Title: Artificial Intelligence-Informed Handheld Breast Ultrasound for Screening: A Systematic Review of Diagnostic Test Accuracy

Authors:

Arianna Bunnell, MS.

Affiliation: Information and Computer Sciences, University of Hawai'i at Mānoa, Honolulu, HI, USA. University of Hawai'i Cancer Center, Honolulu, HI, USA.

Dustin Valdez, MS.

Affiliation: University of Hawai'i Cancer Center, Honolulu, HI, USA

Fredrik Strand, MD.

Affiliation: Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden. Breast Radiology Unit; Medical Diagnostics Karolinska, Karolinska University Hospital, Stockholm, Sweden.

Yannik Glaser, MS.

Affiliation: Information and Computer Sciences, University of Hawai'i at Mānoa, Honolulu, HI, USA

Peter J Sadowski, PhD.

Affiliation: Information and Computer Sciences, University of Hawai'i at Mānoa, Honolulu, HI, USA.

John A Shepherd, PhD.

Affiliations: University of Hawai'i Cancer Center, Honolulu, HI, USA

ABSTRACT

Background

Breast cancer screening programs using mammography have led to significant mortality reduction in high-income countries. However, many low- and middle-income countries lack resources for mammographic screening. Handheld breast ultrasound (BUS) is a low-cost alternative but requires substantial training. Artificial intelligence (AI) enabled BUS may aid in both the detection (perception) and classification (interpretation) of breast cancer, enabling screening use in low-resource contexts.

Purpose.

To investigate whether AI-enhanced BUS is sufficiently accurate to serve as the primary modality in breast cancer screening, particularly in resource-limited environments.

Materials and Methods.

This review (CRD42023493053) is reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) and SWiM (Synthesis Without Meta-analysis) guidelines. PubMed and Google Scholar were searched from January 1, 2016 to December 12, 2023. A meta-analysis was not attempted. Studies are grouped according to their AI task type, application time, and AI task. Study quality is assessed using the QUality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool.

Results.

Of 763 candidate studies, 314 total full texts were reviewed. 34 studies are included. The AI tasks of included studies are as follows: 1 frame selection, 6 detection, 11 segmentation, and 16 classification. In total, 5.7 million BUS images from over 185,000 patients were used for AI training or validation. A single study included a prospective testing set. 79% of studies were at high or unclear risk of bias. Exemplary classification and segmentation AI systems perform with 0.976 AUROC and 0.838 DSC.

Conclusion.

There has been encouraging development of AI for BUS. Despite studies demonstrating high performance across all identified tasks, the evidence supporting AI-enhanced BUS generally lacks robustness. High-quality model validation on geographically external, screening datasets with complete metadata will be key to realizing the potential for AI-enhanced BUS in increasing access to screening in resource-limited environments.

INTRODUCTION

Breast cancer has become the most prevalent cancer in the world with the WHO estimating 2.3 million women diagnosed in 2020 (1, 2). High-income countries have implemented population-wide screening programs using mammography and witnessed an estimated 20% reduction in mortality in women invited for screening since the 1980s (3). Further, regular screening with mammography is widely recommended by professional societies (4-8). However, implementing mammographic screening is resource-intensive. Thus, many low- and middle-income countries have not been able to implement population-wide mammographic screening programs. Handheld breast ultrasound (BUS) is an alternative to mammography that requires less equipment cost, support infrastructure, and training. However, cancer screening with BUS been found to have substantially higher false-positive rates; one exemplary study found a rate of 74/1,000 biopsies per screening exam with BUS alone compared to 8/1,000 with mammography alone (9). AI-assisted BUS may reduce the false-positive and unnecessary biopsy rate. BUS is a highly noisy, complex imaging modality which requires significant training for both image interpretation and performing exams. Importantly, AI-assisted BUS has the potential to alleviate the need for highly trained staff, a radiologist or sonographer, to perform the examination, increasing accessibility in low-resource medical contexts. The tipping point of broad acceptance of AI-enabled BUS has yet to occur.

For a lesion with malignancy-suspicion to be detected, the radiologist must first notice an abnormality in the ultrasound image, a perceptual task, and then assess the probability that this lesion may be cancer, an interpretative task. Therefore, in this systematic review, we ask two questions: **Question 1 - Perception:** How accurate are AI-informed BUS models for frame selection, lesion detection, and segmentation when incorporated into the screening care paradigm? **Question 2 - Interpretation:** How accurate are AI-informed BUS models for cancer classification when incorporated into the screening care paradigm? Questions 1 and 2 are separated due to differences in performance evaluation of task types. Question 2 is concerned only with accuracy in diagnosis of lesions as benign or malignant, while Question 1 evaluates accuracy in lesion location, either alone (perception AI) or in addition to accuracy in diagnosis (perception and interpretation AI). To answer these questions, we evaluate the current literature for potential for bias in the selected studies and attribute the literature to each task-specific question to examine performance.

METHODS

The abstract and full text of this systematic review are reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (see **Supplement**) (10). A protocol for this review was registered as PROSPERO CRD42023493053. We followed much of the methods of Freeman et al.'s review of AI-informed mammography (11). Data extraction templates and results can be requested from the corresponding author.

Data source, eligibility criteria, and search strategy

Data sources, searching, and screening

The search was conducted on PubMed (12) and Google Scholar (13) using the Publish or Perish software (Harzing, version 8) . Only papers published since 2016 in English were considered and our search was updated on December 12, 2023. The search encompassed three themes: breast

cancer, AI, and ultrasound. Exact search strings can be found in the **Supplement**. Evidence on systematic review methodologies suggests the exclusion of non-English studies is unlikely to have affected results (14, 15).

Inclusion and exclusion criteria

We included studies which reported on the performance of AI for the detection, diagnosis, or localization of breast cancer from BUS, on an unseen group of patients. Studies must additionally validate on exams from all task-relevant BI-RADS categories. Furthermore, included studies must report a performance metric which balances sensitivity and specificity. Lastly, studies must work *automatically* from BUS images, with no human-defined features. However, selection of a region of interest (ROI) is acceptable. Studies are additionally excluded if they include/exclude patients based on symptom presence or risk; include procedural imaging; are designed for ancillary tasks (i.e., NAC response); or are opinion pieces, reviews, or meta-analyses.

Data collection and analysis

Data extraction

A single reviewer (A.B.) extracted data, subject to review by a second reviewer (D.V.) with differences resolved through discussion. The following characteristics were extracted from included articles: author(s); journal and year of publication; country of study; reference standard definition; index test definition; characteristics and count of images/videos/patients; inclusion/exclusion criteria; reader study details (if applicable); AI model source (commercial or academic); and AI and/or reader performance.

Data synthesis

Data synthesis is reported in accordance with the Synthesis Without Meta-analysis (SWiM) reporting guideline (see **Supplement**) (16). The synthesis groupings were informed by the clinical paradigm. No meta-analysis was planned for this study as the AI tasks are heterogeneous and not well-suited for intercomparison. We utilize descriptive statistics, tables, and narrative methods. Certainty of evidence is evaluated using the following: number of studies, data split quality (if applicable), and data diversity. Heterogeneity of studies is assessed through comparison of reference standard definitions and dataset characteristics.

Studies were grouped for synthesis by clinical application time, AI task, and AI aid type (perception or interpretation). The clinical application time groups were exam time (AI is applied during BUS examination), processing time (exam recording), and reading time (pre-selected exam frames). The AI task groups and types were frame selection (perception), lesion detection (perception and interpretation), cancer classification (interpretation), and lesion delineation (perception). We can define sub-groups based on the intersections of application time and task. For example, lesion detection AI applied during exam and processing time can be referred to as real-time and offline detection AI, respectively.

The outcome of interest for this review is AI performance. Lesion detection AI is evaluated by average precision (AP) or mean average precision (mAP). Frame selection is evaluated by AUROC in frame selection and/or diagnosis from selected frames. Cancer classification is evaluated by AUROC or sensitivity/specificity. Lesion delineation is evaluated by Dice Similarity Coefficient (DSC) or intersection over union (IOU). No metric conversions were attempted.

Study quality

Study quality was independently assessed by two reviewers (A.B. & D.V.) using the quality assessment of diagnostic accuracy studies-2 (QUADAS-2) tool (17) (see **Supplement**) using criteria adapted from (11). The reviewers resolved differences through discussion. Bias criteria are rated yes, no, unclear, or not applicable. Applicability criteria are rated high, low, or unclear. Studies are classified according to their majority category. If categories are tied, the study is rated as the highest of the tied categories.

Additionally, studies are evaluated based on completeness of reporting on the racial/ethnic, age, breast density, background echotexture, and BMI diversity of their participants, as well as BUS machine types. Age-adjusted breast density, race/ethnicity, and BMI are known risk factors for breast cancer (18-21). BUS machine model reporting is examined to evaluate AI generalizability.
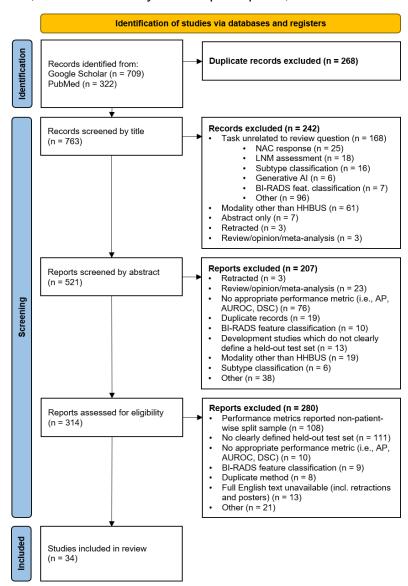
Changes from protocol

The addition of AUROC in diagnosis as an evaluation metric for frame selection AI was done in response to the observation that frames identified for human examination may not be most useful for downstream AI. AUROC and sensitivity/specificity were added as acceptable evaluation metrics for lesion detection AI in response to the literature. Data cleaning method was not extracted, as it was not well-defined for validation studies. Analysis by AI type was not planned but was added to emphasize clinical utility.

RESULTS

Study selection

PubMed and Google Scholar yielded 322 and 709 studies, respectively. After removing

Figure 1. PRISMA 2020 flow diagram. HHBUS = handheld breast ultrasound; NAC = neoadjuvant chemotherapy; LNM = lymph node metastasis.

Table 1. AI Segmentation (top) and Frame Selection (bottom) Systems. Table S3 provides a complete accounting of public datasets (i.e., UDIAT). CV = cross-validation. Unknown values (not reported in study) are indicated with a "?" symbol.

Study	Population	Reference Standard	Index Test	Performance
Byra 2020 (45)	882 images of 882 lesions (? malignancy) from ? patients from a single clinical site. **External Testing:** BUSI UDIAT and OASBUD.	Delineations from a single "medical expert."	Adapted U-Net with additionally dilated convolution blocks replacing traditional convolution blocks.	0.701 mean DSC when not finetuned on external test sets.
Chen 2023 (46)	BUSI and UDIAT (42% malignancy) **External Testing:** STU-Hospital (? malignancy).		Adapted U-Net with attention modules with varying sizes replacing traditional convolution blocks.	0.802 DSC on external test set.
Han 2020 (47)	2,800 images from 2,800 patients (50% malignancy) from a single hospital in China. **External Testing:** UDIAT	Delineations from physicians at department of US.	GAN-based architecture with attention and segmentation mask generator and discriminator networks.	0.78 DSC on external test set
Huang 2022b (48)	2,020 images from ? patients (50.2% malignancy) from UDIAT and a single hospital in China.	Delineations from "experienced radiologist."	Combination CNN and graph convolutional architecture for mask and specific boundary-rendering, respectively.	0.919 DSC on 5-fold CV.
Ning 2022 (49)	UDIAT, BUSI, and ultrasoundcases. **External Testing:** onlinemedicalimages and OASBUD.		Custom U-Net with background/foreground information streams and shape-, edge-, and position-aware fusion units.	0.872 DSC on external test set.
Qu 2020 (50)	980 images from 980 women (60.7% malignancy) from a single university hospital in China and UDIAT.	Delineations from "experts."	Custom ResNet with varying-scale attention modules and upsampling.	0.905 DSC on five-fold CV.
Wang 2021 (51)	3,279 images from 1,154 patients (57.2% malignancy) (ultrasoundcases & BUSI). **External Testing:** BUSI & radiopaedia.		Custom U-Net with ResNet34 encoder and residual feedback.	0.82 DSC on external test set.
Webb 2021 (52)	31,070 images from 851 women (? malignancy) from a single clinic in the USA	Delineations from 3 "experts" (testing) and "research technologists" with US experience (development).	Custom DenseNet264 with added feature pyramid network and ResNet-C input stream pretrained on thyroid US images.	0.832 DSC on internal test set.
Zhang 2023 (53)	1,342 images from ? patients from 5 hospitals in China. **External Testing:** 570 images from ? patients from a single hospital in China & BUSI & onlinemedicalimages.	Delineations from "experienced radiologists."	Combination U-Net and DenseNet backbone from pre-selected ROI.	0.89 mean IOU on external test set
Zhao 2022 (54)	9,836 images from 4,875 patients from ? hospitals in China.	Delineations are from 3 "experienced radiologists."	Custom U-Net architecture with local and de-noising attention.	0.838 DSC on internal test set
Zhuang 2019 (55)	857 images from ? patients from a single hospital in the Netherlands (ultrasoundcases). **External Testing:** STU-Hospital & UDIAT.		Custom attention-based residual U-Net.	0.834 DSC on external test set
Frame Selection				
Huang 2022a (30)	2,606 videos from 653 patients (26.7% malignancy) from 8 hospitals in China	**Keyframe/Location:** Frame and bounding box from "experienced sonographers" **Classification:** Histological results from biopsy or surgery	Reinforcement learning scheme with 3D convolutional BiLSTM with frame-based reward structure based on lesion presence, proximity to labelled frame, and malignancy indicators.	0.793 diagnostic AUROC on internal test set from selected frames

full texts were evaluated. 34 studies are included (**Figure 1**).

Characteristics of included studies

The 34 included studies examined 30 AI models: 3 commercial (21% of studies), 25 academic (74%), and 2 later commercialized (6%). (22) preceded S-Detect for Breast (Samsung Medison Co., Seongnam, Korea) and (23) preceded CADAI-B (BeamWorks Inc., Daegu, Korea). Included studies analyzed a total of 5.7 million BUS images and 3,566 videos from over 185,000 patients. 5.44 million (95%) images and 143,203 patients are contributed by a single article (24). A majority (59%) of studies were conducted in the East Asia region (20 studies; 12 in China). 5 studies used only public datasets (see **Supplement**).

AI Tasks

There were 6 lesion detection studies (23, 25-29), 1 frame selection study (30), 16 classification studies (12 AI models) (22, 24, 31-44), and 11 segmentation studies (45-55). 18 studies use *perception* AI (23, 25-30, 45-55) and 22 studies use *interpretation* AI (22-29, 31-44), with 6 studies (23, 25-29) using AI for both.

Perception

Frame selection (1 study)

Frame selection AI models identify exam frames for downstream examination. See **Table 1** (bottom) for a summary. Huang 2022a (30) develop a reinforcement learning model, rewarded by optimizing identifying frames likely to contain lesions, annotations, and malignancies. Their model increased diagnostic performance of senior and junior readers by 0.03 and 0.01 AUROC, respectively.

Lesion segmentation (11 studies)

Lesion segmentation AI models delineate lesions. See **Table 1** for a summary. Six (55%) and nine (82%) studies train and test on at least partially public data. The most common approach was extending the U-Net (56) architecture (seven studies, 64%). Reported DSC ranges from 0.701 (45) to 0.872 (49) on test datasets ranging from 42 (46) to 1,910 (54) images. The remaining studies develop convolutional (50, 52), graph convolutional (48), and adversarial networks (47). Han 2020 (47) report 0.78 DSC on an external test dataset. Huang 2022b (48) and Qu 2020 (50) report 0.919 and 0.905 DSC on five-fold cross-validation. Webb 2021 (52) report 0.832 DSC on an internal test set of 121 images (85 patients).

Interpretation

Cancer classification (16 studies)

Cancer classification AI models classify lesions/images as either benign or cancerous. See **Table 2** for a summary. Operator involvement required prior to AI use varied: six studies (38%) require ROI selection, three studies require seed point placement (19%), three studies (19%) require image hand-cropping, three studies (19%) apply automatic cropping/segmentation, and one study (6%) is unclear. Choi 2019 (33), Lee 2022 (39), and Park 2019 (41) test S-Detect for Breast (Samsung Medison Co., Seongnam, Korea). Choi 2019 and Lee 2022 find standalone AI to perform with 85% and 86.2% sensitivity and 95.4% and 85.1% specificity, respectively. Park 2019

find AI assistance to increase reader sensitivity by 10.7% and specificity by 8.2%. Han 2017 finetune GoogLeNet (57) and report 0.96 AUROC on an internal dataset. Berg 2021 (31),

Table 2. Lesion classification AI system summary table. Table S3 provides a complete accounting of public datasets (i.e., UDIAT, BUSI, OASBUD). Unknown values (not reported in study) are indicated with a "?" symbol.

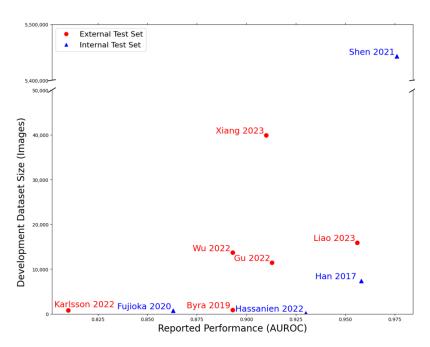| Study	Population	Reference Standard	Index Test	Performance
Berg 2021 (31)	**External Testing:** 638 images of 319 lesions (27.5% malignancy) from ? women from a single health center in the US	Histological results from biopsy with benign follow-up of at least 2 years	Koios DS from pre-selected ROI	0.77 AUROC of AI alone on external test set
Byra 2019 (32)	882 images from 882 patients (23.1% malignancy) from a single health center in California. **External Testing:** UDIAT & OASBUD	Histological results from biopsy with benign follow-up of at least 2 years	SVM from finetuned VGG19 pretrained on ImageNet from pre-selected ROI	0.893 AUROC on external test set
Choi 2019 (33)	**External Testing:** 759 images of 253 lesions from 226 patients (31.6% malignancy) from a single medical center in South Korea	Histological results from biopsy with benign follow-up of ?	S-Detect for Breast	85.0% sensitivity and 95.4% specificity for AI alone
Fujioka 2020 (34)	702 images from 217 patients (48.9% malignancy in testing) from a single health center in Japan	Histological results from biopsy with benign follow-up of at least 1 year	Bidirectional GAN from hand-cropped images	0.863 AUROC on internal test set
Gu 2022 (35)	11,478 images from 4,149 patients (42.7% malignancy) from 30 tertiary-care hospitals in China **External Testing:** 1,291 images from 397 patients (62.1% malignancy) from 2 tertiary-care hospitals in China & BUSI	Histological results from biopsy or surgery	Finetuned VGG19 backbone pretrained on ImageNet from pre-selected ROI	0.913 AUROC on external test set
Guldogan 2023 (36)	**External Testing:** 1,430 orthogonal images of 715 lesions (18.8% malignancy) from 530 women	Histological results from biopsy with benign follow-up of at least 2 years	Koios DS from pre-selected ROI	98.5% sensitivity and 65.4% specificity for AI alone
Han 2017 (22)	7,408 images from 5,151 patients (42.6% malignancy) from a single health center in South Korea	Histological results from biopsy	Finetuned GoogLeNet pretrained on grayscale ImageNet from semi-automatic segmentation	0.958 AUROC on internal test set
Hassanien 2022 (37)	UDIAT		Finetuned SwinTransformer from hand-cropped images	0.93 AUROC on internal test set
Karlsson 2022 (38)	BUSI **External Testing:** 293 images from ? women (90.1% malignancy) from a single university hospital in Sweden		Finetuned ResNet50V2 from hand- and automatically-cropped images	0.81 AUROC on external test set
Lee 2022 (39)	**External Testing:** 492 lesions from 472 women (40.7% malignancy) from a single health center in South Korea	Histological results from biopsy with benign follow-up of at least 2 years	S-Detect for Breast	0.855 AUROC on external test set
Liao 2023 (40)	15,910 images from 6,795 patients (2.56% malignancy) from a single hospital in China **External Testing 1:** 896 images from 391 patients (2.23% malignancy) from a single hospital in China **External Testing 2:** 490 images from 235 patients (2.04% malignancy) from a single hospital in China	Histological results from biopsy with benign follow-up of at least 3 years	80 Dual-branch ResNet50 learners for B-mode and Doppler ensembled into parent model	0.956 AUROC on external test set

Study	Population	Reference Standard	Index Test	Performance
Park 2019 (41)	**External Testing:** 100 video clips of lesions from 91 women (41% malignant) from a single hospital in South Korea	Histological results from biopsy or surgery	S-Detect for Breast	+0.105 difference in AUROC with/without AI for readers on external test set
Shen 2021 (24)	5,442,907 images from 143,203 patients (1.1% malignancy) from >100 hospitals in New York **External Testing**: BUSI	Histological results from biopsy with benign follow-up of at most 15 months (test set); Pathology report (training set)	Deep convolutional network with spatial and scan-wise attention and saliency map concatenation from entire input image set per breast	0.976 AUROC on internal test set
Wanderley 2023 (42)	**External Testing:** 555 lesions from 509 women (40% malignancy) from a single health center in Brazil	Histological results from biopsy	Koios DS from pre-selected ROI	98.2% sensitivity and 39.0% specificity of CAD alone on external test set
Wu 2022 (43)	13,684 images from 3,447 patients (28.7% malignancy) from a single hospital in China **External Testing:** 440 images from 228 patients (54.3% malignancy) from a single hospital in China	Histological results from biopsy or surgery	Finetuned MobileNet from hand-cropped images.	0.893 AUROC on external test set
Xiang 2023 (44)	39,899 images of 8,051 lesions from 7,218 patients (64.1% malignancy) from a single university hospital in China **External Testing 1:** 2,637 images of 777 lesions from 693 patients (47.6% malignancy) from a single hospital in China **External Testing 2:** 957 images of 419 lesions from 382 patients (48.9% malignancy) from a single hospital in China **External Testing 3:** 2,416 images of 648 lesions from 504 patients (25.3% malignancy) from a single hospital in China	Histological results from biopsy or surgery	Custom finetuned DenseNet121 with self-attention averaged over all views of a lesion	0.91 AUROC on external test set

Guldogan 2023 (36), and Wanderley 2023 (42) all validate Koios DS (Koios Medical, Inc., Chicago IL) through reader studies. Berg 2021 find standalone AI performs with 0.77 AUROC. Guldogan 2023 and Wanderley 2023 evaluate binned predictions and find AI alone performs with 98.5% and 98.2% sensitivity and 65.4% and 39% specificity, respectively. The nine remaining studies develop AI models. Reported AUROC values range from 0.81 (38) to 0.98 (24) on test datasets ranging from 33 (37) to 25,000 (24) patients. The most common approach was to finetune and optionally extend an existing architecture from ImageNet (58) weights. Otherwise, studies used generative adversarial networks (34) and custom convolutional architectures (24). All studies except Liao 2023 (40) explicitly work on unenhanced (B-mode) BUS images. **Figure 2** displays reported performance vs. development dataset size. Only two studies developed on datasets with over 20,000 images, performing with 0.91 (44) and 0.976 (24) AUROC.

Perception and interpretation

Lesion detection (6 studies)

Lesion detection AI models perform both lesion detection and cancer classification. See **Table 3** for a summary. Lesion localization precision varied: a single study provides heatmap-style visualizations (23), three studies provide bounding boxes (26-28), and two studies provide delineations (25, 29). Qiu 2023 (29), Meng 2023 (28), and Fujioka 2023 (26) all extend the YOLO family (59) and achieve 0.87 AUROC (no location performance measure) on 278 videos, 0.78 mAP on 647 images, and an increase in per-case sensitivity and specificity of 11.7% and 20.9% (reader study) on 230

Figure 2. Scatter plot showing reported performance (AUROC) for lesion classification models against the size of the development dataset by number of breast ultrasound images. Studies are additionally identified by whether reported performance is on an internal or external testing set.

videos, respectively. Kim 2021b (23) extend the GoogLeNet (57) architecture to achieve 0.9 AUROC and 99% correct localization on an external dataset of 200 images. Lai 2022 (27) evaluate standalone BU-CAD (TaiHao Medical Inc., Taipei City, Taiwan) on 344 images, resulting in a location-adjusted AUROC of 0.84. Bunnell 2023 (25) develop an extension to the Mask RCNN (60) architecture and achieve mAP 0.39 on an internal test dataset of 447 images.

Clinical application time

We define an example care paradigm inclusive of low-resource, teleradiology-exclusive medical scenarios (**Figure 3**). The clinical application time of studies included 5 exam, 2 processing, and 27 reading time studies.

Study quality assessment

Figure 4 displays bias assessment results. 18 (53%) and 9 (27%) studies have high or unclear risk of bias overall. All studies but one are of high applicability concern. Concerns about applicability for Qiu 2023 (29) are attributed to an unclear location reference standard. Generally, studies are at an unclear risk of bias and high applicability concern for patient selection due to incomplete reporting of the participant selection process. All included studies except Liao 2023 (40) and Shen 2021 (24) are of high index test applicability concern due to making image-level predictions only.

Table 3. Lesion detection AI system summary table. Table S3 provides a complete accounting of public datasets (i.e., UDIAT, BUSI, OASBUD). Unknown values (not reported in study) are indicated with a "?" symbol.

Study	Population	Reference Standard	Index Test	Performance
Bunnell 2023 (25)	37,921 images from 2,148 women (24.2% malignancy) from ? clinical sites in the US.	**Location:** Delineations from a single radiologist. **Classification:** Histological results from biopsy with no record of cancer for benign.	Finetuned Mask-RCNN with ResNet-101 backbone and custom heads for BI-RADS mass feature prediction.	0.39 mAP on internal test set.
Fujioka 2023 (26)	88 videos from 45 women (? malignancy) from a single breast surgery department in Japan. **Internal Testing:** 232 videos (40.5% malignancy) from 232 women from a single breast surgery department in Japan.	**Location**: 2 experts (>10 years of BUS experience). A 3rd expert then performed adjudication. **Classification:** Unclear.	Finetuned YOLOv3-tiny combined with edge detection post-processing of regions to isolate lesions.	95.5% sensitivity and 2.2% specificity for AI alone.
Kim 2021b (23)	1,400 images from 971 patients (50% malignancy) from a single university hospital in South Korea. **External Testing:** 200 images from 125 patients (50% malignancy) from a single university hospital in South Korea.	**Location:** Delineations from a single radiologist. **Classification:** Histological results from biopsy with benign follow-up of at least 2 years.	GoogLeNet from hand-cropped images with saliency maps for localization.	0.9 AUROC on external test set.
Lai 2022 (27)	**External Testing:** 344 images from 172 women (37.8% malignancy) from a single hospital in Taiwan.	**Location:** From "expert panel" of 5 radiologists. **Classification:** Histological results from biopsy with benign follow-up of at least 2 years.	BU-CAD (TaiHao Medical Inc., Taipei Taiwan)	0.838 AULROC on external test set.
Meng 2023 (28)	7,040 images from 3,759 women (60.7% malignancy) from ? hospitals in China. **External Testing:** BUSI	**Location:** Delineations from "experienced radiologists." **Classification:** Histological results from biopsy.	Adapted YOLOv3 with added bilateral spatial and global channel attention modules.	0.782 mAP on external test set.
Qiu 2023 (29)	480 video clips (18,122 images) of 480 lesions from 420 women (40.8% malignancy) from a single hospital in China **Prospective Testing:** 292 video clips of 292 lesions from 278 women (42.5% malignancy) from 2 hospitals in China	**Location:** Delineations from 2 "experienced radiologists." **Classification:** Histological results from biopsy.	Finetuned YOLOv5 network with attention	0.87 AUROC on prospective testing set

See **Figure 5** for a complete breakdown of diversity reporting. 35% of included studies failed to report diversity along any axis. The most reported diversity axes were participant age (15 studies) and machine type (18 studies). Classification studies were the most complete, with 11 (69%) reporting along at least one axis.

DISCUSSION

Main Findings

In this systematic review, we evaluated the accuracy of BUS AI models for each identified task. We identified 6 studies performing lesion detection, 1 frame selection study, 16 cancer classification studies, and 11 lesion segmentation studies. 12 studies aid in perceptual tasks, 16 studies aid in interpretative tasks, and 6 studies aid in both. We also examine clinical application time in the screening care paradigm: 5 studies were designed for exam time, 2 for processing time, and 27 for reading time.

Overall, the current state-of-the-art in AI-informed BUS for frame selection, lesion detection, and lesion segmentation (perception) does not yet provide evidence that it performs sufficiently well
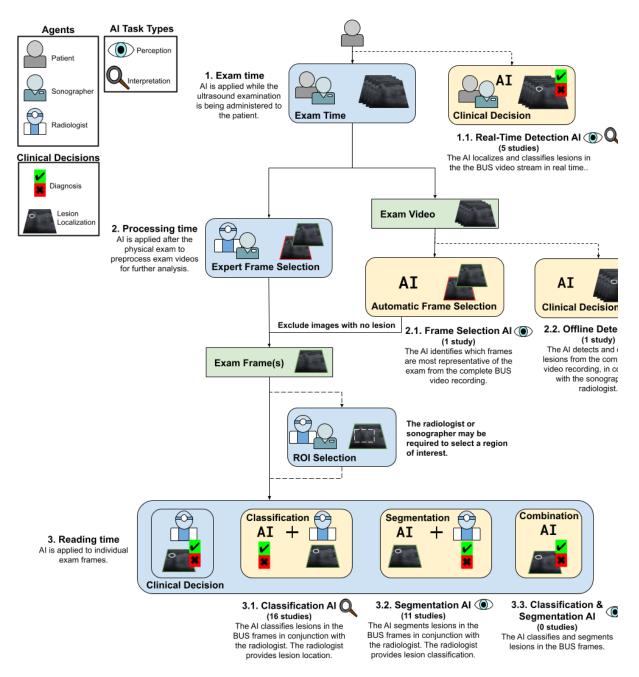
Figure 3. Diagram showing the different opportunities in the care paradigm where AI can be applied.

for integration into breast cancer screening where BUS is the primarily modality, particularly when not supervised at all stages by a radiologist (Question 1). Zhao 2022 provide the highest-quality perceptual evidence, reporting 0.838 DSC on an internal test dataset of 1,910 images. The included studies report high performance, but lack sufficient validation and population reporting and commonly validate on datasets unrepresentative of screening (<3% cancer prevalence). Validation of models on larger datasets containing more normal/benign imaging, as well as unaltered BUS video, would improve evidence supporting these models.

Many more high-quality studies develop cancer classification AI, forming a more robust picture of interpretation AI performance (Question 2). We refer to Shen 2021 (24), Xiang 2023 (44), and
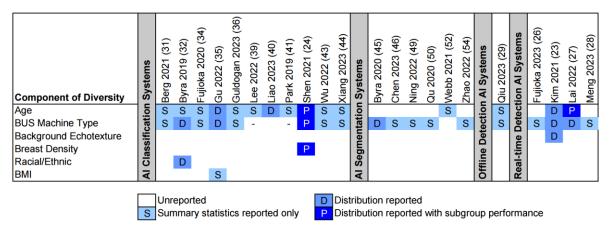
Figure 4. QUADAS-2 bias assessment results.

Study	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
AI Classification Systems							
Berg 2021 (31)	Low	High	Low	High	High	High	Low
Byra 2019 (32)	Unclear	High	Low	High	High	High	Low
Choi 2019 (33)	High	Low	High	High	High	High	High
Fujioka 2020 (34)	High	Unclear	High	High	High	High	High
Gu 2022a (35)	Unclear	Low	High	High	High	High	High
Guldogan 2023 (36)	Low	Low	High	High	High	High	High
Han 2017 (22)	Unclear	Low	High	High	High	High	High
Hassanien 2022 (37)	Unclear	Low	High	High	High	High	Unclear
Karlsson 2022 (38)	Unclear	Low	High	Low	High	High	High
Lee 2022 (39)	Low	Low	Low	High	High	High	Low
Liao 2023 (40)	Low	High	Low	Low	High	High	Low
Park 2019 (41)	Unclear	Low	High	Low	High	High	High
Shen 2021 (24)	Unclear	Low	High	High	High	Low	High
Wanderley 2023 (42)	Unclear	High	High	Low	High	High	High
Wu 2022 (43)	Unclear	High	High	High	High	High	High
Xiang 2023 (44)	Unclear	High	High	Low	High	High	High
AI Segmentation Systems							
Byra 2020 (45)	Unclear	High	High	Unclear	High	High	High
Chen 2023 (46)	Low	Low	High	Unclear	High	High	High
Han 2020b (47)	Unclear	Low	High	Unclear	High	High	High
Huang 2022a (48)	High	Unclear	High	Unclear	High	High	High
Ning 2022 (49)	Unclear	Low	High	Unclear	High	High	High
Qu 2020 (50)	Unclear	Low	High	Unclear	High	High	Unclear
Wang 2021 (51)	Unclear	Low	High	Unclear	High	High	High
Webb 2021 (52)	High	High	High	Unclear	High	High	High
Zhang 2023 (53)	Unclear	Low	High	Unclear	High	High	High
Zhao 2022 (54)	Unclear	Low	Low	Unclear	High	High	Low
Zhuang 2019 (55)	Unclear	Low	High	Unclear	High	High	High
Real-Time Detection AI							
Bunnell 2023 (25)	High	Low	High	High	High	High	High
Fujioka 2023 (26)	Unclear	High	High	High	High	Low	High
Kim 2021 (23)	Unclear	Low	High	High	High	High	High
Lai 2022 (27)	Unclear	Low	Unclear	High	High	High	High
Meng 2023 (28)	Unclear	Low	High	Low	High	High	High
Offline Detection AI							
Qiu 2023 (29)	Unclear	Low	High	Low	High	Low	Unclear
Frame Selection AI							
Huang 2022b (30)	Unclear	Low	H / U	Unclear	High	Low	H / U

Figure 4. QUADAS-2 bias assessment results. Figure is best viewed in color. Reference standard assessments for frame selection studies are reported classification first, frame selection second. H = high; U = unclear.

Liao 2023 (40) as the best examples, showing performances of 0.976, 0.91, and 0.956 AUROC (respectively) on large datasets. We suggest that validation of BUS cancer classification AI on a common dataset with comprehensive patient metadata and containing more normal/benign imaging may facilitate easier comparison between methods, allowing for a more complete picture of the state of the field on subgroups of interest.

Comparison with other studies.

Although others have reviewed AI-informed BUS (61-72), we contribute the first *systematic* review not limited to a single BUS modality, as in (67), and contribute the only QUADAS-2 bias assessment of AI for BUS. (11) serves as a close analog to this work, examining test accuracy in mammography AI. However, (11) excludes all studies which evaluate performance on split sample datasets. This strict validation criteria improves the evidence supporting model performance in new patient populations and represents the highest level of dataset split quality.

Component of Diversity	Berg 2021 (31)	Byra 2019 (32)	Fujioka 2020 (34)	Gu 2022 (35)	Guldogan 2023 (36)	Lee 2022 (39)	Liao 2023 (40)	Park 2019 (41)	Shen 2021 (24)	Wu 2022 (43)	Xiang 2023 (44)	Byra 2020 (45)	Chen 2023 (46)	Ning 2022 (49)	Qu 2020 (50)	Webb 2021 (52)	Zhao 2022 (54)	Qiu 2023 (29)	Fujioka 2023 (26)	Kim 2021 (23)	Lai 2022 (27)	Meng 2023 (28)
Age	S	S	S	D	S	S	D	S	P	S	S					S		S		D	P	
BUS Machine Type	S	D	S	D	S	-		-	P	S	S	D	S	S	S		S	S	S	D	D	S
Background Echotexture																				D		
Breast Density									P													
Racial/Ethnic		D																				
BMI				S																		

Column groups (left to right): AI Classification Systems (Berg 2021 – Xiang 2023), AI Segmentation Systems (Byra 2020 – Zhao 2022), Offline Detection AI Systems (Qiu 2023), Real-time Detection AI Systems (Fujioka 2023 – Meng 2023).

Legend: □ Unreported • S Summary statistics reported only • D Distribution reported • P Distribution reported with subgroup performance

Figure 5. Heatmap showing axes of reported diversity for included studies. Figure is best viewed in color. Studies which fail to report along any of the included axes are omitted from the plot. Studies which only use one kind of ultrasound machine and report on an additional axis are indicated with a – on the above plots.

We remove this restriction due to the relatively early stage of the field of BUS AI development as compared to mammography AI. For example, the FDA approved the first mammography CAD system in 1998 (73), whereas the first BUS CAD system wasn't approved until 2016 (74). In initial stages, more AI models may be developed and validated within a single institution.

Strengths and Limitations

We followed conventional methodology for systematic reviews and applied strict inclusion criteria to ensure the reliability and quality of the included studies. Studies using internal validation on the image-, video-, or lesion-level, or no held-out testing set are at risk of reporting inflated performance and do not provide good evidence of model generalizability. By upholding strict standards for model validation, we attempt to provide a clear picture of AI performance. However, we did not apply exclusion criteria based on dataset size, thus our review is limited in inclusion of studies with small testing sets, which provide poor evidence of generalizability. Lastly, we are limited in that we consider the application of QUADAS-2 guidelines in the manner of (11), but do not evaluate with a bias framework specific for medical AI studies, such as QUADAS-2 for AI (75) or STARD-AI (76), both of which are yet to be published. CONSORT-AI (77) and DECIDE-AI (78) were not applicable as included studies are not clinical trials or evaluated online. This review is limited in that there may be unidentified AI tasks which exist within the screening paradigm. One example of this may be AI designed to verify coverage of the entire breast during BUS scanning.

Conclusions and Recommendations

We conclude that high accuracy can be obtained in both perception and interpretation BUS AI. However, researchers developing AI-informed BUS systems should concentrate their efforts on providing explicit, high-quality model validation on geographically external test sets, representative of screening, with complete metadata. Studies should emphasize the entire clinical workflow. For example, real-time detection methods for low-resource settings must have performance reported on a dataset of complete BUS exam frames from a geographically external set of participants, imaged by non-experts, rather than on curated or randomly-selected frames. Considering the potential for AI-enhanced BUS to improve access to breast cancer screening in low- and middle-income countries in particular, the absence of a radiologist or experienced breast

sonographer to additionally examine all imaging limits the safeguards we can assume are in place in the clinic, adding to the urgency of more complete, high-quality performance and metadata reporting for BUS AI across the clinical paradigm.

REFERENCES

1.   Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021;71(3):209-49.
2.   Organization WH. Global breast cancer initiative implementation framework: assessing, strengthening and scaling-up of services for the early detection and management of breast cancer: World Health Organization; 2023.
3.   Marmot MG, Altman D, Cameron D, Dewar J, Thompson S, Wilcox M. The benefits and harms of breast cancer screening: an independent review. British journal of cancer. 2013;108(11):2205-40.
4.   Mainiero MB, Lourenco A, Mahoney MC, Newell MS, Bailey L, Barke LD, D'Orsi C, Harvey JA, Hayes MK, Huynh PT. ACR appropriateness criteria breast cancer screening. Journal of the American College of Radiology. 2016;13(11):R45-R9.
5.   Schunemann HJ, Lerda D, Quinn C, Follmann M, Alonso-Coello P, Rossi PG, Broeders MJ, Grawingholt A, Saz-Parkinson Z. Breast Cancer Screening and Diagnosis: A Synopsis of the European Breast Guidelines. Annals of internal medicine. 2020;172(1):46-+. doi: 10.7326/M19-2125.
6.   Organization WH. WHO position paper on mammography screening: World Health Organization; 2014.
7.   European breast cancer guidelines - Screening ages and frequencies. European Commission Initiatives on Breast and Colorectal Cancer, 2023.
8.   Breast screening (mammogram): National Health Service; 2021 [September 25, 2023]. Available from: https://www.nhs.uk/conditions/breast-screening-mammogram/when-youll-be-invited-and-who-should-go/.
9.   Berg WA. Combined Screening With Ultrasound and Mammography vs Mammography Alone in Women at Elevated Risk of Breast Cancer. JAMA. 2008;299(18):2151. doi: 10.1001/jama.299.18.2151.
10.   Page MJ, Mckenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, Mcdonald S, Mcguinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Journal of Clinical Epidemiology. 2021;134:178-89. doi: 10.1016/j.jclinepi.2021.03.001.
11.   Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, Taylor-Phillips S. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ. 2021:n1872. doi: 10.1136/bmj.n1872.
12.   PubMed Central (PMC) [Internet] [Internet]. National Library of Medicine (US) National Center for Biotechnology Information.   [cited June 19, 2023]. Available from: https://www.ncbi.nlm.nih.gov/pmc/.
13.   Google Scholar [Internet]. Google LLC.   [cited June 19, 2023]. Available from: https://scholar.google.com/.
14.   Morrison A, Polisena J, Husereau D, Moulton K, Clark M, Fiander M, Mierzwinski-Urban M, Clifford T, Hutton B, Rabb D. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. International journal of technology assessment in health care. 2012;28(2):138-44.

15.   Nussbaumer-Streit B, Klerings I, Dobrescu A, Persad E, Stevens A, Garritty C, Kamel C, Affengruber L, King V, Gartlehner G. Excluding non-English publications from evidence-syntheses did not change conclusions: a meta-epidemiological study. Journal of clinical epidemiology. 2020;118:42-54.

16.   Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, Hartmann-Boyce J, Ryan R, Shepperd S, Thomas J. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. bmj. 2020;368.

17.   Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM, Group* Q-. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36.

18.   Kerlikowske K, Chen S, Golmakani MK, Sprague BL, Tice JA, Tosteson ANA, Rauscher GH, Henderson LM, Buist DSM, Lee JM, Gard CC, Miglioretti DL. Cumulative Advanced Breast Cancer Risk Prediction Model Developed in a Screening Mammography Population. JNCI : Journal of the National Cancer Institute. 2022;114(5):676-85. doi: 10.1093/jnci/djac008.

19.   Liu K, Zhang W, Dai Z, Wang M, Tian T, Liu X, Kang H, Guan H, Zhang S, Dai Z. Association between body mass index and breast cancer risk: Evidence based on a dose–response meta-analysis. Cancer management and research. 2018:143-51.

20.   Maskarinec G, Meng L, Ursin G. Ethnic differences in mammographic densities. International journal of epidemiology. 2001;30(5):959-65. doi: 10.1093/ije/30.5.959.

21.   Maskarinec G, Sen C, Koga K, Conroy SM. Ethnic Differences in Breast Cancer Survival: Status and Determinants. Women's Health. 2011;7(6):677-87. doi: 10.2217/whe.11.67.

22.   Han S, Kang H-K, Jeong J-Y, Park M-H, Kim W, Bang W-C, Seong Y-K. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Physics in Medicine & Biology. 2017;62(19):7714-28. doi: 10.1088/1361-6560/aa82ec.

23.   Kim J, Kim HJ, Kim C, Lee JH, Kim KW, Park YM, Kim HW, Ki SY, Kim YM, Kim WH. Weakly-supervised deep learning for ultrasound diagnosis of breast cancer. Scientific Reports. 2021;11(1). doi: 10.1038/s41598-021-03806-7.

24.   Shen Y, Shamout FE, Oliver JR, Witowski J, Kannan K, Park J, Wu N, Huddleston C, Wolfson S, Millet A, Ehrenpreis R, Awal D, Tyma C, Samreen N, Gao Y, Chhor C, Gandhi S, Lee C, Kumari-Subaiya S, Leonard C, Mohammed R, Moczulski C, Altabet J, Babb J, Lewin A, Reig B, Moy L, Heacock L, Geras KJ. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nature Communications. 2021;12(1). doi: 10.1038/s41467-021-26023-2.

25.   Bunnell A. EARLY BREAST CANCER DIAGNOSIS VIA BREAST ULTRASOUND AND DEEP LEARNING2023.

26.   Fujioka T, Kubota K, Hsu JF, Chang RF, Sawada T, Ide Y, Taruno K, Hankyo M, Kurita T, Nakamura S. Examining the effectiveness of a deep learning-based computer-aided breast cancer detection system for breast ultrasound. Journal of Medical Ultrasonics. 2023:1-10.

27.   Lai Y-C, Chen H-H, Hsu J-F, Hong Y-J, Chiu T-T, Chiou H-J. Evaluation of physician performance using a concurrent-read artificial intelligence system to support breast ultrasound interpretation. The Breast. 2022;65:124-35. doi: 10.1016/j.breast.2022.07.009.

28.   Meng H, Liu X, Niu J, Wang Y, Liao J, Li Q, Chen C. DGANet: A Dual Global Attention Neural Network for Breast Lesion Detection in Ultrasound Images. Ultrasound in medicine & biology. 2023;49(1):31-44. doi: 10.1016/j.ultrasmedbio.2022.07.006.

29.   Qiu S, Zhuang S, Li B, Wang J, Zhuang Z. Prospective assessment of breast lesions AI classification model based on ultrasound dynamic videos and ACR BI-RADS characteristics. Frontiers in Oncology. 2023;13.

30.   Huang R, Ying Q, Lin Z, Zheng Z, Tan L, Tang G, Zhang Q, Luo M, Yi X, Liu P, Pan W, Wu J, Luo B, Ni D. Extracting keyframes of breast ultrasound video using deep reinforcement learning. Medical image analysis. 2022;80:102490-. doi: 10.1016/j.media.2022.102490.

31.   Berg WA, Gur D, Bandos AI, Nair B, Gizienski T-A, Tyma CS, Abrams G, Davis KM, Mehta AS, Rathfon G, Waheed UX, Hakim CM. Impact of Original and Artificially Improved Artificial Intelligence–based Computer-aided Diagnosis on Breast US Interpretation. Journal of breast imaging (Online). 2021;3(3):301-11. doi: 10.1093/jbi/wbab013.

32.   Byra M, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Medical Physics. 2019;46(2):746-55. doi: 10.1002/mp.13361.

33.   Choi JS, Han B-K, Ko ES, Bae JM, Ko EY, Song SH, Kwon M-R, Shin JH, Hahn SY. Effect of a Deep Learning Framework-Based Computer-Aided Diagnosis System on the Diagnostic Performance of Radiologists in Differentiating between Malignant and Benign Masses on Breast Ultrasonography. Korean Journal of Radiology. 2019;20(5):749. doi: 10.3348/kjr.2018.0530.

34.   Fujioka T, Kubota K, Mori M, Kikuchi Y, Katsuta L, Kimura M, Yamaga E, Adachi M, Oda G, Nakagawa T, Kitazume Y, Tateishi U. Efficient Anomaly Detection with Generative Adversarial Network for Breast Ultrasound Imaging. Diagnostics. 2020;10(7):456. doi: 10.3390/diagnostics10070456.

35.   Gu Y, Xu W, Lin B, An X, Tian J, Ran H, Ren W, Chang C, Yuan J, Kang C, Deng Y, Wang H, Luo B, Guo S, Zhou Q, Xue E, Zhan W, Zhou Q, Li J, Zhou P, Chen M, Gu Y, Chen W, Zhang Y, Li J, Cong L, Zhu L, Wang H, Jiang Y. Deep learning based on ultrasound images assists breast lesion diagnosis in China: a multicenter diagnostic study. Insights into Imaging. 2022;13(1). doi: 10.1186/s13244-022-01259-8.

36.   Guldogan N, Taskin F, Icten GE, Yilmaz E, Turk EB, Erdemli S, Parlakkilic UT, Turkoglu O, Aribal E. Artificial Intelligence in BI-RADS Categorization of Breast Lesions on Ultrasound: Can We Omit Excessive Follow-ups and Biopsies? Academic Radiology. 2023.

37.   Hassanien MA, Kumar Singh V, Puig D, Abdel-Nasser M. Transformer-Based Radiomics for Predicting Breast Tumor Malignancy Score in Ultrasonography. Artificial Intelligence Research and Development: IOS Press; 2022. p. 298-307.

38.   Karlsson J, Ramkull J, Arvidsson I, Heyden A, Åström K, Overgaard NC, Lång K, editors. Machine learning algorithm for classification of breast ultrasound images. Medical Imaging 2022: Computer-Aided Diagnosis; 2022: SPIE.

39.   Lee SE, Han K, Youk JH, Lee JE, Hwang J-Y, Rho M, Yoon J, Kim E-K, Yoon JH. Differing benefits of artificial intelligence-based computer-aided diagnosis for breast US according to workflow and experience level. Ultrasonography. 2022;41(4):718-27. doi: 10.14366/usg.22014.

40.   Liao J, Gui Y, Li Z, Deng Z, Han X, Tian H, Cai L, Liu X, Tang C, Liu J, Wei Y, Hu L, Niu F, Liu J, Yang X, Li S, Cui X, Wu X, Chen Q, Wan A, Jiang J, Zhang Y, Luo X, Wang P, Cai Z, Chen L. Artificial intelligence-assisted ultrasound image analysis to discriminate early breast cancer in Chinese population: a retrospective, multicentre, cohort study. eClinicalMedicine. 2023;60:102001. doi: 10.1016/j.eclinm.2023.102001.

41.   Park HJ, Kim SM, La Yun B, Jang M, Kim B, Jang JY, Lee JY, Lee SH. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: added value for the inexperienced breast radiologist. Medicine. 2019;98(3).

42.   Wanderley MC, Soares CMA, Morais MMM, Cruz RM, Lima IRM, Chojniak R, Bitencourt AGV. Application of artificial intelligence in predicting malignancy risk in breast masses on ultrasound. Radiologia Brasileira. 2023;56:229-34.

43.   Wu H, Ye X, Jiang Y, Tian H, Yang K, Cui C, Shi S, Liu Y, Huang S, Chen J, Xu J, Dong F. A Comparative Study of Multiple Deep Learning Models Based on Multi-Input Resolution for Breast Ultrasound Images. Frontiers in oncology. 2022;12:869421-. doi: 10.3389/fonc.2022.869421.

44.   Xiang H, Wang X, Xu M, Zhang Y, Zeng S, Li C, Liu L, Deng T, Tang G, Yan C. Deep learning-assisted diagnosis of breast lesions on us images: A multivendor, multicenter study. Radiology: Artificial Intelligence. 2023;5(5):e220185.

45.   Byra M, Jarosik P, Szubert A, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M. Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. Biomedical signal processing and control. 2020;61:102027. doi: 10.1016/j.bspc.2020.102027.

46.   Chen G, Li L, Dai Y, Zhang J, Yap MH. AAU-net: An Adaptive Attention U-net for Breast Lesions Segmentation in Ultrasound Images. IEEE Transactions on Medical Imaging. 2022:1-. doi: 10.1109/tmi.2022.3226268.

47.   Han L, Huang Y, Dou H, Wang S, Ahamad S, Luo H, Liu Q, Fan J, Zhang J. Semi-supervised segmentation of lesion from breast ultrasound images with attentional generative adversarial network. Computer methods and programs in biomedicine. 2020;189:105275-. Epub 20191212. doi: 10.1016/j.cmpb.2019.105275. PubMed PMID: 31978805.

48.   Huang R, Lin M, Dou H, Lin Z, Ying Q, Jia X, Xu W, Mei Z, Yang X, Dong Y, Zhou J, Ni D. Boundary-rendering network for breast lesion segmentation in ultrasound images. Medical image analysis. 2022;80:102478-. doi: 10.1016/j.media.2022.102478.

49.   Ning Z, Zhong S, Feng Q, Chen W, Zhang Y. SMU-Net: Saliency-Guided Morphology-Aware U-Net for Breast Lesion Segmentation in Ultrasound Image. IEEE transactions on medical imaging. 2022;41(2):476-90. doi: 10.1109/TMI.2021.3116087.

50.   Qu X, Shi Y, Hou Y, Jiang J. An attention-supervised full-resolution residual network for the segmentation of breast ultrasound images. Medical physics (Lancaster). 2020;47(11):5702-14. doi: 10.1002/mp.14470.

51.   Wang K, Liang S, Zhang Y, editors. Residual feedback network for breast lesion segmentation in ultrasound image. Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24; 2021: Springer.

52.   Webb JM, Adusei SA, Wang Y, Samreen N, Adler K, Meixner DD, Fazzio RT, Fatemi M, Alizad A. Comparing deep learning-based automatic segmentation of breast masses to expert interobserver variability in ultrasound imaging. Computers in biology and medicine. 2021;139:104966-. doi: 10.1016/j.compbiomed.2021.104966.

53.   Zhang S, Liao M, Wang J, Zhu Y, Zhang Y, Zhang J, Zheng R, Lv L, Zhu D, Chen H, Wang W. Fully automatic tumor segmentation of breast ultrasound images with deep learning. Journal of Applied Clinical Medical Physics. 2023;24(1). doi: 10.1002/acm2.13863.

54.   Zhao H, Niu J, Wang Y, Li Q, Yu Z, editors. Focal U-Net: A Focal Self-attention based U-Net for Breast Lesion Segmentation in Ultrasound Images2022; Piscataway: The Institute of Electrical and Electronics Engineers, Inc. (IEEE).

55.   Zhuang Z, Li N, Joseph Raj AN, Mahesh VG, Qiu S. An RDAU-NET model for lesion segmentation in breast ultrasound images. PloS one. 2019;14(8):e0221535.

56.   Ronneberger O, Fischer P, Brox T, editors. U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18; 2015: Springer.

57.   Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A, editors. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition; 2015.

58.   Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015;115:211-52.

59.   Jiang P, Ergu D, Liu F, Cai Y, Ma B. A Review of Yolo algorithm developments. Procedia computer science. 2022;199:1066-73.

60.   He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN2017 March 01, 2017:[arXiv:1703.06870   p.].   Available   from: https://ui.adsabs.harvard.edu/abs/2017arXiv170306870H.

61.   Li J, Wang S-R, Li Q-L, Zhu T, Zhu P-S, Chen M, Cui X-W. Diagnostic value of multiple ultrasound diagnostic techniques for axillary lymph node metastases in breast cancer: A systematic analysis and network meta-analysis. Frontiers in Oncology. 2022;12.

62.   Akkus Z, Cai J, Boonrod A, Zeinoddini A, Weston AD, Philbrick KA, Erickson BJ. A survey of deep-learning applications in ultrasound: Artificial intelligence–powered ultrasound for improving clinical workflow. Journal of the American College of Radiology. 2019;16(9):1318-28.

63.   Brunetti N, Calabrese M, Martinoli C, Tagliafico AS. Artificial Intelligence in Breast Ultrasound: From Diagnosis to Prognosis—A Rapid Review. Diagnostics. 2022;13(1):58. doi: 10.3390/diagnostics13010058.

64.   Jahwar AF, Abdulazeez AM, editors. Segmentation and classification for breast cancer ultrasound images using deep learning techniques: A review. 2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA); 2022: IEEE.

65.   Kim J, Kim HJ, Kim C, Kim WH. Artificial intelligence in breast ultrasonography. Ultrasonography. 2021;40(2):183-90. doi: 10.14366/usg.20117.

66.   Kubota K. Breast ultrasound in the age of advanced technology and artificial intelligence. Journal of Medical Ultrasonics. 2021;48(2):113-4. doi: 10.1007/s10396-021-01091-5.

67.   Mao Y-J, Lim H-J, Ni M, Yan W-H, Wong DW-C, Cheung JC-W. Breast Tumour Classification Using Ultrasound Elastography with Machine Learning: A Systematic Scoping Review. Cancers. 2022;14(2):367. doi: 10.3390/cancers14020367.

68.   Villa-Camacho JC, Baikpour M, Chou S-HS. Artificial intelligence for breast US. Journal of Breast Imaging. 2023;5(1):11-20.

69.   Vocaturo E, Zumpano E, editors. Artificial Intelligence approaches on Ultrasound for Breast Cancer Diagnosis2021; Piscataway: IEEE.

70.   Wu G-G, Zhou L-Q, Xu J-W, Wang J-Y, Wei Q, Deng Y-B, Cui X-W, Dietrich CF. Artificial intelligence in breast ultrasound. World Journal of Radiology. 2019;11(2):19-26. doi: 10.4329/wjr.v11.i2.19.

71.   Trepanier C, Huang A, Liu M, Ha R. Emerging uses of artificial intelligence in breast and axillary ultrasound. Clinical Imaging. 2023.

72.   Afrin H, Larson NB, Fatemi M, Alizad A. Deep Learning in Different Ultrasound Methods for Breast Cancer, from Diagnosis to Prognosis: Current Trends, Challenges, and an Analysis. Cancers. 2023;15(12):3139. doi: 10.3390/cancers15123139.

73.   June 26, 1998.

74.   November 9, 2016.

75.   Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, Kahn CE, Esteva A, Karthikesalingam A, Mateen B, Webster D, Milea D, Ting D, Treanor D, Cushnan D, King D, Mcpherson D, Glocker B, Greaves F, Harling L, Ordish J, Cohen JF, Deeks J, Leeflang M, Diamond M, Mcinnes MDF, Mccradden M, Abràmoff MD, Normahani P, Markar SR, Chang S, Liu X, Mallett S, Shetty S, Denniston A, Collins GS, Moher D, Whiting P, Bossuyt PM, Darzi A. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nature Medicine. 2021;27(10):1663-5. doi: 10.1038/s41591-021-01517-0.

76.   Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, Moons K, Collins G, Moher D, Bossuyt PM, Darzi A, Karthikesalingam A, Denniston AK, Mateen BA, Ting D, Treanor D, King D, Greaves F, Godwin J, Pearson-Stuttard J, Harling L, Mcinnes M, Rifai N, Tomasev N, Normahani P, Whiting P, Aggarwal R, Vollmer S, Markar SR, Panch T, Liu X. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open. 2021;11(6):e047709. doi: 10.1136/bmjopen-2020-047709.

77.   Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Ashrafian H, Beam AL, Chan A-W, Collins GS, Deeks ADJ, Elzarrad MK, Espinoza C, Esteva A, Faes L, Ferrante Di Ruffano L, Fletcher J, Golub R, Harvey H, Haug C, Holmes C, Jonas A, Keane PA, Kelly CJ, Lee AY, Lee CS, Manna E, Matcham J, Mccradden M, Monteiro J, Mulrow C, Oakden-Rayner L, Paltoo D, Panico MB, Price G, Rowley S, Savage R, Sarkar R, Vollmer SJ, Yau C. Reporting guidelines for

clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. The Lancet Digital Health. 2020;2(10):e537-e48. doi: 10.1016/s2589-7500(20)30218-1.

78.   Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, Denniston AK, Faes L, Geerts B, Ibrahim M, Liu X, Mateen BA, Mathur P, Mccradden MD, Morgan L, Ordish J, Rogers C, Saria S, Ting DSW, Watkinson P, Weber W, Wheatstone P, Mcculloch P. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022:e070904. doi: 10.1136/bmj-2022-070904.