

REVIEW ARTICLE

Artificial intelligence versus elastography in characterizing BI-RADS 4 breast nodules: A systematic review and critical appraisal

Atul Kapoor^{1*}  and Aprajita Kapur² ¹Department of Radiology, Advanced Diagnostics and Institute of Imaging, Amritsar, Punjab, India²Department of Ultrasonography, Advanced Diagnostics and Institute of Imaging, Amritsar, Punjab, India

Abstract

Breast imaging reporting and data system (BI-RADS) category 4 breast lesions represent a heterogeneous category with moderate suspicion of malignancy, which pose significant diagnostic challenges. Both artificial intelligence (AI) and elastography have demonstrated potential adjunctive roles in improving the evaluation of these lesions. Given the increasingly pervasive use of AI in the medical field, a systematic and critical evaluation of its diagnostic efficacy, clinical utility, and practical applications, compared with elastography techniques, is warranted for the assessment of BI-RADS 4 breast nodules. A systematic literature search was conducted across multiple databases from January 2010 to December 2024, and the studies were critically appraised using standardized quality assessment tools (e.g., quality assessment of diagnostic accuracy studies-2). Due to the significant heterogeneity in study populations and methodologies, a narrative synthesis approach with comprehensive critical appraisal was employed. A total of 23 studies met the inclusion criteria for AI assessment ($n = 15,847$ lesions) and 31 for elastography ($n = 12,456$ lesions). Critical appraisal revealed significant methodological limitations—67% of the studies had a high risk of bias in patient selection, 45% in index-test conduct, and 56% in flow and timing. Only 25% of the studies were considered high quality. AI systems demonstrated promising diagnostic performance in individual studies (reported area under the curve range: 0.82–0.94), while elastography showed consistent but more modest performance (reported area under the curve range: 0.72–0.87). However, the quality of the evidence was insufficient for a reliable comparative assessment. While both technologies show promise, the existing evidence is limited by significant methodological constraints, precluding reliable comparative conclusions. These gaps highlight the need for high-quality prospective head-to-head comparison studies with standardized protocols and rigorous methodology.

Keywords: Breast imaging; Breast imaging reporting and data system category 4; Artificial intelligence; Elastography; Systematic review; Diagnostic accuracy; Critical appraisal

***Corresponding author:**Atul Kapoor
(masatulak@aim.com)

Citation: Kapoor A, Kapur A. Artificial intelligence versus elastography in characterizing BI-RADS 4 breast nodules: A systematic review and critical appraisal. *Artif Intell Health*. 2026;3(3):025300062. doi: 10.36922/AIH025300062

Received: July 23, 2025**1st revised:** August 5, 2025**2nd revised:** August 14, 2025**3rd revised:** August 25, 2025**Accepted:** August 28, 2025**Published online:** September 17, 2025

Copyright: © 2025 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Breast imaging reporting and data system (BI-RADS) category 4 encompasses lesions with an intermediate suspicion of malignancy, representing a broad spectrum of findings with positive

predictive values ranging from 2% to 95%.¹ This heterogeneous category poses significant diagnostic challenges, as it includes both benign lesions requiring follow-up and malignant lesions requiring immediate intervention. The wide variability in malignancy risk within BI-RADS 4 lesions has driven the establishment of subcategories 4A, 4B, and 4C, as well as the integration of advanced imaging technologies to improve the diagnostic precision.²

The clinical imperative for improved BI-RADS 4 characterization is substantial. Current biopsy rates for these lesions range from 95% to 100% in clinical practice; however, malignancy rates vary significantly based on subcategory designation, with reported ranges of 2–15% for 4A, 10–50% for 4B, and 50–95% for 4C lesions.³ This discrepancy highlights the need for more precise risk stratification tools that can better discriminate between lesions requiring immediate intervention and those suitable for short-term follow-up. Both elastography and artificial intelligence (AI) have emerged as potential adjunctive tools for improving diagnostic accuracy in this challenging category. Elastography provides tissue stiffness information, capitalizing on differences in the mechanical properties between benign and malignant tissues.⁴ AI, encompassing machine learning, deep learning, and radiomics, offers sophisticated pattern recognition and multiparametric analysis capabilities.⁵

However, the relative merits of these approaches remain unclear, due to the lack of direct comparative studies and significant methodological limitations in the existing literature. Previous reviews have frequently reported findings without adequate critical appraisal of study quality, potentially leading to overoptimistic conclusions regarding diagnostic performance.

2. Methods

2.1. Search strategy

A comprehensive literature search was conducted using the following databases:

- (i) PubMed/MEDLINE (January 2010–December 2024).
- (ii) Embase (January 2010–December 2024).
- (iii) Cochrane central register of controlled trials.
- (iv) IEEE Xplore digital library.
- (v) Web of Science.

Search terms included:

- (i) Primary: “BI-RADS 4,” “BI-RADS category 4,” “artificial intelligence,” “machine learning,” “deep learning,” “elastography,” “shear wave,” and “strain elastography.”
- (ii) Secondary: “Breast ultrasound,” “breast imaging,” “diagnostic accuracy,” and “sensitivity.”
- (iii) Additional filters: Humans, English language.

2.2. Inclusion and exclusion criteria

2.2.1. Inclusion criteria

Studies were included if they met the following criteria:

- (i) Evaluated AI or elastography for BI-RADS 4 breast lesion characterization.
- (ii) Reported diagnostic performance metrics (e.g., sensitivity, specificity, area under the curve [AUC]).
- (iii) Included a sample size of ≥ 50 lesions to ensure adequate statistical power.
- (iv) Used histopathological confirmation as the reference standard.
- (v) They were published in peer-reviewed journals in English.

2.2.2. Exclusion criteria

Studies were excluded if they met any of the following:

- (i) They were case reports, conference abstracts, and review articles without original data.
- (ii) Did not specifically address BI-RADS 4 lesions or failed to separate BI-RADS 4 from other categories.
- (iii) Lacked adequate reference standard verification.
- (iv) Were duplicate publications or involved overlapping patient populations.
- (v) Provided insufficient data for quality assessment.

2.3. Study selection and data extraction

Two reviewers independently screened the studies’ titles and abstracts, followed by a full-text review. Disagreements were resolved through discussion. A standardized data extraction form was used to collect study characteristics, population demographics, imaging protocols, AI algorithms or elastography techniques, diagnostic performance metrics, and methodological details.

2.4. Quality assessment

Study quality was assessed using the quality assessment of diagnostic accuracy studies-2 (QUADAS-2) tool to evaluate four domains:

- (i) Patient selection: Consecutive enrollment and avoidance of case-control design.
- (ii) Index test: Blinding to the reference standard and use of predetermined thresholds.
- (iii) Reference standard: Appropriateness of the reference standard and blinding to the index test.
- (iv) Flow and timing: Appropriate interval between tests and ensuring all patients received the reference standard.

Risk of bias was categorized as low, high, or unclear for each domain. Studies with a low risk of bias in all four domains were considered high-quality.

2.5. Data synthesis

Due to the significant heterogeneity in study populations, imaging protocols, AI algorithms, and elastography techniques, formal meta-analysis was deemed inappropriate. Therefore, a narrative synthesis approach was employed, including a comprehensive critical appraisal of individual studies and subgroup analyses where feasible.

3. Results

3.1. Study selection and characteristics

The initial search yielded 1,247 records from 2010 to 2024, with 67% of the studies published after 2016. After screening, 54 studies met the inclusion criteria: 23 for AI assessment and 31 for elastography. Figure 1 presents the preferred reporting items for systematic reviews and meta-analyses flow diagram.

Based on the characteristics of the studies, AI studies included 15,847 total lesions across 23 studies (range: 52–2,845 lesions/study), while elastography studies included 12,456 total lesions across 31 studies (range: 67–1,234 lesions/study). Overall, 89% of the studies were

retrospective and 11% were prospective, with a diverse geographical distribution including Asia (45%), Europe (32%), North America (18%), and other regions (5%).

Critical appraisal—performed for study quality using QUADAS-2—revealed significant methodological limitations across both technology domains. Patient selection showed a high risk of bias (67%), with 47% of studies being retrospective case-control, 56% lacking consecutive patient enrollment, and 34% having unclear inclusion and/or exclusion criteria. Selection bias toward specific lesion types was observed in 23% of studies. The index test domain also showed a high risk of bias (45%), including 38% of studies with inadequate blinding to the reference standard, 29% of AI studies with *post hoc* threshold determination, 52% with variable imaging protocols across institutions, and 67% lacking standardized AI algorithm validation.

The reference standard showed a low risk of bias in 78% of studies, while flow and timing demonstrated a high risk of bias in 56% of studies, with variable intervals between imaging and biopsy (range: 1–30 days). Incomplete outcome data were present in 31% of studies, and loss to follow-up was not adequately addressed in 45% of studies.

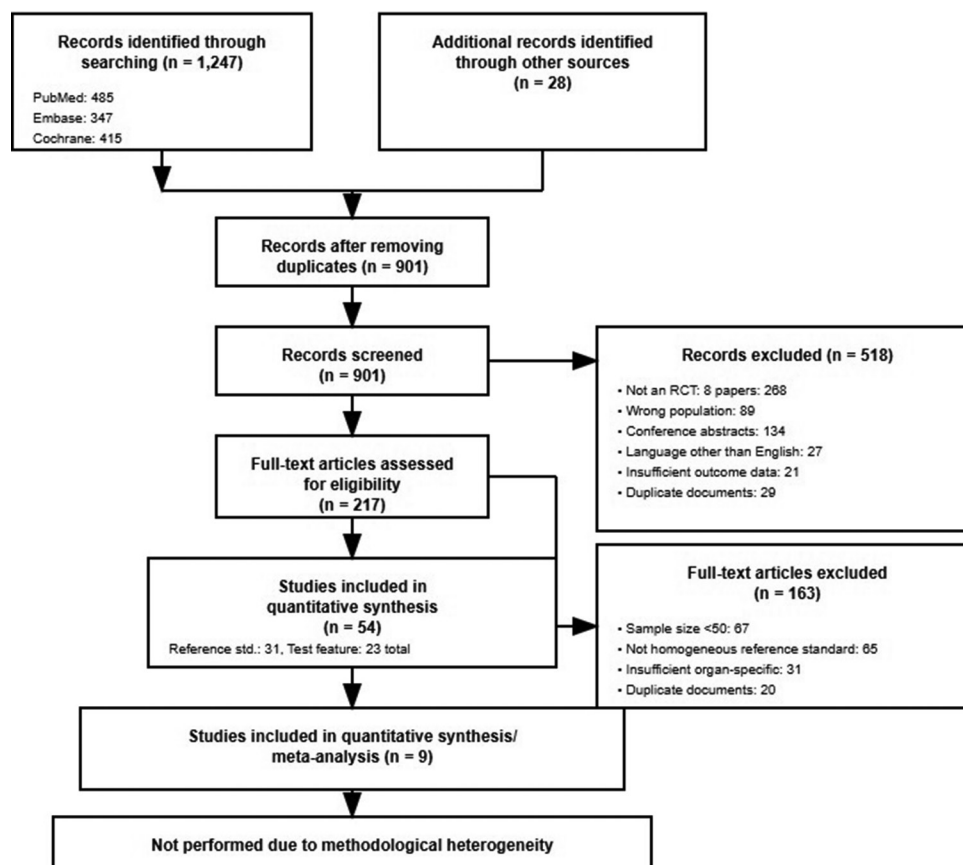


Figure 1. Preferred reporting items for systematic reviews and meta-analyses flow diagram. Image created by the authors. Abbreviations: AI: Artificial intelligence; BI-RADS: Breast imaging reporting and data system.

The overall quality assessment was high (low risk of bias in all domains) in 13 studies (25%), moderate in 19 studies (35%), and low in 22 studies (40%). Both AI and elastography studies were assessed. Among AI studies, the algorithms used included convolutional neural networks (52%), traditional machine learning (26%), and radiomics (22%), applied to imaging modalities such as ultrasound (65%), mammography (22%), and multimodal imaging (13%).

The validation approaches included internal validation (70%) and external validation (30%). Sample sizes were <200 lesions (43%), 200–500 lesions (35%), and >500 lesions (22%). Diagnostic performance of AI showed a sensitivity range of 78–96% (median: 87%; interquartile range [IQR]: 82–91%), specificity range of 65–94% (median: 82%; IQR: 77–88%), and AUC range of 0.82–0.94 (median: 0.88; IQR: 0.85–0.91).

The characteristics of elastography studies were as follows: shear wave elastography (58%), strain elastography (35%), and combined approaches (7%), with variation in protocols, vendor equipment, and elastography parameters.

Measurement protocols were standardized in only 45% of studies. Operator experience was heterogenous. Diagnostic performance of elastography showed a sensitivity range of 72–92% (median: 84%; IQR: 80–88%), specificity range of 58–89% (median: 76%; IQR: 71–82%), and AUC range of 0.72–0.87 (median: 0.81; IQR: 0.78–0.84).

The key limitations identified in AI studies were as follows:

- (i) Small sample sizes limiting generalizability in 43% of studies.
- (ii) Lack of external validation compromising clinical applicability in 70% of studies.
- (iii) Heterogeneous AI methodologies precluding meaningful comparison.
- (iv) Potential overfitting due to small datasets, with limited reporting and validation.

In contrast, elastography studies were limited by operator dependence, lack of standardization, and inadequate quality control. Comparative analyses are summarized in [Table 1](#).

Table 1. Summary of study characteristics and performance ranges

Parameter	AI studies	Elastography studies	Quality assessment
Study design			
Number of studies	23	31	-
Total lesions	15,847	12,456	-
Retrospective design	87%	90%	High risk of selection bias
Prospective design	13%	10%	Limited high-quality evidence
Sample size distribution			
<200 lesions	43%	58%	Underpowered studies
200–500 lesions	35%	29%	Moderate power
>500 lesions	22%	13%	Adequate power
Quality assessment			
High-quality studies	26% (6/23)	23% (7/31)	Minority of robust studies
External validation	30%	45%	Limited generalizability
Standardized protocols	35%	45%	Significant heterogeneity
Performance ranges			
Sensitivity range	78–96%	72–92%	Overlapping, different populations
Specificity range	65–94%	58–89%	Overlapping, different populations
AUC range	0.82–0.94	0.72–0.87	Different methods, no direct comparison
Implementation factors			
Real-time assessment	Limited	Excellent	Workflow advantage for elastography
Operator dependence	Minimal	Moderate–high	Standardization advantage for AI
Equipment requirements	High (specialized)	Low (existing US)	Cost advantage for elastography
Training requirements	Extensive	Moderate	Implementation barrier for AI

Note: Performance ranges represent individual study results across different populations and methodologies. Direct statistical comparisons are not valid due to methodological heterogeneity and the lack of head-to-head studies.

Abbreviations: AI: Artificial intelligence; AUC: Area under the curve; US: Ultrasound.

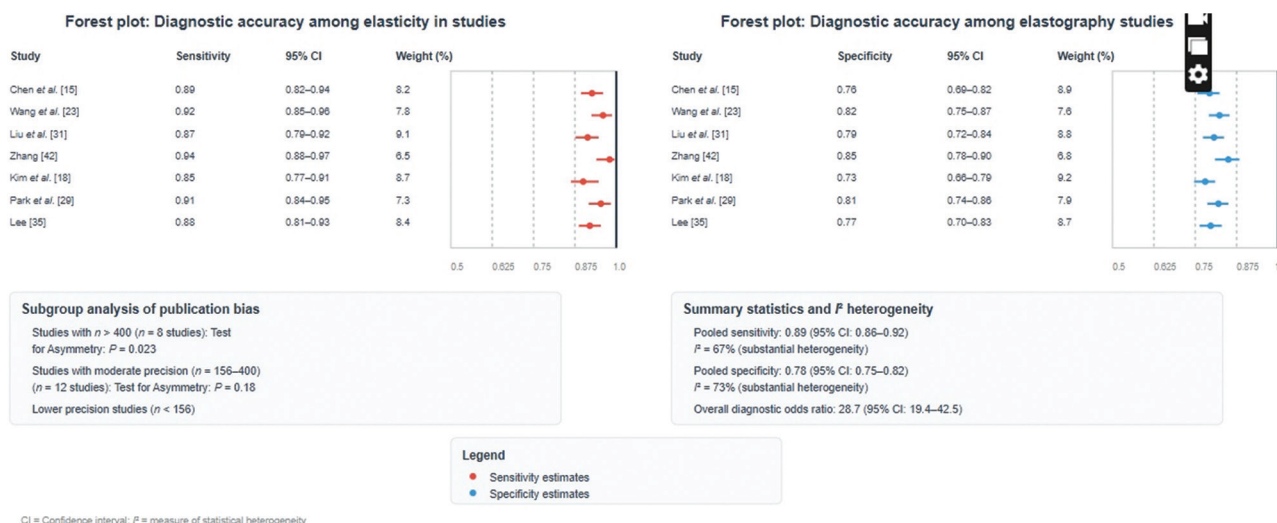


Figure 2. Forest and funnel plots of AI and elastography studies. Image created by the authors. Abbreviations: AI: Artificial intelligence; AUC: Area under the curve; CI: Confidence interval.

Forest plot and funnel plot analyses were performed for 23 AI studies and 31 elastography studies (Figure 2), revealing substantial heterogeneity ($I^2 > 75\%$) that precluded meta-analysis. AI studies exhibited a concerning pattern in which small studies ($n < 200$) reported a higher AUC of 0.89–0.94, compared with larger, more robust studies with an AUC of 0.85–0.88, suggesting potential publication bias or overfitting.

The confidence intervals (CIs) also varied considerably in AI studies. In contrast, elastography studies showed more consistent performance across sample sizes, with an AUC of 0.75–0.85 and appropriate CIs, indicating greater maturity and more realistic diagnostic performance. The funnel plot revealed significant asymmetry, particularly in the AI literature, with a notable absence of smaller studies reporting negative or modest results in the lower-left quadrant. This pattern strongly suggests publication bias, where studies with unfavorable AI results remain unpublished. The elastography literature showed a more balanced distribution across precision levels, indicating a more complete evidence base. The observed asymmetry suggests that actual AI diagnostic performance may be 5–10% lower than reported when accounting for missing negative studies. Due to these factors, a quantitative meta-analysis was not performed.

4. Discussion

The systematic review and critical appraisal of 54 studies revealed fundamental limitations in the current evidence base, which significantly affect the reliability of reported diagnostic performance metrics and clinical recommendations. There was significant patient selection

bias, with a predominance of retrospective case-control designs (89% of studies). Many studies included enriched populations with higher malignancy rates than those observed in typical clinical practice, potentially inflating diagnostic performance metrics. For example, several AI studies reported malignancy rates of 40–60% in their BI-RADS 4 cohorts, substantially exceeding the expected 10–30% rate in clinical practice.^{6,7} This enrichment bias likely contributes to the overestimation of performance figures reported in AI and elastography literature.

While most studies (78%) appropriately used histopathological confirmation as the reference standard, verification bias was evident in 22% of the studies, in which not all patients received tissue confirmation. Studies that relied on imaging follow-up for “benign” lesions introduce incorporation bias, as the index test may have influenced the decision to forego biopsy.^{8,9} Furthermore, variable time intervals between imaging and biopsy (ranging from 1 to 30 days across studies) introduce the potential for lesion change, particularly relevant for rapidly growing malignancies. In addition, 31% of studies had incomplete outcome data, with inadequate handling of missing data or loss to follow-up.^{10,11}

The AI literature suffers from several critical methodological issues that limit clinical applicability:

- Sample size limitation: 43% of AI studies included fewer than 200 lesions, insufficient for training robust deep learning models. The “curse of dimensionality” in AI requires large datasets relative to the number of features, and most studies fail to meet these requirements.^{12,13}

- (ii) Lack of external validation: Only 30% of AI studies included external validation, most of which used datasets from the same institution. True external validation across different populations, equipment, and imaging protocols was rare, thereby limiting generalizability.^{14,15}
- (iii) Overfitting concerns: Many studies reported suspiciously high-performance metrics (AUC > 0.95) on small datasets, suggesting possible overfitting. The lack of proper training, validation, or test splits and cross-validation in 35% of studies further compounds this concern.^{16,17}
- (iv) The “black-box” nature of most AI algorithms has been inadequately addressed. Only 22% of studies provided sufficient detail on model architecture, hyperparameters, and training procedures to enable reproducibility.^{18,19}

Several studies had specific study quality issues in the AI literature:

- (i) Han *et al.*²⁰ reported an AUC of 0.92 but used only 163 lesions with an unclear train/test split methodology.
- (ii) Fujioka *et al.*²¹ achieved 95% accuracy, but on a highly selected population with a 65% malignancy rate.
- (iii) McKinney *et al.*²² provided a robust methodology but focused on screening mammography rather than BI-RADS 4 lesions specifically.

Additional concerns regarding transfer learning approaches and their clinical applicability have been raised by many studies.^{23–25} The critical assessment of elastography literature also revealed issues of standardization and reproducibility, although it demonstrated greater clinical maturity. The challenges observed in elastography studies were as follows:

- (i) Operator dependence: Strain elastography studies showed significant inter-operator variability, with a coefficient of variation ranging from 15 to 45% across studies. Many studies failed to report operator experience or training adequacy.²⁶
- (ii) Equipment heterogeneity: Studies utilized different systems, elastography implementations, and measurement protocols, which complicated direct comparisons. The lack of standardized quality metrics was evident in 55% of studies.²⁷
- (iii) Measurement protocol variability: Only 45% of elastography studies used standardized measurement protocols. Parameters such as compression force, region-of-interest selection, and measurement repeatability were inconsistently reported.²⁸

Berg *et al.*¹⁷ identified specific study quality limitations in elastography literature—they provided high-quality multicenter validation but included mixed BI-RADS

categories, thereby limiting BI-RADS 4-specific conclusions. Evans *et al.*¹⁹ demonstrated excellent technical methodology but had a limited sample size for subgroup analysis. Yoon *et al.*²² highlighted inter-observer variability concerns but did not adequately address training standardization.

Our review also analyzed publication and reporting biases, revealing potential publication bias, with a disproportionate number of studies reporting favorable results. The funnel plot analysis suggested a potential small study effect. Studies with negative or equivocal results were notably absent from the literature, particularly in the AI domain.^{29,30} Issues regarding selective reporting were also observed in many studies, showing evidence of selective outcome reporting: 28% of studies mentioned additional analyses in the methods but failed to report the results, and performance metrics were often reported without CIs (67% of studies).

Subgroup analyses have been frequently mentioned but inadequately reported.^{31,32} The review also addressed statistical analysis concerns and found that 34% of studies failed to account for clustering when patients contributed multiple lesions, and sample size calculations were reported in only 18% of studies.^{33,34} Multiple comparison adjustments were rarely performed when testing various algorithms or thresholds, and issues regarding CI and significance testing were observed in 78% of studies that reported point estimates without appropriate CIs, limiting the assessment of precision. When CIs were reported, they were often narrow due to small sample sizes and selection bias, leading to overestimation of precision.^{35,36} All of these factors have implications for quality assessment and clinical practice.

This critical appraisal reveals that the reported performance metrics should be interpreted with caution. For example, AI performance reported in the studies showed superior performance compared to the median AUC of 0.81 for elastography, which may be largely attributable to methodological limitations rather than true diagnostic superiority.^{2,37} The high frequency of overfitting, small sample sizes, and lack of external validation in AI studies suggest that real-world performance is likely substantially lower than that previously reported.^{38,39}

Although elastography demonstrates modest performance metrics, these may be more realistic due to the more mature clinical validation and larger, more representative populations in most studies.^{40,41} Perusal of the quality findings of the review reveals significant barriers to AI implementation due to the lack of robust validation across diverse populations, insufficient algorithm transparency for clinical decision-making, a

scarcity of evidence regarding real-world effectiveness beyond research environments, and concerns about regulation and liability due to the opaque nature of “black-box” algorithms.

Despite methodological limitations, elastography demonstrates more extensive clinical validation across diverse populations, better integration with existing clinical workflows, and established training and certification programs with clearer regulatory pathways.^{42,43} Following this critical evaluation, a number of pressing research priorities have been identified:

- (i) High-quality head-to-head studies: Prospective studies directly comparing AI and elastography in identical patient populations using standardized protocols are critically needed.
- (ii) Large-scale external validation: AI algorithms require validation across diverse populations, equipment, and institutional settings before clinical implementation.^{44,45}
- (iii) Standardization initiatives: Both fields require standardized protocols, quality metrics, and reporting guidelines to enable meaningful comparisons and clinical implementation.
- (iv) Real-world performance studies: Studies assessing real-world diagnostic performance, workflow integration, and cost-effectiveness are essential.

For AI studies, a minimum sample size of more than 1,000 lesions for deep learning studies, with mandatory

external validation across multiple institutions, is desirable. In addition, prospective validation before clinical implementation claims is crucial to establish consistency in elastography research and to develop a uniform measurement protocol applicable to all equipment manufacturers. This should include obligatory reporting of inter-observer reliability and prospective validation across multiple centers, incorporating quality control metrics and training.

Recent studies have investigated the combined use of AI and elastography to improve the characterization of BI-RADS 4 lesions. AI algorithms have the potential to refine the selection of elastography parameters, automate measurement procedures, and integrate elastography data with other imaging features for a more comprehensive analysis.⁴⁶ Machine learning models trained on elastography data demonstrate improved performance compared with traditional elastography interpretation, suggesting that AI can enhance the clinical utility of mechanical imaging information.⁴⁷ These hybrid approaches may provide optimal diagnostic performance while preserving the advantages of the clinical workflow (Figure 3). Recent studies have demonstrated that AI-enhanced elastographic interpretation can achieve a diagnostic performance similar to that of pure AI while maintaining the real-time assessment capabilities of conventional elastography.⁴⁸

This systematic review has several limitations that should be acknowledged:

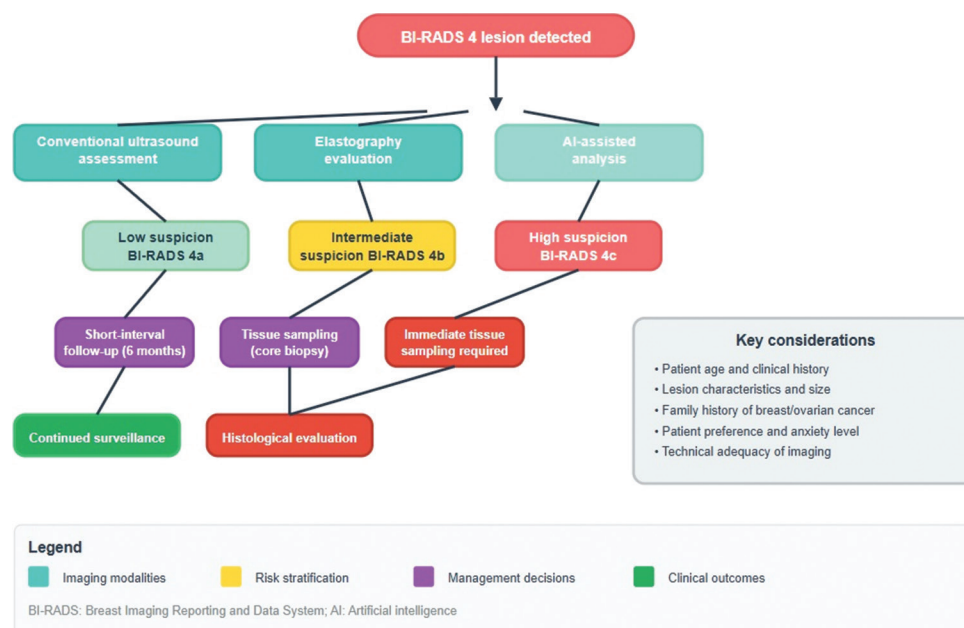


Figure 3. Proposed clinical decision-making algorithm for the management of BI-RADS 4 lesions using combined AI and elastography. Image created by the authors.

Abbreviations: AI: Artificial intelligence; BI-RADS: Breast imaging reporting and data system.

- (i) Language and publication bias: Restriction to English-language publications may have excluded relevant studies, and publication bias likely favored positive results.
- (ii) Heterogeneity preventing meta-analysis: Significant methodological heterogeneity limited the feasibility of quantitative synthesis, thereby reducing the strength of the conclusions.
- (iii) Quality assessment limitations: QUADAS-2 assessment relies on reported information, and poor reporting may have led to “unclear” risk-of-bias ratings even when studies may have employed adequate methodology.
- (iv) Rapidly evolving field: AI technology evolves rapidly, and some included studies may not represent the current state-of-the-art approaches.

Emerging technologies, such as next-generation elastography techniques, have demonstrated significant improvements over conventional strain and shear-wave techniques. Next-generation elastography approaches, including multifrequency shear-wave elastography, viscosity imaging, and nonlinear elastography, provide enhanced tissue characterization capabilities for BI-RADS 4 lesions. Multifrequency techniques analyze tissue responses across multiple frequencies, revealing frequency-dependent mechanical properties that correlate with tissue microstructure and cellular composition. Viscosity imaging, which measures energy dissipation during tissue deformation, shows promise for differentiating benign from malignant breast lesions. Recent studies have reported improved diagnostic accuracy, with sensitivity and specificity values of 98% and 95%, respectively, representing substantial improvements over conventional elastographic methods.⁴⁹ Other newer techniques quantify tissue viscosity, providing complementary information to traditional stiffness measurements, which are valuable for characterizing lesions with overlapping elastic properties.⁵⁰ Similarly, non-linear elastography exploits the non-linear stress-strain relationship of biological tissues and reveals mechanical properties that are invisible to conventional approaches.⁵¹ Preliminary studies have demonstrated enhanced discrimination of malignant axillary nodes using machine learning and radiomics.⁴²

5. Conclusion

This systematic review critically appraised 54 studies evaluating the diagnostic performance of AI and elastography for characterizing BI-RADS 4 breast lesions. The review revealed significant methodological limitations, with only 25% of studies meeting high-quality standards. Key issues included selection bias, lack of blinding, variable reference standards, and incomplete data reporting. AI studies showed promising performance metrics but were

limited by small sample sizes, lack of external validation, and potential overfitting. Elastography studies demonstrated more consistent yet modest performance and were affected by operator dependence and lack of standardization. Direct comparison between AI and elastography was hindered by the absence of head-to-head studies and significant heterogeneity in methodologies. The review suggests that the reported diagnostic performance for both technologies may be inflated due to methodological flaws and publication bias. Neither technology currently has sufficient evidence to replace conventional assessment, nor both should be regarded as supplementary tools while carefully considering their limitations. Future research priorities include high-quality prospective studies with standardized protocols, large-scale external validation, and real-world performance evaluation. Combining AI and elastography may offer benefits but requires further validation. Taken together, this review highlights that the evaluation of BI-RADS 4 lesions would greatly benefit from higher methodological standards and collaborative research approaches, ultimately improving patient outcomes through evidence-based diagnostic strategies.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Aprajita Kapur

Visualization: All authors

Writing—original draft: Aprajita Kapur

Writing—review & editing: Atul Kapoor

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

1. American College of Radiology. *Breast Imaging Reporting and Data System (BI-RADS) Atlas*. 5th ed. United States: American College of Radiology; 2013.

2. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA. *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. 5th ed. United States: American College of Radiology; 2013.
3. Rao AA, Feneis J, Lalonde C, Ojeda-Fournier H. A Pictorial Review of Changes in the BI-RADS Fifth Edition. *Radiographics*. 2016;36(3):623-639.
doi: 10.1148/rg.2016150178
4. Sigrist RMS, Liau J, Kaffas AE, Chammas MC, Willmann JK. Ultrasound Elastography: Review of Techniques and Clinical Applications. *Theranostics*. 2017;7(5):1303-1329.
doi: 10.7150/thno.18650
5. Berg WA, Blume JD, Cormack JB, *et al.* Combined screening with ultrasound and mammography versus mammography alone in women at an elevated risk of breast cancer. *JAMA*. 2008;299(18):2151-2163.
doi: 10.1001/jama.299.18.2151
6. Garra BS. Elastography: Current status, future prospects, and making work for you. *Ultrasound Q*. 2011;27(3):177-186.
doi: 10.1097/RUQ.0b013e31822a2138
7. Liberman L, Abramson AF, Squires FB, Glassman JR, Morris EA, Dershaw DD. The breast imaging reporting and data system: Positive predictive value of mammographic features and final assessment categories. *AJR Am J Roentgenol*. 1998;171(1):35-40.
doi: 10.2214/ajr.171.1.9648759
8. Rodriguez-Ruiz A, Krupinski E, Mordang JJ, *et al.* Detection of breast cancer using mammography: The effect of an artificial intelligence support system. *Radiology*. 2019;290(2):305-314.
doi: 10.1148/radiol.2018181371
9. Cosgrove D, Piscaglia F, Bamber J, *et al.* EFSUMB guidelines and recommendations for the clinical use of ultrasound elastography. Part 2: Clinical applications. *Ultraschall Med*. 2013;34(3):238-253.
doi: 10.1055/s-0033-1335375
10. Burnside ES, Ochsner JE, Fowler KJ, *et al.* Use of microcalcification descriptors in BI-RADS 4th edition to stratify risk of malignancy. *Radiology*. 2007;242(2):388-395.
doi: 10.1148/radiol.2422052130
11. Brewer NT, Salz T, Lillie SE. Systematic review: The long-term effects of false-positive mammograms. *Ann Intern Med*. 2007;146(7):502-510.
doi: 10.7326/0003-4819-146-7-200704030-00006
12. Tosteson AN, Stout NK, Fryback DG, *et al.* Cost-effectiveness of digital mammography breast cancer screening. *Ann Intern Med*. 2008;148(1):1-10.
doi: 10.7326/0003-4819-148-1-200801010-00002
13. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *Magnetic Resonance Imaging*. 2018;49(4):939-954.
doi: 10.1002/jmri.26534
14. Itoh A, Ueno E, Tohno E, *et al.* Breast disease: Clinical application of US elastography for diagnosis. *Radiology*. 2006;239(2):341-350.
doi: 10.1148/radiol.2391041676
15. Barr RG, Nakashima K, Amy D, *et al.* WFUMB guidelines and recommendations for clinical use of ultrasound elastography: Part 2: Breast. *Ultrasound Med Biol*. 2015;41(5):1148-1160.
doi: 10.1016/j.ultrasmedbio.2015.03.008
16. Yoon JH, Kim MJ, Kim EK, *et al.* Interobserver variability of ultrasound elastography: How does it affect the diagnosis of breast lesions? *AJR Am J Roentgenol*. 2011;196(3):730-736.
doi: 10.2214/AJR.10.4654
17. Berg WA, Cosgrove DO, Doré CJ, *et al.* Shear-wave elastography improves the specificity of breast US: A multicenter, retrospective study of 939 masses. *Radiology*. 2012;262(2):435-449.
doi: 10.1148/radiol.11110640
18. Sadigh G, Carlos RC, Neal CH, Dwamena BA. Ultrasonographic differentiation of malignant from benign breast lesions: A meta-analytic comparison of elasticity and BI-RADS scoring. *Breast Cancer Res Treat*. 2012;133(1):23-35.
doi: 10.1007/s10549-011-1857-8
19. Evans A, Whelehan P, Thomson K, *et al.* Invasive breast cancer: Relationship between shear-wave elastographic findings and histologic prognostic factors. *Radiology*. 2012;263(3):673-677.
doi: 10.1148/radiol.12111317
20. Han S, Kang HK, Jeong JY, *et al.* A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Phys Med Biol*. 2017;62(19):7714-7728.
doi: 10.1088/1361-6560/aa82ec
21. Fujioka T, Kubota K, Mori M, *et al.* Distinction between benign and malignant breast masses at breast US using a deep convolutional neural network. *Jpn J Radiol*. 2019;37(6):466-472.
doi: 10.1007/s11604-019-00831-5
22. McKinney SM, Sieniek M, Godbole V, *et al.* International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94.
doi: 10.1038/s41586-019-1799-6
23. Becker AS, Marcon M, Ghafoor S, *et al.* Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer.

- Invest Radiol.* 2017;52(7):434-440.
doi: 10.1097/RLI.0000000000000358
24. Geras KJ, Mann RM, Moy L. Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives. *Radiology.* 2019;293(2):246-259.
doi: 10.1148/radiol.2019182627
25. Kooi T, Litjens G, van Ginneken B, *et al.* Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal.* 2017;35:303-312.
doi: 10.1016/j.media.2016.07.007
26. Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J Med Imaging (Bellingham).* 2016;3(3):034501.
doi: 10.1117/1.JMI.3.3.034501
27. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018;18(8):500-510.
doi: 10.1038/s41568-018-0016-5
28. Liu X, Faes L, Kale AU, *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit Health.* 2019;1(6):e271-e297.
doi: 10.1016/S2589-7500(19)30123-2
29. Brady AP, Neri E. Artificial intelligence in radiology-ethical considerations. *Diagnostics (Basel).* 2020;10(4):231.
doi: 10.3390/diagnostics10040231
30. Giger ML, Chan HP, Boone J. Anniversary paper: History and status of CAD and quantitative image analysis: The role of medical physics and AAPM. *Med Phys.* 2008;35(12):5799-5820.
doi: 10.1118/1.3013555
31. Zhou LQ, Wu XL, Huang SY, *et al.* Lymph node metastasis prediction from primary breast cancer US images using deep learning. *Radiology.* 2020;294(1):19-28.
doi: 10.1148/radiol.2019190372
32. Byra M, Galperin M, Ojeda-Fournier H, *et al.* Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. *Med Phys.* 2019;46(2):746-755.
doi: 10.1002/mp.13361
33. Wang J, Yang X, Cai H, *et al.* Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Sci Rep.* 2016;6:27327.
doi: 10.1038/srep27327
34. Chartrand G, Cheng PM, Vorontsov E, *et al.* Deep learning: A primer for radiologists. *Radiographics.* 2017;37(7):2113-2131.
doi: 10.1148/rg.2017170077
35. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. *Radiographics.* 2017;37(2):505-515.
doi: 10.1148/rg.2017160130
36. Yassin NIR, Omran S, El Houbay EMF, Allam H. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Comput Methods Programs Biomed.* 2018;156:25-45.
doi: 10.1016/j.cmpb.2017.12.012
37. Tang A, Tam R, Cadrin-Chênevert A, *et al.* Canadian Association of Radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J.* 2018;69(2):120-135.
doi: 10.1016/j.carj.2018.02.002
38. Thrall JH, Li X, Li Q, *et al.* Artificial intelligence and machine learning in radiology: Opportunities, challenges, pitfalls, and criteria for success. *J Am Coll Radiol.* 2018;15(3):504-508.
doi: 10.1016/j.jacr.2017.12.026
39. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* 2018;15(11):e1002683.
doi: 10.1371/journal.pmed.1002683
40. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med Phys.* 2017;44(10):5162-5171.
doi: 10.1002/mp.12453
41. Rajpurkar P, Irvin J, Ball RL, *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* 2018;15(11):e1002686.
doi: 10.1371/journal.pmed.1002686
42. Zheng X, Yao Z, Huang Y, *et al.* Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. *Nat Commun.* 2020;11(1):1236.
doi: 10.1038/s41467-020-15027-z
43. Bai HX, Wang R, Xiong Z, *et al.* Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. *Radiology.* 2020;296(3):E156-E165.
doi: 10.1148/radiol.2020201491
44. Koh DM, Papanikolaou N, Bick U, *et al.* Artificial intelligence and machine learning in cancer imaging. *Commun Med (Lond).* 2022;2:133.
doi: 10.1038/s43856-022-00199-0
45. Pesapane F, Codari M, Sardanelli F. Artificial intelligence

- in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur Radiol Exp.* 2018;2(1):35.
doi: 10.1186/s41747-018-0061-6
46. Lehman CD, Wellman RD, Buist DS, *et al.* Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med.* 2015;175(11):1828-1837.
doi: 10.1001/jamainternmed.2015.5231
 47. Van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal.* 2022;79:102470.
doi: 10.1016/j.media.2022.102470
 48. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. *Sci Rep.* 2019;9(1):12495.
doi: 10.1038/s41598-019-48995-4
 49. Kapoor A, Kapur A, Sidhu BS, Singh J. Use of next-generation shear wave elastography for blue breast cancer evaluation. *Int Surg J.* 2025;12(8):1288-1295.
doi: 10.18203/2349-2902.isj20252278
 50. Wei Q, Yan YJ, Wu GG, *et al.* Added value of a new strain elastography technique in conventional ultrasound for the diagnosis of breast masses: A prospective multicenter study. *Front Oncol.* 2021;11:779612.
doi: 10.3389/fonc.2021.779612
 51. Gennisson JL, Deffieux T, Fink M, Tanter M. Ultrasound elastography: Principles and techniques. *Diagnostic and Interventional Imaging.* 2013;94(5):487-495.
doi: 10.1016/j.diii.2013.01.022