Clinical Radiology 79 (2024) e1403–e1413

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

Deep learning-based computer-aided detection of ultrasound in breast cancer diagnosis: A systematic review and meta-analysis

H. Li [a], J. Zhao [a,b,*], Z. Jiang [c]

[a] *Department of Ultrasound, Changzheng Hospital, Naval Medical University (Second Medical University), No.415, Fengyang Rd, Shanghai, China*
[b] *Department of Ultrasound, Shanghai Fourth People's Hospital, School of Medicine, Tongji University, No.1279, Sanmen Rd, Shanghai, China*
[c] *School of Health Science and Engineering, University of Shanghai for Science and Technology, No.516, Jungong Rd, Shanghai, China*

ARTICLE INFORMATION

Article history:
Received 19 February 2024
Received in revised form
5 July 2024
Accepted 1 August 2024

PURPOSE: The aim of this meta-analysis was to assess the diagnostic performance of deep learning (DL) and ultrasound in breast cancer diagnosis. Additionally, we categorized the included studies into two subgroups: B-mode ultrasound diagnostic subgroup and multimodal ultrasound diagnostic subgroup, and compared the performance differences of DL algorithms in breast cancer diagnosis using only B-mode ultrasound or multimodal ultrasound.

METHODS: We conducted a comprehensive search for relevant studies published from January 01, 2017 to July 31, 2023 in the MEDLINE and EMBASE databases. The quality of the included studies was evaluated using the QUADAS-2 tool and radiomics quality scores (RQS). Meta-analysis was performed using R software. Inter-study heterogeneity was assessed by I^2 values and Q-test P-values, with sources of heterogeneity analyzed through a random effects model based on test results. Summary receiver operating characteristics (SROC) curves were used for meta-analysis across multiple trials, while combined sensitivity, specificity, and AUC were calculated to quantify prediction accuracy. Subgroup analysis and sensitivity analyses were also conducted to identify potential sources of study heterogeneity. Publication bias was assessed using the funnel plot method. (PROSPERO identifier: CRD42024545758).

RESULTS: The 20 studies included a total of 14,955 cases, with 4197 cases used for model testing. Among these cases were 1582 breast cancer patients and 2615 benign or other breast lesions. The combined sensitivity, specificity, and AUC values across all studies were found to be 0.93, 0.90, and 0.732, respectively. In subgroup analysis, the multimodal subgroup demonstrated superior performance with combined sensitivity, specificity, and AUC values of 0.93, 0.88, and 0.787, respectively; whereas the combined sensitivity, specificity, and AUC value for the model B subgroup was at a level of 0.92, 0.91, and 0.642, respectively.

CONCLUSIONS: The integration of DL with ultrasound demonstrates high accuracy in the adjunctive diagnosis of breast cancer, while the fusion of DL and multimodal breast ultrasound exhibits superior diagnostic efficacy compared to B-mode ultrasound alone.

* Guarantor and correspondent: J. Zhao, Department of Ultrasound, Shanghai Fourth People's Hospital, School of Medicine, Tongji University, No.1279, Sanmen Rd, Shanghai, People's Republic of China.
E-mail addresses: 2392707695@qq.com (H. Li), ultrasoundczzjq@163.com (J. Zhao).

https://doi.org/10.1016/j.crad.2024.08.002

Introduction

Breast cancer is the most prevalent malignancy worldwide and the leading cause of cancer-related mortality in women.[1] Early detection and treatment are crucial for reducing breast cancer mortality rates. Ultrasound is an invaluable diagnostic tool for detecting breast cancers that may not be easily detected on mammography due to dense breast tissue. Additionally, ultrasound serves as the ideal imaging technique for guiding subsequent biopsy procedures, further enhancing its utility in diagnosing breast cancer.[2] Ultrasound, a cost-effective screening tool, offers painless imaging with no ionizing radiation exposure and dynamic features. However, accurate classification of breast nodules according to breast imaging reporting and data system (BI-RADS) criteria and distinguishing between benign and malignant lesions heavily relies on the expertise of imaging physicians. Over the past decade, significant advancements have been made in ultrasound-based artificial intelligence technology for diagnosing breast tumors. The workflow of ultrasound radiomics encompasses data collection, delineation of regions of interest (ROI), feature extraction, and model development.[3] As a specific form of machine learning, deep learning (DL) has many applications in image segmentation and diagnostic model training. When the amount of data is sufficient, DL will show better performance than traditional machine learning. Therefore, DL methods have been widely used in the diagnosis and molecular typing of breast cancer, as well as the development of computer-aided diagnosis (CAD) systems. Rory Wilding *et al.*[4] developed classification and segmentation algorithms based on ultrasonic images by using machine learning, and classified them according to the BI-RADS standard with the help of radiology experts. The accuracy rate of the classification algorithm to distinguish healthy breast tissue images from abnormal tissue images was 96%, and the accuracy rate to distinguish benign and malignant images was 85%. More accurate diagnosis of breast cancer can contribute to better planning of treatment strategies. Luo Xiao *et al.*[5] used an optimized 3D CNN model to conduct automatic detection of automated breast ultrasound (ABUS) images of 397 female patients, and evaluated the sensitivity and false positives of the detection effect. The network achieved sensitivities of 93.8%, 97.2%, and 100% for volume, lesion, and patient-based evaluations, respectively, with an average of only 1.9 false positives per volume. The results showed high sensitivity and low false positives, indicating that DL technology can automatically detect ABUS lesions.

Therefore, this review aims to comprehensively analyze and evaluate the effectiveness of DL CAD methods for breast cancer diagnosis. By comparing the diagnostic performance of B-mode and multimodal subgroups, it seeks to identify the shortcomings and limitations of ultrasound DL-based CAD methods for breast cancer diagnosis, providing references for future studies.

Method

The reporting items of this study were selected in accordance with the guidelines for Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA) to ensure optimal reporting, aligning with the requirements of academic rigor and suitability for this study.[6] The literature retrieval, data extraction, and quality evaluation were independently conducted by two individuals. In the event of any discrepancies during the data verification process, a third party was engaged to facilitate resolution and reach a consensus.

Eligibility criteria

We included studies that met the following inclusion criteria: 1) Diagnostic studies employed DL algorithms; 2) Pathological assessment of tissues resected after breast cancer surgery served as the reference standard; 3) All patients underwent ultrasound examination of their breast masses, including B-mode imaging, color Doppler flow imaging (CDFI), pulse Doppler (PW), elastography, or one of the ABUS, with B-mode imaging being mandatory; 4) The study population must exclusively comprise female patients; 5) The language of the publication was English; 6) The study's publication period spans from January 1, 2017 to July 31, 2023; 7) Complete data reporting diagnostic information, including sensitivity, specificity, and AUC values, was available for meta-analysis; 8) The sample size ($n>30$). Exclusion criteria were as follows: 1) Review articles, systematic evaluations, comments, in vitro experiments, and other non-original research documents were excluded; 2) Studies with incomplete or inaccurate data reported in the diagnostic tests were also excluded.

Study selection

We independently conducted a comprehensive literature search using the DL approach for ultrasound studies. Both the PUBMED and EMBASE databases were searched, and the literature included in our analysis was published from January 1, 2017 to July 31, 2023. The following terms were utilized for conducting the literature search:((("Malignant and Benign" [Title]) AND ("Breast" [Title])) OR ("Breast Neoplasms" [Title]) OR ("Breast Cancer" [Title]) OR ("Breast Carcinoma" [Title]) OR ("Mammary Cancer" [Title]) OR ("Breast Neoplasms" [Mesh])) AND (("Deep Learning" [Mesh]) OR ("Deep Learning" [Title]))AND (("Ultrasonography" [Mesh])OR("Ultrasonography" [Title])OR ("Ultrasonics" [Title]) OR ("Ultrasound" [Title])).

Data extraction

The data were extracted by two researchers. The literature was screened based on the inclusion and exclusion criteria, and essential information was collected, including PMID, title, author(s), first author, journal name, year of publication, PMCID, DOI, and the number of included patients. Through statistical analysis, additional parameters such as the number of patients in both positive and control groups, the best model selected, accuracy rate, sensitivity rate, specificity rate, positive predictive value (PPV), negative predictive value (NPV), true positives (TP), false positives (FP), false negatives (FN), true negatives (TN) were obtained along with area under curve (AUC) values. The included studies were divided into B-mode subgroup and multimodal subgroup according to the patterns of ultrasound images. In this study, multimodal ultrasound scanning includes B-mode imaging, CDFI, PW, elastography, or one of the ABUS.

Quality assessment

For all the included studies, the QUADAS-2[7] tool was used for the quality evaluation. The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool can be utilized to evaluate the methodological quality of included studies in systematic reviews or meta-analyses of diagnostic accuracy research. It comprises four fundamental domains: patient selection, index test, reference standard, and flow and timing. Within the patient selection domain, the tool assesses whether the study explicitly states the inclusion and exclusion criteria, if patients were consecutively or randomly recruited, and identifies potential biases in patient selection.

The QUADAS-2 tool offers a structured and standardized approach for assessing the methodological quality of diagnostic accuracy studies, enabling reviewers to evaluate both the risk of bias and applicability of study results. By utilizing this tool, reviewers can ensure that only studies with robust methodological quality are incorporated into their systematic review or meta-analysis, thereby enhancing the reliability and validity of the findings. The quality evaluation comprises 10 assessment items, with questions within each item being rated as "yes" (2), "no" (0), or "unclear" (1). The QUADAS score ranges from 0 to 20, where a score greater than or equal to 15 indicates good-quality literature.

In this meta-analysis, radiomics quality scores (RQS) was employed to assess the methodological rigor of the included studies. Serving as a practical reference tool for evaluating radiomics research, the RQS plays an indispensable role. The RQS encompasses six domains (image protocols, radiomics feature extraction, data analysis and statistics, model validation, clinical validity, and open science) with a total of 16 criteria. Following Philippe Lambin *et al.*'s approach,[8] each criterion is evaluated to determine a final score ranging from -8 to 36 which can be converted into a percentage (scores below 0 are considered as 0%, while a score of 36 corresponds to 100%).

Meta-analysis

The meta-analysis was conducted using R 4.3.0 software, and the I^2 statistics and Q-test statistics were employed to evaluate the heterogeneity among individual studies. A value of 0 for I^2 indicates no observed heterogeneity between studies. Generally, an I^2 value below 50% suggests a lower degree of heterogeneity among studies. The Q test is commonly performed at a significance level of 0.1, and if the corresponding P-value exceeds this threshold, it suggests small between-study heterogeneity.[29] The heterogeneity analysis was also conducted and incorporated into this meta-analysis, while the origin of heterogeneity was investigated through sensitivity analysis. The summary receiver operating characteristic (SROC)[30] curve was employed to depict the diagnostic performance of all studies and subgroups, while calculating the combined AUC, combined sensitivity, and combined specificity for both overall studies and subgroups. The funnel plot[31] was employed to assess publication bias in all included studies. A symmetrical distribution of the funnel plot indicates the absence of publication bias, while an obvious asymmetry suggests its presence. The "meta4diag" software package, based on Bayesian theory, is primarily utilized for diagnostic test accuracy (DTA) in meta-analysis. To combine effect values from individual studies, we applied the integrated nested Laplace approximation (INLA) method. Additionally, the "meta" package was used for calculating I^2 statistics, conducting Q tests, and constructing funnel plots. The "robvis" package is utilized for conducting quality evaluation using Quadas2 methodology. The forest diagram optimization is achieved through the utilization of the "forestploter" and "ggplot2″ packages. Additionally, the "meta" package facilitates I^2 calculations, Q tests, and the construction of funnel plots.

Result

Collection and selection of literature

The literature screening process and its outcomes are summarized in Fig 1. Initially, a total of 79 articles were identified. However, three articles with duplicate records were excluded, followed by the exclusion of 44 articles with low correlation after reviewing their abstracts. Furthermore, 12 articles were excluded upon reading the full text, while five articles had missing data performance and seven more lacked clear diagnostic criteria upon thorough examination. Finally, a total of 20 relevant literatures were included in this study encompassing a cohort of 14, 955 patients for conducting a meta-analysis.

Baseline characteristics of the studies

The 20 studies included a total of 14,955 patients, with 6,261 (41.87%) diagnosed with breast cancer. Among the cases analyzed, a test set consisting of 4,197 patients was used for model testing. This set comprised 1,582 patients with breast cancer and 2,615 patients with benign or other

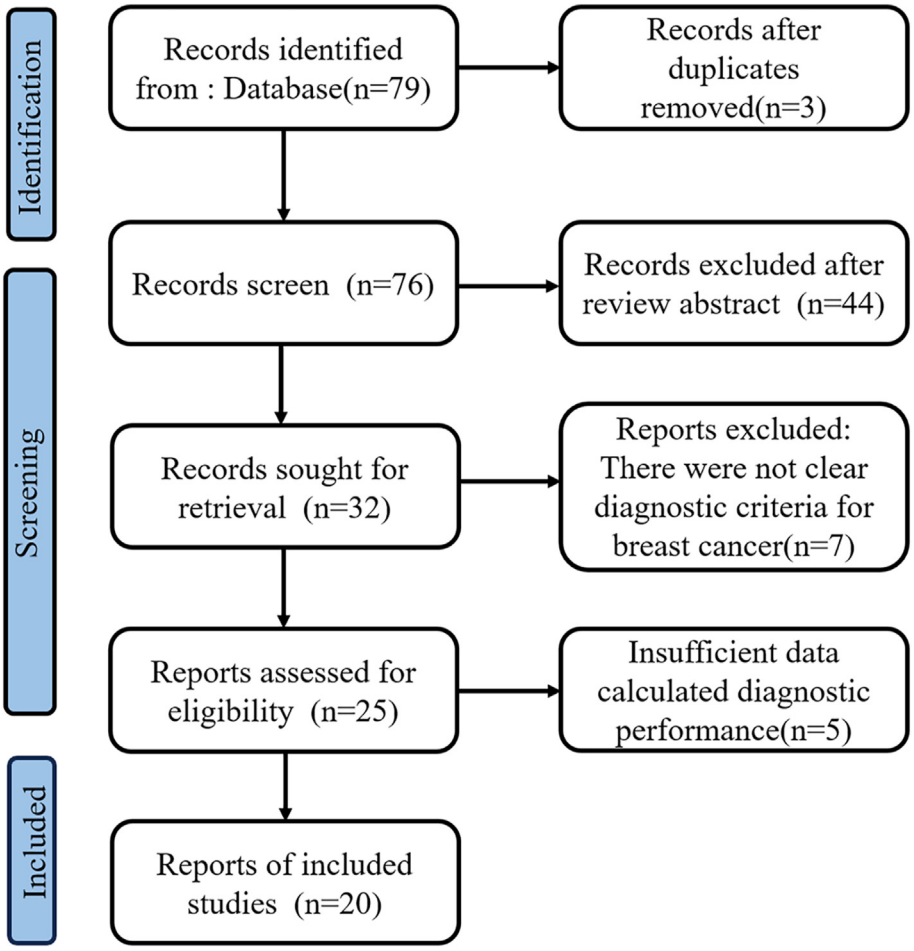

Figure 1 Depicts a flow chart that includes inclusion and exclusion studies.

breast lesions. Patient ages were not reported in five studies; however, the remaining studies covered a wide age range from 11 to 91 years old. Model training in these studies involved B-mode ultrasound images in 11 cases and multimodal ultrasound images in the other nine cases. The studies were categorized into B-mode subgroup and multimode subgroup based on ultrasonic image patterns. All the included studies utilized DL models, out of which seventeen (references [9–17,20,21–27]) employed CNN models to achieve optimal learning performance. Three studies incorporated transfer learning algorithms ([12,25,16]), reducing the cost of model training. VGG-Net (5 times) and Res-Net (5 times) demonstrated superior performance among convolutional layer architectures examined within this study.

Two of the twenty studies were multicenter ([17,23]), while eighteen were single-center investigations. One study had a prospective design ([18]), whereas others adopted retrospective designs. Five studies underwent external validation ([9,10,17,18,22]), while fifteen underwent internal validation procedures. The baseline characteristics of all twenty included studies are presented in Table 1 and Table 2.

All the USS machine vendors and manufacturers include: Invenia ABUS (Automated Breast Ultrasound System, GE Healthcare) , Volusone E8 color Doppler ultrasound imaging system , RS80A system (Samsung Medison Co., Ltd.) , EUP-L54MA 9.75-MHz linear probe (Hitachi Medical Systems, Tokyo, Japan) , Esaote My Lab (Esaote, Italy), GE LOGIQ E9 (General Electric Co., USA), Hitachi (Hitachi, Ltd., Japan), Mindray Resona 7 (Shenzhen Mindray Bio-Medical Electronics Co., Ltd., China), Samsung (Samsung Medison Co., Ltd. Korea), Siemens (Siemens Healthcare GmbH, USA), Sonoscape (SonoScape Medical Corp., China), Supersonic (Supersonic Imagine, France), and Toshiba (Aplio 500, Aplio i900, CANON Medical Systems Corporation, Japan) systems.

Methodological quality

The QUADAS scores of all included studies ranged from 15 to 18, as presented in Table 2, indicating a high level of reliability for the included studies.

The quality assessment of all 20 included studies is presented in Fig 2, demonstrating that all studies exhibited a low overall risk of bias. To address the second question, "Was a case-control design avoided?" Fifteen studies identified confirmed cases of malignant and benign tumors, indicating a high risk, while two did not provide such identification, resulting in an unclear classification. Regarding the fifth question, "If a threshold was used, was it

Table 1
Baseline characteristics of the included studies (1)

Study	Numbers of patients	Rate of maligancy	Data source	Stand reference	Age(Mean,Sd)	Image	Optimal model	Test set
Ji Soo Choi-2019[9]	253	31.62%	Single institution	Surgical pathology	(47[IQR],42.0 − 53.5)	Only B mode	CNN-GoogLeNet	Internal test
Tomoyuki Fujioka-2019[10]	357	59.94%	Single institution	Surgical pathology	21 ~ 84(55.0 ± 13.1)	Only B mode	CNN-GoogleNet Inception v2	Internal test
Anton S Becker-2017[11]	632	12.97%	Single institution	Surgical pathology	15 ~ 91(53 ± 15)	Only B mode	DLS	Internal test
Michal Byra-2019[12]	882	23.13%	Single institution	Surgical pathology	18 ~ 90(51 ± 15)	Only B mode	CNN-VGG-19 (Transfer Learning)	Internal test
Zhijin Zhao-2022[13]	479	24.40%	Single institution	Surgical pathology	16 ~ 90(44.3 ± 13.1)	Only B mode	CNN-MobileNet	Internal test
Woo Kyung Moon-2020[14]	1225	43.51%	Single institution	Surgical pathology	17 ~ 85(45.59 ± 9.62)	Only B mode	CNN-WA(VGGNet, ResNet, and DenseNet)	External test
Shuai Zhang-2023[15]	745	31.36%	Single institution	Surgical pathology	20 ~ 85	Only B mode	CNN-DenseNet	External test
Jaeil Kim-2021[16]	1110	61.98%	Single institution	Surgical pathology	16 ~ 83	Only B mode	CNN-VGG-16	Internal test
Xianyu Zhang-2021[17]	2523	65.27%	Multiple institutions	Surgical pathology	NA	Only B mode	CNN-Xception	External test
Tomoyuki Fujioka-2023[18]	277	28.88%	Single institution	Surgical pathology	NA	Only B mode	CADe	External test
Weichang Ding-2022[19]	197	50.76%	Single institution	Surgical pathology	61.25 ± 12.53	Only B mode	DenseNet121	Internal test
Qiucheng Wang-2022[20]	769	59.03%	Single institution	Surgical pathology	27 ~ 79	multimodel	CNN-ResNet101 v2	Internal test
Xuehua Xiao-2022[21]	120	53.33%	Single institution	Surgical pathology	22 ~ 71(47.23 ± 12.67)	multimodel	DNN	Internal test
Xuejun Qian-2021[22]	775	26.97%	Single institution	Surgical pathology	23 ~ 73 (46.6)	multimodel	CNN-ResNet-18+SENet	External test
Teng-Fei Yu-2021[23]	3623	32.54%	Multiple institutions	Surgical pathology	NA	multimodel	CNN-ResNet50	Internal test
Min Young Kim-2021[24]	146	6.41%	Single institution	Surgical pathology	23 ~ 74(46 ± 10)	multimodel	CNN-GoogLeNet	Internal test
Yi Wang-2020[25]	316	42.72%	Single institution	Surgical pathology	11 ~ 85(39.0 ± 24.0)	multimodel	CNN-Inception-v3 (Transfer Learning)	Internal test
Wen-Xuan Liao-2020[26]	141	34.67%	Single institution	Surgical pathology	28 ~ 76(51.4 ± 9.9)	multimodel	CNN-VGG-19 (Transfer Learning)	Internal test
Chunxiao Lai-2023[27]	264	51.14%	Single institution	Surgical pathology	NA	multimodel	ResNet-GAP	Internal test
Xiaoyan Fei-2021[28]	121	40.53%	Single institution	Surgical pathology	NA	multimodel	DPN	Internal test

pre-specified?" None of the studies explicitly explained the threshold employed, thus rendering this aspect as unclear.

The median RQS was 47.22% (IQR = 9.72), with quality ratings conducted by two or more readers for the 20 systematic reviews, and a summary of each review's results is presented in Table 2.

Publication bias

We conducted a funnel plot analysis to assess publication bias among the included studies and found no evidence of such bias [p=0.8672 (>0.05)]. The results of the evaluation are shown in Fig 3. However, due to the limited number of studies in our group analyses, we acknowledge that the efficiency of this test may be low and therefore did not perform it for these analyses.

Meta-analysis

The heterogeneity of the DL model in diagnosed breast cancer patients was analyzed using Mantel-Haenszel and DerSimonian-Laird methods, revealing an I^2 value of 48.7% [13.9%, 69.5%] (< 50%). Additionally, the Q test yielded a P-value of 0.0078 (< 0.1). These findings suggest the presence of heterogeneity among the included studies. However, it is not statistically significant. Figure 6 illustrates the SROCs curves for all studies and subgroups, demonstrating an overall combined AUC of 0.732, combined sensitivity of 0.93 (95%CI [0.90,0.95]), and combined specificity of 0.90 (95%CI [0.87,0.93]), indicating a high diagnostic efficacy level. The sensitivity analysis results for all studies and groups are summarized in Fig 4.

To account for potential heterogeneity, we conducted a grouping analysis of the included studies to investigate their sources. Based on the image source, we categorized the studies into two subgroups: B-mode subgroup and multimode subgroup, as image acquisition method and quality may impact diagnostic outcomes. Eleven studies were assigned to the B-mode subgroup while nine belonged to the multimode subgroup. The combined AUC, sensitivity, and specificity of the B-mode subgroup were 0.642, 0.920, and 0.910, respectively; whereas those of the multimodal subgroup were 0.787, 0.930, and 0.880, respectively. In order to depict both overall study performance and individual subgroup performance comprehensively, separate SROC curves were constructed for the population as well as each respective subgroup for comparison purposes. Furthermore, we assessed heterogeneity within each group separately. Notably, both B-mode and multimode subgroups exhibited low levels of heterogeneity with I^2 values of 29.6% [0%, 65.4%] (< 50%) (P = 0.1635 (> 0.1)) for B-mode subgroup; I^2 = 42.8% [0%, 73.6%], P = 0.0823 (< 0.1) for multimode subgroup. The results indicated no significant heterogeneity in the B-mode subgroup while some non-significant heterogeneity was observed in the multimode subgroup. Thus, the source of heterogeneity across all studies might be attributed to differences between these two subgroups. Further details regarding all study findings are summarized in Fig 5.

Table 2
Baseline characteristics of the included studies (2)

Study	Optimal model	AUC	Quadas scores	RQS(%)	Test data	EET	CET	TP	FP	FN	TN
Ji Soo Choi-2019	CNN-GoogLeNet	0.919	17	23(63.89%)	253	80	173	72	17	8	156
Tomoyuki Fujioka-2019	CNN-GoogleNet Inception v2	0.913	17	16(44.44%)	120	72	48	69	6	3	42
Anton S Becker-2017	DLS	0.84	17	18(50.00%)	192	20	172	17	34	3	138
Michal Byra-2019	CNN-VGG-19	0.936	17	19(52.78%)	150	34	116	29	12	5	104
Zhijin Zhao-2022	CNN-MobileNet	0.897	15	16(44.44%)	195	46	149	43	15	3	134
Woo Kyung Moon-2020	CNN-WA(VGGNet, ResNet, and DenseNet)	0.9711	15	19(52.78%)	697	210	487	194	21	16	466
Shuai Zhang-2023	CNN-DenseNet	0.991	17	17(47.22%)	130	40	90	38	0	2	90
Jaeil Kim-2021	CNN-VGG-16	0.96	15	19(52.78%)	200	100	100	100	3	0	97
Xianyu Zhang-2021	CNN-Xception	0.96	18	18(50.00%)	210	93	117	76	18	17	99
Tomoyuki Fujioka-2023	CADe	0.7726	17	15(41.67%)	133	80	53	76	7	4	46
Weichang Ding-2022	DenseNet121	0.915	17	15(41.67%)	100	97	88	13	12	84	100
Qiucheng Wang-2022	CNN-ResNet101 v2	0.85	15	13(36.11%)	172	103	69	85	23	18	46
Xuehua Xiao-2022	DNN	0.838	15	10(27.78%)	120	64	56	63	4	1	52
Xuejun Qian-2021	CNN-ResNet-18+SENet	0.955	17	24(66.67%)	152	44	108	39	3	5	105
Teng-Fei Yu-2021	CNN-ResNet50	0.96	17	20(55.56%)	114	55	59	50	13	5	46
Min Young Kim-2021	CNN-GoogLeNet	0.803	17	14(38.89%)	156	10	146	9	37	1	109
Yi Wang-2020	CNN-Inception-v3	0.9468	17	16(44.44%)	316	135	181	120	22	15	159
Wen-Xuan Liao-2020	CNN-VGG-19	0.98	15	16(44.44%)	199	69	130	65	11	4	119
Chunxiao Lai-2023	ResNet-GAP	0.936	16	17(47.22%)	135	129	129	18	6	111	135
Xiaoyan Fei-2021	DPN	0.961	15	15(41.67%)	92	135	90	8	2	127	92

AUC, Aea Under Curve; RQS, Radiomics Quality Score; EET, Experimental Events Total; CET, Control Events Total. TP, true positive; FP, false positive; TN, true negative; FN, false negative.

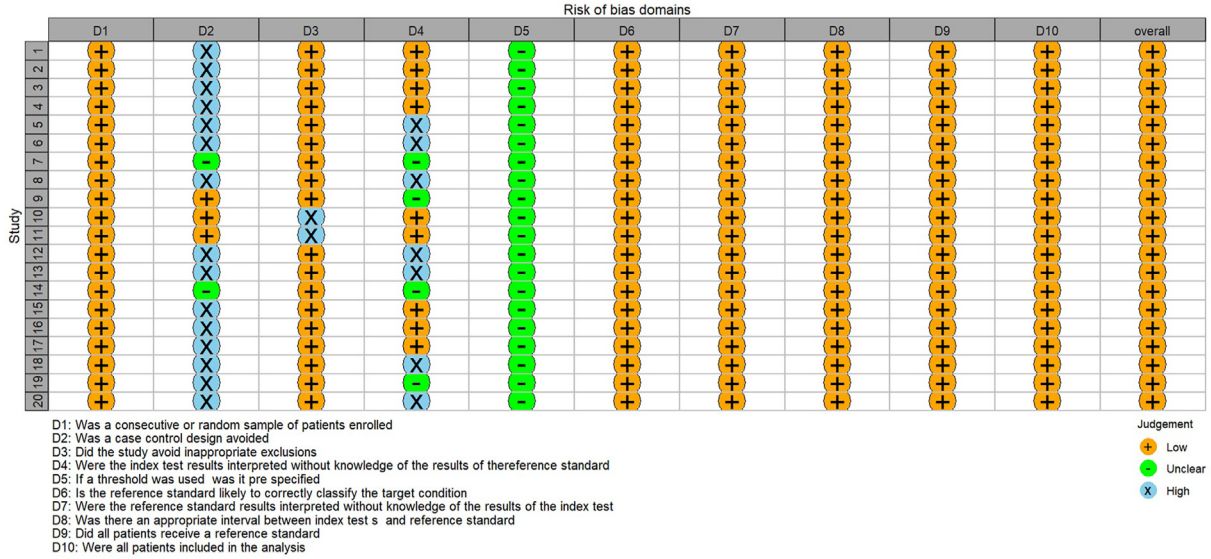

Figure 2 Methodological quality assessment of the included studies was conducted based on Quadas-2. The color orange indicates low risk, while green circles represent unclear criteria, and blue is used to indicate high risk deviation.

Discussion

The field of DL, a subfield of machine learning, encompasses algorithms that excel in representation learning from data.[32] Currently, there exist several DL frameworks including deep neural networks, convolutional neural networks, deep belief networks, recurrent neural networks, etc. Among these frameworks, convolutional neural network stands out as the most extensively employed technique in image processing due to its exceptional performance in handling large-scale image datasets.[33] In this meta-analysis, convolutional neural networks were employed in 17 studies.

The diagnostic performance of ultrasound-based DL for breast cancer was assessed through this meta-analysis. QUADAS quality assessment revealed that all included studies scored above 15 points, indicating their high reliability. Two multicenter studies were included in this analysis, enhancing the objectivity of the conclusions drawn. Out of these studies, five had external validation results and fifteen had internal validation results, thereby increasing the overall reliability of the diagnostic performance. RQS provided a more objective assessment of the quality of imaging omics studies included in this research, thereby indicating consistent literature quality.

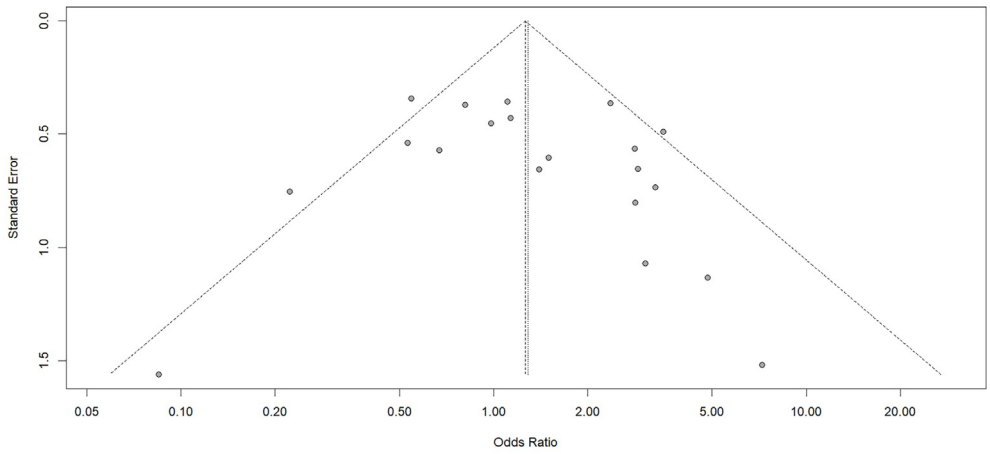

Figure 3 The funnel plot based on diagnostic odds ratio (OR) included 20 studies, with 3 falling outside the skew zone and 17 within it. The symmetry of the funnel plot is indicative of low heterogeneity, suggesting good data quality.

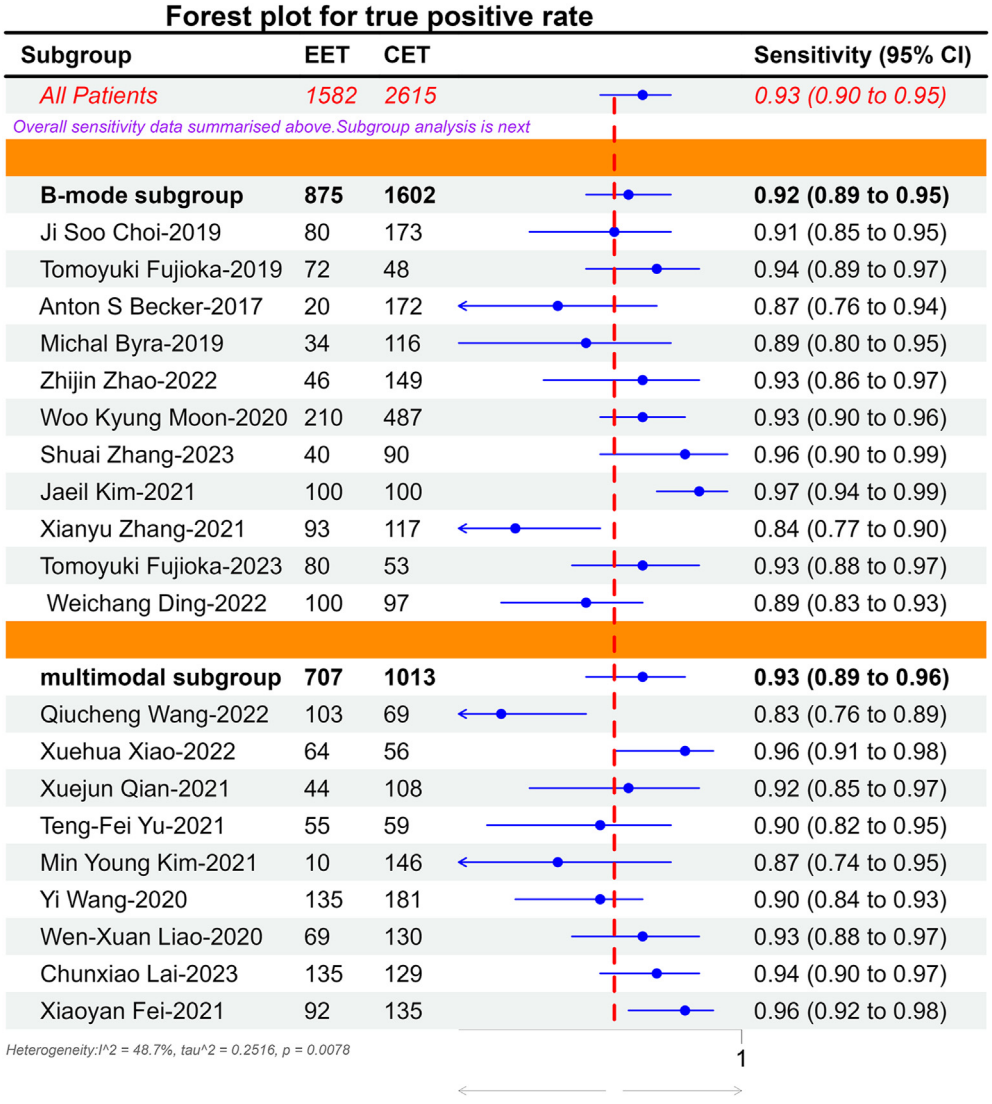

Figure 4 The analysis in this study utilized sensitive forest maps. The pooled sensitivity of all studies was 0.93, with a 95% confidence interval (CI) of [0.90, 0.95]. In the B-mode subgroup, the combined sensitivity was found to be 0.92, with a 95% confidence interval (CI) of [0.89, 0.95]. For the multimodal subgroup, the combined sensitivity was determined to be 0.93, with a 95% confidence interval (CI) of [0.89, 0.96]. EET, Experimental Events Total; CET, Control Events Total.

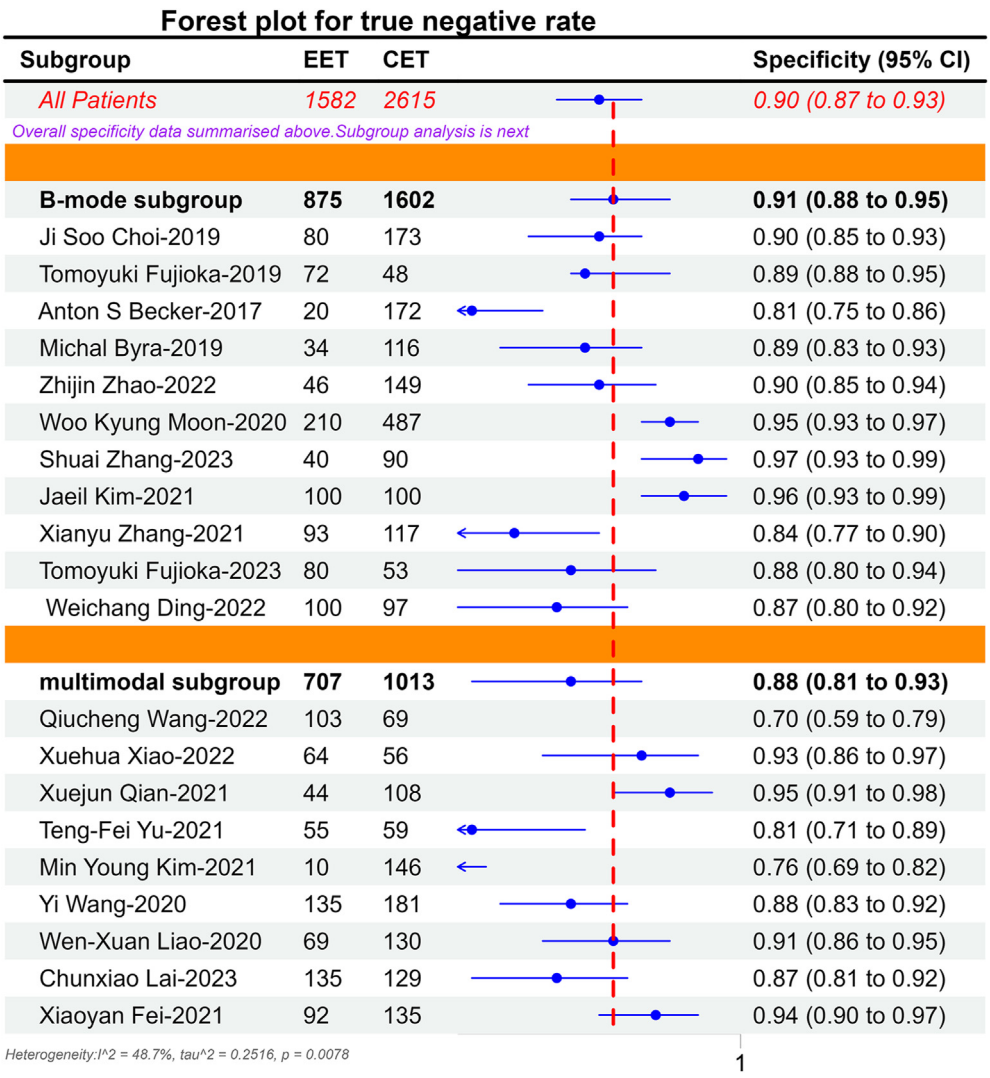

Figure 5 Illustrates the specific forest maps utilized for group analysis in the included studies. The pooled specificity across all studies was determined to be 0.90, with a 95% confidence interval (CI) of 0.87–0.93. Within the B-mode subgroup, the combined specificity was found to be 0.91, with a corresponding CI of 0.88–0.95. In the multimodal subgroup, the comprehensive specificity was determined as 0.88, and its associated CI ranged from [0.81, 0.93].

In the heterogeneity analysis, all included studies exhibited a low degree of heterogeneity ($I^2 = 48.7\% < 50\%$), while the Q test P value was 0.0078 (< 0.1), indicating some superficial heterogeneity. We hypothesized that this heterogeneity may be attributed to variations in ultrasound image acquisition methods. Ultrasound images obtained through B-mode differ in diagnostic significance from those collected using other modes such as CDFI, PW, ABUS, etc., thus necessitating subgroup discussions. The subgroup analysis revealed significant differences between groups and reduced the heterogeneity within the B-mode group ($I^2 = 29.6\%$). Furthermore, potential sources of heterogeneity include tumor histology and molecular subtypes, specific DL algorithms employed, different verification methods utilized, as well as varying research standards across institutions.

In the subgroup analysis, we categorized the included studies into two groups: B-mode group and multimode group. The B-mode group solely utilized B-mode for ultrasound imaging of breast lesions, while the multimode group employed a combination of B-mode, CDFI, PW, SWE, ABUS, etc., along with DL algorithms to assist in diagnosing breast cancer patients. Notably, three of the included studies incorporated CAD systems for image analysis. Currently, numerous CAD systems based on DL algorithms have been applied in clinical breast cancer diagnosis.[32] Our meta-analysis comprised 11 studies in the B-mode group and 9 studies in the multimodal group. Regarding model performance within each group, the multimodal approach demonstrated superior performance (AUC = 0.787), which we hypothesized was attributed to its inclusion of more diagnostic information-rich images. Conversely, the B-mode group exhibited relatively lower performance (AUC = 0.642), potentially due to limitations in image quality.

Compared with previous reviews, we pay more attention to the significant improvement in the accuracy

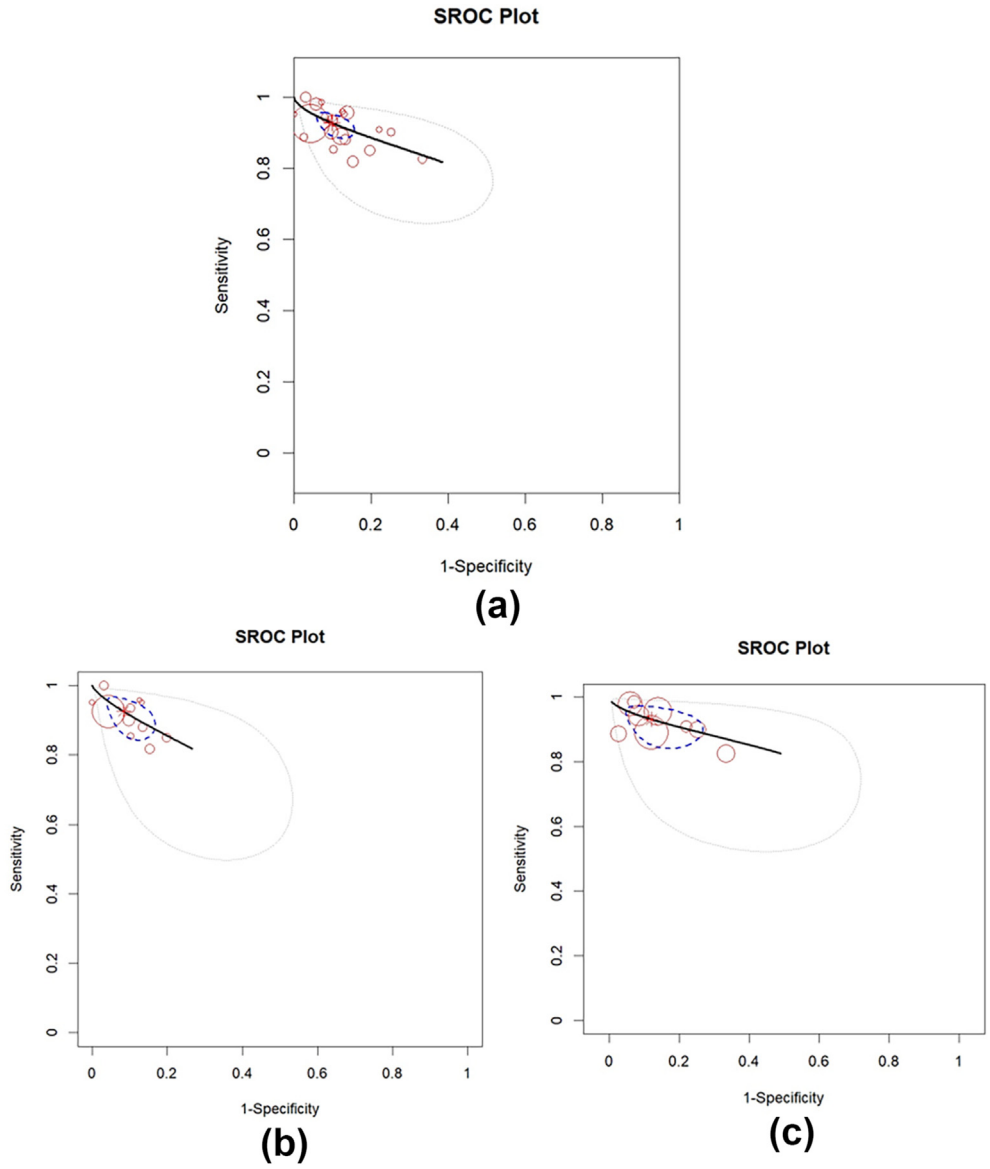

Figure 6 Illustrates the receiver operating characteristic curve (SROC) for all studies and subgroups, with A representing the comprehensive SROC, B denoting the B-mode subgroup, and C indicating the multimode subgroup. The area under the curve was calculated as 0.732 for all studies, 0.642 for the B-mode subgroup, and 0.787 for the multimode subgroup. Each red circle represents a study, and indicates its sensitivity and specificity. The size of the circles is proportional to the number of individuals included in the study. The solid black line represents the SROC curve, the stars represent the summary points, the blue dashed line represents the 95% confidence interval, and the gray dashed line encloses the 95% predictive interval.

of artificial intelligence diagnosis of breast cancer brought about by the progress of DL algorithm technology. Because, compared with traditional machine learning, DL has better performance, especially when the sample size is large. Moreover, unlike previous reviews, we divided the patients according to the way they were examined by ultrasound and divided them into subgroups of B-mode and multimode, and then compared their performance, which has the potential to guide the development of ultrasound-based artificial intelligence diagnosis of breast cancer in the future. For example, Dan *et al.*[34] included 16 studies comparing the accuracy of DL in diagnosing breast cancer in the US clinical environment with that of human readers. However, this conclusion is insufficient to prove that DL is superior to human readers or improves the accuracy of breast ultrasound diagnosis.

Our meta-analysis still has some limitations. First, the majority of included studies were retrospective, with only one prospective study, which may have impacted the model's predictive performance. Second, there was no standardized protocol for ultrasound imaging across all cases, leading to potential heterogeneity among studies. Third, the multimodal group had a small number of included studies and wide confidence intervals resulting in reduced reliability of our findings.

With the increasing prevalence of breast cancer, early diagnosis has become increasingly crucial. Ultrasound imaging technology has emerged as a pivotal tool in breast cancer detection. However, current diagnostic methods heavily rely on subjective judgment and the limited expertise of physicians, leading to restricted reliability and accuracy. Consequently, there is a growing research interest in leveraging DL algorithms for automated breast cancer diagnosis and detection. Over the past three decades, CAD has witnessed remarkable advancements with numerous clinical and commercial products successfully employing CAD systems for efficient breast cancer identification and diagnosis.[35] Despite the fact that CAD has been widely used in breast cancer detection and diagnosis, there are still many challenges in this field, such as the need for standardized images for predictive model construction and large clinical cases. In our included studies, they all achieved good results in breast cancer detection and diagnosis. Compared with traditional machine learning, their studies optimized the DL algorithm to achieve better predictive performance. Some studies included multicenter data, making the predictive model more generalizable. Some studies used transfer learning algorithm models to reduce the cost of model training. In the future, we will focus on establishing standardized ultrasound image acquisition protocols, including more cases, and combining studies from multiple centers to establish better DL models that can be continuously optimized to achieve better AI diagnosis and prediction effects.

In conclusion, our findings highlight the potential of DL-based CAD for breast cancer ultrasound diagnosis. However, an analysis of the limitations present in this study and proposed feasible directions for future research have been conducted. First, the accuracy of ultrasound detection may be compromised by factors such as operator experience, equipment quality, and image resolution. Second, DL models can be sensitive to overfitting, requiring careful optimization and validation to ensure their generalizability and reliability. Moreover, the implementation of these techniques in clinical practice requires extensive training and collaboration between healthcare professionals and artificial intelligence experts.

With further research and optimization, these technologies have the potential to become valuable tools for healthcare professionals in the fight against breast cancer. The integration of ultrasound and DL offers a promising approach to improving early detection, diagnosis, and treatment of breast cancer, ultimately enhancing the survival and quality of life for patients worldwide.

Ethics statement

The authors used publicly available, non-identifiable aggregated data in this "negligible risk" study, so no formal ethical review was required because of the minimal potential harm involved. The participants included in the study had already provided sufficient consent during the main study.

CRediT authorship contribution statement

1. Guarantor of integrity of the entire study– **Hongtao Li**
2. Study concepts and design– **Jiaqi Zhao**
3. Literature research– **Zhuoyun Jiang**
4. Clinical studies–N/A
5. Experimental studies/data analysis– **Hongtao Li**
6. Statistical analysis– **Zhuoyun Jiang**
7. Manuscript preparation– **Jiaqi Zhao**
8. Manuscript editing– **Hongtao Li**

Conflict of interest

The authors declare no conflict of interest.

Acknowledgment

This research was funded by Military Medical Talent Project of “Pyramid Talent Program” of Three-year Action Plan for Talent Construction of The Second Affiliated Hospital of Naval Medical University (Second Military Medical University) (1009), Start-up Scientific Research Project of Shanghai Fourth People’s Hospital Affiliated to Tongji University (SYKYQD06101), Medical Research Project of Health Commission of Shanghai Hongkou District (HW2302-26), and Clinical Key Supporting Project of Health Commission of Shanghai Hongkou District (HKLCFC202404).

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020;**70**:7–30. https://doi.org/10.3322/caac.21590.
2. Sood R, Rositch AF, Shakoor D, *et al*. Ultrasound for breast cancer detection globally: a systematic review and meta-analysis. *J Glob Oncol* 2019:1–17. https://doi.org/10.1200/JGO.19.00127.
3. Zhang D, Liu J, Ni X. Application progress of ultrasomics in diagnosis and treatment of breast Neoplasms. *Chin J Med Imaging* 2021;**29**:1256–60.
4. Wilding R, Sheraton VM, Soto L, *et al*. Deep learning applied to breast imaging classification and segmentation with human expert intervention. *J Ultrasound* 2022;**25**:659–66. https://doi.org/10.1007/s40477-021-00642-3.
5. PhD XL, Xu M, Tang G, *et al*. The lesion detection efficacy of deep learning on automatic breast ultrasound and factors affecting its efficacy: a pilot study. *Br J Radiol* 2022;**95**:20210438, https://doi.org/10.1259/bjr.20210438.
6. McInnes MDF, Moher D, Thombs BD, *et al*. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* 2018;**319**:388–96. https://doi.org/10.1001/jama.2017.19163.
7. Whiting PF, Rutjes AWS, Westwood ME, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;**155**:529–36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
8. Lambin P, Leijenaar RTH, Deist TM, *et al*. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017;**14**:749–62. https://doi.org/10.1038/nrclinonc.2017.141.
9. Choi JS, Han B-K, Ko ES, *et al*. 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 J Radiol* 2019;**20**:749. https://doi.org/10.3348/kjr.2018.0530.
10. Fujioka T, Kubota K, Mori M, *et al*. Distinction between benign and malignant breast masses at breast ultrasound using deep learning

method with convolutional neural network. *Jpn J Radiol* 2019;**37**:466–72. https://doi.org/10.1007/s11604-019-00831-5.
11. Becker AS, Mueller M, Stoffel E, *et al*. Classification of breast cancer from ultrasound imaging using a generic deep learning analysis software: a pilot study. *Br J Radiol* 2017:20170576, https://doi.org/10.1259/bjr.20170576.
12. Byra M, Galperin M, Ojeda-Fournier H, *et al*. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. *Med Phys* 2019;**46**:746–55. https://doi.org/10.1002/mp.13361.
13. Zhao Z, Hou S, Li S, *et al*. Application of deep learning to reduce the rate of malignancy among BI-rads 4A breast lesions based on ultrasonography. *Ultrasound Med Biol* 2022;**48**:2267–75. https://doi.org/10.1016/j.ultrasmedbio.2022.06.019.
14. Moon WK, Lee Y-W, Ke H-H, *et al*. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Comput Methods Programs Biomed* 2020;**190**:105361, https://doi.org/10.1016/j.cmpb.2020.105361.
15. Zhang S, Liao M, Wang J, *et al*. Fully automatic tumor segmentation of breast ultrasound images with deep learning. *J Appl Clin Med Phys* 2023;**24**, https://doi.org/10.1002/acm2.13863.
16. Kim J, Kim HJ, Kim C, *et al*. Weakly-supervised deep learning for ultrasound diagnosis of breast cancer. *Sci Rep* 2021;**11**:24382, https://doi.org/10.1038/s41598-021-03806-7.
17. Zhang X, Li H, Wang C, *et al*. Evaluating the accuracy of breast cancer and molecular subtype diagnosis by ultrasound image deep learning model. *Front Oncol* 2021;**11**:623506, https://doi.org/10.3389/fonc.2021.623506.
18. Fujioka T, Kubota K, Hsu JF, *et al*. Examining the effectiveness of a deep learning-based computer-aided breast cancer detection system for breast ultrasound. *J Med Ultrason* 2023;**50**:511–20. https://doi.org/10.1007/s10396-023-01332-9.
19. Ding W, Wang J, Zhou W, *et al*. Joint localization and classification of breast cancer in B-mode ultrasound imaging via collaborative learning with elastography. *IEEE J Biomed Health Inform* 2022;**26**:4474–85. https://doi.org/10.1109/JBHI.2022.3186933.
20. Wang Q, Chen H, Luo G, *et al*. Performance of novel deep learning network with the incorporation of the automatic segmentation network for diagnosis of breast cancer in automated breast ultrasound. *Eur Radiol* 2022;**32**:7163–72. https://doi.org/10.1007/s00330-022-08836-x.
21. Xiao X, Gan F, Yu H. Tomographic ultrasound imaging in the diagnosis of breast tumors under the guidance of deep learning algorithms. *Comput Intell Neurosci* 2022;**2022**:1–7. https://doi.org/10.1155/2022/9227440.
22. Qian X, Pei J, Zheng H, *et al*. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nat Biomed Eng* 2021;**5**:522–32. https://doi.org/10.1038/s41551-021-00711-2.
23. Yu T-F, He W, Gan C-G, *et al*. Deep learning applied to two-dimensional color Doppler flow imaging ultrasound images significantly improves diagnostic performance in the classification of breast masses: a multicenter study. *Chin Med J (Engl)* 2021;**134**:415–24. https://doi.org/10.1097/CM9.0000000000001329.
24. Kim MY, Kim S-Y, Kim YS, *et al*. Added value of deep learning-based computer-aided diagnosis and shear wave elastography to b-mode ultrasound for evaluation of breast masses detected by screening ultrasound. *Medicine (Baltimore)* 2021;**100**:e26823, https://doi.org/10.1097/MD.0000000000026823.
25. Wang Y, Choi EJ, Choi Y, *et al*. Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning. *Ultrasound Med Biol* 2020;**46**:1119–32. https://doi.org/10.1016/j.ultrasmedbio.2020.01.001.
26. Liao W-X, He P, Hao J, *et al*. Automatic identification of breast ultrasound image based on supervised block-based region segmentation algorithm and features combination migration deep learning model. *IEEE J Biomed Health Inform* 2020;**24**:984–93. https://doi.org/10.1109/JBHI.2019.2960821.
27. Li C, Zhang H, Chen J, *et al*. Deep learning radiomics of ultrasonography for differentiating sclerosing adenosis from breast cancer. *Clin Hemorheol Microcirc* 2023;**84**:153–63. https://doi.org/10.3233/CH-221608.
28. Fei X, Zhou S, Han X, *et al*. Doubly supervised parameter transfer classifier for diagnosis of breast cancer with imbalanced ultrasound imaging modalities. *Pattern Recognit* 2021;**120**:108139, https://doi.org/10.1016/j.patcog.2021.108139.
29. Higgins JPT, Thompson SG, Deeks JJ, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**:557–60. https://doi.org/10.1136/bmj.327.7414.557.
30. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? *Stat Med* 2009;**28**:2653–68. https://doi.org/10.1002/sim.3631.
31. Peters JL, Sutton AJ, Jones DR, *et al*. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *J Clin Epidemiol* 2008;**61**:991–6. https://doi.org/10.1016/j.jclinepi.2007.11.010.
32. Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. *Front Med* 2020;**7**:27. https://doi.org/10.3389/fmed.2020.00027.
33. Chan H-P, Samala RK, Hadjiiski LM. CAD and AI for breast cancer—recent development and challenges. *Br J Radiol* 2020;**93**:20190580, https://doi.org/10.1259/bjr.20190580.
34. Dan Q, Xu Z, Burrows H, *et al*. Diagnostic performance of deep learning in ultrasound diagnosis of breast cancer: a systematic review. *Npj Precis Oncol* 2024;**8**:21. https://doi.org/10.1038/s41698-024-00514-z.
35. Burt JR, Torosdagli N, Khosravan N, *et al*. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. *Br J Radiol* 2018:20170545, https://doi.org/10.1259/bjr.20170545.