POLITECNICO DI TORINO
Repository ISTITUZIONALE

Artificial intelligence in breast cancer imaging: a systematic review of detection, segmentation, explainability, and clinical translation

Original
Artificial intelligence in breast cancer imaging: a systematic review of detection, segmentation, explainability, and clinical translation / Iqbal Khan, I., Shah, S.A.H., Tsipourakis, A., Malavolta, M., Aleesha, A., Baqir Hussain Shah, S., Hajiani, M., Buccoliero, A., Panagiotopoulos, K., Shah, S.T.H., Deriu, M.A.. - In: FRONTIERS IN IMAGING. - ISSN 2813-3315. - 5:(2026), pp. 1-25. [10.3389/fimag.2026.1910013]

Availability:
This version is available at: 11583/3015131 since: 2026-09-03T10:32:06Z

Publisher:
Frontiers

Published
DOI:10.3389/fimag.2026.1910013

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

27 November 2026

frontiers in Imaging

TYPE Systematic Review
PUBLISHED 03 September 2026
DOI 10.3389/fimag.2026.1910013

Check for updates

OPEN ACCESS

EDITED BY
Marco Arnesano,
University of eCampus, Italy

REVIEWED BY
Stephan Rogalla,
Stanford University, United States
Istiak Ahmed,
Research Lead, Bangladesh
Hima Deepthi Vankayalapati,
SVKM's Narsee Monjee Institute of Management Studies, India

*CORRESPONDENCE
Syed Adil Hussain Shah
✉ syedadilhussain.shah@gpi.it
Syed Taimoor Hussain Shah
✉ taimoor.shah@polito.it

RECEIVED 15 June 2026
REVISED 05 August 2026
ACCEPTED 11 August 2026
PUBLISHED 03 September 2026

CITATION
Khan II, Shah SAH, Tsipourakis A, Malavolta M, Aleesha A, Shah SBH, Hajiani M, Buccoliero A, Panagiotopoulos K, Shah STH and Deriu MA (2026) Artificial intelligence in breast cancer imaging: a systematic review of detection, segmentation, explainability, and clinical translation. *Front. Imaging* 5:1910013. doi: 10.3389/fimag.2026.1910013

COPYRIGHT
© 2026 Khan, Shah, Tsipourakis, Malavolta, Aleesha, Shah, Hajiani, Buccoliero, Panagiotopoulos, Shah and Deriu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Artificial intelligence in breast cancer imaging: a systematic review of detection, segmentation, explainability, and clinical translation

Iqra Iqbal Khan[1,2], Syed Adil Hussain Shah[3,4]*, Alexandra Tsipourakis[3], Marta Malavolta[3], Aleesha Aleesha[5], Syed Baqir Hussain Shah[6,7], Mozhdeh Hajiani[8,9], Andrea Buccoliero[4], Konstantinos Panagiotopoulos[3], Syed Taimoor Hussain Shah[3]* and Marco Agostino Deriu[3]

[1]Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan, [2]Department of Computing and Emerging Technologies, Emerson University Multan, Multan, Pakistan, [3]PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin, Italy, [4]Department of Research and Development (R&D), GPI SpA, Trento, Italy, [5]Department of Computer Science, Ravensbourne University, London, United Kingdom, [6]Department of Computer Science, COMSATS University Islamabad (CUI), Wah, Pakistan, [7]Future Information Innovation Academy, Fudan University, Shanghai, China, [8]Department of Control and Computer Engineering (DAUIN), Politecnico di Torino, Torino, Italy, [9]7HC SRL, Rome, Italy

Introduction: Breast cancer remains the most commonly diagnosed malignancy in women worldwide, with approximately 2.3 million new cases and 670,000 deaths in 2022. Artificial intelligence (AI) is increasingly being applied across breast imaging for detection, characterization, segmentation, risk stratification, explainability, and clinical translation.

Methods: This systematic review synthesized 284 eligible peer-reviewed publications published between 2015 and 2025, identified through a PRISMA 2020-compliant search of four indexed databases. A supplementary relevance-ranked Google Scholar screen was reported separately. A targeted May 2026 narrative update of recent prospective and implementation studies was also conducted without adding these records to the PRISMA denominator.

Results: The reviewed evidence encompassed classical machine learning and radiomics, convolutional neural networks, YOLO-family detectors, vision transformers and hybrid CNN-Transformer architectures, U-Net variants, Mask R-CNN, SAM, and MedSAM across mammography, tomosynthesis, MRI, ultrasound, contrast-enhanced mammography, and emerging photoacoustic imaging. The evidence highlights advances in explainable AI, multimodal and radiogenomic fusion, federated learning, and open-source deployment, while persistent challenges include data heterogeneity, class imbalance, domain shift, bias, incomplete calibration, and limited prospective validation. Retrospective benchmark findings were derived from the 284-study systematic corpus, whereas prospective clinical utility and real-world implementation were informed by the separate 2026 narrative update.

Discussion: AI shows substantial potential to enhance breast cancer imaging, but broader clinical translation requires robust external and prospective validation, improved calibration, assessment of generalizability and bias, and integration into clinical workflows. Supplementary Data Sheet S1 should be

interpreted as a provenance-tagged evidence map rather than a formal study-level risk-of-bias assessment, with 38 entries verified from full text, 48 based on abstracts, and 198 based on DOI/source pages or other web-accessible records.

KEYWORDS

artificial intelligence, breast cancer detection, clinical translation, explainable AI, medical imaging, deep learning, mammography, image segmentation

1 Introduction

Breast cancer remains the most commonly diagnosed malignancy among women globally, with approximately 2.3 million new cases and 670,000 deaths in 2022 (Bray et al., 2024). The burden continues to grow because of population aging, changing reproductive patterns, and late-stage diagnosis in settings with limited imaging infrastructure (Ginsburg et al., 2020). Delayed diagnosis allows tumors to progress from localized disease to regional or distant metastasis, where treatment becomes more complex and survival declines substantially. 5-year relative survival approaches 99% for localized disease but falls to 86% for regional disease and 28% for distant disease, making early detection central to improving outcomes (SEER Cancer Statistics Review, 1975-2018).

Human interpretation remains another major source of diagnostic error. Perceptual errors account for an estimated 60%−80% of missed diagnoses in retrospective audits (Majid et al., 2003), and fatigue during high-volume reporting can reduce radiologist performance (Krupinski et al., 2010; Qureshi et al., 2024). Satisfaction-of-search bias, subtle architectural distortion, and fine microcalcifications further increase diagnostic difficulty. Inter-reader variability, with Cohen's kappa around 0.4–0.6 among experienced radiologists (Elmore et al., 1994), affects recall and biopsy thresholds. The broader breast imaging ecosystem adds modality-specific complexity: tomosynthesis reduces tissue superimposition; dynamic contrast-enhanced MRI offers high sensitivity and moderate specificity, at high cost (Mann et al., 2019); ultrasound supports biopsy guidance and supplementary screening but is operator-dependent; and contrast-enhanced mammography, molecular breast imaging, and photoacoustic imaging add functional or vascular information (Litjens et al., 2017).

Artificial intelligence, and deep learning in particular, has emerged as a transformative augmentation technology across this entire diagnostic spectrum. The methodological landscape spans multiple architectural paradigms that have evolved in rapid succession over the past decade. Classical machine learning and radiomics remain relevant in data-scarce contexts and for interpretable radiogenomic biomarker discovery (Grimm et al., 2015; Saha et al., 2018; van Griethuysen et al., 2017). Convolutional neural networks, which exploit the spatial structure of image data through hierarchical local feature extraction, have dominated breast imaging AI since the ImageNet deep-learning breakthrough (Krizhevsky et al., 2012), consistent with the broader expansion of deep learning in medical image analysis (Litjens et al., 2017; Shah et al., 2024; Shen et al., 2017). Single-stage YOLO detection frameworks formulate lesion localization as a direct regression problem and enable fast inference, suitable for real-time detection workflows (Redmon et al., 2016). Vision transformers apply global self-attention to sequences of image patches and have motivated hybrid CNN-Transformer approaches for medical image analysis (Chen et al., 2021; Dosovitskiy et al., 2020; Liu et al., 2021). Specialized encoder-decoder segmentation architectures, including U-Net and transformer-enhanced variants, have become foundational for lesion boundary delineation across medical imaging tasks (Chen et al., 2024; Ronneberger et al., 2015).

The translational potential of these AI advances has increasingly moved from research demonstration toward prospective clinical evaluation and early implementation. Several AI-enabled imaging systems have received regulatory clearance or are listed among AI/ML-enabled medical devices by the U.S. Food and Drug Administration (Almarie et al., 2025; Health C. for D and R., 2026). External evaluations of commercial mammography AI systems have also shown that AI can support screening interpretation, although prospective utility, workflow effects, and deployment context remain essential determinants of clinical value (Elías-Cabot et al., 2026; Salim et al., 2020; Yilmaz et al., 2025). Recent evidence has further strengthened this translational direction: the final MASAI randomized trial reported non-inferior interval cancer rates with higher sensitivity and unchanged specificity for AI-supported mammography screening (Gommers et al., 2026); the AI-STREAM prospective multicenter cohort in South Korea reported a 13.8% increase in cancer detection without a significant recall-rate increase (Chang et al., 2025); and the AITIC paired noninferiority trial showed that partially autonomous AI triage reduced radiologist workload by 63.6% while increasing cancer detection, although recall rates increased in the overall cohort (Elías-Cabot et al., 2026). Despite regulatory progress, clinical adoption remains heterogeneous, constrained by PACS and electronic health record integration challenges, unclear reimbursement pathways, and persisting radiologist concerns about interpretability and generalizability (Goh et al., 2025).

Explainability approaches, including saliency maps, lesion-localization outputs, segmentation masks, attention visualization, and case-based interpretability, have consequently become important components of breast-imaging AI evaluation, supporting spatial transparency, failure analysis, and radiologist review, in the context of high-risk AI regulation and medical-device assessment (Adadi and Berrada, 2018; Barnett et al., 2021; Byra et al., 2020; Douglas et al., 2023; Huo et al., 2021; Ma et al., 2024; Regulation (EU), 2024; Ribli et al., 2018; Sani et al., 2024; Selvaraju et al., 2017; Vakanski et al., 2020). At the same time,

algorithmic fairness has emerged as a regulatory and ethical imperative, following demonstrations that non-representative medical imaging datasets can produce biased classifiers (Larrazabal et al., 2020), while federated learning frameworks address the complementary privacy challenge of building generalizable models across institutions without centralizing sensitive patient data (Kaissis et al., 2020; Rieke et al., 2020).

Open-access resources have also accelerated the field. Public datasets including CBIS-DDSM, INbreast, VinDr-Mammo, BUSI, Duke Breast Cancer MRI, and EMBED enable reproducible benchmarking and fairness auditing (Jeong et al., 2023; Lee et al., 2017; Nguyen et al., 2023; RSNA, 2026; Saha et al., 2018). Open-source systems such as Mirai, nnU-Net, and MedSAM allow local validation (Ma et al., 2024; Nowakowska et al., 2023; Yala et al., 2021), while Streamlit, Flask, Gradio, and TensorFlow.js support lightweight or browser-based deployment (Cohen et al., 2019; Smilkov et al., 2019). These resources democratize breast imaging AI, although most studies remain retrospective and inconsistent in external validation, fairness assessment, explainability, and workflow evaluation.

The remainder of this review is organized as follows: Section 2 describes the PRISMA-guided methodology, PICOS framework, search strategy, study selection, data extraction, and synthesis approach. Section 3 presents the results and evidence synthesis, including clinical rationale, benchmark datasets, imaging modalities, AI model families from classical machine learning through vision transformers and segmentation architectures, explainability methods, multimodal fusion, web deployment platforms, and the structured evidence-quality and methodological-limitations assessment. Section 4 discusses challenges, review limitations, and future research priorities. Section 5 concludes the review.

2 Methods

2.1 Reporting guideline and protocol registration

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021). The protocol was not prospectively registered, and this is acknowledged as a review limitation. To maximize reproducibility despite the absence of registration, the revised manuscript provides the complete database-specific search strings, search dates, record counts, screening flow, PICOS eligibility criteria, data-extraction fields, source-provenance rules, and predefined evidence-quality and methodological-limitations criteria in the main text and Supplementary Material. Only peer-reviewed original journal articles and full conference papers reporting quantitative breast-imaging AI results were eligible for the systematic evidence base of 284 included studies. Protocols without results, systematic or narrative reviews, dataset or resource descriptions without predictive-model evaluation, preprints, deployment frameworks, regulatory listings, general algorithms, and non-breast-specific methodological sources are cited only as contextual or technical background, are not counted in the PRISMA denominator, and are listed separately in Supplementary Table S9 where applicable. The PRISMA 2020 flow diagram is presented in Figure 1.

2.2 PICOS framework

The eligibility criteria were structured according to the PICOS framework as follows. Population: adult women (age 18 years and above) undergoing breast imaging for screening, diagnosis, or treatment monitoring; studies on pediatric populations were excluded, as detailed in Table 1. Intervention: any artificial intelligence, machine learning, or deep learning method applied to breast imaging data for detection, classification, segmentation, risk prediction, or explainability. Comparator: human radiologist performance, conventional computer-aided detection systems, classical machine learning baselines, or no comparator; single-arm studies were included but interpreted as lower-level evidence. Outcomes: primary outcomes were diagnostic accuracy metrics including area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and F1 score; secondary outcomes were segmentation performance (Dice coefficient, Hausdorff distance, and IoU), risk prediction performance (C-index, calibration), and clinical utility metrics (workload reduction, reading time). Setting: any clinical or research imaging setting, including single-institution retrospective studies, multi-institution prospective trials, and grand challenge benchmarks.

2.3 Search strategy and study selection

A systematic search was performed across PubMed/MEDLINE, Scopus, IEEE Xplore, and Web of Science Core Collection for publications released from January 2015 to December 2025, using terms related to breast cancer, mammography, MRI, ultrasound, deep learning, machine learning, YOLO, vision transformers, segmentation, explainable AI, radiomics, and computer-aided diagnosis. Each database was queried independently using database-specific syntax and field tags; the complete verbatim search strings, search dates, and retrieved record counts are provided in Supplementary Table S7. Persistent access links or identifiers for the public datasets are listed in Supplementary Table S3. The four indexed database searches were last run on 15 December 2025 and identified 8,580 records before deduplication. Because Google Scholar does not support stable Boolean export or reliable deduplicated retrieval, the first 300 relevance-ranked Google Scholar records were screened only as a supplementary source to identify potentially missed studies and were not added to the main PRISMA database denominator. After removal of 2,668 duplicate records, 5,912 records were screened; 4,065 records were excluded at title/abstract screening; 1,847 full texts were sought and assessed; and 284 peer-reviewed studies met the eligibility criteria for inclusion. To directly address whether the supplementary Google Scholar screen contributed to the quantitative corpus, none of the 300 relevance-ranked Google Scholar records yielded an eligible study that was added to the 284-study quantitative corpus. The 284 included studies were drawn

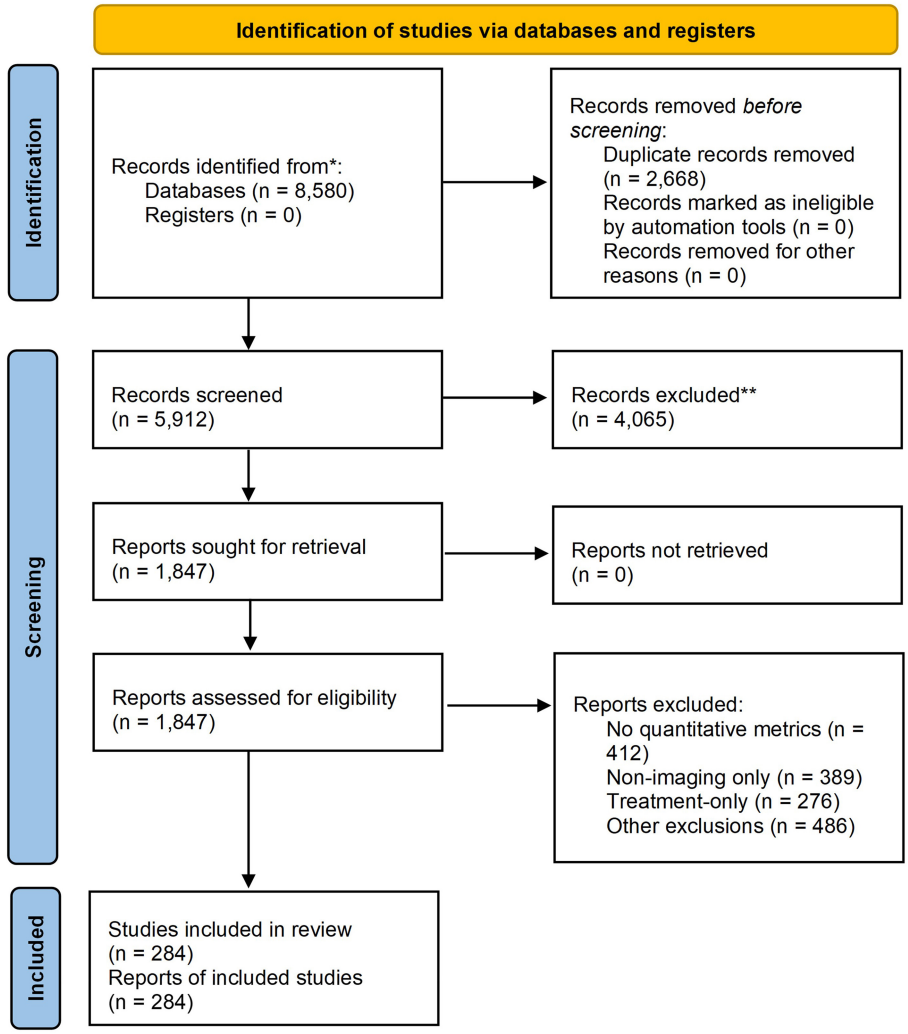

FIGURE 1
PRISMA 2020 flow diagram for systematic literature search and study selection.

exclusively from the four indexed databases, and the Google Scholar screen functioned only as a completeness check for potentially missed records; any items it surfaced were either already retrieved by the indexed-database searches or, where non-indexed, retained solely for qualitative or narrative contextualization and listed among the context-only sources (Supplementary Table S9).

During revision, as a *post hoc* quality-control assessment, two reviewers independently reassessed a computer-generated random sample of 591 of the 5,912 title and abstract records (10.0%) and 185 of the 1,847 full-text reports (10.0%) using the predefined eligibility criteria, with each reviewer blinded to the other's initial decisions. Inter-reviewer agreement was calculated from the initial inclusion and exclusion decisions recorded before consensus discussion. The reviewers agreed on 443 of the 591 title and abstract records (75.0%) and on 157 of the 185 full-text reports (84.9%). All extraction discrepancies were resolved by consensus before the final synthesis. Fully independent duplicate screening and extraction across the entire corpus, as recommended for Cochrane-style systematic reviews, was not operationally feasible for this single-team review given the volume of records handled (8,580 identified; 5,912 screened after deduplication; 1,847 full texts assessed). Primary screening was therefore performed by one reviewer, and a blinded, computer-generated random dual-screening sample (10% at each of the title/abstract and full-text stages) was used to quantify inter-reviewer reliability rather than to duplicate the entire screening process. This is a deliberate, resource-driven deviation from full dual-independent screening and is acknowledged as a limitation of this review (see limitations of this review); it is also why the appraisal is described as a structured evidence-quality assessment rather than a formal risk-of-bias evaluation.

No automated screening or AI-assisted triage tool was used. The targeted narrative update was searched on 20 May 2026 using the same concept blocks restricted to recent prospective, multicentre, and implementation studies; these records were narratively integrated in the clinical-translation discussion and were not added to the PRISMA denominator.

TABLE 1 PICOS eligibility framework for systematic literature search.

PICOS element	Inclusion criteria	Exclusion criteria
Population	Adult women (age >= 18) undergoing breast imaging (screening, diagnosis, monitoring)	Pediatric populations; non-breast cancers; studies focused exclusively on male breast cancer
Intervention	Any AI/ML/DL method applied to breast imaging: classification, detection, segmentation, risk scoring, and XAI	Non-imaging AI methods; purely genomic or clinical prediction without imaging
Comparator	Radiologist performance; classical CAD; ML baselines; no comparator (single-arm)	No exclusion based solely on comparator; studies without an eligible quantitative outcome were excluded
Outcomes	AUC, sensitivity, specificity, F1 score, Dice, IoU, C-index, Hausdorff distance, and workload reduction	Qualitative only; opinion or editorial studies
Study design	Original peer-reviewed journal articles and full conference papers reporting quantitative breast-imaging AI results	Protocols without results; systematic, scoping, or narrative reviews; dataset/resource papers without model evaluation; case reports; editorials; opinion pieces; letters; non-English publications

2.4 Data extraction and synthesis

For each included study, the review recorded publication year, imaging modality, dataset/source, sample size, model architecture, task type, validation strategy, comparator or reference standard, primary performance metrics, explainability method, open-source/code availability, deployment status, regulatory-clearance status, and methodological limitations. Full reports were used to determine eligibility for the 284 included studies. The full-text eligibility review and the later expansion of Supplementary Data Sheet S1 were distinct processes. Supplementary Data Sheet S1 is therefore now described as a study-level evidence map and extraction-verification log, rather than as a complete full-text extraction dataset.

The evidence map retains the source used to verify each row and now includes an explicit full-text verification status. Of the 284 entries, 38 were verified from full text, 48 were based on abstracts, and 198 were based on DOI/source pages or other web-accessible records. For entries not verified from a full report, the spreadsheet now states that the information is source-limited. In those rows, "not reported" means only that the item was not identified in the consulted source; it is not interpreted as evidence that the feature was absent from the full article. Such entries were not assigned a definitive study-level methodological-quality judgment.

A second reviewer independently checked 57 of the 284 included studies (20.1%) across the predefined core fields, including imaging modality, dataset, model architecture, task, validation design, comparator, performance metrics, and methodological-limitations information. All discrepancies identified during this quality-control assessment were resolved by consensus. This inter-reviewer reliability check is distinct from the per-study full-text verification status reported above: the former quantifies agreement between two reviewers on a sample of studies, whereas the latter records the type of source consulted for each evidence-map row (full text, abstract, or source record). Because a validated QUADAS-2 or PROBAST instrument was not applied study by study to all 284 full reports, this review no longer describes the appraisal as a formal risk-of-bias assessment. Instead, evidence was contextualized using predefined review-specific criteria covering population and participant selection, index-model development, reference standard, validation independence, flow and timing, reporting transparency, calibration and subgroup reporting, and clinical applicability. The operational criteria are provided in Supplementary Table S1. Owing to substantial heterogeneity across modalities, datasets, architectures, endpoints, validation protocols, and reporting quality, quantitative meta-analysis was not performed. Findings were narratively synthesized and weighted by evidence level, with prospective and externally validated studies interpreted more strongly than single-institution retrospective benchmarks.

3 Results and evidence synthesis

Unless otherwise noted, performance values reported in the model-family sections represent retrospective, dataset-specific benchmarks rather than evidence of clinical readiness; a non-pooled quantitative range summary is provided in Supplementary Table S4. Representative study-level evidence-map entries are shown in Supplementary Table S8, and a minimum reporting checklist for future studies is provided in Supplementary Table S6. To make the constraints of each cited work explicit, the principal methodological limitation and evidence-quality context for studies appearing in the model-family literature tables are reported in Supplementary Table S2. Supplementary Data Sheet S1 provides provenance-tagged evidence-map fields for all 284 studies and identifies which rows were verified from full text.

3.1 Epidemiology and clinical rationale

According to GLOBOCAN 2022, breast cancer accounted for 11.6% of all new cancer cases worldwide, with approximately 2.3 million cases and 670,000 deaths in 2022 (Bray et al., 2024). Incidence rates are highest in Australia and New Zealand, Northern America, and Northern Europe, yet mortality-to-incidence ratios are substantially higher in sub-Saharan Africa and parts of South Asia, reflecting stark global inequities in access to early detection and treatment infrastructure (Ginsburg et al., 2020). The economic burden is estimated at USD 88 billion annually in direct healthcare costs (Chen et al., 2018). Population-based

screening has reduced mortality by 15%−25% in randomized trials (Broeders et al., 2012), but performance is substantially degraded in dense breasts, where sensitivity falls to 48% (Sprague et al., 2014). False-positive rates of 10%−12% per round generate unnecessary procedures and significant patient anxiety (Lehman et al., 2017), while inter-reader agreement among experienced radiologists ranges only from 0.4 to 0.6 as Cohen's kappa (Elmore et al., 1994). Prospective and implementation studies have demonstrated that AI-assisted screening workflows can reduce radiologist workload while maintaining or improving selected screening performance metrics, although the magnitude of benefit depends on study design, threshold setting, and clinical workflow (Chang et al., 2025; Elías-Cabot et al., 2026; Gommers et al., 2026; Salim et al., 2020; Yilmaz et al., 2025).

3.2 Publicly available benchmark datasets for breast cancer AI

The reproducibility and generalizability of breast cancer AI research depends critically on well-curated, openly accessible benchmark datasets. Standardized public datasets enable consistent evaluation, facilitate cross-study comparison, and allow researchers in resource-limited settings to develop and validate AI systems without the time, cost and high complexity of prospective data collection. This section provides a comprehensive review of the major publicly available datasets organized by imaging modality, followed by a summary table for rapid reference.

3.2.1 Mammography datasets

The Digital Database for Screening Mammography (DDSM; Heath et al., 1998) contains 2,620 digitized-film mammography studies, corresponding to approximately 10,480 images, with mass and calcification annotations and was a foundational CAD benchmark. CBIS-DDSM (Sawyer-Lee et al., 2016) reformatted a curated subset of DDSM into standardized DICOM with verified ROIs, BI-RADS assessments, and pathology labels, becoming one of the most widely used mammography AI datasets. INbreast (Moreira et al., 2012) provides 115 full-field digital mammography cases with 410 images and high-quality XML annotations. VinDr-Mammo (Nguyen et al., 2023) includes 5,000 four-view examinations, corresponding to 20,000 images, with BI-RADS consensus labels. The RSNA 2022 mammography dataset (Trivedi et al., 2026) contains 54,706 screening mammography images. EMBED (Jeong et al., 2023), with approximately 3.4 million mammograms and linked demographic metadata, is central for fairness research, while CMMD (Cui et al., 2021) adds biopsy-confirmed Chinese cases.

3.2.2 Breast MRI and ultrasound datasets

For breast MRI research, the Duke Breast Cancer MRI dataset (Saha et al., 2022), accessible through TCIA, provides 922 pre-treatment dynamic contrast-enhanced MRI examinations with expert tumor segmentations, clinical metadata, and linked genomic data, making it the primary benchmark for both lesion classification and radiogenomic association studies. The ISPY1 and ISPY2 trial datasets (Li et al., 2022), also available through TCIA, offer longitudinal DCE-MRI from neoadjuvant chemotherapy trials with pathological complete response labels, enabling treatment response prediction research. In breast ultrasound, the Breast Ultrasound Images dataset (BUSI; Al-Dhabyani et al., 2020), comprising 780 images across normal, benign, and malignant classes from Women's Imaging Center at Cairo University, with pixel-level segmentation masks, has emerged as the standard ultrasound benchmark and is freely available on Kaggle. The UDIAT Benchmark Dataset (Thomas et al., 2023), comprising 163 B-mode ultrasound images with expert segmentation masks from the UDIAT Diagnostic Center in Spain, and the STU Hospital dataset (soso, 2026) of 733 ultrasound images with binary lesion masks, provide additional segmentation benchmarks; both are available through institutional requests or on GitHub repositories.

3.2.3 Histopathology and multi-modal datasets

For histopathological image analysis, the BreakHis dataset (Spanhol et al., 2016), comprising 7,909 microscopy images across eight tumor subtypes at four magnification factors from the P&D Laboratory in Brazil, is publicly available from the Federal University of Parana and provides the standard multi-class classification benchmark for breast histopathology. The Breast Cancer Semantic Segmentation dataset (BCSS; Amgad et al., 2019), hosted on TCIA under a Creative Commons license, provides pixel-level annotations across five tissue classes for over 20,000 patches extracted from TCGA whole-slide images. The CAMELYON16 and CAMELYON17 grand challenge datasets (Golden, 2017; Litjens et al., 2018) provide whole-slide images of lymph node sections for metastasis detection and patient-level classification, with detailed annotations from the grand challenge platform. For multi-modal and radiogenomic research, the TCGA-BRCA dataset (The Cancer Genome Atlas Network, 2012) within The Cancer Genome Atlas provides linked genomic, proteomic, clinical, and imaging data for 1,098 breast cancer patients, enabling integrated analysis of imaging phenotypes and molecular tumor profiles. The comprehensive catalog of these datasets is summarized in Table 2.

3.3 Imaging modalities and AI augmentation

A diverse ecosystem of imaging modalities is used across the breast cancer diagnostic pathway, and each modality creates different opportunities for AI augmentation. Digital mammography remains the cornerstone of population-based screening and is commonly targeted by CNNs, YOLO-family detectors, vision transformers, and risk-prediction models for lesion detection, classification, localization, and screening triage (Al-Masni et al., 2018; Aly et al., 2021; McKinney et al., 2020; Yala et al., 2021). Digital breast tomosynthesis extends mammography into quasi-three-dimensional imaging, where AI

TABLE 2 Publicly available benchmark datasets for breast cancer AI research (classification and segmentation).

Dataset	Modality	Task	Size	Access/license	Key characteristics
CBIS-DDSM (Sawyer-Lee et al., 2016)	Mammography	Classif. + Detection	1,644 abnormality cases	TCIA/Public	Curated DICOM subset; 753 calcification and 891 mass cases; ROI masks, BI-RADS, and pathology labels
INbreast (Moreira et al., 2012)	Mammography	Classif. + Detection	115 cases (410 imgs)	Public (request)	High-quality FFDM; XML annotations; masses, calcifications, asymmetry, distortion
VinDr-Mammo (Nguyen et al., 2023)	Mammography	Classif. + Detection	5,000 exams (20,000 images)	PhysioNet/CC BY 4.0	Vietnamese population; BI-RADS 1–5; radiologist consensus; cross-population benchmark
RSNA MMG 2022 (Trivedi et al., 2026)	Mammography	Classification	54,706 images	Kaggle/CC0	Screening MMG; binary cancer labels; multiple scanner types; largest public MMG dataset
DDSM (Heath et al., 1998)	Mammography	Classif. + Detection	2,620 studies (~10,480 images)	Public (UF)	Original digitized film MMG; foundational benchmark; superseded by CBIS-DDSM for DL
EMBED (Jeong et al., 2023)	Mammography	Classification	3.4 M images	AWS Open Data	Racially diverse (4 groups); FDA AI comparison; primary fairness benchmark
CMMD (Cui et al., 2021)	Mammography	Classification	1,775 patients	TCIA/Public	Chinese population; biopsy-confirmed; mass + calcification types; clinical metadata
Duke Breast MRI (Saha et al., 2022)	Breast MRI (DCE)	Classif. + Segmentation	922 patients	TCIA/Public	Pre-treatment DCE-MRI; tumor segmentations; clinical + genomic metadata
ISPY1/ISPY2 (Li et al., 2022)	Breast MRI	Segmentation + pCR	237/719 patients	TCIA/Public	Longitudinal MRI; neoadjuvant treatment response; genomic data linked
BUSI (Al-Dhabyani et al., 2020)	Ultrasound	Classif. + Segmentation	780 images	Kaggle/Public	Normal/benign/malignant; pixel masks; most cited US benchmark
UDIAT (Thomas et al., 2023)	Ultrasound	Segmentation	163 images	Public (request)	B-mode US; expert masks; used for U-Net benchmark comparisons
STU Hospital (soso, 2026)	Ultrasound	Segmentation	733 images	GitHub/Public	Diverse lesion morphologies; binary masks; code available
BCSS (Amgad et al., 2019)	Histopathology	Segmentation	20,000+ patches	TCIA/CC BY 3.0	WSI patches; 5 tissue classes; pixel-level; TCGA slides
BreakHis (Spanhol et al., 2016)	Histopathology	Classification	7,909 images	Public (UFPR)	8 subtypes; 4 magnifications (40x-400x); benign/malignant split
CAMELYON16/17 (Golden, 2017; Litjens et al., 2018)	Histopathology	Detection + Segmentation	1,399 WSIs (combined release)	Grand Challenge/CC0	Combined CAMELYON16/17 release; lymph-node metastasis WSIs from multiple centers
TCGA-BRCA (The Cancer Genome Atlas Network, 2012)	Multi-modal	Classification + Genomics	1,098 patients	GDC/Open	Genomic + imaging + clinical; intrinsic subtype labels; radiogenomics benchmark

can support slice aggregation, volumetric lesion detection, and reading-time reduction (Conant et al., 2019; Elías-Cabot et al., 2026; Gommers et al., 2026). Dynamic contrast-enhanced MRI provides high sensitivity for invasive cancer detection and is well suited to radiomics, kinetic enhancement analysis, CNN classification, and multimodal fusion, although cost and protocol variability limit broad deployment (Hu et al., 2020; Mann et al., 2019; Saha et al., 2018). Breast ultrasound is central for supplementary screening and biopsy guidance, with AI applications focused on lesion classification, margin assessment, real-time detection, and U-Net-based segmentation (Al-Dhabyani et al., 2020; Lu et al., 2025; Mostafa et al., 2025;

TABLE 3 Breast imaging modalities, AI integration strategies, and representative performance benchmarks.

Modality	AI application	Best reported AUC	Sensitivity	Key limitations
Digital mammography	CNN detection, YOLO localization, ViT classification, and risk scoring	0.97 in selected evaluation (McKinney et al., 2020)	Algorithm-dependent (McKinney et al., 2020; Salim et al., 2020; Shen et al., 2019)	Limited sensitivity in dense tissue; 2D superimposition artifact
Digital breast tomosynthesis	3D CNN, slice aggregation, and volumetric detection/segmentation	0.96 in selected DBT evaluation (Conant et al., 2019)	91.8% in selected DBT evaluation (Conant et al., 2019)	Higher radiation dose than 2D mammography; longer reading time without assistance
Breast MRI (DCE)	Radiomics, CNN classification, kinetic modeling, and cross-modal fusion	High but protocol-dependent (Mann et al., 2019)	~90%–95% sensitivity reported for MRI (Mann et al., 2019)	High cost; moderate specificity; scanner and protocol variability
Breast ultrasound	CNN classification, YOLO detection, and U-Net segmentation	Varies by dataset; commonly benchmarked on BUSI/UDIAT (Al-Dhabyani et al., 2020; Yap et al., 2018)	Operator- and dataset-dependent (Lu et al., 2025; Yap et al., 2018)	Operator dependency; low specificity; probe-pressure variability
Contrast-enhanced MMG	CNN-radiomics hybrids, dual-energy analysis	Evidence still emerging	Evidence still emerging	Contrast agent reactions; limited clinical availability
PET/CT	Metabolic staging support and radiomics	Not a primary screening benchmark	Context-dependent	High cost; radiation exposure; not for primary screening
Molecular breast imaging	Density-adjusted scoring and supplemental screening support	Limited public AI benchmarks	Context-dependent	Radiation dose; limited availability; not standard of care
Photoacoustic imaging	Emerging deep models and feature extraction	Limited large-scale validation	Limited large-scale validation	Experimental; limited clinical validation

Yap et al., 2018). Contrast-enhanced mammography, molecular breast imaging, PET/CT, and photoacoustic imaging are less established as public AI benchmarks but provide complementary functional, metabolic, or vascular information that may support future multimodal decision-support systems (Litjens et al., 2017). Table 3 summarizes the major imaging modalities, representative AI applications, performance ranges, and clinical limitations.

3.4 Classical machine learning and radiomics

Classical machine learning and radiomics remain important in breast imaging AI because they provide structured, feature-based representations that are often more interpretable than end-to-end deep learning models. Representative classical machine learning and radiomics studies are summarized in Table 4. Radiomic pipelines usually extract first-order intensity, shape, texture, and higher-order transform features before applying classifiers such as logistic regression, support vector machines, random forests, or gradient-boosting models. van Griethuysen et al. (2017) introduced PyRadiomics, an open-source Python package for standardized radiomic feature extraction. Although PyRadiomics is not a breast cancer-specific diagnostic model, it provides reproducible extraction of radiomic features across MRI, mammography, ultrasound, CT, and histopathology, supporting transparent feature-level interpretation and reproducible radiomics research. It does not report diagnostic AUC, sensitivity, or specificity because it is a feature-extraction framework rather than a trained classifier.

Saha et al. (2018) developed a machine-learning radiogenomic analysis using 922 patients with invasive breast cancer and pre-operative dynamic contrast-enhanced breast MRI. Their algorithm extracted 529 imaging features from the tumor and surrounding tissue, trained models on 461 patients, and evaluated them on the remaining 461 patients to predict molecular subtype, ER, PR, HER2 status, and Ki-67 proliferation status. The study demonstrated moderate associations between MRI-derived quantitative features and molecular/genomic biomarkers, supporting the feasibility of non-invasive radiogenomic assessment. Its strength is the large cohort and public TCIA-linked dataset, but it should not be presented as a clinically deployed diagnostic system; no web application or real-time clinical interface was provided.

Grimm et al. (2015) investigated computational radiogenomics using routine pre-operative breast MRI from 275 breast cancer patients. The study used computer vision-derived MRI features to examine associations with molecular subtypes, particularly luminal

TABLE 4 Classical machine learning and radiomics studies for breast cancer classification and risk prediction.

Study/ resource [Ref]	Method	Dataset	Task	Performance/ main finding	XAI/ interpretability	Open source/Web app
van Griethuysen et al. (2017)	PyRadiomics standardized feature extraction	Multi-modality imaging	Radiomic feature computation	Not a diagnostic classifier; provides reproducible first-order, shape, texture, and filtered radiomic features	Transparent hand-crafted features	Open-source Python package; no breast-specific web app
Saha et al. (2018)	Machine-learning radiogenomics using 529 DCE-MRI features	922 breast cancer patients; Duke/TCIA-linked MRI cohort	Prediction of molecular subtype, ER, PR, HER2, and Ki-67 status	Reported moderate associations between MRI features and molecular/genomic biomarkers; no single pooled AUC should be assigned	Feature-level radiogenomic interpretation	Public TCIA-linked dataset; no clinical web app
Grimm et al. (2015)	Computer vision-based MRI radiogenomics	275 pre-operative breast MRI cases	Association of MRI features with luminal A/B molecular subtypes	Demonstrated subtype-related imaging phenotype associations; not a deployed diagnostic classifier	Feature-level biological interpretation	No web app or deployed model reported
I-SPY/Duke MRI/TCGA-BRCA resources (Grimm et al., 2015; Hylton et al., 2012; Saha et al., 2018; The Cancer Genome Atlas Network, 2012)	Radiomics and imaging-genomic linkage	Breast MRI, clinical, genomic, and treatment-response data	Subtype exploration, treatment-response prediction, radiogenomics	Useful for reproducible model development; performance varies by preprocessing and validation design	Feature-based interpretation possible	Public or controlled-access repositories

Study-specific methodological limitations for these entries are summarized in Supplementary Table S2, with provenance-tagged evidence-map fields provided in Supplementary Data Sheet S1.

A and luminal B disease. Their findings showed that breast MRI imaging phenotypes were associated with molecular subtype-related tumor biology, supporting the rationale for radiogenomics in breast cancer. However, this work should be described as an association-based radiogenomic study rather than a standalone diagnostic classifier; therefore, a single AUC, sensitivity, specificity, or Dice score should not be assigned unless taken directly from the original paper. No clinical web application or open-source deployment was reported.

Clinical trial datasets and public repositories such as Duke Breast Cancer MRI, I-SPY, and TCGA-BRCA provide the strongest basis for reproducible breast radiomics and radiogenomics (Grimm et al., 2015; Hylton et al., 2012; Saha et al., 2018; The Cancer Genome Atlas Network, 2012). These resources allow imaging features to be linked with molecular subtype, treatment response, and clinical outcome. However, cross-study comparison remains difficult because radiomic performance depends heavily on acquisition protocols, segmentation quality, feature harmonization, model selection, and validation strategy. Therefore, radiomics studies should be interpreted cautiously unless they include external validation, transparent feature-selection procedures, and publicly available code.

3.5 Convolutional neural networks

CNNs have dominated breast imaging AI since AlexNet (Krizhevsky et al., 2012). Their use of local receptive fields, weight sharing, and hierarchical features makes them well suited for spatial pattern recognition, while transfer learning from ImageNet remains a common strategy when labeled medical datasets are limited. Representative CNN studies are summarized in Table 5.

McKinney et al. (2020) developed a deep-learning mammography system and evaluated it retrospectively on independent screening cohorts from the United Kingdom and United States. Compared with historical clinical readings, the system produced absolute reductions in false positives of 1.2% in the UK and 5.7% in the US, and absolute reductions in false negatives of 2.7 and 9.4%, respectively. In an independent reader study involving six radiologists, the AI system achieved an AUC of 0.871 compared with a mean reader AUC of 0.750. A simulated UK double-reading workflow maintained non-inferior performance while reducing second-reader workload by 88%. These results provide strong retrospective external evidence, but the study was not a prospective clinical deployment trial, and the training code and model were not publicly released.

TABLE 5 CNN-based studies for breast cancer detection, classification, and risk prediction.

Study [Ref]	Architecture	Dataset	Task	Performance	XAI	Open source/ web/clinical
McKinney et al. (2020)	Deep-learning mammography system	Independent UK + US screening cohorts	Screening cancer prediction	AUC 0.871 in reader study; FP and FN reductions; 88% simulated second-reader workload reduction	Localization analysis	Retrospective external evaluation; model not public
Shen et al. (2019)	End-to-end all-convolutional CNN	DDSM + INbreast	Cancer detection/ classification	AUC 0.91 on DDSM; AUC 0.98 on INbreast	Lesion-level localization output	Implementation resources released; retrospective benchmarks
Yala et al. (2021)	Multi-view deep-learning risk model (Mirai)	MGH + Sweden + Taiwan cohorts	1- to 5-year risk prediction	C-index 0.76–0.81 across held-out cohorts	No dedicated lesion-level XAI	GitHub, MIT license; retrospective external validation
Ribli et al. (2018)	Faster R-CNN (VGG-16)	DDSM + INbreast	Lesion detection	AUC 0.95; FROC +10%	Bbox explanation	GitHub (open source)
Lu et al. (2025)	MTL-OCA/Res-UNet + object contextual attention	OASBUD + UDIAT breast ultrasound	Joint segmentation + classification	OASBUD: Dice 83.75%, IoU 72.03%, Acc 91.67%; UDIAT: Dice 84.85%, IoU 73.68%, Acc 94.79%	Grad-CAM heatmaps; segmentation masks	No standalone clinical web app reported
Hu et al. (2020)	Pretrained CNN feature extraction + SVM fusion	927 lesions from 616 women; DCE-MRI + T2w MRI	Benign/malignant lesion classification	AUC: DCE 0.85; T2w 0.78; image fusion 0.85; feature fusion 0.87; classifier fusion 0.86	Not reported	No clinical web app or prospective deployment reported
Leong et al. (2022)	Transfer learning CNNs: ResNet50, ResNet34, VGG16, AlexNet	Mammographic ROI images with microcalcifications	Microcalcification discrimination/ classification	ResNet50 Acc 97.58%; ResNet34 Acc 97.35%; VGG16 Acc 96.97%; AlexNet Acc 83.06%	Not reported; confusion matrices used	No clinical web app reported
Lotter et al. (2021)	ResNet (annot-efficient)	Private MMG + DBT	Detection (few labels)	AUC 0.94 MMG/0.93 DBT	Grad-CAM	Nat Med supp. code
Calisto et al. (2020)	BreastScreening multimodal visual interface	31 clinicians; 566 images; mammography, ultrasound, MRI	Multimodal breast imaging diagnosis support	Reduced time per image in multimodality workflow; usability/workload and BI-RADS-related evaluation reported; no AUC/sensitivity/specificity because not an AI classifier	Not reported	Research interface/prototype; supporting materials and datasets available through project repository
Ha et al. (2019)	ResNet-50	Private MRI (216 pts)	Subtype prediction	AUC 0.87 (LumA)	Activation maps	No public code

Study-specific methodological limitations for these entries are summarized in Supplementary Table S2, with provenance-tagged evidence-map fields provided in Supplementary Data Sheet S1.

Shen et al. (2019) developed an end-to-end all-convolutional network for breast cancer detection on screening mammography. On an independent DDSM test set, four-model averaging achieved a per-image AUC of 0.91, sensitivity of 86.1%, and specificity of 80.1%. After transfer and fine-tuning on INbreast, the corresponding AUC was 0.98, with sensitivity of 86.7% and specificity of 96.1%. The authors released implementation resources, supporting reproducibility, although the evidence remains retrospective and benchmark-based rather than prospective clinical validation.

Yala et al. (2021) introduced Mirai, a multi-view deep-learning model designed to predict breast cancer risk at multiple future time points from screening mammograms. The model was developed using Massachusetts General Hospital data and evaluated on

held-out cohorts from Massachusetts General Hospital, Karolinska in Sweden, and Chang Gung Memorial Hospital in Taiwan, obtaining C-indices of 0.76, 0.81, and 0.79, respectively. Mirai outperformed established clinical risk models where direct comparison was available and was released as an open-source implementation under an MIT license. However, the study was retrospective, dedicated lesion-level explanations were not a central component, and local calibration and prospective workflow evaluation remain necessary before clinical use.

Ribli et al. (2018) applied Faster R-CNN to the combined DDSM and INbreast datasets for lesion detection and malignancy classification, using a VGG-16 backbone pretrained on ImageNet, achieving an AUC of 0.95 and improving FROC performance by 10% over the previous state of the art; bounding box predictions served as implicit spatial explanations and the complete codebase with pretrained weights was released on GitHub, enabling one of the earliest reproducible deep learning benchmarks in breast imaging and facilitating dozens of subsequent comparative studies; no dedicated web application was developed, but the GitHub implementation has been adapted for clinical evaluation in several research settings.

Lu et al. (2025) proposed a multi-task learning network with object contextual attention (MTL-OCA) for automatic joint segmentation and classification of breast ultrasound images. The model used a Res-UNet backbone with an object contextual attention module to refine lesion segmentation masks and a classification branch to predict normal, benign, and malignant categories. The method was evaluated using 5-fold cross-validation on the OASBUD dataset and further tested on the UDIAT dataset. On OASBUD, MTL-OCA achieved a Dice score of 83.75%, IoU of 72.03%, and classification accuracy of 91.67%. On UDIAT, it achieved a Dice score of 84.85%, IoU of 73.68%, and classification accuracy of 94.79%. For UDIAT classification, the model reported benign-class precision, sensitivity, specificity, and F1-score of 0.958, 0.964, 0.913, and 0.961, respectively, and malignant-class precision, sensitivity, specificity, and F1-score of 0.931, 0.943, 0.962, and 0.937. Grad-CAM heatmaps were used to support interpretability, but no standalone clinical web application was reported.

Hu et al. (2020) developed a deep transfer-learning computer-aided diagnosis framework for multiparametric breast MRI using DCE-MRI and T2-weighted MRI sequences from 927 lesions in 616 women. A pretrained CNN was used for feature extraction, and support vector machine classifiers were trained, to distinguish benign from malignant lesions using single-sequence and multiparametric fusion strategies. The study reported AUC values of 0.85 for DCE-MRI alone, 0.78 for T2-weighted MRI alone, 0.85 for image fusion, 0.87 for feature fusion, and 0.86 for classifier fusion, showing that feature-level integration of multiparametric MRI information improved diagnostic performance compared with single-sequence analysis. No Grad-CAM++, CBAM module, public clinical web application, or prospective deployment was reported.

Leong et al. (2022) investigated microcalcification discrimination in mammography using adaptive transfer learning with deep convolutional neural networks. The study applied image filtering to region-of-interest mammogram images containing microcalcifications to reduce artifacts and noise before model training. Several CNN architectures were compared, including ResNet50, ResNet34, VGG16, and AlexNet, with hyperparameter tuning performed for model optimization. The best-performing model was the fine-tuned ResNet50, which achieved an accuracy of 97.58%, followed by ResNet34 with 97.35%, VGG16 with 96.97%, and AlexNet with 83.06%. Confusion matrices were used for model comparison, but no explicit clinical web application, public deployment interface, or XAI method such as Grad-CAM was reported. This study is pertinent to early breast cancer detection because microcalcifications are small, spatially dispersed abnormalities that are often difficult to identify in screening mammography.

Lotter et al. (2021) demonstrated an annotation-efficient ResNet-based detection approach for both mammography and digital breast tomosynthesis, showing that competitive detection performance (AUC 0.94 for mammography, AUC 0.93 for DBT) could be achieved with substantially fewer per-lesion annotations by leveraging case-level supervision; Grad-CAM was reported for detection localization; the approach was described in detail in Nature Medicine with supplementary code, enabling reproduction; clinical applicability was discussed in the context of scaling AI development to centers without large annotation budgets.

Calisto et al. (2020) developed BreastScreening, a multimodal interface integrating mammography, ultrasound, and MRI for diagnostic workflow support. Evaluated with 31 clinicians and 566 images, it focused on usability, visualization, workload, reading time, and BI-RADS-related decision support rather than autonomous classification. Therefore, AUC, sensitivity, specificity, Dice, and F1-score were not applicable.

Ha et al. (2019) fine-tuned ResNet-50 on a private MRI dataset of 216 patients to predict molecular subtype, reporting AUC 0.87 for luminal A vs. other subtypes. Activation maps were visualized, but formal XAI, public data release, and code sharing were not reported. Volumetric breast imaging may benefit from 3D CNNs, but public DBT benchmarks, external validation, and standardized volumetric explainability remain limited.

3.6 YOLO-based real-time detection

The You Only Look Once (YOLO; Redmon et al., 2016) family of single-stage object-detection architectures is attractive for breast imaging because it combines lesion localization and classification in a fast inference pipeline. However, the evidence should be presented conservatively: verified breast-imaging applications are strongest for mammographic mass detection/classification and more recent ultrasound segmentation, whereas claims about every successive YOLO version should be avoided unless the specific breast-imaging article is independently verified, as detailed in Table 6.

Al-Masni et al. (2018) proposed one of the earliest YOLO-based CAD systems for simultaneous breast mass detection and benign/malignant classification in digital mammograms. Using 600 original DDSM mammograms and 2,400 augmented mammograms, the system achieved 99.7% mass-detection accuracy and 97.0% benign/malignant classification accuracy under five-fold cross-validation. This study demonstrated the feasibility of a

TABLE 6 Verified YOLO-family studies relevant to breast cancer imaging.

Study [Ref]	YOLO approach	Modality/ dataset	Task	Reported performance	Clinical caution
Al-Masni et al. (2018)	YOLO-based CAD	Digital mammography/DDSM; 600 original and 2,400 augmented mammograms	Mass detection + benign/malignant classification	Detection accuracy 99.7%; classification accuracy 97.0%	Retrospective, augmented single-dataset CAD study; the 99.7% accuracy reflects internal DDSM benchmark performance with high overfitting risk and does not equate to clinical sensitivity/specificity; no external or prospective screening validation
Aly et al. (2021)	YOLOv1–YOLOv3 comparison; YOLOv3 with anchor optimization	Full-field digital mammography/ INbreast	Mass detection + classification	Mass detection 89.4%; AP benign 94.2%; AP malignant 84.6%; ResNet/InceptionV3 classification accuracy 91.0%/95.5%	Benchmark study; external clinical validation not reported
Karaca Aydemir et al. (2025)	YOLOv5-CAD with transfer learning	CBIS-DDSM, VinDr-Mammo, INbreast, MIAS, private cohort	Mass detection + classification	CBIS-DDSM→INbreast: mAP 0.843, precision 0.855, recall 0.774, 14.8 ms; VinDr-Mammo→INbreast: mAP 0.840, precision 0.829, recall 0.787, 16.54 ms; external MIAS/private mAP 0.756/0.816	Useful transfer-learning evidence; still retrospective
Lan et al. (2024)	Improved YOLOv8-GHOST-P2	MIAS and DDSM	Breast mass lesion-area detection	F1 scores 98.38%, 98.80%, 98.57% across improved variants; parameter reductions 42.9%, 46.64%, 2.8%	Efficient benchmark model; needs independent clinical validation
Mostafa et al. (2025)	Adapted/optimized YOLOv8	Breast ultrasound/BUSI	Tumor segmentation	Adapted YOLOv8 improved from F1 93.44% to 95.29% and Dice 82.10% to 84.40% with separate benign/malignant training	Research prototype; not equivalent to deployed diagnostic software

Very high accuracies reported on a single benchmark or on heavily augmented data [for example, 99.7% detection accuracy on augmented DDSM (Al-Masni et al., 2018)] are internal, dataset-specific results that carry a substantial risk of overfitting and dataset-specific optimism. They should be interpreted as benchmark performance only, not as clinical sensitivity or specificity in prospective screening workflows

unified YOLO-based detection-classification pipeline, although it remained a retrospective benchmark study rather than prospective screening evidence.

Aly et al. (2021) extended YOLO-based CAD to full-field digital mammograms and compared YOLOv1, YOLOv2, and YOLOv3 configurations on INbreast. Their best YOLOv3-based setting detected 89.4% of masses and achieved average precision values of 94.2% for benign masses and 84.6% for malignant masses. When YOLO's classification network was replaced with ResNet and InceptionV3 feature extractors, the overall classification accuracy reached 91.0 and 95.5%, respectively. This work supported the usefulness of anchor optimization and training-set augmentation for mammographic mass detection, but external clinical validation was not reported.

Karaca Aydemir et al. (2025) proposed a YOLOv5-CAD framework using transfer learning across CBIS-DDSM, VinDr-Mammo, INbreast, MIAS, and a private cohort. In five-fold cross-validation, transfer learning from CBIS-DDSM to INbreast achieved mAP 0.843, precision 0.855, recall 0.774, and 14.8 ms inference time. Transfer learning from VinDr-Mammo to INbreast achieved mAP 0.840, precision 0.829, recall 0.787, and 16.54 ms inference time. When tested on external datasets, the best model achieved mAP 0.756 on MIAS and mAP 0.816 on the private cohort, showing promising but still retrospective generalization.

Lan et al. (2024) proposed an improved YOLOv8-GHOST-P2 model for breast mass lesion-area detection on MIAS and DDSM. Compared with standard YOLOv8, the improved models achieved F1 scores of 98.38, 98.80, and 98.57%, while reducing model

parameters by 42.9, 46.64, and 2.8%, respectively, depending on the model variant. These results suggest computational advantages for efficient mammographic lesion detection, but the evidence remains dataset-specific.

Mostafa et al. (2025) evaluated an adapted YOLOv8 model for breast tumor segmentation in ultrasound images using the BUSI dataset (Al-Dhabyani et al., 2020). The study compared combined-class training with separate benign/malignant training and reported that the adapted YOLOv8 improved from F1-score 93.44%−95.29% and from Dice coefficient 82.10%−84.40% when trained separately by lesion class. This work illustrates the extension of YOLO-family models from mammographic detection to ultrasound segmentation, although it remains a research prototype rather than deployed clinical software.

Accordingly, this review treats YOLO-family methods as fast detection and localization frameworks with promising retrospective evidence, but not as uniformly validated clinical systems. Reported performance should be interpreted in relation to dataset size, augmentation strategy, lesion type, annotation quality, imaging modality, and external validation status. In particular, the very high classification and detection accuracies reported on single-benchmark or heavily augmented datasets [for example, the 99.7% mass-detection accuracy on augmented DDSM images (Al-Masni et al., 2018)] should be read as internal, dataset-specific benchmark values that carry a substantial risk of overfitting and optimistic bias. Such figures do not represent clinical sensitivity or specificity and cannot be assumed to transfer to prospective, multi-vendor screening populations without external validation.

3.7 Vision transformers and hybrid architectures

The Vision Transformer (ViT), introduced by Dosovitskiy et al. (2020), reformulated image analysis by dividing images into fixed-size patches and processing them as token sequences through self-attention. This design is well suited to breast imaging, since long-range spatial relationships, bilateral asymmetry, multifocal disease, and volumetric context can be difficult to capture using only local convolutional filters. However, ViT itself should be treated as an architecture-level reference rather than a breast cancer-specific diagnostic model, because its original validation was performed on general computer-vision benchmarks rather than breast imaging datasets, as summarized in Table 7.

Liu et al. (2021) introduced the Swin Transformer, which improved the scalability of transformer-based image analysis through hierarchical feature maps and shifted-window self-attention. This architecture is particularly relevant to high-resolution medical imaging, as it reduces the computational cost of global self-attention while preserving long-range contextual modeling. In this review, the Swin Transformer is therefore cited as a foundational architecture motivating later breast-imaging adaptations, rather than as direct evidence of clinical breast cancer diagnostic performance, as detailed in Table 7.

Chen et al. (2021, 2024) developed TransUNet, a hybrid CNN-Transformer segmentation architecture that integrates convolutional feature extraction with transformer-based global context modeling. The original TransUNet work introduced transformer encoders for medical image segmentation, while the later Medical Image Analysis paper expanded the framework by rethinking both encoder and decoder design through self-attention and cross-attention. These studies are relevant to breast imaging because they provide a validated design template for lesion segmentation. However, their reported performance should not be interpreted as breast-specific, unless the model is evaluated directly on breast datasets.

Hatamizadeh et al. (2022) proposed Swin UNETR, a 3D transformer-based segmentation framework using a hierarchical Swin Transformer encoder connected to a convolutional decoder through skip connections. The model was developed for brain tumor MRI segmentation and ranked among top-performing methods in BraTS 2021 validation experiments. It is useful as a volumetric medical segmentation reference for 3D imaging tasks such as breast MRI or digital breast tomosynthesis, but it should not be described as a breast-specific model unless externally validated on breast imaging data.

Zaidkilani et al. (2025) proposed CoAtUNet, a symmetric encoder-decoder architecture for breast ultrasound tumor segmentation. The model combines convolutional layers for local lesion-boundary extraction with transformer-style attention for long-range contextual modeling. It was evaluated on the UDIAT and BUSI public breast ultrasound datasets and achieved Dice coefficients of 79.12 and 76.21%, respectively. The study is directly relevant to breast imaging because it applies a hybrid CNN-Transformer design to breast ultrasound segmentation, and the implementation is reported as publicly available.

Nantogmah et al. (2025) proposed a cross-attention multimodal fusion framework for breast cancer diagnosis by integrating mammography features with clinical variables. The study compared feature concatenation, co-attention, and cross-attention strategies and reported that the cross-attention model achieved AUC-ROC 0.98, accuracy 0.96, F1-score 0.94, precision 0.92, and recall 0.95 using public datasets including TCGA and CBIS-DDSM. Because this work is currently an arXiv preprint, it should be cited cautiously as emerging evidence rather than as peer-reviewed clinical validation.

Overall, transformer and hybrid CNN-Transformer models are best presented as promising architectures for global context modeling, volumetric segmentation, and multimodal fusion, rather than as uniformly superior replacements for CNNs. Their clinical value depends on breast-specific validation, external testing, calibration, interpretability, and workflow integration.

3.8 Segmentation architectures

Precise lesion segmentation is a prerequisite for quantitative radiomic feature extraction, surgical planning, treatment response monitoring, and radiotherapy target delineation. Segmentation in breast imaging presents challenges including indistinct lesion margins, heterogeneous background parenchyma, and the need for both semantic and instance-level delineation across volumetric modalities. The architectural evolution from standard U-Net through attention mechanisms, transformer integration, and

TABLE 7 Vision transformer and hybrid CNN-transformer architectures for breast cancer imaging.

Study [Ref]	Architecture	Dataset/ modality	Task	Performance/ main finding	XAI/ Interpretability	Open source/ web/clinical
Chen et al. (2021, 2024)	TransUNet/CNN-Transformer U-Net	Medical segmentation benchmarks	Encoder-decoder segmentation	Validated medical segmentation framework; do not report as breast-specific unless tested on breast data	Attention maps/cross-attention	Open-source implementations reported
Hatamizadeh et al. (2022)	Swin UNETR	Brain tumor MRI/BraTS 2021	3D semantic segmentation	Top-performing BraTS 2021 validation method; not breast-specific	Volumetric attention modeling	MONAI-related implementation; useful 3D template
Zaidkilani et al. (2025)	CoAtUNet	UDIAT + BUSI breast ultrasound	Breast tumor segmentation	Dice 79.12% on UDIAT; Dice 76.21% on BUSI	Hybrid CNN-attention feature modeling	Public implementation reported; no clinical web app
Nantogmah et al. (2025)	Cross-attention multimodal fusion	Mammography + clinical data; TCGA and CBIS-DDSM	Breast cancer diagnosis/ classification	AUC-ROC 0.98; accuracy 0.96; F1 0.94; precision 0.92; recall 0.95	Cross-attention + explainability	arXiv preprint; not clinical deployment

Study-specific methodological limitations for these entries are summarized in Supplementary Table S2, with provenance-tagged evidence-map fields provided in Supplementary Data Sheet S1. A clear distinction should be drawn between general-purpose benchmark architectures and models evaluated directly on breast imaging. ViT [Dosovitskiy et al. (2020)], the Swin Transformer [Liu et al. (2021)], TransUNet [Chen et al. (2021, 2024)], and Swin UNETR [Hatamizadeh et al. (2022)] are foundational architectures whose original validation used general computer-vision or non-breast medical benchmarks (for example, ImageNet, multi-organ CT, ACDC cardiac MRI, and BraTS brain MRI), and they are cited here as architecture-level references only. By contrast, CoAtUNet [Zaidkilani et al. (2025)] was developed and evaluated on breast ultrasound (UDIAT and BUSI), and the cross-attention fusion model [Nantogmah et al. (2025)] was evaluated on mammography with clinical data; these two entries constitute the breast-specific evidence in this table, with the latter reported only as an arXiv preprint

foundation model-based promptable segmentation reflects a consistent progression toward globally informed, annotation-efficient segmentation systems, as detailed in Table 8.

Ronneberger et al. (2015) introduced U-Net with its symmetric encoder-decoder structure and skip connections enabling precise boundary preservation through direct feature map concatenation, originally applied to histological cell segmentation achieving Dice of 0.92; the original codebase released on GitHub has been forked over 10,000 times and serves as the foundational template for virtually all subsequent medical segmentation architectures; its direct clinical relevance to breast imaging has been demonstrated across hundreds of subsequent studies applying U-Net to mammography, ultrasound, MRI, and histopathology tasks.

Vakanski et al. (2020) augmented U-Net with spatial and channel attention gates applied to the BUSI ultrasound dataset, achieving Dice of 0.87 and Hausdorff distance of 5.8 mm for breast lesion boundary delineation; the attention mechanism selectively amplified lesion features while suppressing the speckle noise artifacts characteristic of ultrasound imaging; Grad-CAM-adapted attention visualizations were provided and the code was released on GitHub with detailed training instructions; the authors provided a systematic analysis of failure cases involving posterior acoustic shadowing and posterior acoustic enhancement, offering clinically relevant guidance for AI-assisted ultrasound interpretation.

Douglas et al. (2023) evaluated 2D and 3D U-Net architectures for DCE-MRI breast lesion segmentation using 994 lesions from 689 patients. Five-fold cross-validation compared U-Net segmentations with fuzzy c-means and radiologist segmentations. The authors found that 2D U-Net outperformed 3D U-Net for center-slice and volumetric segmentation and outperformed fuzzy c-means for center-slice segmentation. The study did not provide a clinical web application or formal XAI, but the masks support downstream radiomics and CAD.

Sani et al. (2024) proposed a grouped Mask Region Convolutional Neural Network, combining Mask R-CNN with group convolution to improve breast cancer segmentation in mammography images. The model was evaluated on the INbreast and MIAS mammography datasets, focusing on lesion segmentation rather than whole-image classification. By incorporating group convolution, the architecture aimed to improve feature extraction efficiency and segmentation quality while preserving the instance-segmentation advantages of Mask R-CNN. The study reported a Dice coefficient of 86.63% and a Jaccard index of 87.76%, indicating strong overlap between predicted lesion masks and reference annotations. No Detectron2 implementation, public GitHub repository, clinical web application, or formal XAI module was reported. This paper is a safer citation for Mask R-CNN-like mammographic segmentation than unsupported claims involving CBIS-DDSM or Detectron2 deployment.

Chen et al. (2021, 2024) developed TransUNet, which integrates transformer-encoded global context with the U-Net encoder-decoder framework using a hybrid CNN-Transformer architecture. The original TransUNet study introduced transformer encoders for medical image segmentation, while the later Medical Image Analysis paper further refined U-Net design through transformer-based encoder and decoder components (Ronneberger et al., 2015). These studies provide an important architectural basis for transformer-enhanced segmentation, but their reported performance should not be presented as

TABLE 8 Segmentation architectures for breast lesion delineation across imaging modalities.

Study [Ref]	Architecture	Dataset	Task	Dice/IoU	XAI	Open source/ web/clinical
Ronneberger et al. (2015)	U-Net	ISBI 2012 (histology)	Cell segmentation	Dice 0.92	None	GitHub (10K+ forks); foundational
Vakanski et al. (2020)	Attention U-Net	BUSI + private	US lesion segmentation	Dice 0.87; HD 5.8mm	Attn. visualization	GitHub; failure case analysis
Douglas et al. (2023)	2D U-Net and 3D U-Net	994 DCE-MRI breast lesions from 689 patients	Breast lesion segmentation on DCE-MRI	2D U-Net significantly outperformed 3D U-Net for center-slice and volumetric segmentation; Dice and Hausdorff distance reported; exact values depend on comparison setting	No formal XAI; segmentation masks provide spatial interpretability	No clinical web app reported; relevant for radiomics/CAD pipelines
Sani et al. (2024)	Grouped Mask R-CNN/Mask R-CNN + group convolution	INbreast + MIAS mammography	Breast lesion segmentation	Dice 86.63%; Jaccard 87.76%	Not reported; segmentation masks provide spatial interpretability	No Detectron2/ GitHub/web app reported; research segmentation model
Chen et al. (2021, 2024)	TransUNet/CNN-Transformer U-Net	Medical segmentation benchmarks	Encoder-decoder medical image segmentation	Validated medical segmentation framework; do not report as breast-specific unless tested on breast datasets	Attention/cross-attention visualization	Open-source implementations reported; architectural reference for breast-imaging adaptations
Xiong et al. (2023)	Mammo-SAM; SAM adapted for mammography	Whole mammograms; CBIS-DDSM and INbreast	Automatic breast mass segmentation	Improved over zero-shot SAM and conventional baselines; report exact Dice/IoU from original paper if available	Prompt-based segmentation visualization	No clinical web app or prospective validation reported
Ma et al. (2024)	MedSAM	1.5M medical images	Universal medical	Dice 0.88 (breast)	Prompt viz.	GitHub + interactive web demo
Huo et al. (2021)	nnU-Net	100 3D fat-suppressed breast DCE-MRI scans	Whole-breast and fibroglandular tissue segmentation	Dice: 0.968 ± 0.017 whole breast; 0.877 ± 0.081 FGT. Avg surface distance: 0.201 ± 0.080 mm whole breast; 0.310 ± 0.043 mm FGT. IoU not reported as primary metric	Not reported	No clinical web app reported; supports automated breast-density and FGT quantification
Byra et al. (2020)	Selective kernel U-Net (SK-U-Net) with conventional + dilated convolutions	882 breast mass US images for development; 893 external US images from three datasets	Breast mass segmentation in ultrasound	Internal test Dice 0.826 vs. U-Net 0.778; benign Dice 0.820; malignant Dice 0.842; external Dice 0.646–0.780; fine-tuning improved Dice by ~6%; IoU not reported as primary metric	Selective-kernel attention/receptive-field analysis; Spearman correlation 0.7 with mass size	GitHub implementation + pretrained weights; no clinical web app reported

Study-specific methodological limitations for these entries are summarized in Supplementary Table S2, with provenance-tagged evidence-map fields provided in Supplementary Data Sheet S1.

breast-specific unless the model is evaluated directly on breast imaging datasets.

Xiong et al. (2023) proposed Mammo-SAM, a breast-specific adaptation of the foundation Segment Anything Model (SAM) for automatic breast mass segmentation in whole mammograms. The authors first observed that zero-shot SAM was insufficient for reliable mammographic mass segmentation because breast masses often have low contrast, blurred boundaries, and heterogeneous appearances. To address this, they adapted SAM to the mammography domain using task-specific optimization for whole-mammogram mass segmentation. The model was evaluated on public mammography benchmarks, including CBIS-DDSM and INbreast, and was compared with both zero-shot SAM and conventional segmentation baselines. Mammo-SAM improved breast mass segmentation performance over unadapted SAM, showing that foundation segmentation models require domain adaptation before clinical use in mammography. The study is important because it demonstrates the feasibility of adapting promptable foundation models for breast lesion segmentation; however, it remains a methodological MLMI conference study rather than a clinically deployed system, and no standalone web application or prospective clinical validation was reported.

Ma et al. (2024) released MedSAM, fine-tuned on 1.5 million medical image-mask pairs spanning 10 imaging modalities including mammography and breast MRI, achieving Dice of 0.88 on the breast imaging subset; the complete code and pre-trained weights are freely available on GitHub and an interactive web demonstration enables clinical evaluation without any installation; the foundation model approach dramatically reduces the annotation burden compared to task-specific training, and the promptable segmentation interface is designed to integrate naturally into clinical reading workstation workflows.

Huo et al. (2021) applied nnU-Net to segment whole breast and fibroglandular tissue in 100 three-dimensional fat-suppressed breast DCE-MRI scans. Five-fold cross-validation achieved Dice 0.968 for whole breast and 0.877 for FGT, with strong correlations between automatic and manual density measurements. No explicit XAI or clinical web application was reported.

Byra et al. (2020) proposed SK-U-Net for ultrasound breast mass segmentation using selective kernel blocks to adapt receptive field size. Developed on 882 images and tested on 893 external images, it achieved internal Dice 0.826 vs. 0.778 for U-Net and external Dice 0.646–0.780. Fine-tuning improved external performance, and the implementation was released on GitHub.

3.9 Explainable AI

Explainable AI has become essential for breast imaging applications because diagnostic models must not only provide accurate predictions but also support radiologist trust, error analysis, and regulatory transparency. The U.S. FDA's AI-enabled medical device guidance and the EU AI Act emphasize transparency, risk management, and human oversight for high-risk medical AI systems (Health C. for D and R., 2026; Regulation (EU), 2024). In breast cancer imaging, explainability is particularly important because clinically meaningful decisions depend on whether the model attends to lesions, calcifications, parenchymal texture, enhancement patterns, or irrelevant confounders such as scanner marks, image borders, or acquisition artifacts. The explainability approaches most relevant to the reviewed breast-imaging studies include saliency-map methods such as Grad-CAM and Grad-CAM++, lesion-localization outputs from detection models, segmentation masks, attention or cross-attention visualization in transformer-based and multimodal architectures, and case-based interpretability. Broader XAI methods such as SHAP, LIME, Integrated Gradients, Score-CAM, TCAV, and LRP remain important in medical AI, but they are not emphasized here unless explicitly applied in the reviewed breast cancer imaging studies. The explainability approaches reported across the reviewed studies are summarized in Table 9.

In this review, interpretability refers to inherently understandable model structure or outputs, whereas explainability refers mainly to *post-hoc* methods. Saliency maps usually reflect image-gradient attribution, Grad-CAM weights convolutional feature maps using class gradients, and Grad-CAM++ modifies this weighting to improve localization when multiple discriminative regions are present.

Selvaraju et al. (2017) introduced Grad-CAM, which generates class-discriminative heatmaps by weighting convolutional feature maps using gradient information from the target class. In breast imaging studies, Grad-CAM or saliency-style visualization is commonly used to verify whether CNN predictions are spatially aligned with suspicious masses, calcifications, enhancement regions, or ultrasound lesions. McKinney et al. (2020) used saliency-style visualizations to support interpretation of mammography screening predictions, while Shen et al. (2019) and Lotter et al. (2021) used localization-oriented outputs to connect CNN decisions with mammographic lesion regions. Lu et al. (2025) used Grad-CAM heatmaps together with segmentation masks in a multi-task ultrasound framework. These applications show that Grad-CAM-type methods are useful for qualitative clinical review, although they should not be treated as proof of causal reasoning.

Chattopadhyay et al. (2017) introduced Grad-CAM++, an extension of Grad-CAM designed to improve localization of discriminative image regions. In this review, Grad-CAM++ is treated as a method-level reference rather than direct breast-imaging evidence unless a reviewed breast-imaging study explicitly applies it. Grad-CAM++-style visualization may be relevant to breast imaging because multifocal lesions, bilateral findings, and spatially distributed microcalcifications can make standard heatmaps difficult to interpret; however, it should not be presented as independent clinical evidence unless directly evaluated in a breast-imaging model.

Object-detection and segmentation models provide additional forms of spatial interpretability. YOLO and Faster R-CNN-based systems produce bounding boxes or lesion-localization outputs that show where a suspicious finding was detected, as reported in mammographic lesion-detection studies such as Ribli et al. (2018), Al-Masni et al. (2018), Aly et al. (2021), and Karaca Aydemir et al. (2025). Similarly, segmentation models generate lesion, tissue, or breast-region masks that can support visual verification, radiomic feature extraction, treatment-response assessment, and

TABLE 9 Explainability approaches reported in reviewed breast cancer imaging AI studies.

Explainability approach	Representative reviewed studies	How it was used	Clinical relevance
Grad-CAM/saliency-style heatmaps (Selvaraju et al., 2017)	McKinney et al. (2020); Shen et al. (2019); Lotter et al. (2021); Lu et al. (2025)	Highlighted image regions contributing to CNN predictions in mammography, DBT, or ultrasound models	Helps radiologists assess whether model attention overlaps with suspicious lesions, calcifications, or enhancement regions
Bounding-box or lesion-localization outputs	Ribli et al. (2018); Al-Masni et al. (2018); Aly et al. (2021); Karaca Aydemir et al. (2025)	Object-detection models produced lesion-level bounding boxes or localization outputs	Provides spatial interpretability by showing where the model detected a suspicious lesion
Segmentation masks as spatial explanation	Vakanski et al. (2020); Douglas et al. (2023); Sani et al. (2024); Xiong et al. (2023); Ma et al. (2024); Huo et al. (2021); Byra et al. (2020)	Segmentation models produced lesion, tissue, or breast-region masks	Supports lesion delineation, radiomics, treatment-response assessment, and visual verification
Attention or cross-attention visualization	Zaidkilani et al. (2025); Nantogmah et al. (2025)	Hybrid CNN-Transformer and multimodal models used attention mechanisms to model global context or imaging-clinical feature interactions	Can help interpret long-range image context or multimodal feature interactions
Case-based interpretability	Barnett et al. (2021)	Predictions were supported by comparison with visually similar mammographic mass cases	Clinically intuitive because radiologists often reason by comparison with prior cases

quantitative density analysis (Byra et al., 2020; Douglas et al., 2023; Huo et al., 2021; Ma et al., 2024; Sani et al., 2024; Vakanski et al., 2020; Xiong et al., 2023). These outputs are clinically useful because they show the spatial basis of model predictions, but localization maps and segmentation masks should not be interpreted as full mechanistic explanations.

Attention and cross-attention mechanisms provide another form of interpretability in transformer-based and multimodal models. In breast ultrasound, CoAtUNet used hybrid convolutional and transformer-style attention for tumor segmentation (Zaidkilani et al., 2025), while cross-attention fusion has been proposed to model interactions between mammographic features and clinical variables (Nantogmah et al., 2025). These approaches can help visualize or analyze global image context and multimodal feature interactions, but attention weights should be interpreted cautiously because they do not always provide faithful explanations of model causality.

Case-based interpretability offers a more clinically intuitive alternative to pixel-level heatmaps. Barnett et al. (2021) developed a case-based interpretable model for mammographic mass classification, showing how predictions can be supported by comparison with visually similar training examples. This form of explanation is relevant to radiology because clinicians often reason by comparison with prior cases, but its reliability depends on representative training examples, careful validation, and transparent presentation of similar cases.

Overall, explainability in breast imaging should be interpreted as a support tool for model validation, failure analysis, and radiologist review, not as proof that a model is clinically safe. The most clinically useful explanation strategy is likely multimodal, combining lesion-localization outputs, segmentation masks, CNN heatmaps, attention or cross-attention analysis, case-based examples, and uncertainty estimates for cases requiring mandatory radiologist review.

3.10 Multimodal fusion, radiogenomics, and federated learning

No single imaging modality captures the full biological complexity of breast cancer. Mammography provides population-level screening information, ultrasound contributes lesion morphology and margin assessment, MRI provides vascular and kinetic enhancement patterns, and clinical or genomic variables add biological and prognostic context. For this reason, recent AI research increasingly focuses on multimodal fusion, radiogenomics, and privacy-preserving learning. Fusion strategies are generally organized into three categories: early fusion, where raw or preprocessed inputs are concatenated before model training; intermediate fusion, where modality-specific feature representations are combined inside the network; and late fusion, where predictions from separate modality-specific models are combined using ensemble weighting or meta-classifiers.

Nantogmah et al. (2025) proposed a cross-attention multimodal fusion framework for breast cancer diagnosis by integrating mammography-derived information with clinical variables. The study compared feature concatenation, co-attention, and cross-attention strategies, showing that the cross-attention model achieved AUC-ROC 0.98, accuracy 0.96, F1-score 0.94, precision 0.92, and recall 0.95 using public datasets including TCGA and CBIS-DDSM. The use of cross-attention is relevant because it allows the model to

learn how imaging and clinical variables interact rather than treating each source as an independent predictor. However, because this work is currently an arXiv preprint, it should be presented as emerging evidence rather than fully validated clinical evidence.

A complementary and more comprehensive example of attention-driven multimodal design is OncoVision, a recent end-to-end pipeline that integrates mammographic imaging with structured clinical data for breast cancer diagnosis (Ahmed et al., 2025). Using an attention-based encoder-decoder backbone, OncoVision jointly segments four regions of interest (masses, calcifications, axillary findings, and breast tissue) and predicts ten structured clinical descriptors, including mass morphology, calcification type, ACR breast density, and BI-RADS category, before combining imaging and clinical evidence through two late-fusion strategies. The system is operationalized as a secure web application that generates structured, BI-RADS-oriented reports with dual-confidence scoring and attention-weighted visualizations, positioning it as an illustrative bridge between lesion-level imaging models and actionable, report-level clinical decision support. As with the cross-attention model above, OncoVision is currently an arXiv preprint rather than a peer-reviewed, externally validated study; it is therefore cited here as emerging multimodal evidence, is not included in the 284-study quantitative corpus, and is retained as a representative context source (Supplementary Table S9).

Radiogenomics provides another route for multimodal AI by linking quantitative imaging phenotypes with molecular or genomic tumor characteristics. Saha et al. (2018) developed a machine-learning radiogenomic analysis using 922 breast cancer patients and 529 DCE-MRI features, evaluating associations between MRI-derived imaging features and molecular markers including ER, PR, HER2, Ki-67, and molecular subtype. Grimm et al. (2015) similarly investigated associations between routine breast MRI features and luminal A/luminal B molecular subtypes, showing that computer vision-derived imaging phenotypes can reflect underlying tumor biology. These studies support the potential of breast MRI radiogenomics for non-invasive biological characterization, but they should be interpreted as association-based radiogenomic studies rather than deployed diagnostic systems.

Federated learning addresses a different but equally important translational barrier: the difficulty of training generalizable AI models across multiple institutions without centralizing sensitive patient data. Rieke et al. (2020) described federated learning as a framework in which institutions collaboratively train models by sharing model updates rather than raw imaging data. Kaissis et al. (2020) further emphasized secure and privacy-preserving machine learning strategies for medical imaging, including federated learning and privacy-protection mechanisms. In breast imaging, these approaches are especially relevant because robust AI systems require multi-institutional, multi-vendor, and demographically diverse data, while privacy regulations and data-governance constraints often limit direct data sharing. Federated learning therefore represents a promising pathway for improving generalizability, fairness, and external validation without requiring centralized aggregation of patient images.

3.11 Open-access web applications and clinical deployment

A major translational challenge in breast imaging AI is moving from retrospective model development to tools that can be accessed, tested, and integrated into clinical or research workflows. Open-source code, browser-based interfaces, and web-deployable inference frameworks can reduce technical barriers for researchers and clinicians, especially in settings without dedicated AI workstations or local machine-learning expertise. However, deployment claims should be interpreted cautiously: a public repository or browser demo is not equivalent to regulatory approval, prospective validation, or routine clinical implementation. Table 10 summarizes representative open-access and web-oriented systems relevant to breast imaging AI.

Yala et al. (2021) released the Mirai mammography-based breast cancer risk model as an open-source system with code and trained model resources. Mirai predicts future breast cancer risk from mammograms and has been externally evaluated across different populations, including cohorts from the United States, Sweden, and Taiwan. Its open-source availability makes it one of the strongest examples of reproducible breast imaging AI for risk prediction. Nevertheless, Mirai should be described as a risk-stratification model rather than a general diagnostic detector, and its clinical use still requires local validation, calibration, and workflow integration.

Ma et al. (2024) released MedSAM, a medical adaptation of the Segment Anything Model trained on large-scale medical image-mask pairs. MedSAM provides public code, model weights, and an interactive demonstration for prompt-guided segmentation. Although it is not breast-specific, it is relevant to breast imaging because promptable segmentation can support lesion, tissue, or organ delineation workflows in mammography, MRI, and ultrasound. Its main clinical value is as an accessible segmentation framework, but it should not be presented as a validated autonomous breast cancer diagnostic system.

Cohen et al. (2019) implemented Chester, a browser-based chest X-ray disease prediction system using TensorFlow.js. Chester is not a breast imaging tool, but it is relevant as a deployment template because it demonstrates local browser-based inference in which images can remain on the user's device. This client-side design is important for privacy-sensitive medical imaging applications. TensorFlow.js itself provides the technical framework for running trained machine-learning models directly inside web browsers through JavaScript and WebGL acceleration (Smilkov et al., 2019). These technologies may support future browser-based breast imaging tools, but they should be cited as deployment frameworks rather than breast-specific diagnostic evidence.

Calisto et al. (2020) developed BreastScreening, a multimodal interface for breast imaging diagnosis that integrates mammography, ultrasound, and MRI information into a single visual workflow. The system was evaluated with clinicians and focused on usability, multimodal visualization, diagnostic workflow, and BI-RADS-related decision support rather than autonomous AI classification. It is therefore best presented as a human-computer interaction and clinical interface study, not as an AI classifier with AUC or sensitivity metrics. This work is

TABLE 10 Open-access and web-based AI deployment frameworks and platforms relevant to breast cancer imaging.

Platform/tool [Ref]	Stack	Modality/task	Access	XAI	Clinical applicability/notes
Mirai (Yala et al., 2021)	Deep learning risk model	Mammography/5-year breast cancer risk prediction	Open-source code and model resources	Risk-score output; multi-task prediction	Strong reproducible example for risk stratification; requires local calibration and clinical workflow validation
MedSAM (Ma et al., 2024)	Foundation segmentation model/SAM adaptation	Multi-modal medical imaging; potentially mammography, MRI, ultrasound	Public code, weights, and interactive demo	Prompt-based segmentation visualization	Useful for accessible segmentation research; not an autonomous breast cancer diagnostic system
Chester (Cohen et al., 2019)	TensorFlow.js browser-based inference prototype	Chest X-ray/medical image classification template	Browser-based open prototype	Grad-CAM-style visualization reported	Not breast-specific; useful privacy-preserving client-side deployment template
BreastScreening (Calisto et al., 2020)	Multimodal clinical interface/HCI system	Mammography, ultrasound, MRI diagnostic workflow	Research prototype	Not an XAI method; supports multimodal visual review	Clinician-facing workflow-support interface; not an autonomous AI classifier
Ribli et al. detection model (Ribli et al., 2018)	Faster R-CNN mammography detector	Mammography/lesion detection and classification	Public code reported	Bounding-box localization	Open research model for reproducibility; not a validated clinical web platform

relevant because clinical deployment of breast imaging AI will require usable interfaces that support radiologist interaction with multimodal data.

Ribli et al. (2018) released code for a Faster R-CNN-based mammographic lesion detection model. Although the original work was not primarily presented as a web application, its public implementation and lesion-level bounding box outputs make it relevant to open research deployment and reproducibility. It should be described as an open-source detection model rather than a validated clinical web platform.

Overall, open-access deployment in breast imaging AI should be evaluated using three separate criteria: public availability of code or model weights, usability of the interface or inference pipeline, and level of clinical validation. A GitHub repository, browser demo, or Flask/Streamlit prototype can support reproducibility and education, but it should not be described as clinically deployable unless it has undergone external validation, regulatory review, and real-world workflow testing.

3.12 Recent prospective and implementation evidence (2025–2026)

The 2025–2026 literature strengthens the clinical-translation component of this review. MASAI included more than 105,000 women and showed that AI-supported screening achieved a non-inferior interval cancer rate, higher sensitivity, unchanged specificity, and 12% interval-cancer reduction vs. standard double reading (Gommers et al., 2026). AI-STREAM evaluated 24,543 women and reported 13.8% higher cancer detection without a significant recall-rate increase (Chang et al., 2025). AITIC reported 63.6% workload reduction and 15.2% increased cancer detection, but recall increased and did not meet the prespecified noninferiority criterion (Elías-Cabot et al., 2026). Differences in recall burden probably reflect threshold setting, whether AI was used as triage or decision support, and local workflow design. Supplementary Table S5 compares trial populations, endpoints, recall, workload, and implementation barriers. Other recent work emphasizes lesion-level evaluation, threshold recalibration, drift monitoring, fairness surveillance, and prospective calibration for risk prediction (Kelly et al., 2026; Lowry et al., 2026; Taib et al., 2025).

3.13 Structured evidence-quality and methodological-limitations assessment

A formal study-by-study QUADAS-2 or PROBAST risk-of-bias assessment was not performed across all 284 reports. Instead, the review used a structured, review-specific framework to

contextualize the strength and limitations of the evidence. The predefined domains were population and participant selection, index-model development and thresholding, reference standard or labeling process, validation independence, flow and timing, reporting transparency, calibration and subgroup or fairness reporting, and clinical applicability. These criteria were used to guide interpretation of the narrative synthesis and were not treated as validated risk-of-bias ratings. When an entry was supported only by an abstract, DOI page, or another source record, domains requiring full methodological detail were marked as not assessable rather than assigned a low, moderate, or high-risk judgment. Supplementary Table S1 provides the operational criteria, Supplementary Table S2 summarizes the principal limitations of influential studies for which sufficient information was available, and Supplementary Data Sheet S1 records source provenance, full-text verification status, and source-supported methodological notes. Across the evidence base, the most frequent concerns were retrospective single-institution development, internal-only train-test splits, limited independent external validation, variable reference standards, and incomplete reporting of confidence intervals, calibration, decision-curve analysis, subgroup performance, and failure cases. Consequently, prospective, multicentre, externally validated, and workflow-tested evidence was weighted more strongly than internally validated retrospective benchmarks.

Two considerations explain why a formal, validated risk-of-bias instrument could not be applied uniformly across the corpus. First, per-study QUADAS-2 or PROBAST scoring requires full-report methodological detail for every study, whereas only 38 of the 284 evidence-map entries (13.4%) were verified from full text; the remaining 246 entries were source-limited (48 abstract-based and 198 based on DOI or source records), so assigning definitive domain-level risk judgments to them would have manufactured a precision that the underlying sources do not support. Second, the corpus spans highly heterogeneous modalities, tasks, endpoints, and reporting conventions, for which a single tabular instrument would force non-comparable studies onto a common scale. Where source detail was sufficient, the most frequent methodological concerns were retrospective single-institution development, internal-only train/test splits with possible patient- or lesion-level leakage, limited or absent independent external and multicentre validation, scanner- and vendor-related domain shift, implicit or *post hoc* decision-threshold optimization, variable reference standards, and incomplete reporting of calibration, decision-curve analysis, subgroup and fairness performance, and failure cases. These recurring patterns, together with the source-verification composition of the evidence map, are summarized in Figure 2.

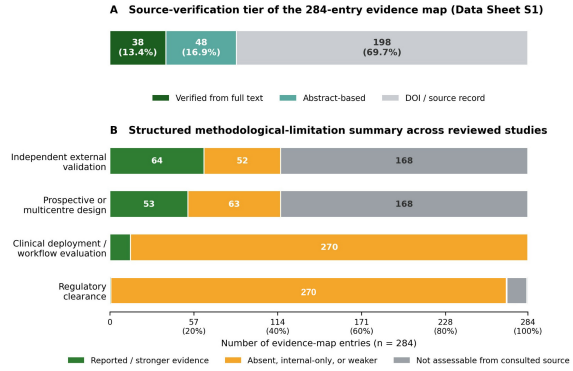

FIGURE 2
Structured methodological-limitation summary of the 284-study evidence base (derived from Supplementary Data Sheet S1). **(A)** Source-verification tier of the evidence map: 38 entries (13.4%) were verified from full text, 48 (16.9%) were abstract-based, and 198 (69.7%) were based on DOI or source records. **(B)** Distribution of key methodological dimensions across all 284 entries, classified by a reproducible rule over the evidence-map fields. Green denotes stronger or adequately reported evidence, amber denotes an absent, internal-only, or weaker condition, and gray denotes dimensions that were not assessable from the consulted source. For the deployment and regulatory-clearance dimensions, the amber category reflects research-stage status (no deployment or clearance reported) rather than methodological bias. Commercial or pre-commercial mentions without an explicitly identified regulatory clearance were classified as not assessable. This is a structured overview and not a formal QUADAS-2 or PROBAST risk-of-bias rating.

3.14 Evidence hierarchy and interpretation of study-level limitations

To improve clinical interpretability, the revised synthesis separates benchmark performance from stronger translational evidence. Studies with prospective design, multicentre recruitment, external or temporal validation, radiologist comparison, calibration assessment, subgroup/fairness reporting, and workflow evaluation were treated as higher-level evidence. Public-dataset benchmarks, augmented datasets, and single-institution internal validations were retained because they are methodologically informative, but their conclusions were weighted more cautiously. The main literature tables now emphasize clinical caution and study-specific methodological limitations, while Supplementary Data Sheet S1 provides a provenance-tagged study-level evidence map and identifies whether each row was verified from a full report.

4 Discussion

4.1 Challenges and limitations

The most pervasive challenge in breast imaging AI research is the limited generalizability of models trained on single-institution or single-vendor datasets. External validation studies and recent implementation-focused evaluations have shown that breast imaging AI performance may decline when models are transferred across institutions, scanner vendors, imaging protocols, and population groups, emphasizing the need for multicenter validation, local threshold calibration, and prospective monitoring before clinical deployment (Kelly et al., 2026; Salim et al., 2020; Taib et al., 2025). The EMBED dataset (Jeong et al., 2023), with 3.4 million demographically diverse mammograms, and VinDr-Mammo (Nguyen et al., 2023), representing a non-Western population, are beginning to address this through more realistic cross-population evaluation platforms. Class imbalance, with cancer-positive cases comprising only a small fraction of screening

examinations, can bias models toward the majority negative class; therefore, future studies should report class distribution, cancer prevalence, sensitivity, specificity, positive predictive value, calibration, and threshold-selection strategy rather than relying only on accuracy. Algorithmic fairness studies have shown that non-representative medical imaging datasets can produce biased classifiers, while breast-screening implementation studies increasingly emphasize subgroup performance assessment across race, age, breast density, scanner vendor, and site before deployment (Jeong et al., 2023; Kelly et al., 2026; Larrazabal et al., 2020). Regulatory and implementation barriers persist despite the growing number of FDA-listed AI-enabled medical devices, including issues related to transparency, post-deployment monitoring, workflow integration, reimbursement, and radiologist trust (Almarie et al., 2025; Goh et al., 2025; Health C. for D and R., 2026; Regulation (EU), 2024).

Calibration and uncertainty quantification remain underreported but are central to safe deployment. A model with high AUC can still be miscalibrated, producing overconfident false negatives or unnecessary recalls. Future studies should report calibration plots, Brier score, expected/observed ratios, and threshold-selection rationale, and should evaluate uncertainty using approaches such as ensembles, Monte Carlo dropout, test-time augmentation variance, or evidential prediction. In screening, uncertain or out-of-distribution cases should trigger mandatory radiologist review rather than autonomous triage.

4.2 Limitations of this review

This review has limitations. First, the protocol was not prospectively registered, although the revised manuscript provides complete search strings, screening counts, eligibility criteria, and a provenance-tagged 284-study evidence map to improve reproducibility. Second, meta-analysis was not performed because modalities, datasets, architectures, validation strategies, and outcome definitions were highly heterogeneous. Third, independent duplicate screening and extraction were not performed for the entire corpus. During revision, inter-reviewer reliability was assessed through *post hoc* independent reassessment of random 10% samples at the title and abstract and full-text screening stages, and a second reviewer checked the extraction of 20.1% of the included studies. These procedures provide quantitative quality-control information but do not replace complete duplicate screening and extraction. Fourth, the expanded evidence map was not based on full-text re-extraction for every included study: 38 entries were verified from full text, 48 were abstract-based, and 198 were based on DOI/source pages or other web-accessible records. Accordingly, the spreadsheet should not be interpreted as a complete full-text extraction dataset, and the absence of a feature in a source-limited row does not establish that it was absent from the full report. Source-limited rows were not used for definitive study-level methodological-quality judgments. Fifth, no formal study-by-study QUADAS-2 or PROBAST assessment was completed across all 284 full reports; the review instead used a structured evidence-quality and methodological-limitations framework, and this distinction may limit comparability with reviews using validated tools. Sixth, English-only inclusion may introduce language bias. Seventh, publication bias likely favors high-performing systems, whereas negative validations, external failures, calibration problems, and subgroup disparities are underreported. Eighth, private datasets, unavailable code, and proprietary commercial algorithms limited reproducibility assessment. Finally, although the May 2026 narrative update captures major recent prospective and implementation studies, the field is evolving rapidly, and the review should be interpreted as a current evidence synthesis rather than a definitive ranking of algorithms.

4.3 Future directions and research priorities

Future breast imaging AI research should move beyond retrospective benchmark performance toward prospective, externally validated, calibrated, and clinically monitored systems. The highest priority is to demonstrate reproducible benefit in real screening and diagnostic workflows rather than only improved AUC, accuracy, Dice, or F1 score on retrospective datasets.

Multimodal and longitudinal learning should integrate mammography, DBT, ultrasound, MRI, contrast-enhanced mammography, clinical variables, pathology, genomics, and prior examinations. Such models may better capture disease biology and temporal change, but must report ablation analyses showing whether each modality adds clinically meaningful value.

Self-supervised, weakly supervised, and foundation-model approaches are promising because they can learn from large volumes of unlabeled breast images and reduce dependence on scarce expert annotations. Breast-specific adaptation of promptable segmentation and vision foundation models should be evaluated with external cohorts, lesion-level localization, calibration, and subgroup fairness analysis.

Synthetic data generation using generative adversarial networks, diffusion models, simulation, and physics-informed augmentation may help address class imbalance and rare lesion presentations. However, synthetic images should be used only with strict train-test separation, realism assessment by experts, and evaluation on real external clinical data to avoid inflated performance.

Large language models and vision-language models may support BI-RADS-consistent report structuring, prior-report summarization, multimodal retrieval, explainable decision support, and patient-facing communication. Their use in breast imaging should be constrained by factual grounding, audit trails, privacy safeguards, hallucination testing, and clear radiologist oversight.

Future studies should routinely report calibration, uncertainty, subgroup fairness, lesion-level localization, external validation, workflow impact, recall burden, reading time, and post-deployment drift monitoring. The most clinically relevant next step is not another architecture-only comparison, but multicentre implementation research that demonstrates safe, equitable, and reproducible benefit across populations and imaging systems (Chang et al., 2025; Elías-Cabot et al., 2026; Gommers et al., 2026; Kelly et al., 2026; Larrazabal et al., 2020; Lowry et al., 2026; Regulation (EU), 2024; Yilmaz et al., 2025).

5 Conclusion

This systematic review synthesized evidence from 284 peer-reviewed studies in the primary PRISMA-guided review, supplemented by a separate targeted 2026 narrative update of recent prospective and implementation studies. The review covered the major methodological landscape of AI-assisted breast cancer imaging, including classical machine learning and radiomics, convolutional neural networks, YOLO-family detection models, vision transformers, segmentation architectures, explainable AI, multimodal fusion, federated learning, and open-access deployment frameworks. Several AI systems have reported diagnostic performance comparable to expert readers in selected retrospective, external-validation, and prospective settings; however, generalizability, calibration, external validation, algorithmic fairness, and workflow integration remain major determinants of clinical utility before broad clinical adoption can be justified. Public availability, open-source code, or browser-based deployment should not be equated with regulatory clearance or clinical readiness. Realizing the full potential of AI in breast cancer imaging will require interdisciplinary collaboration, prospective multicentre validation, transparent reporting, complete full-text study-level extraction in future evidence syntheses, continuous post-deployment monitoring, and careful evaluation of sensitivity, specificity, recall burden, calibration, lesion localization, and equity. The accompanying Supplementary Data Sheet S1 should be interpreted as a provenance-tagged evidence map rather than a complete full-text extraction or formal study-level risk-of-bias appraisal across all 284 studies: 38 entries were verified from full reports, 48 from abstracts, and 198 from DOI/source pages or other web-accessible records.

Future research should prioritize prospective multicenter trials, lesion-level evaluation, subgroup fairness analysis, uncertainty quantification, and calibration across diverse populations and imaging systems. Foundation models, multimodal fusion, and federated learning are promising directions, but they require breast-specific validation, transparent reporting, and workflow-centered implementation before routine clinical adoption. As a minimum reporting standard, future breast imaging AI studies should report external validation, calibration, subgroup performance by breast density and demographic group, lesion-level localization accuracy where applicable, and workflow impact.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

IK: Conceptualization, Data curation, Methodology, Writing – original draft, Writing – review & editing, Validation. SAS: Conceptualization, Data curation, Methodology, Validation, Writing – original draft, Writing – review & editing. AT: Data curation, Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing. MM: Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing. AA: Data curation, Visualization, Writing – original draft, Writing – review & editing. SBS: Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. MH: Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. AB: Project administration, Supervision, Writing – original draft, Writing – review & editing. KP: Data curation, Formal analysis, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. STS: Conceptualization, Data curation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. MD: Funding acquisition, Supervision, Writing – review & editing, Project administration.

Funding

The author(s) declared that financial support was received for this work and/or its publication. The GALATEA project has received funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101183057 (https://galateahe.eu/). Additional support was provided by Progetto PNC 0000001 – D34Health, funded by the Italian National Plan for Complementary Investments to the National Recovery and Resilience Plan, Intervention No. 1, "Research initiatives for innovative technologies and pathways in the health and care sector"; CUP B83C22006120001 (https://d34health.it/).

Acknowledgments

The authors gratefully acknowledge Politecnico di Torino for providing the academic and research environment that supported the development of this work. The authors also acknowledge GPI SpA for its support and interest in digital health, biomedical innovation, and the translation of artificial intelligence tools toward clinically relevant applications.

Conflict of interest

SAS and AB were employed by GPI SpA. MH was employed by 7HC SRL.

The remaining authors declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author MD declared that they were an editorial board member of Frontiers at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. The author(s) used ChatGPT by OpenAI and Claude by Anthropic to support language editing, structural refinement, and formatting. All scientific content, citations, interpretations, and final wording were reviewed, verified, and approved by the author(s).

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimag.2026.1910013/full#supplementary-material

References

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Ahmed, I., Ahmed, G., Sanjid, K. S., Hossain, Md. T., Khan, Md. N., Khan, Md. M., et al. (2025). *OncoVision: Integrating Mammography and Clinical Data through Attention-Driven Multimodal AI for Enhanced Breast Cancer Diagnosis*. doi: 10.48550/arXiv.2511.19667. [Epub ahead of print 2025].

Al-Dhabyani, W., Gomaa, M., Khaled, H., and Fahmy, A. (2020). Dataset of breast ultrasound images. *Data Brief* 28:104863. doi: 10.1016/j.dib.2019.104863

Almarie, B., Gonzalez-Gonzalez, L. F., Dos Santos Barbosa, L. A., Lutz, A., Grosse, U., and Fregni, F. (2025). Machine learning-enabled medical devices authorized by the US Food and Drug Administration in 2024: regulatory characteristics, predicate lineage, and transparency reporting. *Biomedicines* 13:3005. doi: 10.3390/biomedicines13123005

Al-Masni, M. A., Al-Antari, M. A., Park, J.-M., Gi, G., Kim, T.-Y., Rivera, P., et al. (2018). Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Comput. Methods. Programs Biomed*. 157, 85–94. doi: 10.1016/j.cmpb.2018.01.017

Aly, G. H., Marey, M., El-Sayed, S. A., and Tolba, M. F. (2021). YOLO based breast masses detection and classification in full-field digital mammograms. *Comput. Methods Programs Biomed*. 200:105823. doi: 10.1016/j.cmpb.2020.105823

Amgad, M., Elfandy, H., Hussein, H., Atteya, L. A., Elsebaie, M. A. T., Abo Elnasr, L. S., et al. (2019). Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* 35, 3461–3467. doi: 10.1093/bioinformatics/btz083

Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., et al. (2021). A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nat. Mach. Intell.* 3, 1061–1070. doi: 10.1038/s42256-021-00423-x

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., et al. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 74, 229–263. doi: 10.3322/caac.21834

Broeders, M., Moss, S., Nyström, L., Njor, S., Jonsson, H., Paap, E., et al. (2012). The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies. *J. Med. Screen.* 19, 14–25. doi: 10.1258/jms.2012.012078

Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., et al. (2020). Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. *Biomed. Signal Process. Control* 61:102027. doi: 10.1016/j.bspc.2020.102027

Calisto, F. M., Nunes, N., and Nascimento, J. C. (2020). "BreastScreening: on the use of multi-modality in medical imaging diagnosis," In: *Proceedings of the International Conference on Advanced Visual Interfaces* (Salerno Italy: ACM), pp. 1–5. doi: 10.1145/3399715.3399744

Chang, Y. W., Ryu, J. K., An, J. K., Choi, N., Park, Y. M., Ko, K. H., et al. (2025). Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study. *Nat. Commun.* 16:2248. doi: 10.1038/s41467-025-57469-3

Chattopadhyay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2017). Grad-CAM++: improved visual explanations for deep convolutional networks. *arXiv.org*. Epub ahead of print 30 October 2017.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*. Epub ahead of print 2021.

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., et al. (2024). TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* 97:103280. doi: 10.1016/j.media.2024.103280

Chen, S., Kuhn, M., Prettner, K., and Bloom, D. E. (2018). The macroeconomic burden of noncommunicable diseases in the United States: estimates and projections. *PLoS ONE* 13:e0206702. doi: 10.1371/journal.pone.0206702

Cohen, J. P., Bertin, P., and Frappier, V. (2019). *Chester: A Web Delivered Locally Computed Chest X-Ray Disease Prediction System*. doi: 10.48550/arXiv.1901.11210. [Epub ahead of print 2019].

Conant, E. F., Toledano, A. Y., Periaswamy, S., Fotin, S. V., Go, J., Boatsman, J. E., et al. (2019). Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiol. Artif. Intell.* 1:e180096. doi: 10.1148/ryai.2019180096

Cui, C., Li, L., Cai, H., Fan, Z., Zhang, L., Dan, T., et al. (2021). *The Chinese Mammography Database (CMMD): An Online Mammography Database with Biopsy Confirmed Types for Machine Diagnosis of Breast*. doi: 10.7937/TCIA.EQDE-4B16. [Epub ahead of print 2021].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. et al. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. Epub ahead of print 2020.

Douglas, L., Bhattacharjee, R., Fuhrman, J., Drukker, K., Hu, Q., Edwards, A., et al. (2023). U-Net breast lesion segmentations for breast dynamic contrast-enhanced magnetic resonance imaging. *J. Med. Imag.* 10:4502. Epub ahead of print 20 November 2023. doi: 10.1117/1.JMI.10.6.064502

Elías-Cabot, E., Romero-Martín, S., Raya-Povedano, J. L., Rodríguez-Ruiz, A., and Álvarez-Benito, M. (2026). AI-based triage and decision support in mammography and digital tomosynthesis for breast cancer screening: a paired, noninferiority trial. *Nat. Med.* 32, 1296–1305. doi: 10.1038/s41591-026-04277-x

Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., and Feinstein, A. R. (1994). Variability in radiologists' interpretations of mammograms. *N. Engl. J. Med.* 331, 1493–1499. doi: 10.1056/NEJM199412013312206

Ginsburg, O., Yip, C. H., Brooks, A., Cabanes, A., Caleffi, M., Dunstan Yataco, J. A., et al. (2020). Breast cancer early detection: a phased approach to implementation. *Cancer* 126, 2379–2393. doi: 10.1002/cncr.32887

Goh, S. S. N., Ng, Q. X., Chan, F. J. H., Goh, R. S. J., Jagmohan, P., Ali, S. H., et al. (2025). Radiologists' perspectives on AI integration in mammographic breast cancer screening: a mixed methods study. *Cancers.* 17:3491. doi: 10.3390/cancers17213491

Golden, J. A. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 318:2199. doi: 10.1001/jama.2017.14580

Gommers, J., Hernström, V., Josefsson, V., Sartor, H., Schmidt, D., Hjelmgren, A., et al. (2026). Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *Lancet* 407, 505–514. doi: 10.1016/S0140-6736(25)02464-X

Grimm, L. J., Zhang, J., and Mazurowski, M. A. (2015). Computational approach to radiogenomics of breast cancer: luminal A and luminal B molecular subtypes are

associated with imaging features on routine breast MRI extracted using computer vision algorithms. *Magn. Reson. Imaging* 42, 902–907. doi: 10.1002/jmri.24879

Ha, R., Mutasa, S., Karcich, J., Gupta, N., Pascual Van Sant, E., Nemer, J., et al. (2019). Predicting breast cancer molecular subtype with MRI dataset utilizing convolutional neural network algorithm. *J. Digit. Imaging* 32, 276–282. doi: 10.1007/s10278-019-00179-2

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., and Xu, D., (2022). *Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images*. Epub ahead of print 2022. doi: 10.1007/978-3-031-08999-2_22

Health C. for D and R. (2026). *Artificial Intelligence-Enabled Medical Devices*. FDA. Available online at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices [Accessed May 25, 2026].

Heath, M., Bowyer, K., Kopans, D., Philip Kegelmeyer Jr, W., Moore, R., Chang, K., et al. (1998). "Current status of the digital database for screening mammography," in *Digital Mammography*, eds N. Karssemeijer, M. Thijssen, J. Hendriks, et al. (Dordrecht: Springer Netherlands), pp. 457–460. doi: 10.1007/978-94-011-5318-8_75

Hu, Q., Whitney, H. M., and Giger, M. L. (2020). A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI. *Sci. Rep.* 10:10536. doi: 10.1038/s41598-020-67441-4

Huo, L., Hu, X., Xiao, Q., Gu, Y., Chu, X., and Jiang, L. (2021). Segmentation of whole breast and fibroglandular tissue using nnU-Net in dynamic contrast enhanced MR images. *Magn. Reson. Imaging* 82, 31–41. doi: 10.1016/j.mri.2021.06.017

Hylton, N. M., Blume, J. D., Bernreuter, W. K., Pisano, E. D., Rosen, M. A., Morris, E. A., et al. (2012). Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy—results from ACRIN 6657/I-SPY TRIAL. *Radiology* 263, 663–672. doi: 10.1148/radiol.12110748

Jeong, J. J., Vey, B. L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., et al. (2023). The EMory BrEast imaging dataset (EMBED): a racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiol. Artif. Intell.* 5:e220047. doi: 10.1148/ryai.220047

Kaissis, G. A., Makowski, M. R., Rückert, D., and Braren, R. F., (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* 2, 305–311. doi: 10.1038/s42256-020-0186-1

Karaca Aydemir, B. K., Telatar, Z., Güney, S., and Dengiz, B., (2025). Detecting and classifying breast masses via YOLO-based deep learning. *Neural Comput. Applic.* 37, 11555–11582. doi: 10.1007/s00521-025-11153-1

Kelly, C. J., Wilson, M., Warren, L. M., Sidebottom, R., Halling-Brown, M., Yang, L., et al. (2026). Diagnostic accuracy, fairness and clinical implementation of AI for breast cancer screening: results of multicenter retrospective and prospective technical feasibility studies. *Nat. Cancer* 7, 494–506. doi: 10.1038/s43018-026-01127-0

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Available online at: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html [Accessed May 25, 2026].

Krupinski, E. A., Berbaum, K. S., Caldwell, R. T., Schartz, K. M., and Kim, J. (2010). Long radiology workdays reduce detection and accommodation accuracy. *J. Am. Coll. Radiol.* 7, 698–704. doi: 10.1016/j.jacr.2010.03.004

Lan, Y., Lv, Y., Xu, J., Zhang, Y., and Zhang, Y., (2024). Breast mass lesion area detection method based on an improved YOLOv8 model. *era* 32, 5846–5867. doi: 10.3934/era.2024270

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl. Acad. Sci. USA* 117, 12592–12594. doi: 10.1073/pnas.1919012117

Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., and Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* 4:170177. doi: 10.1038/sdata.2017.177

Lehman, C. D., Arao, R. F., Sprague, B. L., Lee, J. M., Buist, D. S., Kerlikowske, K., et al. (2017). National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology* 283, 49–58. doi: 10.1148/radiol.2016161174

Leong, Y. S., Hasikin, K., Lai, K. W., Mohd Zain, N., and Azizan, M. M. (2022). Microcalcification discrimination in mammography using deep convolutional neural network: towards rapid and early breast cancer diagnosis. *Front. Public Health* 10:875305. doi: 10.3389/fpubh.2022.875305

Li, W., Newitt, D. C., Gibbs, J., Wilmes, L. J., Jones,E. F., Arasu, V. A., et al. (2022). *I-SPY 2 Breast Dynamic Contrast Enhanced MRI (I-SPY2 TRIAL)*. doi: 10.7937/TCIA.D8Z0-9T85 [Epub ahead of print 2022].

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., et al. (2018). 1399 HandE-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*. 7:giy065. doi: 10.1093/gigascience/giy065

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. Epub ahead of print 2021. doi: 10.1109/ICCV48922.2021.00986

Lotter, W., Diab, A. R., Haslam, B., Kim, J. G., Grisot, G., Wu, E., et al. (2021). Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat. Med.* 27, 244–249. doi: 10.1038/s41591-020-01174-9

Lowry, K. P., Jeong, H. E., Kim, K. H., Hughes, K. S., Lee, C. I., Yala, A., et al. (2026). Current state of mammography-based artificial intelligence for future breast cancer risk prediction: a systematic review. *JNCI J. Nat. Cancer Instit.* 118, 392–403. doi: 10.1093/jnci/djag002

Lu, Y., Sun, F., Wang, J., and Yu, K. (2025). Automatic joint segmentation and classification of breast ultrasound images via multi-task learning with object contextual attention. *Front. Oncol.* 15:1567577. doi: 10.3389/fonc.2025.1567577

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. (2024). Segment anything in medical images. *Nat. Commun.* 15:654. doi: 10.1038/s41467-024-44824-z

Majid, A. S., de Paredes, E. S., Doherty, R. D., Sharma, N. R., and Salvador, X. (2003). Missed breast carcinoma: pitfalls and pearls. *RadioGraphics* 23, 881–895. doi: 10.1148/rg.234025083

Mann, R. M., Cho, N., and Moy, L. (2019). Breast MRI: state of the art. *Radiology* 292, 520–536. doi: 10.1148/radiol.2019182947

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94. doi: 10.1038/s41586-019-1799-6

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., and Cardoso, J. S. (2012). INbreast: toward a full-field digital mammographic database. *Acad. Radiol.* 19, 236–248. doi: 10.1016/j.acra.2011.09.014

Mostafa, A. M., Alaerjan, A. S., Aldughayfiq, B., Allahem, H., Mahmoud, A. A., Said, W., et al. (2025). Optimized YOLOv8 for enhanced breast tumor segmentation in ultrasound imaging. *Discov. Onc.* 16:1152. doi: 10.1007/s12672-025-02889-2

Nantogmah, M. T., Alhassan, A-B., and Alhassan, S. (2025). *Cross-Attention Multimodal Fusion for Breast Cancer Diagnosis: Integrating Mammography and Clinical Data with Explainability*. Epub ahead of print 2025.

Nguyen, H. T., Nguyen, H. Q., Pham, H. H., Lam, K., Le, L. T., Dao, M., et al. (2023). VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Sci. Data* 10:277. doi: 10.1038/s41597-023-02100-7

Nowakowska, S., Borkowski, K., Ruppert, C. M., Landsmann, A., Marcon, M., Berger, N., et al. (2023). Generalizable attention U-Net for segmentation of fibroglandular tissue and background parenchymal enhancement in breast DCE-MRI. *Insights Imaging* 14:185. doi: 10.1186/s13244-023-01531-5

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021:n71. doi: 10.1136/bmj.n71

Qureshi, S. A., Aziz-ul-Rehman, Hussain, L., Shah, S. T. H., Mir, A. A., et al. (2024). Breast cancer detection using mammography: image processing to deep learning. *IEEE Access*, 13, 60776–60801. doi: 10.1109/ACCESS.2024.3523745

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). *You Only Look Once: Unified, Real-Time Object Detection*. Epub ahead of print 9 May 2016. doi: 10.1109/CVPR.2016.91

Regulation (EU) (2024). 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Available online at: http://data.europa.eu/eli/reg/2024/1689/oj [Accessed May 25, 2026].

Ribli, D., Horváth, A., Unger, Z., Pollner, P., and Csabai, I. (2018). Detecting and classifying lesions in mammograms with Deep Learning. *Sci. Rep.* 8:4165. doi: 10.1038/s41598-018-22437-z

Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., et al. (2020). The future of digital health with federated learning. *NPJ Digit. Med.* 3:119. doi: 10.1038/s41746-020-00323-1

Ronneberger, O., Fischer, P., and Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Epub ahead of print 2015. doi: 10.1007/978-3-319-24574-4_28

RSNA (2026). *RSNA Screening Mammography Breast Cancer Detection*, Available online at: https://kaggle.com/rsna-breast-cancer-detection [Accessed May 25, 2026].

Saha, A., Harowicz, M. R., Grimm, L. J., Kim, C. E., Ghate, S. V., Walsh, R., et al. (2018). A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *Br. J. Cancer* 119, 508–516. doi: 10.1038/s41416-018-0185-8

Saha, A., Harowicz, M. R., Grimm, L. J., Weng, J., Cain, E. H., Kim, C. E., et al. (2022). *Dynamic Contrast-Enhanced Magnetic Resonance Images of Breast Cancer Patients with Tumor Locations*. doi: 10.7937/TCIA.e3sv-re93. [Epub ahead of print 31 August 2022].

Salim, M., Wåhlin, E., Dembrower, K., Azavedo, E., Foukakis, T., Liu, Y., et al. (2020). External evaluation of 3 commercial artificial intelligence algorithms

for independent assessment of screening mammograms. *JAMA Oncol.* 6:1581. doi: 10.1001/jamaoncol.2020.3321

Sani, Z., Prasad, R., and Hashim, E. K. M. (2024). Grouped mask region convolution neural networks for improved breast cancer segmentation in mammography images. *Evol. Syst. (Berl).* 15, 25–40. doi: 10.1007/s12530-023-09527-8

Sawyer-Lee, R., Gimenez, F., Hoogi, A., Miyake, K. K. , Gorovoy, M., and Rubin, D. L. (2016). *Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM)*. doi: 10.7937/K9/TCIA.2016.7O02S9CY. [Epub ahead of print 2016].

SEER Cancer Statistics Review (1975-2018). *SEER*. Available online at: https://seer.cancer.gov/csr/1975_2018/index.html [Accessed May 25, 2026].

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., et al. (2017). "Grad-cam: visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, 618–626. doi: 10.1109/ICCV.2017.74

Shah, S. T. H., Shah, S. A. H., Khan, I. I., Imran, A., Shah, S. B. H., Mehmood, A., et al. (2024). Data-driven classification and explainable-AI in the field of lung imaging. *Front. Big. Data* 7, 1–18. doi: 10.3389/fdata.2024.1393758

Shen, D., Wu, G., and Suk, H-I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., and Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Sci. Rep.* 9:12495. doi: 10.1038/s41598-019-48995-4

Smilkov, D., Thorat, N., Assogba, Y., Yuan, A., Kreeger, N., Yu, P., et al. (2019). *TensorFlow.js: Machine Learning for the Web and Beyond*. doi: 10.48550/arXiv.1901.05350. [Epub ahead of print 2019].

Soso (2026). *xbhlk/STU-Hospital*. Available online at: https://github.com/xbhlk/STU-Hospital [Accessed May 26, 2026].

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (2016). A dataset for breast cancer histopathological image classification. *IEEE Trans. Biomed. Eng.* 63, 1455–1462. doi: 10.1109/TBME.2015.2496264

Sprague, B. L., Gangnon, R. E., Burt, V., Trentham-Dietz, A., Hampton, J. M., Wellman, R. D., et al. (2014). Prevalence of mammographically dense breasts in the United States. *J. Natl. Cancer Inst.* 106:dju255. doi: 10.1093/jnci/dju255

Taib, A. G., Partridge, G. J. W., Yao, L., Darker, I., and Chen, Y. (2025). The evaluation of artificial intelligence in mammography-based breast cancer screening: is breast-level analysis enough? *Eur. Radiol.* 35, 8230–8243. doi: 10.1007/s00330-025-11733-8

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Thomas, C., Byra, M., Marti, R., Yap, M. H., and Zwiggelaar, R. (2023). BUS-Set: a benchmark for quantitative evaluation of breast ultrasound segmentation networks with public datasets. *Med. Phys.* 50, 3223–3243. doi: 10.1002/mp.16287

Trivedi, H. M., Vazirabad, M., Kitamura, F. C., Chen, Y., and Frazer, H. (2026). Open-source dataset for the RSNA screening mammography cancer detection challenge. *Radiol. Artif. Intell.* 8:e250375. doi: 10.1148/ryai.250375

Vakanski, A., Xian, M., and Freer, P. E. (2020). Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. *Ultrasound Med. Biol.* 46, 2819–2833. doi: 10.1016/j.ultrasmedbio.2020.06.015

van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 77, e104–e107. doi: 10.1158/0008-5472.CAN-17-0339

Xiong, X., Wang, C., Li, W., and Li, G., (2023). "Mammo-SAM: adapting foundation segment anything model for automatic breast mass segmentation in whole mammograms," in Machine Learning in Medical Imaging, eds X. Cao, X. Xu and I. Rekik, et al. (Cham: Springer Nature Switzerland), pp. 176–185. doi: 10.1007/978-3-031-45673-2_18

Yala, A., Mikhael, P. G., Strand, F., Lin, G., Smith, K., Wan, Y-L., et al. (2021). Toward robust mammography-based models for breast cancer risk. *Sci. Transl. Med.* 13: eaba4373. doi: 10.1126/scitranslmed.aba4373

Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., et al. (2018). Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE J. Biomed. Health Inform.* 22, 1218–1226. doi: 10.1109/JBHI.2017.2731873

Yilmaz, E., Seker, M. E., Guldogan, N., Turk, E. B., Erdemli, S., Koyluoglu, Y. O., et al. (2025). Clinical application of AI in mammography: insights from a prospective study. *Acad. Radiol.* 32, 5016–5027. doi: 10.1016/j.acra.2025.05.025

Zaidkilani, N., Garcia, M. A., and Puig, D. (2025). CoAtUNet: A symmetric encoder-decoder with hybrid transformers for semantic segmentation of breast ultrasound images. *Neurocomputing* 629:129660. doi: 10.1016/j.neucom.2025.129660