AACR Annual Meeting. April 5-10, 2024 | San Diego, California | Abstract #3449

Is AI-enhanced breast ultrasound ready for breast cancer screening in low-resource environments? A systematic review

Arianna Bunnell[1,2], Dustin Valdez[1,2], Fredrik Strand[3], Yannik Glaser[2], Peter Sadowski[2], John Shepherd[1]

University of Hawai'i Cancer Center, HI, USA[1], University of Hawai'i at Mānoa, HI, USA[2], Karolinska Institutet, Stockholm, Sweden[3]

UNIVERSITY OF HAWAI'I CANCER CENTER

ARTIFICIAL INTELLIGENCE PRECISION HEALTH INSTITUTE
UNIVERSITY OF HAWAI'I

Purpose

A systematic review was performed to investigate whether artificial intelligence (AI) algorithms, using breast ultrasound (BUS) imaging to perform a variety of tasks, are sufficiently accurate for use in early detection programs in low-resource areas. The BCSC defines the acceptable ranges for sensitivity and specificity for screening mammography to be >75% and 88-95%, respectively [1, 2]. No a priori benchmarks exist for frame selection or lesion segmentation.

Background

High-income countries have implemented population-wide breast cancer screening programs using mammography and seen a significant reduction in mortality in screened women. However, many low- and middle-income countries, and rural areas, lack the access, workforce, and/or infrastructure necessary for implementing such programs. Integration of AI may reduce the false-positive rate of handheld BUS, reduce the training needed to perform exams, and make early detection programs viable where none exists presently.

Methods

Two reviewers independently assessed articles from PubMed and Google Scholar (1/1/2016 to 8/6/2023) and performed QUADAS-2 [3] bias rating. Studies developing AI for BUS for diagnosis of breast cancer which report performance on unseen women met the inclusion criteria. Studies were evaluated on dataset size and composition, task-specific AI performance (AUC, AP, DSC), and clinical application.

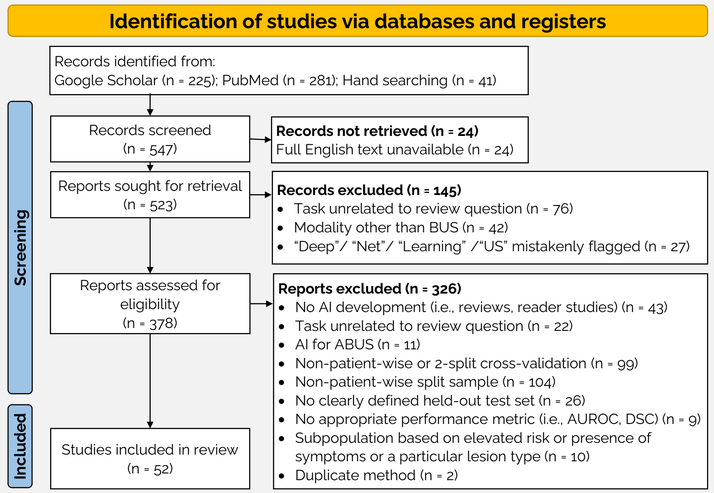
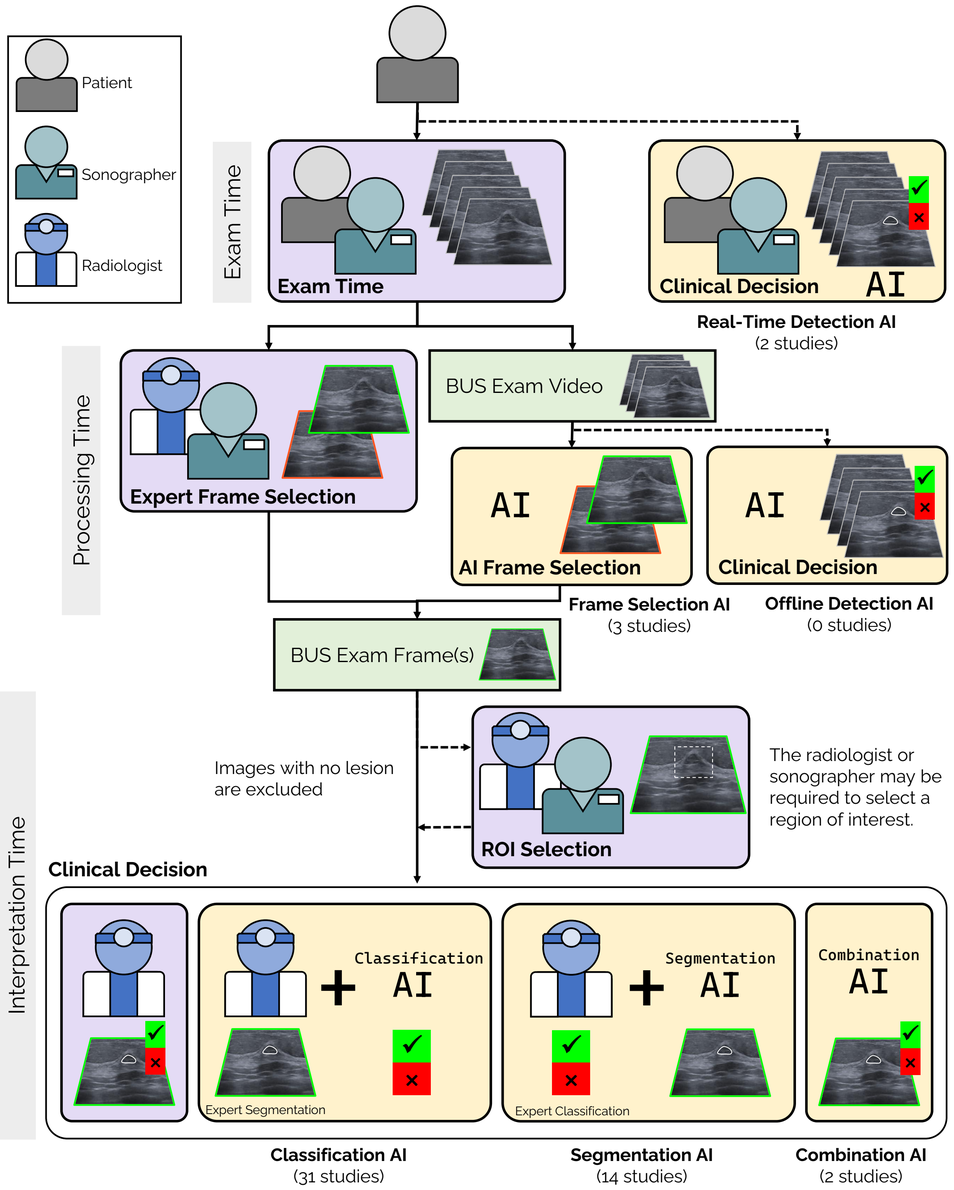


Figure 1: PRISMA 2020 [4] flow diagram for this review. 547 records were screened based on title/abstract for inclusion into the review. 378 total full texts were reviewed against the inclusion/exclusion criteria.



Methods

Figure 1 shows a PRISMA [4] flowchart of the inclusion process. Studies were evaluated via narrative data synthesis. The strength of evidence for clinical AI performance is characterized by the quality, size, and integrity of the testing set.

Results

- **Figure 2** shows the number of studies identified at each clinical application time.
- Real-time detection models localize and classify lesions ≥75% of the time.
- Frame selection models report diagnostic AUC ≥0.85 from AI-selected frames.
- Performance on complete and normal exams and outside of the development population are unknown for real-time detection and frame selection AI models.
- Classification-only models report generally high performance (see **Figure 3**), mitigated by enriched prevalence and small datasets.
- 14 total classification-only studies (45%) meet the BCSC guidelines. The most well-validated (Shen 2021) reports AUC 0.976.
- Segmentation-only models report DSC generally >0.8 (64%) but tend to validate on public datasets with poor metadata.

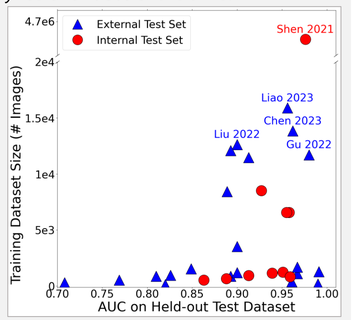


Figure 3: Scatter plot showing performance of classification-only models against the size of the training dataset by number of BUS images.

Conclusion

The classification- and segmentation-only AI task categories comprised 90% of the sample. Classification-only models showed AUC values rivaling mammography AI [5]. However, strength of reported evidence was lacking across all tasks due to limited testing dataset size and diversity. Validation of models on larger, geographically distinct datasets containing normal and benign imaging, comprehensive reporting of patient and image metadata, as well as whole breast exams, are needed to support the broad adaptation of AI-informed BUS for screening and early detection programs.

References

[1] Lehman CD, Arao RF, Sprague BL, et al. Radiology 2017; 283(1): 49-58. doi: 10.1148/radiol.2016161174
[2] Rosenberg RD, Yankaskas BC, Abraham LA, et al. Radiology 2006; 241(1): 55-66. doi: 10.1148/radiol.2411051504
[3] Whiting, Penny F., et al. Annals of internal medicine 155.8 (2011):. doi: 10.7326/0003-4819-155-8-201110180-0000
[4] Page, Matthew J., et al. BMJ 372 (2021). doi: 10.1136/bmj.n71
[5] Yoon J. H., et. al. Radiology 2023; 307.5; p.e222639. doi: 10.1148/radiol.222639
(Liao 2023) Liao J, Gui Y, Li Z, et al. eClinicalMedicine 2023; **60**: 102001. doi: 10.1016/j.eclinm.2023.102001.
(Shen 2021) Shen Y, Shamout FE, Oliver JR, et al. Nature Communications 2021; 12(1). doi: 10.1038/s41467-021-26023-2
(Chen 2023) Chen J, Jiang Y, Huang Z, et al. doi: 10.2139/ssrn.4409592
(Gu 2022) Gu Y, Xu W, Liu T, et al.. Eur Radiol 33, 2954–2964 (2023). doi: 10.1007/s00330-022-09263-8.
(Liu 2022) Liu T et al.. MICCAI 2022. Lecture Notes in Computer Science, vol 13433. doi: 10.1007/978-3-031-16437-8_45

Figure 2: Diagram showing the different opportunities in the care paradigm where AI can be applied. AI systems are classified according to both where in the care paradigm they fall (exam time, processing time, and interpretation time) as well as the AI task. Note that frame selection may occur in conjunction with the exam, moving this step from processing time to exam time. If an AI system both classifies and segments breast lesions during interpretation time, it is placed in a separate, combination category.