REVIEW **Open access**

Deep Learning for Breast Cancer Detection in Medical Imaging (Mammography, Ultrasound, MRI): A Critical Review

Ahmed El-Kholy[1*], Nour Abdelrahman[1], Karim Hassan[2]

Abstract

Breast cancer remains a leading cause of cancer-related mortality among women worldwide, underscoring the importance of effective screening strategies for early detection and improved survival. Although conventional modalities such as mammography reduce mortality, they are limited by false positives, false negatives, and overdiagnosis, particularly in dense breast tissue and diverse populations. Deep learning, especially convolutional neural networks (CNNs), has shown promise in improving diagnostic accuracy and reducing inter-reader variability; however, its translation into routine clinical practice requires critical evaluation beyond reported performance metrics. This critical review evaluates CNN-based deep learning applications for breast cancer detection across mammography, ultrasound, and MRI, with emphasis on training strategies and barriers to clinical deployment. A targeted literature search identified peer-reviewed studies focusing on CNN architectures, transfer learning, and implementation challenges. Findings indicate that models such as ResNet, DenseNet, and EfficientNet perform well in controlled settings, supported by transfer learning and data augmentation approaches. However, these results often fail to translate into consistent clinical performance, particularly across imaging modalities and real-world workflows. Limitations including demographic bias, insufficient external validation, and weak evidence of outcome or cost-effectiveness highlight a substantial gap between experimental success and clinical readiness. The review concludes that while deep learning in breast imaging is promising, its adoption should remain cautious and evidence-driven until robust clinical benefit is clearly demonstrated.

Keywords Deep learning, Breast cancer detection, Convolutional neural networks, Transfer learning, Clinical deployment, Health disparities

*Correspondence:
Ahmed El-Kholy
ahmed.elkholy@outlook.com
[1] Department of Healthcare Systems Engineering, Faculty of Medicine, Alexandria University, Alexandria, Egypt
[2] Department of Clinical AI Analytics, Faculty of Medicine, Ain Shams University, Cairo, Egypt

Introduction

Breast cancer imposes a substantial global burden, with screening mammography serving as the cornerstone of early detection efforts despite its inherent limitations. High false-positive rates contribute to patient anxiety and unnecessary interventions, while false negatives risk delayed diagnoses, and overdiagnosis leads to overtreatment of indolent cases [1, 2]. These issues are exacerbated in populations with dense breast tissue or from underrepresented demographics. Consequently, there is an urgent need for innovative tools to enhance screening efficacy without compounding existing problems.

Deep learning, particularly convolutional neural networks, has been positioned as a transformative solution for breast cancer detection, with several commercial systems receiving FDA approvals for adjunctive use in

mammography [3, 4]. Vendors such as those behind iCAD and similar tools claim substantial improvements in detection rates and workflow efficiency. Nevertheless, the evidence supporting these claims often stems from retrospective analyses that may not capture the complexities of real-world clinical environments. This creates a tension between technological promise and practical implementation.

Critical questions remain about whether these AI systems deliver equitable performance across diverse populations, demonstrate cost-effectiveness in varied healthcare settings, and integrate smoothly into radiologist workflows without introducing new forms of error or fatigue [5, 6]. For instance, do variations in performance by age, ethnicity, or imaging manufacturer undermine their broad applicability? Moreover, the potential for AI to alter radiologist decision-making processes raises concerns about long-term skill maintenance. Addressing these issues is essential for responsible adoption.

This critical review systematically examines the literature on deep learning applications for breast cancer detection in medical imaging modalities from 2017 to 2022. It focuses on CNN architectures, associated training strategies, reported performance metrics, and barriers to clinical deployment while maintaining an evaluative and questioning perspective. The roadmap includes detailed methodological description, architectural and training analysis, performance critique, and forward-looking recommendations. By doing so, the review contributes to a more nuanced understanding of AI's role in healthcare systems.

Figure 1 illustrates the review's central analytical framework, showing how modality-specific imaging conditions, CNN design choices, and training strategies culminate in reported performance claims that are subsequently constrained by translational and clinical deployment barriers.

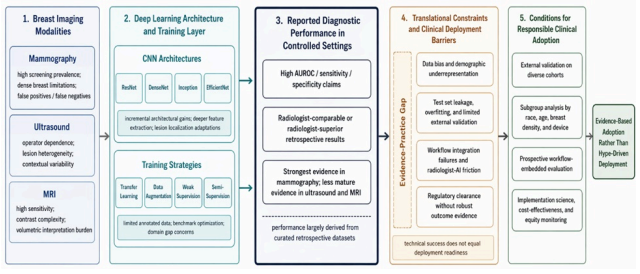

Figure 1. From Algorithmic Performance to Clinical Readiness in Deep Learning for Breast Cancer Detection: A Cross-Modality Critical Appraisal Framework

Materials and Methods

Search strategy

The search strategy for this critical review involved comprehensive queries across major databases including PubMed, IEEE Xplore, arXiv, and Google Scholar. Targeted search strings such as "deep learning" breast cancer detection mammography, "CNN architecture" breast imaging, and "transfer learning" medical imaging breast were employed to identify relevant literature from 2017 to 2022 [7, 8]. This time window was chosen to capture the rapid evolution of deep learning applications following early breakthroughs in medical imaging. Additional terms focusing on clinical deployment, FDA approvals, radiologist-AI interaction, and demographic bias ensured coverage of practical implementation aspects.

Boolean operators and filters for peer-reviewed publications were applied to maintain rigor. The strategy yielded a focused set of 31 publications that form the evidentiary basis for the subsequent critical analysis [9]. This deliberate narrowing avoided dilution by unrelated or low-quality sources. The approach prioritized depth over breadth to enable meaningful evaluative synthesis.

Inclusion and exclusion criteria

Inclusion criteria required original research articles utilizing CNN-based deep learning for breast cancer detection or classification in mammography, ultrasound, or MRI. Studies had to be published between 2017 and 2022 in English and report empirical performance metrics or deployment insights [10, 11]. This ensured relevance to the core topics of architectures, training, and barriers. Only peer-reviewed sources with sufficient methodological transparency were considered eligible.

Exclusion criteria eliminated reviews, editorials, non-CNN machine learning approaches, studies on non-breast cancers, and publications lacking sufficient methodological detail for critical appraisal. Non-peer-reviewed preprints were also excluded unless subsequently published. These criteria narrowed the scope to high-quality, focused contributions [12]. The resulting selection process

emphasized clinical applicability over purely technical novelty.

Screening and selection

Screening and selection were performed independently by the reviewer with dual verification against the inclusion criteria to minimize selection bias. Titles and abstracts were first screened, followed by full-text assessment for eligibility [13]. Emphasis was placed on studies addressing not only technical performance but also potential clinical implications. This step-by-step process reduced the risk of overlooking key limitations in the literature.

Discrepancies in selection were resolved through critical discussion of relevance to the review's objectives. The process resulted in the selection of 31 key references that collectively represent the state of the field while allowing for evaluative synthesis [14]. Such rigor ensures the review avoids cherry-picking optimistic results. The final corpus supports a balanced critique grounded in available evidence.

Data extraction

Data extraction focused on key elements including CNN architecture details, training strategies such as transfer learning and data augmentation, performance metrics like AUROC and sensitivity/specificity, and any mentions of clinical deployment or barriers [15, 16]. Information on dataset demographics, validation methods, and bias considerations was also recorded to facilitate critical assessment. Standardization prevented subjective interpretation during synthesis. This structured extraction highlighted recurring patterns across modalities.

Extraction was standardized using a predefined template to enable consistent comparison across studies. Special attention was given to claims versus evidence regarding real-world applicability [17]. The process uncovered frequent overstatements in performance reporting. Overall, it laid the foundation for identifying systemic weaknesses in the field.

Critical appraisal framework

The critical appraisal framework adapted elements of QUADAS-2 for diagnostic accuracy studies, with additional criteria for deployment readiness including regulatory status, workflow fit, and equity considerations [18]. Bias assessment emphasized patient demographic representation and generalizability to underrepresented groups. This extension moved beyond standard diagnostic tools to incorporate implementation science lenses. The framework proved essential for exposing gaps in translational readiness.

Studies were evaluated not only on methodological quality but also on the gap between reported technical success and practical utility. This framework highlighted overoptimism in performance claims and underreporting of limitations [19]. By prioritizing clinical relevance, the appraisal revealed how many studies fall short of addressing real healthcare system needs. Such evaluation underscores the necessity of moving past benchmark-focused research.

CNN architectures

Evolution of architectures

The evolution of CNN architectures for breast imaging has progressed from basic convolutional networks to sophisticated models like ResNet, DenseNet, Inception, and EfficientNet, enabling deeper feature extraction for detection tasks in mammography, ultrasound, and MRI [6, 20]. Early models focused on classification, while later iterations incorporated detection and segmentation capabilities to localize lesions more precisely. This progression reflects broader advances in computer vision adapted to medical domains. However, the pace of architectural innovation has not always aligned with clinical demands.

Architectures have been iteratively refined to handle the unique challenges of breast imaging, such as varying lesion sizes and tissue heterogeneity. Yet incremental gains from newer models often come at the expense of increased computational demands without proportional clinical benefits [7, 21]. The field has seen a shift toward models that balance accuracy and efficiency. Critically, this evolution risks prioritizing novelty over robustness in diverse clinical scenarios.

Breast-specific modifications

Breast-specific modifications to standard CNNs include multi-view aggregation strategies that combine craniocaudal and mediolateral oblique mammographic views for improved context [22, 23]. Attention mechanisms have been integrated to enhance lesion localization, while global context modules address the holistic nature of breast

tissue analysis. These adaptations aim to mimic radiologist reasoning more closely across modalities. Such customizations have extended to ultrasound and MRI where volumetric data poses additional hurdles.

These modifications have shown promise in improving localization accuracy in complex cases. Nevertheless, they raise questions about reproducibility and whether they justify the added complexity in deployment [24]. Many adaptations appear dataset-specific rather than universally transferable. A critical view questions if these tweaks truly bridge the gap to routine clinical use or merely optimize for specific test conditions.

Critical assessment

A critical assessment of CNN architectures reveals that differences between models like ResNet and DenseNet are rarely clinically significant in breast cancer detection, often overshadowed by data quality and training regimes [25, 26]. Diminishing returns on architectural innovation suggest that overparameterization may lead to unnecessary complexity without enhancing real-world performance. This pattern indicates a need to prioritize practical utility over benchmark chasing in medical imaging research. Moreover, comparative studies frequently fail to control for confounding factors adequately.

Many studies overstate the advantages of specific architectures while underplaying the role of ensemble methods or simple baselines [27, 28]. The critical view is that architectural tweaks alone cannot overcome fundamental issues in medical imaging data, such as label noise and distribution shifts [29]. Future directions should emphasize hybrid systems informed by domain expertise. Without such shifts, architectural progress risks remaining an academic exercise rather than a clinical asset.

Training strategies

Transfer learning

Transfer learning from ImageNet pre-trained weights has become the dominant strategy for CNNs in breast imaging due to the scarcity of large annotated medical datasets [5, 20]. This approach accelerates convergence and improves performance on mammography and ultrasound tasks by leveraging natural image features [8]. However, the domain gap between photographic and medical images limits its effectiveness in some cases, particularly for MRI contrast variations. Critics note that reliance on non-medical pre-training may embed irrelevant biases.

While beneficial for initial feature extraction, transfer learning's limitations become apparent when models encounter real-world protocol differences [9]. Critical evaluation shows that fine-tuning strategies vary widely, often leading to inconsistent generalizability across studies [10]. Overreliance on this method may hinder the development of truly medical-specific representations. A more tailored pre-training paradigm could mitigate these shortcomings in future work.

Data augmentation

Data augmentation techniques, including geometric transformations, intensity variations, and adversarial methods, have been widely employed to mitigate overfitting in breast cancer detection models [11, 12]. Synthetic mammogram generation via generative approaches has also been explored to expand training sets [13]. These strategies enhance model robustness to variations in imaging protocols and patient anatomy. Nevertheless, their effectiveness depends heavily on the underlying data quality.

In ultrasound and MRI, augmentation must account for modality-specific artifacts, yet many studies apply generic techniques without sufficient validation [15]. A critical perspective questions whether augmented performance truly reflects real clinical variability or merely inflates internal metrics. Without rigorous external testing, augmentation risks creating models that perform well only in simulated conditions. This highlights the need for augmentation strategies grounded in clinical realism.

Weak and semi-supervision

Weak and semi-supervised learning approaches, such as multiple instance learning for image-level labels, have reduced the annotation burden in large-scale breast imaging datasets [16, 30]. These methods enable training on partially labeled data, which is common in screening mammography where only cancer cases are explicitly annotated [17]. Semi-supervision further leverages unlabeled images for better feature learning across modalities. However, they introduce uncertainties in localization tasks critical for clinical decision-making.

Despite their efficiency, these techniques may amplify errors in low-prevalence screening settings [18]. Critical

analysis indicates that weak supervision trades annotation savings for potential reductions in model reliability. The approach works best when combined with strong validation protocols. Without such safeguards, semi-supervised gains may not translate to trustworthy clinical tools.

Critical assessment

Critical assessment of training strategies underscores that while transfer learning and augmentation dominate, they may not be optimal for addressing the unique challenges of breast imaging data [4, 25]. The prevalence of these methods reflects data scarcity more than innovation, potentially masking underlying dataset biases [26]. Diminishing returns suggest a need for alternative paradigms like self-supervised learning tailored to medical domains. Overdependence on established techniques stifles exploration of more robust alternatives.

Augmentation and weak supervision cannot fix fundamentally biased or unrepresentative training data, leading to models that perform well in lab settings but falter clinically [27, 28]. This review highlights the risk of over-optimizing for benchmark metrics at the expense of robustness [29]. A more critical approach requires rethinking training pipelines with equity and deployment in mind from the outset. Until these issues are resolved, training advances will remain limited in their translational value.

Table 1 provides a structured comparison of dominant training strategies, emphasizing how their technical advantages are frequently offset by limitations that hinder reliable clinical translation.

Table 1. Critical Comparison of Training Strategies in Deep Learning for Breast Cancer Detection and Their Translational Limitations

Training Strategy	Reported Technical Advantage	Common Implementation Approach	Underly Limitat
Transfer Learning	Accelerates convergence and improves performance with limited data	Fine-tuning ImageNet pre-trained CNNs on breast imaging datasets	Doma mismat betwee natural a medic image inconsis
			fine-tun protoc
Data Augmentation	Enhances robustness and reduces overfitting	Geometric transformations, intensity shifts, adversarial augmentation	Synthe variability not reflec clinica diversi
Weak Supervision	Enables learning from image-level labels with minimal annotation	Multiple instance learning for mammography datasets	Impreci localizat and propagati labeling r
Semi-Supervised Learning	Utilizes large volumes of unlabeled data to improve feature learning	Combination of labeled and unlabeled datasets with pseudo-labeling	Risk c reinforc incorre predicti during tra
Synthetic Data Generation	Expands dataset size and diversity	GAN-based mammogram or lesion synthesis	Limited re and pote distributi artifac
Modality-Specific Augmentation	Attempts to adapt models to ultrasound or MRI variability	Custom preprocessing and augmentation pipelines	Lack c standardiz across stu and institu
Ensemble Training	Improves predictive performance by combining multiple models	Aggregation of CNN outputs across architectures	Increas computat comple without c clinical be
Self-Supervised Learning (Emerging)	Reduces dependence on labeled data	Pretext tasks tailored to medical imaging	Limite validatio breast ima contex

Reported performance

Discrimination metrics

Reported discrimination metrics for deep learning models in breast cancer detection typically range from AUROC values of 0.85 to 0.99 across mammography, ultrasound, and MRI studies [1, 7, 21]. Several investigations demonstrate AI systems outperforming or matching radiologists in retrospective reader studies, with improvements in sensitivity and specificity [2, 22]. These results have fueled enthusiasm for AI-assisted screening programs. Comparative analyses against traditional computer-aided detection systems further emphasize deep learning superiority in controlled environments [8].

However, performance varies considerably by modality, with mammography showing the most mature applications and ultrasound and MRI presenting additional challenges due to operator dependency and complexity [23, 31]. Metrics are often derived from curated test sets that may not reflect routine clinical practice. This variability calls into question the generalizability of high AUROC scores. Without modality-specific benchmarking, claims of broad applicability remain premature.

Critical assessment of performance claims

A critical assessment of performance claims reveals substantial discrepancies between laboratory results and potential clinical performance, with retrospective designs prone to test set leakage and overfitting [24, 25]. Prospective validations are scarce, and when available, they show attenuated benefits compared to internal evaluations [9, 26]. Concerns about data contamination and optimistic reporting bias undermine confidence in many published AUROC figures. Such methodological flaws inflate perceived efficacy and hinder informed deployment decisions.

False positive and negative rates, though improved over traditional CAD, remain problematic in low-prevalence screening contexts, potentially leading to increased workload or missed cancers [10, 11]. The critical lens exposes how vendor-sponsored studies may inflate claims, while independent evaluations reveal limitations in generalization [12, 13]. Overall, performance metrics must be interpreted cautiously in light of methodological shortcomings. Until prospective, diverse testing becomes standard, reported gains risk misleading clinical adoption.

Clinical deployment barriers

Regulatory and reimbursement

Regulatory pathways for breast imaging AI have accelerated with FDA clearances for several deep learning systems, yet these approvals often rest on limited retrospective data that fail to guarantee real-world safety or efficacy [3, 4]. The 510(k) process, while efficient, frequently clears devices as substantially equivalent to older CAD systems without mandating large-scale prospective trials or long-term outcome studies [28]. Reimbursement remains fragmented, with CPT codes for AI-assisted mammography lagging behind technological claims and creating financial disincentives for widespread adoption. This regulatory-reimbursement mismatch risks premature integration of tools whose clinical value has not been fully established.

Critics argue that FDA oversight emphasizes technical performance over downstream patient outcomes, leaving health systems to navigate uncertain cost-effectiveness on their own [6, 29]. CE marking in Europe follows a similar pattern, with self-certification pathways that have drawn scrutiny for insufficient post-market surveillance. Coverage gaps persist particularly for underserved populations, where AI tools could theoretically address disparities but instead exacerbate access inequities. Without aligned incentives, deployment barriers will continue to stall meaningful translation.

Workflow integration

Workflow integration of AI into breast cancer screening demands seamless PACS compatibility and intuitive radiologist-AI interaction modes, yet most systems remain siloed from existing clinical infrastructure [8, 9]. Second-reader, triage, or concurrent reading paradigms each introduce unique challenges, including alert fatigue from excessive false positives that could erode radiologist trust and increase burnout. Studies highlight inconsistent performance when AI outputs disrupt established reading protocols rather than augment them. The lack of standardized integration protocols further complicates adoption across diverse hospital settings.

Radiologist-AI collaboration studies reveal variable acceptance, with some clinicians reporting improved confidence while others experience eroded diagnostic autonomy [10, 11]. Real-time feedback loops and explainability features are often underdeveloped, limiting

the ability of AI to serve as a true decision support tool. Workflow simulations in the reviewed literature rarely account for high-volume screening environments where time pressures amplify integration failures. Until these human factors are rigorously addressed, AI risks becoming another underutilized technology layer.

Health disparities and bias

Health disparities in breast cancer AI arise primarily from training datasets dominated by white, European or North American populations, leading to documented performance drops in non-White, older, or high-density breast cohorts [20, 22, 30]. Demographic bias manifests as reduced sensitivity in underrepresented groups, potentially widening existing outcome gaps rather than narrowing them. Generalization failures across imaging manufacturers and protocols compound these issues, rendering many models unreliable for global screening programs. The literature consistently underreports subgroup analyses, obscuring the true extent of inequity.

Critical evaluation shows that without deliberate debiasing strategies, AI systems risk perpetuating systemic healthcare injustices under the guise of technological neutrality [12, 13]. Performance disparities by race, age, and breast density remain insufficiently quantified in most studies, limiting informed deployment decisions. This bias not only affects diagnostic accuracy but also erodes trust among diverse patient populations. Addressing these disparities requires more than post-hoc fixes; it demands upstream changes in data curation and validation.

Implementation science gaps

Implementation science gaps are evident in the scarcity of prospective studies evaluating AI-assisted breast screening in live clinical environments, leaving critical questions about real-world effectiveness unanswered [9, 26]. User acceptance research is sparse, with most evidence drawn from small-scale pilots that fail to capture organizational and cultural barriers. Cost-effectiveness analyses are virtually absent, despite high development and maintenance expenses that could strain resource-limited systems. Unintended consequences, such as over-reliance on AI outputs, receive minimal attention in the current evidence base.

The reviewed literature reveals a persistent disconnect between technical validation and implementation readiness, with few studies employing frameworks from implementation science [14, 15]. This gap hinders scalable deployment and exposes patients to unproven technologies. Stakeholder engagement with radiologists, administrators, and patients remains underdeveloped. Bridging these gaps is essential before AI can transition from experimental to standard-of-care status in breast cancer screening.

Critical synthesis

The evidence-practice gap

The evidence-practice gap in deep learning for breast cancer detection is stark, as published performance metrics far outpace demonstrated clinical readiness across the 31 reviewed studies [24, 25]. Most systems remain confined to retrospective evaluations, with FDA approvals frequently granted on surrogate endpoints that do not reliably predict improved patient outcomes [16, 17]. This mismatch between laboratory promise and deployment reality undermines claims of transformative impact. The field must confront the uncomfortable truth that technical success does not equate to healthcare system integration.

Table 2 provides a cross-modality analytical comparison showing that the apparent success of deep learning differs substantially across mammography, ultrasound, and MRI because each modality couples distinct technical advantages with distinct translational vulnerabilities.

Table 2. Cross-Modality Analytical Comparison of Deep Learning Utility and Translational Fragility in Breast Cancer Imaging

Analytical dimension	Mammography	Ultrasound	
Clinical role in the literature	Most mature and most commonly studied screening modality	Frequently positioned as adjunctive and diagnostically variable	C s o

Typical deep learning advantage claimed	Improved lesion detection, reduced reader variability, stronger retrospective discrimination	Improved pattern recognition in heterogeneous lesion appearance		Likely consequence of premature deployment	Misleading confidence in broadly deployable mammography AI	Unstable performance in real-world ultrasound practice	
Data structure and acquisition characteristics	Standardized views support model training, but device and population shifts remain consequential	Strong operator dependence introduces instability in image quality and lesion representation		Most defensible research priority	Diverse external validation and subgroup reporting	Standardization-aware validation and clinically realistic augmentation	
Principal source of performance optimism	Large curated retrospective datasets and reader-study comparisons	Small or selective datasets with inflated internal validity					
Principal source of translational fragility	Dense breast limitations, demographic underrepresentation, manufacturer variation	Operator variability, limited reproducibility, contextual inconsistency					
External validation vulnerability	Moderate to high, especially across institutions and devices	High, because acquisition variability weakens portability					

Critical synthesis further exposes how selective reporting and optimistic interpretations inflate perceived readiness while downplaying persistent limitations [18, 19]. Prospective data are insufficient to support broad claims of superiority over traditional workflows. The gap persists because research incentives favor benchmark achievements over messy implementation challenges. Closing it requires a fundamental shift in study design priorities.

Who benefits and who is left behind?

Who benefits from current AI developments in breast imaging remains unclear, as performance disparities suggest that advantaged populations gain the most while marginalized groups risk being left behind [20, 22, 30]. Demographic underrepresentation in training data systematically disadvantages non-White and low-resource populations, potentially widening rather than reducing health inequities [12, 13]. The reviewed evidence indicates that without targeted validation, AI could reinforce existing biases in screening access and outcomes. This raises ethical questions about equitable technology distribution in oncology.

Critical assessment reveals that benefits accrue unevenly, with high-income settings better positioned to absorb integration costs and mitigate risks [26, 27]. Vulnerable subgroups face higher false-negative risks that could delay diagnoses disproportionately. The literature offers limited guidance on mitigating these harms, highlighting a collective failure to prioritize fairness. Future progress must center inclusive design to ensure AI serves all patients rather than a privileged subset.

Recommendations

For researchers

Researchers should prioritize reporting subgroup performance stratified by race, age, breast density, and imaging device to expose hidden biases and improve generalizability [20, 22, 30]. External validation on independent, diverse cohorts must become non-negotiable rather than optional, moving beyond internal test sets that inflate results [24, 25]. Prospective silent-mode studies embedded in real workflows would provide the missing link between bench and bedside. Such practices would elevate the field from incremental technical gains to clinically meaningful contributions.

Collaboration with implementation scientists and clinicians from the study outset is essential to embed deployment considerations early [9, 26]. Open sharing of code, datasets, and negative results would accelerate collective learning and reduce publication bias. Researchers must also engage directly with affected communities to ensure relevance and equity. Only through these rigorous standards can the literature support trustworthy AI advancement.

For journal editors and reviewers

Journal editors and reviewers should reject manuscripts lacking robust external validation or subgroup analyses, enforcing higher evidentiary thresholds for deep learning studies in medical imaging [10, 11]. Mandatory discussion of deployment barriers, including bias and workflow implications, would discourage purely technical papers that ignore clinical context [12, 13]. Requiring prospective elements or clear pathways to implementation would align publications more closely with healthcare needs. This gatekeeping role is critical to raising overall research quality.

Editors must also demand transparency in dataset demographics and conflict-of-interest disclosures to combat hype-driven narratives [14, 15]. Special issues or dedicated sections on implementation science could incentivize the missing translational work. Peer review should explicitly evaluate equity and fairness dimensions alongside technical merit. Such policies would reshape incentives and accelerate progress toward deployable solutions.

For regulatory bodies (FDA, EMA)

Regulatory bodies like the FDA and EMA should mandate diverse validation populations representing global demographics before granting clearances for breast imaging AI [3, 4, 28]. Post-market surveillance requirements must include ongoing bias monitoring and real-world performance tracking to catch degradation over time or across subgroups [16, 17]. Transparency in model updates and training data provenance would enable independent scrutiny and build public confidence. These measures would shift regulation from initial approval to sustained accountability.

Clearer guidance on reimbursement-linked evidence standards would align market incentives with clinical utility [6, 29]. International harmonization of requirements could reduce duplication while elevating safety benchmarks worldwide. Regulatory frameworks must evolve to incorporate implementation science metrics rather than relying solely on diagnostic accuracy. Only then can regulators fulfill their mandate to protect patients while fostering innovation.

For clinicians and hospital administrators

Clinicians and hospital administrators should demand local validation of vendor AI tools against their specific patient populations and imaging protocols before adoption [18, 19]. Questioning performance claims with reference to prospective evidence rather than marketing materials would prevent costly missteps in procurement. Phased deployment with built-in audit mechanisms allows iterative assessment of impact on workflow and outcomes. This cautious stance protects both patients and institutional resources.

Administrators must integrate AI procurement into broader digital health strategies that address training, liability, and equity implications [8, 9]. Multidisciplinary oversight committees involving radiologists, informaticians, and ethicists can guide responsible implementation. Ongoing

monitoring for alert fatigue and bias drift should be standard operating procedure. By exercising informed skepticism, frontline stakeholders can drive vendors toward genuinely useful technologies.

Research gaps

Prospective deployment studies

Prospective deployment studies remain critically underrepresented, with randomized trials comparing AI-assisted versus unaided reading sorely needed to establish true clinical benefit [9, 26]. Workflow integration research must examine real-time effects on radiologist performance, reading time, and diagnostic confidence across varied settings. Current evidence relies too heavily on retrospective designs that cannot capture dynamic clinical interactions. Filling this gap is essential for evidence-based policy on AI adoption.

Large-scale, multi-center trials incorporating cost-effectiveness and patient-reported outcomes would provide the definitive data currently lacking [14, 15]. Silent-mode implementations could ethically test systems without risking patient harm during evaluation. Such studies would clarify which use cases deliver net benefit versus added burden. Until completed, deployment decisions will continue to rest on incomplete foundations.

Bias mitigation

Bias mitigation strategies require urgent attention, with algorithmic fairness techniques and domain adaptation methods needing rigorous testing in breast imaging contexts [20, 22, 30]. Diverse training data initiatives, including federated learning across institutions, offer promising avenues but lack sufficient validation [12, 13]. Research must move beyond detection of disparities to proactive correction and continuous monitoring frameworks. This gap, if unaddressed, threatens the ethical foundation of AI in healthcare.

Comparative effectiveness studies of different debiasing approaches would guide best practices for developers and regulators [16, 17]. Longitudinal analyses of bias evolution post-deployment are virtually absent from the literature. Interdisciplinary work combining computer science, epidemiology, and health equity expertise is necessary to generate actionable solutions. Closing this research gap represents one of the most pressing priorities for responsible AI development.

Conclusion

This critical review demonstrates that while CNN architectures and training strategies for breast cancer detection in mammography, ultrasound, and MRI have advanced considerably between 2017 and 2022, substantial deployment barriers persist and undermine claims of clinical transformation. Transfer learning, data augmentation, and weak supervision have enabled impressive laboratory performance, yet these gains rarely survive translation to diverse real-world environments. The synthesis reveals a field still maturing technically but lagging in implementation readiness and equity considerations. Overall, deep learning shows technical promise that has not yet matured into reliable healthcare system integration.

Performance claims continue to outpace clinical evidence, with retrospective AUROC values masking prospective limitations, demographic biases, and workflow incompatibilities. Regulatory approvals exist but do not equate to proven effectiveness or cost-effectiveness, leaving health systems to navigate uncertain value propositions. Bias concerns remain largely unresolved despite growing awareness, risking exacerbation of existing disparities. The reviewed literature therefore urges caution against premature scaling.

The field must prioritize external validation, subgroup analyses, and implementation science to bridge the persistent evidence-practice gap. Researchers, journals, regulators, and clinicians each have defined roles in elevating standards and ensuring accountability. Without these shifts, AI risks remaining an academic curiosity rather than a patient-centered tool. Sustained critical evaluation will be essential as the technology evolves.

Ultimately, this critical review calls for evidence-based adoption grounded in technology assessment before widespread deployment of deep learning systems for breast cancer screening. The potential benefits are real but conditional on addressing the identified gaps with rigor and equity at the forefront. Responsible advancement demands that clinical utility, not technical novelty, drive future progress in AI for healthcare systems. Only through such disciplined scrutiny can the field deliver on its promise to improve breast cancer outcomes globally.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 Mar 2022 Revised: 06 Jun 2022 Accepted: 16 Jul 2022
Published online: 20 January 2023

References

McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.

Shen L. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. arXiv. 2017;arXiv:1711.05775.

Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol. 2017;52(7):434-40.

Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 2019;292(1):60-6.

Ayana G, Dese K, Choe SW. Transfer learning in breast cancer diagnoses via ultrasound imaging. Cancers (Basel). 2021;13(4):738.

Yap MH, Pons G, Marti J, Ganau S, Sentis M, Zwiggelaar R, et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform. 2018;22(4):1218-26.

Hamed G, Marey MA, Amin SE, Tolba MF. Deep learning in breast cancer detection and classification. In: International Conference on Artificial Intelligence and Computer Vision; 2020. p. 322-33.

Bharati S, Podder P, Mondal M. Artificial neural network based breast cancer screening: a comprehensive review. arXiv. 2020;arXiv:2006.01767.

Jaglan P, Dass R, Duhan M. Breast cancer detection techniques: issues and challenges. J Inst Eng India Ser B. 2019;100(4):379-86.

Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. Sci Rep. 2019;9(1):12495.

Yoon JH, Kim EK. Deep learning-based artificial intelligence for mammography. Korean J Radiol. 2021;22(8):1225-39.

Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, et al. Risk factors and preventions of breast cancer. Int J Biol Sci. 2017;13(11):1387-97.

Bhushan A, Gonsalves A, Menon JU. Current state of breast cancer diagnosis, treatment, and theranostics. Pharmaceutics. 2021;13(5):723.

Amanova N, Martin J, Elster C. Explainability for deep learning in mammography image quality assessment. Mach Learn Sci Technol. 2022;3(2):025015.

Keating NL, Pace LE. Breast cancer screening in 2018: time for shared decision making. JAMA. 2018;319(17):1814-5.

Seely JM, Alhassan T. Screening for breast cancer in 2018-what should we be doing today? Curr Oncol. 2018;25(Suppl 1):S115-S124.

Wei B, Han Z, He X, Yin Y. Deep learning model based breast cancer histopathological image classification. In: 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA); 2017. p. 348-353.

Brentnall AR, Cuzick J, Buist DS, Bowles EJ. Long-term accuracy of breast cancer risk assessment combining classic risk factors and breast density. JAMA Oncol. 2018;4(9):e180174.

Zanotel M, Bednarova I, Londero V, Linda A, Lorenzon M, Girometti R, et al. Automated breast ultrasound: basic principles and emerging clinical applications. Radiol Med. 2018;123(1):1-12.

Wakili MA, Shehu HA, Sharif MH, Sharif MH, Umar A, Kusetogullari H, et al. Classification of breast cancer histopathological images using DenseNet and transfer learning. Comput Intell Neurosci. 2022;2022:8904768.

Ha R, Chang P, Karcich J, Mutasa S, Van Sant EP, Liu MZ, et al. Convolutional neural network based breast cancer risk stratification using a mammographic dataset. Acad Radiol. 2019;26(4):544-9.

Liu H, Chen Y, Zhang Y, Wang L, Luo R, Wu H, et al. A deep learning model integrating mammography and clinical factors facilitates the malignancy prediction of BI-RADS 4 microcalcifications in breast cancer screening. Eur Radiol. 2021;31(8):5902-12.

Qian X, Pei J, Zheng H, Xie X, Yan L, Zhang H, et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat Biomed Eng. 2021;5(6):522-32.

Kretz T, Müller KR, Schaeffter T, Elster C. Mammography image quality assurance using deep learning. IEEE Trans Biomed Eng. 2020;67(12):3317-26.

Madani M, Behzadi MM, Nabavi S. The role of deep learning in advancing breast cancer detection using different imaging modalities: a systematic review. Cancers (Basel). 2022;14(21):5334.

Duffy SW, Tabár L, Yen AM, Dean PB, Smith RA, Jonsson H, et al. Mammography screening reduces rates of advanced and fatal breast cancers: results in 549,091 women. Cancer. 2020;126(13):2971-9.

Geras KJ, Wolfson S, Shen Y, Wu N, Kim S, Kim E, et al. High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv. 2017;arXiv:1703.07047.

Lehman CD, Mercaldo S, Lamb LR, King TA, Ellisen LW, Specht M, et al. Deep learning vs traditional breast cancer risk models to support risk-based mammography screening. J Natl Cancer Inst. 2022;114(10):1355-63.

De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24(9):1342-50. https://doi.org/10.1038/s41591-018-0107-6.

Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clin Cancer Res. 2018;24(23):5902-9.

Witowski J, Heacock L, Reig B, Kang SK, Lewin A, Pysarenko K, et al. Improving breast cancer diagnostics with deep learning for MRI. Sci Transl Med. 2022;14(664):eabo4802.