imbalance, duplication or dependence between training and test samples, or insufficient control of data leakage.

Interpretation was further limited by incomplete and inconsistent reporting of diagnostic parameters. Some studies reported sensitivity without specificity, while others reported only AUC, F1-score, or accuracy. Differences in lesion prevalence, case selection, reference standards, ultrasound modalities, classification thresholds, and diagnostic objectives also prevented direct numerical comparison. Moreover, models designed for image classification, molecular subtype prediction, radiomics, structured reporting, and multimodal synthesis addressed related but clinically distinct tasks.

Discussion

Findings of the present systematic review demonstrate that AI-assisted ultrasound has achieved remarkable diagnostic performance across a wide range of clinical settings. Most studies reported that deep learning algorithms significantly improved lesion classification accuracy while reducing observer variability compared with conventional ultrasound interpretation. The increasing adoption of CNNs, hybrid learning models, transformer architectures, and explainable AI reflects the rapid evolution of computer-aided diagnosis (CAD) systems for breast imaging [7,11].

One of the most consistent findings across the included studies was the superior capability of deep learning algorithms to differentiate benign from malignant breast lesions. Kim *et al.* developed a weakly supervised deep learning model that accurately localized suspicious lesions while minimizing annotation requirements, thereby facilitating large-scale clinical implementation [7]. Likewise, Shen *et al.* demonstrated that AI-assisted ultrasound interpretation substantially reduced false-positive findings without compromising cancer detection rates, suggesting that AI could reduce unnecessary biopsies and improve clinical decision-making [8]. Similar improvements in diagnostic discrimination were also reported by Ru *et al.*, who proposed a hybrid spatial geometric learning framework capable of maintaining high diagnostic accuracy across multicenter datasets [13].

Beyond lesion classification, several studies explored AI applications for predicting breast cancer biological characteristics. Ye *et al.* successfully developed a CNN capable of identifying triple-negative breast cancer directly from ultrasound images, whereas Ferre *et al.* applied machine learning algorithms to classify triple-negative and HER2-positive molecular subtypes using quantitative ultrasound features [9,15].Furthermore, Liu *et al.* introduced an ultrasound-based machine learning model capable of predicting tumor-infiltrating lymphocytes (TILs), highlighting the growing potential of AI for non-invasive tumor characterization and precision oncology [16]. These developments indicate that AI is evolving beyond simple image classification toward comprehensive radiogenomic prediction and personalized treatment planning.

Recent advances also emphasize improvements in AI interpretability and clinical integration. Traditional deep learning models have frequently been criticized as "black-box" algorithms with limited transparency. To address this limitation, Chen *et al.* introduced BIScreener, an interpretable deep learning framework designed to enhance diagnostic confidence while maintaining high predictive accuracy [15]. Similarly, Latha *et al.* integrated EfficientNet-B7 with explainable AI techniques, allowing clinicians to visualize image regions contributing to diagnostic decisions [20]. Al-Tam *et al.* further proposed an interpretable ensemble transformer architecture that combines transformer learning with explainable attention mechanisms, improving both diagnostic performance and model transparency [21]. Collectively, these studies indicate that explainability has become an essential requirement for the clinical adoption of AI-assisted breast ultrasound.

Another emerging trend identified in this review is the integration of multimodal artificial intelligence and large language models into breast ultrasound workflows. Liu *et al.* compared GPT-4 based structured reporting systems with conventional reporting approaches and demonstrated that large language models could substantially improve reporting consistency and

workflow efficiency [12]. Likewise, Azhar *et al.* proposed an AI-powered multimodal reporting system integrating radiologist annotations with deep learning image analysis to generate standardized breast ultrasound reports [22]. These findings suggest that AI may extend beyond image interpretation to facilitate radiology reporting, communication, and clinical documentation, thereby improving overall diagnostic efficiency.

Another important finding of this systematic review is the continuous evolution of artificial intelligence from conventional lesion classification toward more comprehensive and clinically applicable diagnostic systems. Liu *et al.* demonstrated that integrating large language models GPT-4 into breast ultrasound reporting significantly improved report standardization and workflow efficiency, suggesting a future role for AI beyond image interpretation alone [12]. Buzatto *et al.* further showed that combining ultrasonographic features with clinical variables through machine learning algorithms enhanced the prediction of breast lesion malignancy, supporting the value of multimodal data integration in clinical decision-making [14]. Likewise, Zhang *et al.* proposed the REAF framework, which combines region of interest extraction with adaptive feature fusion to improve lesion localization and classification accuracy, particularly in heterogeneous ultrasound datasets [19]. Image quality enhancement has also emerged as a promising research direction, with Jiménez Gaona *et al.* demonstrating that generative adversarial networks (GANs) effectively reduced ultrasound image noise while preserving diagnostically relevant anatomical features, thereby potentially improving subsequent AI-based classification performance [25]. Furthermore, Güler introduced a Dirichlet distribution-based ensemble learning strategy integrated with transfer learning, achieving robust classification performance by leveraging complementary predictive models and reducing overfitting [23]. Finally, Prananda and Frannita emphasized that the incorporation of XAI is essential for improving clinician trust, facilitating transparent decision-making, and promoting the safe clinical implementation of AI-assisted breast cancer diagnosis [24]. Collectively, these studies indicate that recent advances in AI are not limited to improving diagnostic accuracy but also encompass enhanced image quality, model interpretability, workflow optimization, and clinical usability, thereby strengthening the potential integration of AI into routine breast ultrasound practice.

Despite these promising findings, several limitations were consistently observed across the included studies. Most investigations were retrospective and single-center studies with relatively limited sample sizes, raising concerns regarding selection bias and model generalizability [10,17,18,26]. Considerable heterogeneity also existed among AI architectures, ultrasound equipment, patient populations, image acquisition protocols, validation strategies, and reported performance metrics. Although many studies reported excellent accuracy and area under the receiver operating characteristic AUC, direct comparison between algorithms remains difficult because standardized evaluation frameworks have not yet been universally adopted. Consequently, external multicenter validation remains necessary before widespread clinical implementation.

Conclusion

This systematic review synthesized evidence from 18 studies published between 2021 and 2026 evaluating the application of AI in breast ultrasound for early breast cancer detection. The findings demonstrate that AI-assisted ultrasound has consistently shown high diagnostic performance in differentiating benign from malignant breast lesions, reducing false-positive findings, and improving diagnostic consistency. Recent advances in deep learning, transformer-based models, XAI, ensemble learning, and large language models have further expanded the role of AI beyond lesion classification to include image enhancement, molecular subtype prediction, structured reporting, and clinical decision support. Commonly, AI-assisted breast ultrasound represents a highly promising adjunctive tool for the early detection of breast cancer. With continued technological advancement and rigorous clinical validation, AI has the potential to enhance diagnostic accuracy, improve radiologist's workflow efficiency, reduce unnecessary biopsies, and support precision medicine. Future research should focus on developing robust, interpretable, and clinically generalizable AI systems to facilitate their safe and effective integration into routine breast imaging practice.

Acknowledgments

The authors sincerely thank Universitas Muhammadiyah Aceh for its academic support during the completion of this systematic review. We also appreciate the valuable contributions of all researchers whose published studies formed the basis of this review.

Competing Interests

There is no conflict of interest in the present study for any of the authors.

Funding

This research received no external funding.

Underlying Data

All underlying data have been presented in the article

Declaration Of Artificial Intelligence Use

The authors used OpenAI ChatGPT to assist with language editing, grammar improvement, and manuscript organization during the preparation of this article. All AI-assisted processes were critically reviewed by the authors, and the final decisions and interpretations presented in this article were made exclusively by the authors.

How to cite

Utami WA, Wahyuni S, Zamharira M, Akele RY. Diagnostic accuracy of artificial intelligence-assisted early detection of breast cancer using ultrasound imaging: A systematic review. Narra Rev 2026; 2 (2): e37 - http://doi.org/10.52225/narrarev.v2i2.37.

References

1. Freihat O, Sipos D. Global burden and projections of breast cancer incidence and mortality to 2050: A comprehensive analysis of GLOBOCAN data. Front Public Health 2025;13:1622954.
2. Dan Q, Xu Z, Burrows H, *et al.* Diagnostic performance of deep learning in ultrasound diagnosis of breast cancer: A systematic review. npj Precis Onc 2021;8:21.
3. Li H, Zhao J, Jiang Z. Deep learning-based computer-aided detection of ultrasound in breast cancer diagnosis: A systematic review and meta-analysis. Clin Radiol 2024;79(11):e1403-e1413.
4. Page MJ, McKenzie JE, Bossuyt PM, *et al.* The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
5. Bunnell A, Valdez D, Strand F, *et al.* Artificial intelligence-enhanced handheld breast ultrasound for screening: A systematic review of diagnostic test accuracy. PLOS Digit Health 2025;4(9):e0001019.
6. Salameh JP, Bossuyt PM, Mcgrath TA, *et al.* Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): Explanation, elaboration, and checklist accuracy. BMJ 2020;370:m2632.
7. Kim J, Kim HJ, Kim C, *et al.* Weakly - supervised deep learning for ultrasound diagnosis of breast cancer. Sci Rep 2021;11:24382.
8. 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):5645.
9. Ye H, Hang J, Zhang M, *et al.* Automatic identification of triple negative breast cancer in ultrasonography using a deep convolutional neural network. Sci Rep 2021;11:20474.
10. Luo Xiao, Xu M, Tang G, *et al.* The lesion detection efficacy of deep learning on automatic breast ultrasound and factors affecting its efficacy: A pilot study. Br J Radiol 2022;95(1130):20210438.
11. Ferre R, Elst J, Senthilnathan S, *et al.* Machine learning analysis of breast ultrasound to classify triple negative and HER2 + breast cancer subtypes. Breast Dis 2023;42(1):59-66.
12. Liu C, Wei M, Qin Y, *et al.* Harnessing large language models for structured reporting in breast ultrasound: A comparative study of open AI (GPT-4.0) and Microsoft Bing (GPT-4). Ultrasound Med Biol 2024;50(11):1697-1703.
13. Ru J, Zhu Z, Shi J. Spatial and geometric learning for classification of breast tumors from multi-center ultrasound images: A hybrid learning approach. BMC Med Imaging 2024;24(1):133.

14. Buzatto IPC, Recife SA, Miguel L, *et al.* Machine learning can reliably predict malignancy of breast lesions based on clinical and ultrasonographic features. Breast Cancer Res Treat 2025;211(3):581-593.
15. Chen Y, Wang P, Ouyang J, *et al.* BIScreener: Enhancing breast cancer ultrasound diagnosis through integrated deep learning with interpretability. Anal Methods 2025;17(27):5704-5713.
16. Liu B, Gu X, Xie D, *et al.* An ultrasound-based machine learning model for predicting tumor-in fi ltrating lymphocytes in breast cancer. Technol Cancer Res Treat 2025;24:15330338251334453.
17. Baek J, O'Connell AM, Parker KJ. Improving breast cancer diagnosis by incorporating raw ultrasound parameters into machine learning. Mach Learn Sci Technol 2022;3(4):045013.
18. Hossain ABMA, Nisha JK, Johora F. Breast cancer classification from ultrasound images using VGG16 model based transfer learning. Int J Image Graph Signal Process 2023;15(1):12-22.
19. Zhang Z, Lim JW, Zheng Y, *et al.* REAF: ROI Extraction and Adaptive Fusion for Breast Cancer Diagnosis in Ultrasound Images. 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Istanbul: Türkiye; 2023.
20. Latha M, Kumar PS, Chandrika RR, *et al.* Revolutionizing breast ultrasound diagnostics with EfficientNet-B7 and Explainable AI. BMC Med Imaging 2024;24(1):230.
21. Al-Tam RM, Al-Hejri AM, Hashim FA, *et al.* An interpretable ensemble transformer framework for breast cancer detection in ultrasound images. Diagnostics 2026;16(4);622.
22. Azhar K, Lee BD, Byon SS, *et al.* AI-Powered synthesis of structured multimodal breast ultrasound reports integrating radiologist annotations and deep learning analysis. Bioengineering 2024;11(9):890.
23. Güler O. A dirichlet distribution-based complex ensemble approach for breast cancer classification from ultrasound images with transfer learning and multiphase spaced repetition method. J Imaging Inform Med 2026;39(1):202-228.
24. Prananda AR, Frannita EL. Toward better analysis of breast cancer diagnosis: Interpretable AI for breast cancer classification. IT J Res Dev 2023;7(2):220-227.
25. Jiménez-Gaona Y, Rodríguez-Alvarez MJ, Escudero L. Ultrasound breast images denoising using generative adversarial networks (GANs). Intell Data Anal 2024;28:1661-1678.
26. Kim YS, Lee SE, Chang JM, *et al.* Ultrasonographic morphological characteristics determined using a deep learning-based computer-aided diagnostic system of breast cancer. Medicine 2022;101(3):e28621.