Proceedings of CONF-MLA 2026 Symposium: Explainable Computing, Modeling & Data Science in Complex Systems
DOI: 10.54254/2755-2721/2026.GF35159

Artificial Intelligence in Breast Cancer Imaging: From Mammography to MRI

Jiayu Yu[1], Fan Yu[2*]

[1]*Guangdong Experimental High School (Liwan Campus), Guangzhou, China*
[2]*Medical School, Shenzhen University, Shenzhen, China*
**Corresponding Author. Email: realleodyx@163.com*

Abstract. Artificial intelligence (AI) in breast imaging has progressed from retrospective algorithm testing to prospective and real-world evaluation. In full-field digital mammography (FFDM), randomized, prospective, and implementation studies show that AI-supported reading maintains or increases cancer detection while reducing radiologist workload when thresholds, arbitration rules, and audit procedures are prespecified. Evidence for digital breast tomosynthesis (DBT) is expanding, but remains weighted toward reader studies and retrospective external validation rather than population-level deployment. Mammography-derived risk models extend AI from lesion detection to longitudinal risk stratification, provided that absolute-risk calibration and subgroup monitoring are performed locally. In breast MRI, AI triage, lesion classification, ultrafast acquisition, and reconstruction methods address access and workload constraints, although safety depends on conservative operating points and interval cancer surveillance. This paper draws the conclusion with a five-layer translation framework linking technical validity, workflow fit, safety monitoring, fairness, and economic sustainability, which offers references for further research in the applications of artificial intelligence in breast cancer imaging.

Keywords: Breast cancer, mammography, digital breast tomosynthesis, MRI, artificial intelligence, deep learning, risk prediction

1. Introduction

Breast cancer remains the most commonly diagnosed cancer in women worldwide. Mammography screening reduces mortality through earlier detection, but routine interpretation is constrained by false positives, interval cancers, reader variability, and limited specialist capacity. AI systems trained on large mammography datasets now perform at or above reader level in selected test sets and have entered randomized, prospective, real-world, and interval cancer evaluations [1-7]. DBT improves lesion conspicuity but increases interpretation time, making decision support and case prioritization clinically relevant [8-10]. Image-based risk prediction reframes screening as a longitudinal pathway in which routine mammograms inform 1- to 5-year risk estimates [11-13]. Breast MRI provides high sensitivity for dense-breast and high-risk populations, yet scanner time, reading time, and biopsy burden limit broader use. AI applications in MRI focus on triage, prioritization, lesion classification, and acquisition acceleration [14-18].

DOI: 10.54254/2755-2721/2026.GF35159

This review makes three contributions. First, it reorganizes breast-imaging AI around clinical workflows rather than algorithm classes, linking each modality to endpoints that matter in deployment. Second, it separates evidence maturity across FFDM, DBT, risk prediction, and MRI, so that mature screening evidence is not generalized to less validated tasks. Third, it proposes a five-layer clinical translation framework that connects operating points, external validation, interval cancer surveillance, fairness audits and economic evaluation.

2. Literature search and evidence selection

This narrative review adopts a workflow-centered perspective rather than a systematic-review design. Searches were conducted in PubMed, Web of Science, Scopus, and Google Scholar. The search covered studies published up to May 31, 2026. Search terms included mammography, full-field digital mammography, digital breast tomosynthesis, breast MRI, artificial intelligence, deep learning, screening, detection, risk prediction, triage, reconstruction, fairness, domain shift, and economic evaluation.

Randomized trials, prospective studies, real-world implementation studies, multicenter studies, external validation cohorts, and high-quality reader studies were prioritized. Conference abstracts, non-peer-reviewed reports, studies without breast-imaging relevance, purely technical papers without clinical evaluation, and small single-center studies were excluded or down-weighted unless they addressed an underdeveloped workflow, such as DBT interpretation or MRI reconstruction (Table 1). This review does not claim exhaustive evidence capture and does not include PRISMA screening, risk-of-bias scoring, or quantitative meta-analysis.

Table 1. Evidence status, key endpoints, clinical uses and unresolved risks across breast-imaging AI workflows

Workflow	Evidence maturity	Representative studies	Key endpoints	Translation gap
FFDM screening	Strongest: randomized, prospective, interval cancer, and real-world evidence	McKinney; MASAI; ScreenTrustCAD; AI-STREAM; PRAIM	CDR, recall, PPV, workload, interval cancer	Workflow dependence, automation bias, subgroup drift
DBT interpretation	Moderate: reader studies and limited multicenter validation	Park et al.; Bae commentary; Alashban two-stage model	AUC, sensitivity, reading time, lesion-level errors	Vendor shift, dense-breast error profile, limited prospective endpoints
Risk prediction	Moderate: multi-institutional and external cohort validation	Mirai; AsymMirai; Mexican Mirai validation	Calibration, C-index, interval cancer, adjunct imaging burden	False reassurance, over-imaging, ancestry and density calibration
Breast MRI	Emerging: triage and classification evidence; reconstruction less mature	DENSE MRI triage; high-risk triage; Witowski DCE-MRI; ultrafast MRI	Missed cancer, dismissal rate, benign biopsy, reading time	Protocol drift, artifacts, enhancement kinetics, audit burden

3. Mammography, DBT, and risk-adapted screening

3.1. FFDM detection and reader workflows

International evaluation first established the feasibility of high-performing screening AI on curated mammography datasets, including absolute reductions in false positives of 5.7% in the US test set

DOI: 10.54254/2755-2721/2026.GF35159

and 1.2% in the UK test set, with corresponding false-negative reductions of 9.4% and 2.7% [1]. Subsequent clinical studies shifted the question from accuracy to deployment. In the MASAI randomized safety analysis, AI-supported screening detected 6.1 cancers per 1,000 screened participants versus 5.1 per 1000 with standard double reading, while reducing screen-reading workload by 44.3% and keeping the false-positive rate at 1.5% in both groups [2]. The 2026 MASAI interval cancer analysis then reported non-inferior interval cancer rates, higher sensitivity, and unchanged specificity, addressing the major safety endpoint that the interim analysis had not yet resolved [3].

Prospective implementation data further refine this picture. In ScreenTrustCAD, one radiologist plus AI detected 261 cancers compared with 250 cancers under two-radiologist double reading, a non-inferior relative proportion of 1.04; AI alone detected 246 cancers and also met non-inferiority [4]. In AI-STREAM, breast radiologists using AI-CAD detected 140 screening cancers versus 123 without AI-CAD, increasing CDR from 5.01 to 5.70 per 1,000 examinations without increasing recall [6]. PRAIM extends the evidence to routine German population screening: among 463,094 women, AI-supported double reading achieved a CDR of 6.7 per 1000 versus 5.7 per 1000 without AI, a 17.6% adjusted increase, while recall remained non-inferior at 37.4 versus 38.3 per 1,000 [7].

Figure 1 presents a clinical FFDM/DBT workflow in which the AI score is used for routing rather than final diagnosis. Low-suspicion examinations undergo expedited or single-reader handling, discordant examinations proceed to arbitration, and high-suspicion examinations are prioritized for diagnostic work-up. This framing clarifies that the primary intervention is not the algorithm alone but the combination of threshold selection, radiologist override, audit trail, and outcome monitoring.

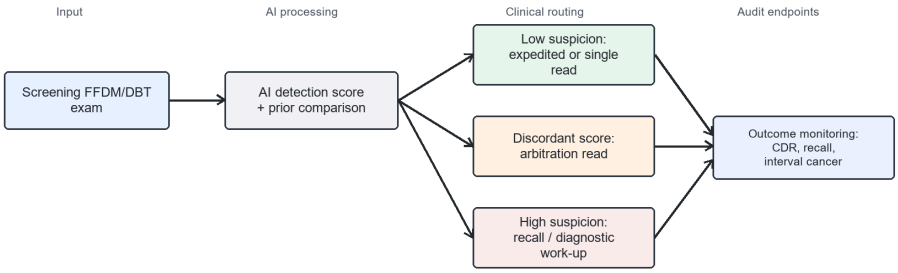

Figure 1. AI-integrated mammography screening workflow with threshold-based routing, arbitration, and outcome monitoring

Clinical interpretation. The mammography evidence base is now sufficiently mature for carefully governed implementation studies, especially in organized screening programs with double reading. The remaining uncertainty is not whether AI detects mammographic signals, but whether local workflows preserve interval cancer safety while translating workload reduction into measurable service capacity.

3.2. Digital breast tomosynthesis

DBT increases the number of image slices reviewed per examination and therefore changes the efficiency problem faced by radiologists. In a multicenter DBT reader study, the stand-alone AI system achieved AUCs of 0.93 in the validation dataset and 0.92 in the reader-study dataset. With AI-assisted interpretation, radiologists' overall AUC increased from 0.90 to 0.92, sensitivity increased from 85.4% to 87.7%, and mean reading time fell from 54.4 to 48.5 seconds [8]. Related commentary and reader-study evidence indicate gains in cancer detection and reading efficiency, but

DOI: 10.54254/2755-2721/2026.GF35159

these studies remain closer to controlled interpretation experiments than to population-screening trials [9]. Two-stage DBT detection and classification models illustrate continuing technical progress, yet their clinical role is constrained by heterogeneous datasets, scanner protocols, and limited standardized external testing [10].

Clinical interpretation. DBT AI has a weaker deployment evidence base than FFDM AI despite encouraging reader-study results. Programs adopting DBT decision support should treat reading-time reduction, lesion-level false negatives, dense-breast subgroup performance, and external scanner validation as co-primary implementation metrics rather than secondary analyses.

3.3. Mammography-based risk prediction

Mammography-based risk models estimate future cancer risk directly from routine images. Mirai was validated on 128,793 mammograms from 62,185 patients across seven sites in five countries; 3815 examinations were followed by a cancer diagnosis within 5 years, and site-level C-indices ranged from 0.75 to 0.84 [11]. AsymMirai added interpretability by emphasizing bilateral asymmetry while preserving predictive performance [12]. External validation in Mexican women produced a lower overall C-index of 0.63, supporting transfer after calibration but also demonstrating that absolute-risk estimates must be tested against local prevalence, scanner mix, and follow-up completeness [13].

Figure 2 converts risk prediction into an operational screening pathway. Routine mammograms and clinical variables enter a calibrated AI model; women are then assigned to low-, average- or elevated-risk pathways with different screening intervals or adjunct imaging. The workflow explicitly includes recalibration and outcome feedback, because risk prediction becomes clinically meaningful only when false reassurance, over-imaging, patient communication, and subgroup equity are actively managed.

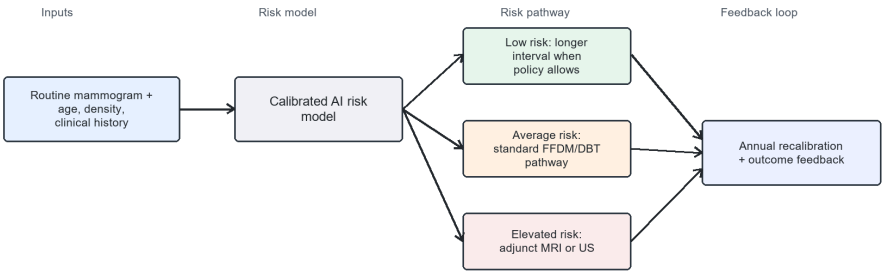

Figure 2. Risk-adapted screening workflow using mammography-derived AI risk estimates and local recalibration

Clinical interpretation. Risk prediction is conceptually distinct from detection and should not be evaluated only with diagnostic metrics. The decisive endpoints are calibration, uptake, downstream imaging burden, interval cancer, and whether risk-adapted pathways improve benefit-harm balance across age, density, and ancestry groups.

4. Breast MRI workflows

4.1. Triage and prioritization

MRI triage models trained on screening cohorts identify normal or very-low-suspicion examinations with high confidence, allowing a subset of examinations to be deprioritized or dismissed under

DOI: 10.54254/2755-2721/2026.GF35159

conservative thresholds [14, 15]. High-risk screening cohorts confirm that low-suspicion subsets exist, but also show that triage thresholds are cohort-dependent and require local validation. In one external validation cohort, an ensemble model dismissed 159 of 1441 screening MRI examinations (11%) as extremely low suspicion without missing a cancer; in the reader-study test set, 19.15% of examinations were triaged without missed cancers [16]. The clinical question is therefore threshold governance: dismissal rates, false-negative tolerance, random audit sampling, and rapid escalation pathways must be defined before deployment.

4.2. Lesion classification and biopsy stewardship

For DCE-MRI, deep-learning classifiers have approached reader-level performance and reduced benign biopsy burden in decision-analytic simulations [17]. In the internal test set of 3936 examinations, one model achieved AUROC 0.924 and AUPRC 0.720; averaging radiologist and AI predictions improved AUPRC by 0.07, and decision-curve analysis suggested that up to 20% of benign biopsies among BI-RADS 4 patients could be avoided at clinically relevant thresholds [17]. This use case differs from screening triage because it acts after an abnormality has already been identified. Evaluation should therefore report lesion-level sensitivity, histologic subtype, background parenchymal enhancement, benign biopsy reduction, and radiologist-AI disagreement patterns.

4.3. Acquisition acceleration and reconstruction

Ultrafast MRI combined with deep learning shortens acquisition and interpretation by using early dynamic information to exclude normal examinations under stringent thresholds [18]. Reconstruction and super-resolution methods operate even further upstream: they change the image that both radiologists and downstream algorithms interpret. Their validation must therefore include physics-aware quality assurance, comparison with fully sampled references, scanner-specific testing, and assessment of altered enhancement kinetics or hallucinated structures.

Figure 3 integrates triage, prioritization, and acquisition acceleration into a single MRI workflow. Referral indications lead to abbreviated or ultrafast acquisition, AI triage and classification route cases into dismissal, standard reading, or prioritized work-up, and reconstruction quality assurance runs in parallel. This design emphasizes that MRI AI should be evaluated as a service redesign rather than as an isolated classifier.

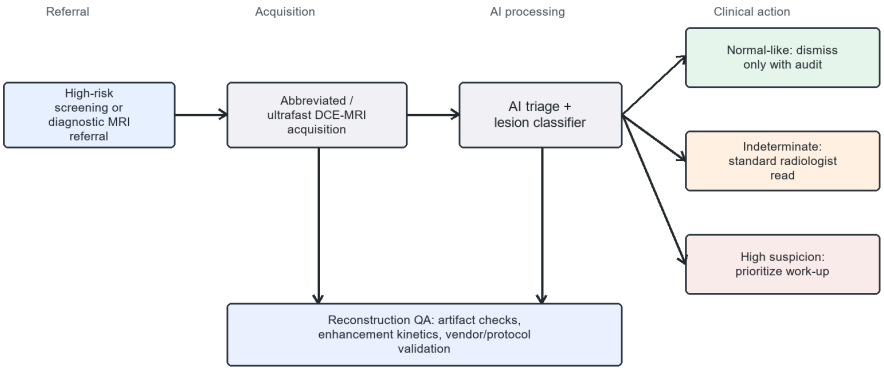

Figure 3. MRI triage, prioritization, and acceleration workflow with reconstruction quality assurance and audit endpoints

DOI: 10.54254/2755-2721/2026.GF35159

Clinical interpretation. MRI AI has substantial operational value because MRI is highly sensitive but resource-intensive. However, the safety margin is narrower than in workload-only applications; missed enhancing lesions, protocol variation, and reconstruction artifacts require stricter monitoring than aggregate AUC reporting provides.

5. Five-layer clinical translation framework

A five-layer framework clarifies the steps between algorithm performance and clinical value (Figure 4). Layer 1 is technical validity: the model must show reproducible discrimination, calibration, external test performance, and failure-mode analysis on held-out data. Layer 2 is workflow fit: thresholds, reader roles, arbitration rules, reporting interfaces, and override mechanisms must be specified before deployment. Layer 3 is safety monitoring: CDR, recall, PPV, benign biopsy, missed cancer, interval cancer, and time-to-work-up must be monitored longitudinally. Layer 4 is generalizability and fairness: external validation must cover site, vendor, density, age, indication, race or ethnicity where available, and subgroup drift. Layer 5 is economic and regulatory sustainability: implementation should include staffing effects, cost assumptions, liability boundaries, post-market monitoring, model-update control, and rollback criteria.

Screening AI should be judged by population-level endpoints, diagnostic AI by lesion-level consequences, risk models by calibration and longitudinal outcomes, and reconstruction models by image fidelity and downstream diagnostic integrity. This endpoint alignment explains why the same AI score does not have the same meaning across modalities. In FFDM, the score changes reader allocation and arbitration; in DBT, it changes search efficiency across volumetric stacks; in risk prediction, it changes screening interval or adjunct imaging; and in MRI, it changes acquisition, triage, prioritization, and reconstruction quality assurance.

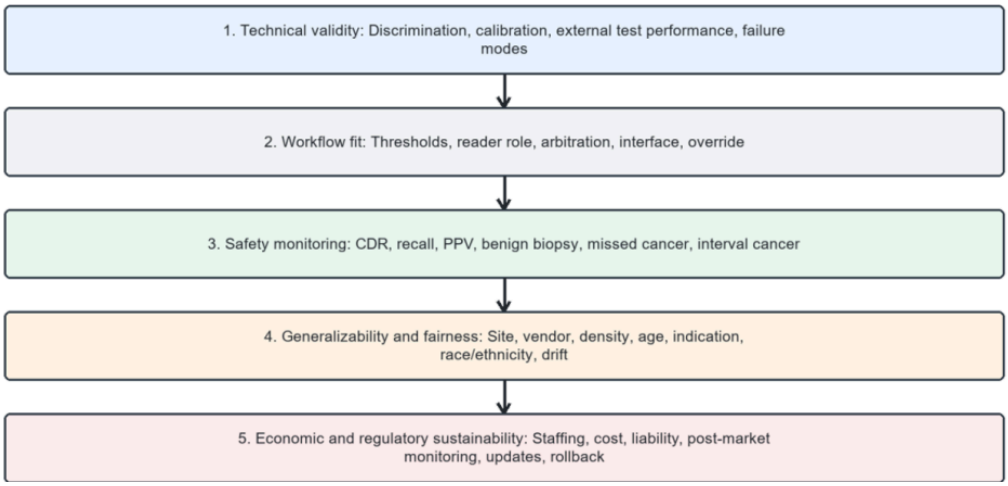

Figure 4. Five-layer clinical translation framework for breast-imaging AI

6. Cross-cutting safeguards: validation, domain shift, fairness, and economics

Foundation and self-supervised models improve sample efficiency and transfer, but their clinical use requires documented training provenance, objective design, and calibration after fine-tuning. Multimodal fusion with clinical variables, pathology, genomics, or treatment data increases task relevance, especially for treatment response prediction, but also increases missing-data and

DOI: 10.54254/2755-2721/2026.GF35159

confounding risk. Uncertainty methods, including temperature scaling, isotonic regression, and conformal prediction, are useful only when connected to explicit escalation rules.

External validation remains the central safeguard against inflated performance. Domain shift arises from scanner vendors, acquisition parameters, compression, positioning, population prevalence, and radiologist practice patterns. Evidence from broader medical imaging suggests that AI models may encode sensitive attributes and acquisition-related shortcuts that are not obvious to human readers [19-21]. Fairness evaluation therefore requires prespecified subgroup reporting by age, density, race or ethnicity where available, site, vendor, and clinical indication. Economic evaluation must measure more than reading time; it should include recall work-up, arbitration staffing, biopsy burden, scanner throughput, software cost, medico-legal governance, and the opportunity cost of false reassurance. A 2025 economic model using mammography challenge data estimated that delegation to high-performing AI reduced expected cost per mammogram by 30.1% under low algorithm and litigation costs, but only 17.5% when both costs were high [22].

Clinical interpretation. The field has moved beyond proof of concept, but implementation science has not kept pace with model development. A clinically credible AI study should report the operating point, external validation cohort, subgroup performance, interval cancer monitoring plan, and economic assumptions as prominently as sensitivity and AUC.

7. Conclusion

AI is becoming a clinically testable component of breast-imaging workflows rather than a stand-alone diagnostic novelty. The strongest evidence supports the governed use in mammography screening, where prospective and real-world studies show maintained or improved cancer detection with reduced workload. DBT and MRI applications are technically compelling but require stronger external validation and endpoint reporting before broad deployment. Risk prediction offers a route toward personalized screening, provided that calibration, patient communication, and interval cancer safeguards are built into the pathway. Across modalities, clinical benefit depends on explicit thresholds, radiologist override, subgroup audits, economic evaluation, and continuous monitoring. Breast-imaging programs that treat AI as workflow infrastructure rather than as a replacement reader are best positioned to convert algorithmic performance into earlier detection, fewer unnecessary procedures, and more efficient use of scarce expertise.

Several limitations define the next phase of breast-imaging AI. First, interval cancer remains underreported or immature in many prospective studies, yet it is the most important safety endpoint for screening triage and risk-adapted intervals. Second, external validation is uneven: FFDM has the strongest prospective and real-world evidence, whereas DBT, ultrasound-adjacent pathways, MRI reconstruction, and some risk models still rely heavily on retrospective or single-region studies. Third, domain shift from vendor, protocol, prevalence, and workflow variation threatens transportability. Fourth, fairness and shortcut learning remain material risks because models infer protected attributes and acquisition artifacts from images. Fifth, economic evaluation is insufficiently granular; workload reduction does not automatically increase throughput unless scheduling, arbitration, technologist workflows, and follow-up capacity are redesigned together.

Future studies should use prospective, multicenter designs with preregistered operating points and primary outcomes that include interval cancers, recall rate, cancer detection rate, benign biopsy rate, reading time, and patient-centered consequences. Shared reference datasets spanning vendors and countries would support comparative benchmarking. Risk-prediction research should combine imaging-derived risk with longitudinal clinical and genetic variables only after demonstrating calibration and equity across subgroups. MRI research should pair acceleration and reconstruction

DOI: 10.54254/2755-2721/2026.GF35159

methods with lesion-level safety analysis. Regulatory oversight should evolve toward lifecycle monitoring, including post-market drift detection, model-update change control, rollback criteria, and public reporting of safety-threshold breaches.

References

[1] McKinney S.M., Sieniek M., Godbole V., et al. (2020) International evaluation of an AI system for breast cancer screening. Nature, 577(7788): 89-94.

[2] Lång K., Josefsson V., Larsson A.M., et al. (2023) Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomized, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol, 24(7): 936-944.

[3] Gommers J., Hernström V., Josefsson V., et al. (2026) Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomized, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. Lancet, 407(10527): 505-514.

[4] Dembrower K., Crippa A., Colón E., Eklund M., Strand F. (2023) ScreenTrustCAD Trial Consortium. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health, 5(10): e703-e711.

[5] Ng A.Y., Oberije C.J.G., Ambrózay É., et al. (2023) Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. Nat Med, 29(12): 3044-3049.

[6] Chang Y.W., 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.

[7] Eisemann N., Bunk S., Mukama T., et al. (2025) Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. Nat Med., 31(3): 917-924.

[8] Park E.K., Kwak S.Y., Lee W., Choi J.S., Kooi T., Kim E.K. (2024) Impact of AI for digital breast tomosynthesis on breast cancer detection and interpretation time. Radiol Artif Intell., 6(3): e230318.

[9] Bae M.S. (2024) AI improves cancer detection and reading time of digital breast tomosynthesis. Radiol Artif Intell., 6(3): e240219.

[10] Alashban Y. (2025) Breast cancer detection and classification with digital breast tomosynthesis: a two-stage deep learning approach. Diagn Interv Radiol., 31(3): 206-214.

[11] Yala A., Mikhael P.G., Strand F., et al. (2022) Multi-institutional validation of a mammography-based breast cancer risk model. J Clin Oncol., 40(16): 1732-1740.

[12] Donnelly J., Moffett L., Barnett A.J., et al. (2024) AsymMirai: interpretable mammography-based deep learning model for 1-5-year breast cancer risk prediction. Radiology, 310(3): e232780.

[13] Avendaño D., Marino M.A., Bosques-Palomo B.A., et al. (2024) Validation of the Mirai model for predicting breast cancer risk in Mexican women. Insights Imaging, 15(1): 244.

[14] Verburg E., van Gils C.H., van der Velden B.H.M., et al. (2022) Deep learning for automated triaging of 4581 breast MRI examinations from the DENSE trial. Radiology, 302(1): 29-36.

[15] Verburg E., van Gils C.H., van der Velden B.H.M., et al. (2023) Validation of combined deep learning triaging and computer-aided diagnosis in 2901 breast MRI examinations from the second screening round of the DENSE trial. Invest Radiol, 58(4): 293-298.

[16] Bhowmik A., Monga N., Belen K., et al. (2023) Automated triage of screening breast MRI examinations in high-risk women using an ensemble deep learning model. Invest Radiol, 58(10): 710-719.

[17] Witowski J., Heacock L., Reig B., et al. (2022) Improving breast cancer diagnostics with deep learning for MRI. Sci Transl Med., 14(664): eabo4802.

[18] Jing X., Wielema M., Cornelissen L.J., et al. (2022) Using deep learning to safely exclude lesions with only ultrafast breast MRI to shorten acquisition and reading time. Eur Radiol., 32(12): 8706-8715.

[19] Gichoya J.W., Banerjee I., Bhimireddy A.R., et al. (2022) AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health, 4(6): e406-e414. doi: 10.1016/S2589-7500(22)00063-2.

[20] Brown A., et al. (2023) Detecting shortcut learning for fair medical AI using shortcut testing. Nat Commun., 14: 4314. doi: 10.1038/s41467-023-39902-7.

[21] Lotter W. (2024) Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. Nat Commun., 15: 7465.

DOI: 10.54254/2755-2721/2026.GF35159

[22] Ahsen M.E., Ayvaci M.U.S., Mookerjee R., et al. (2025) Economics of AI and human task sharing for decision making in screening mammography. Nat Commun., 16: 2289.