Original Article

Using the GoogLeNet deep-learning model to distinguish between benign and malignant breast masses based on conventional ultrasound: a systematic review and meta-analysis

Jinli Wang[1#], Jin Tong[1#], Jun Li[1], Chunli Cao[1], Sirui Wang[1], Tianyu Bi[2], Peishan Zhu[1], Linan Shi[1], Yaqian Deng[1], Ting Ma[1], Jixue Hou[1], Xinwu Cui[3]

[1]Department of Ultrasound, the First Affiliated Hospital of Medical College, Shihezi University, Shihezi, China; [2]School of Business Administration, Lanzhou University of Finance and Economics, Lanzhou, China; [3]Department of Medical Ultrasound, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Contributions: (I) Conception and design: J Wang, P Zhu; (II) Administrative support: J Li, X Cui; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: C Cao, J Wang, S Wang, L Shi, Y Deng; (V) Data analysis and interpretation: J Wang, T Bi, S Wang, T Ma, J Tong, J Hou; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Jun Li, MD, PhD. Department of Ultrasound, the First Affiliated Hospital of Medical College, Shihezi University, No. 107 North 2nd Road, Shihezi 832008, China. Email: 1287424798@qq.com; Xinwu Cui, MD, PhD. Department of Medical Ultrasound, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No. 288 Xintian Avenue, Caidian District, Wuhan 430101, China. Email: cuixinwu@live.cn.

Background: Breast cancer is one of the most common malignancies in women worldwide, and early and accurate diagnosis is crucial for improving treatment outcomes. Conventional ultrasound (CUS) is a widely used screening method for breast cancer; however, the subjective nature of interpreting the results can lead to diagnostic errors. The current study sought to estimate the effectiveness of using a GoogLeNet deep-learning convolutional neural network (CNN) model to identify benign and malignant breast masses based on CUS.

Methods: A literature search was conducted of the Embase, PubMed, Web of Science, Wanfang, China National Knowledge Infrastructure (CNKI), and other databases to retrieve studies related to GoogLeNet deep-learning CUS-based models published before July 15, 2023. The diagnostic performance of the GoogLeNet models was evaluated using several metrics, including pooled sensitivity (PSEN), pooled specificity (PSPE), the positive likelihood ratio (PLR), the negative likelihood ratio (NLR), the diagnostic odds ratio (DOR), and the area under the curve (AUC). The quality of the included studies was evaluated using the Quality Assessment of Diagnostic Accuracy Studies Scale (QUADAS). The eligibility of the included literature were independently searched and assessed by two authors.

Results: All of the 12 studies that used pathological findings as the gold standard were included in the meta-analysis. The overall average estimation of sensitivity and specificity was 0.85 [95% confidence interval (CI): 0.80–0.89] and 0.86 (95% CI: 0.78–0.92), respectively. The PLR and NLR were 6.2 (95% CI: 3.9–9.9) and 0.17 (95% CI: 0.12–0.23), respectively. The DOR was 37.06 (95% CI: 20.78–66.10). The AUC was 0.92 (95% CI: 0.89–0.94). No obvious publication bias was detected.

Conclusions: The GoogLeNet deep-learning model, which uses a CNN, achieved good diagnostic results in distinguishing between benign and malignant breast masses in CUS-based images.

Keywords: GoogLeNet; deep learning; meta-analysis; breast mass; ultrasound (US)

Submitted Apr 01, 2024. Accepted for publication Aug 19, 2024. Published online Sep 26, 2024.
doi: 10.21037/qims-24-679
View this article at: https://dx.doi.org/10.21037/qims-24-679

Introduction

Breast cancer recently surpassed lung cancer as the most commonly diagnosed cancer in women, and in 2022, approximately 2.3 million new breast cancer cases were diagnosed (1-3). Breast cancer seriously affects women's health and quality of life, and is also the fifth major cause of cancer-related death worldwide (2,3). However, with early cancer screening, early diagnosis, and prompt treatment, patient survival and quality of life can be significantly improved. Therefore, it is essential to detect breast cancer in its early stages and initiate treatment to minimize mortality (4-6).

Medical imaging technology can improve the diagnostic accuracy of breast cancer and reduce unnecessary biopsy times (7,8). Ultrasound (US) images and X-ray mammography are commonly used to identify cancers. US is a widely used, non-ionizing imaging technique that is low cost, radiation free, can be observed in real time, and can also be used as an adjunct to X-ray mammography, especially for dense breasts (2,3,5-7,9). A large number of breast US images are generated every day in the daily practice of healthcare organizations. The accurate assessment of these images depends heavily on physicians' ability to recognize and identify image features, which requires sonographers to be experienced in image analysis to ensure reliable diagnostic conclusions. However, the subjectivity of a physician's experience can lead to misdiagnosis or delayed diagnosis breast tumors. Therefore, the accuracy of breast US diagnosis by sonographers urgently needs to be improved (10).

Computer-aided diagnosis (CAD) systems using breast US images hold significant research value and have shown promising application prospects (11). In recent decades, there have been efforts to enhance breast US analysis using computer-aided technology. Extensive research has been conducted on breast US image analysis and intelligent diagnosis to accurately differentiate between benign and malignant breast masses (12,13). One approach uses traditional machine learning and algorithms to define and extract image features, after which a specific classifier is used to classify breast tumors according to thethe meaningful characteristics extracted. Another approach uses deep learning to achieve diagnosis by training models on relatively large data sets.

Traditional classification methods involve manually segmenting an image and then using a classifier or one trained by a shallow neural computer to identify each segment and classify the image (2,14). However, methods used to build and improve classifiers are time consuming and computationally heavy, and the diagnostic performance of such systems largely depends on the quality of the features extracted by the algorithm (15). To overcome these early machine-learning limitations, researchers began using deep learning to identify images. With deep learning, artificial neural networks can extract data-driven and self-optimized feature graphs of the most discriminative features from the input images, and provide corresponding answers or predictions. Therefore, feature detection and selection are not commonly required (16,17). As deep learning can accurately extract meaningful characteristics from images and autonomously calculate inference and decision making, this learning method can diagnose images and is not dependent on the experience of radiologists (18-21).

Recently, convolutional neural networks (CNNs) have gained popularity for their image pattern identification and artificial intelligence (AI) strategies. CNNs are a method of deep learning and are inspired by the structure and function of the brain. In processing data using artificial neural networks containing concealed layers, CNNs imitate the visual cortex of mammals. CNNs are powerful visualization models that can generate hierarchical structures of features (15-18). Research studies have shown that CNNs, through an end-to-end, pixel-to-pixel process, surpass state-of-the-art semantic segmentation techniques in medical image recognition (22-24). These findings, based on multiple research studies conducted in this field, highlight the superior performance of CNNs.

GoogLeNet is a widely used CNN algorithm model for breast US image classification. It performed exceptionally well in the ImageNet ILSVRC14 detection and classification challenge (25). The model provides the top five most likely classification results for a given image, ranked by confidence level. The test accuracy for these top five results is 93.3% (26). The main innovation of GoogLeNet lies in its efficient use of computational

resources. It introduces and incorporates a structure called the "inception module" to approximate sparse connections between activation functions. This module helps eliminate redundancy and correlation between activation functions, thereby improving efficiency in memory usage and execution time without compromising accuracy (27).

Inception V1 and Inception V3 are different versions of the Inception architecture, and were both developed by the Google Research team (28,29). The main difference between these versions is the architectural design and performance improvements, which means that the depth and complexity of these versions differ. Inception V1, also known as GoogLeNet, was launched in 2014 (26). It consists of multiple Inception modules that capture information at different scales by using parallel convolution layers with different filter sizes, with 1×1 convolution for dimensionality reduction. Inception V1 has a relatively simple architecture compared to later versions. Inception V3 was released in 2015 and is a more advanced version of the Inception architecture (30). It was designed to improve on the limitations of Inception V1. It has various improvements over Inception V1; for example, it uses factoring convolution to reduce computational costs, and batch normalization to help with training stability. Inception V3 also has a deeper network structure that allows it to capture more complex features. However, as Inception V1 has a simple network structure and fewer parameters, it is easier to implement and deploy in the case of limited computing resources. Compared to Inception V3, it has higher computational efficiency and faster learning speed (28,30). Therefore, the choice of which version of Inception to use should be evaluated based on specific needs and conditions. If computational resources are limited, datasets are small, or there are limitations on model size, Inception V1 may be a good choice. However, with large datasets, higher accuracy, and better performance requirements, Inception V3 may be a better choice (28-30).

Using CNN models trained on non-medical ImageNet data for medical image analysis is a trend that has recently emerged (31). The review article by Morid *et al.* noted that this transfer learning approach is very common in medical image analysis, with the Inception-V3 CNN model being the most commonly used (32). In various types of medical image analyses, including X-ray, endoscopic, and US image analyses, the GoogLeNet model is the most commonly used, accounting for 19% of all models applied in these analyses. Moreover, GoogLeNet, which has been used in 50% of breast-related research, is the most commonly used model (32). Given the broad application and high effectiveness of the GoogLeNet model in breast US image classification, and the absence of a previous meta-analysis on this topic, it was chosen as the reference model for this meta-analysis. To enhance the stability and reliability of our research findings and to avoid selection bias, we specifically chose the GoogLeNet model as the subject of our study to evaluate its performance in the task of classifying breast tumors as benign or malignant.

Currently, several studies have shown that the deep-learning US GoogLeNet model can effectively differentiate between benign and malignant breast masses, improving the accuracy of diagnoses (6,15,33-35). However, the sensitivity of these models varies among different studies. Six publications used the Inception deep-learning V1 model (6,33,35-38), while another six publications used the Inception V3 deep-learning model (15,34,39-42). The study by Kriti *et al.* reported a sensitivity of 97% for the Inception V1 model (38), while that Ali *et al.* reported a sensitivity of only 74% for the Inception V3 model (39). The sensitivity of GoogLeNet deep-learning US-based models for diagnosing breast lesions varies substantially, and no meta-analysis on this topic appears to have been conducted. Therefore, this meta-analysis sought to assess the efficacy of the conventional ultrasound (CUS)-based GoogLeNet deep-learning model in discriminating between and diagnosing the nature of breast masses to assist sonographers to make more accurate diagnoses. We present this article in accordance with the PRISMA-DTA reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-679/rc).

Methods

This meta-analysis was registered on the PROSPERO website (registration number CRD42023459415).

Search strategy

A literature search of databases, such as Embase, PubMed, Web of Science, Wanfang, China National Knowledge Infrastructure (CNKI), and Cochrane Library databases, was conducted to retrieve all relevant studies published before July 15, 2023. The following broad keywords were used: "Deep learning" or "DL" or "Neural network" or "GoogLeNet" or "Inception V1" or "Inception V3" and "ultrasonography" or "ultrasound" or "ultrasonic" or "diagnostic imaging", and "breast nodules" or "breast

masses". To achieve a comprehensive and accurate literature search, this study employed a comprehensive search approach, including the use of subject terms and free terms for the online retrieval, as well as a manual search to supplement and refine the retrieved relevant literature. This approach ensured the acquisition of a broad range of research resources.

Study selection

To be eligible for inclusion in this meta-analysis, the articles had to meet the following inclusion criteria: (I) employ the GoogLeNet deep-learning model to perform the discriminative diagnosis of benign and malignant breast masses; (II) inclusively analyze the diagnosis of breast masses by CUS; (III) include raw data that either directly demonstrated or could be used to calculate sensitivity and specificity, and included data that provided information on true positives (TPs), false positives (FPs), false negatives (FNs), and true negatives (TNs); (IV) use pathological examination as the gold standard for the diagnosis of breast masses and clearly detailed the number of samples included and the respective counts of the different types of samples; and (V) have collected either test set data or validation set data; if both were collected, the test set data were chosen for the analysis.

Articles were excluded from the meta-analysis if they met any of the following exclusion criteria: (I) were not related to the GoogLeNet deep-learning model of CNN; (II) did not have data available that could be used to calculate the TPs, FPs, FNs, and TNs; (III) comprised review articles, letters, editorials, commentaries, theses, case reports, and conference articles, and so on; and/or (IV) were related to duplicate studies.

Data extraction and quality assessment

In this study, the titles and abstracts were reviewed independently by two authors to identify the qualifying articles. The full texts of the articles were then read to identify articles for inclusion in the meta-analysis. For each study, relevant information was independently extracted, including the first author, year of publication, country, sample volume in the training set, sample volume in the test set, quadruple table data (TPs, FPs, FNs, and TNs), sensitivity, specificity and the type of GoogLeNet model used. Studies were excluded if quadruple table data could not be extracted from thearticle.

The same two observers used the Quality Assessment of Diagnostic Accuracy Studies Scale (QUADAS-2) to assess the quality of the included articles. RevMan 5.4 (Cochrane Collaboration) was used to output the result of QUADAS-2.

Statistical analysis

After data extraction, we evaluated the pooled sensitivity (PSEN) and pooled specificity (PSPE) using bivariate models, and summary receiver operating characteristic (SROC) curves were plotted with the areas under the curve (AUCs). Moreover, publication bias was assessed using Deeks' funnel plots, and the risk of study bias was evaluated using the QUADAS-2 criteria. Further, post-test probabilities were calculated and represented using Fagan's plot. All the data analyses were conducted and all the graphs were generated using Stata 17 (StataCorp LLC) and RevMan 5.4 (Cochrane Collaboration) software.

Based on study results, we assessed heterogeneity quantitatively. If the Q-test results met the criteria of $P>0.1$ and $I^2 \leq 50\%$, a fixed-effects model was used; otherwise, a random-effects model was used. Meta-regression and subgroup analyses were conducted to assess the reasons for clinical heterogeneity. A P value <0.05 was considered statistically significant.

Results

Literature searches results

As of July 2023, we initially retrieved 302 original articles based on our search criteria. A meticulous review of the titles and abstracts of these articles resulted in the preliminary selection of 197 articles. These were articles further subjected to a rigorous assessment based on predefined inclusion and exclusion criteria, culminating in the selection of 12 articles that met the requirements of our meta-analysis. A detailed workflow of the literature selection process is shown in *Figure 1*.

Characteristics of the eligible studies

Table 1 presents the main features of and general information about the 12 articles included in this meta-analysis. The studies were published between 2017 and 2023. Six articles employed the Inception V1 deep-learning model (6,33,35-38). The training and testing datasets

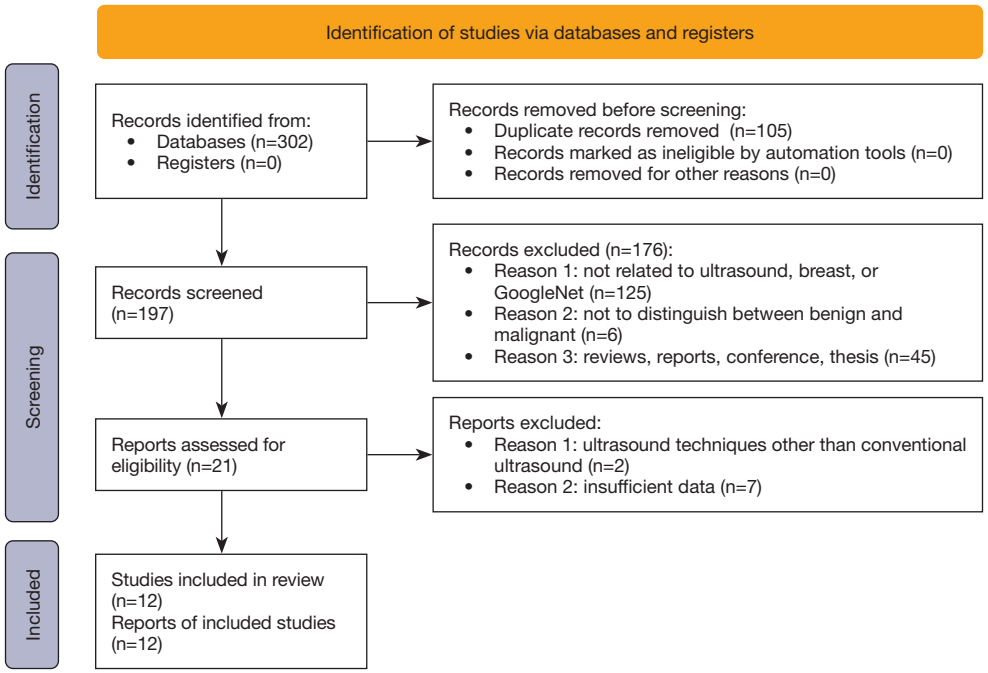

Figure 1 Study flow chart detailing the reasons for exclusion of studies and the total number (n=12) of studies included.

Table 1 Characteristics of the included studies

Author	Year	Country	Training database			Test database			SE	SP	TP	FP	FN	TN	Precision	F1 score	GoogLeNet (type)
			N	B	M	N	B	M									
Han *et al.* (33)	2017	Korea	–	3,765	2,814	–	489	340	0.83	0.95	282	24	58	465	0.92	0.87	Inception V1
Xiao *et al.* (40)	2018	China	–	1,233	619	–	137	69	0.77	0.89	53	15	16	122	0.78	0.77	Inception V3
Zhang *et al.* (41)	2020	China	–	2,500	2,500	–	788	219	0.86	0.82	188	142	31	646	0.86	0.69	Inception V3
Kim *et al.* (37)	2021	Korea	–	2,140	1,440	–	100	100	0.88	0.76	88	24	12	76	0.88	0.83	Inception V1
Yu *et al.* (34)	2022	China	–	500	500	–	48	52	0.81	0.9	42	5	10	43	0.81	0.85	Inception V3
Assari *et al.* (6)	2022	Iran	–	256	226	–	79	77	0.88	0.91	68	7	9	72	0.88	0.89	Inception V1
Sirjani *et al.* (42)	2023	Iran	–	597	319	–	71	79	0.75	0.73	59	19	20	52	0.75	0.75	Inception V3
Kriti *et al.* (38)	2020	India	–	3,982	3,751	–	21	30	0.97	0.9	29	21	1	9	0.97	0.73	Inception V1
Ali *et al.* (39)	2023	Saudi Arabia	–	3,500	3,500	–	1,000	1,000	0.74	0.93	740	73	260	927	0.74	0.82	Inception V3
Masud *et al.* (35)	2022	Saudi Arabia	532	2,148	1,440	133	537	360	0.94	0.88	338	64	22	473	0.94	0.88	Inception V1
Tsai *et al.* (15)	2022	China	93	341	147	27	97	42	0.95	0.96	40	4	2	93	0.95	0.93	Inception V3
Alhussan *et al.* (36)	2023	Saudi Arabia	4,000	4,000	4,000	77	244	105	0.78	0.86	82	34	23	210	0.78	0.74	Inception V1

SE, sensitivity; SP, specificity; M, Malignant; B, Benign; N, Normal; TP, true positive; FP, false positive; FN, false negative; TN, true negative.

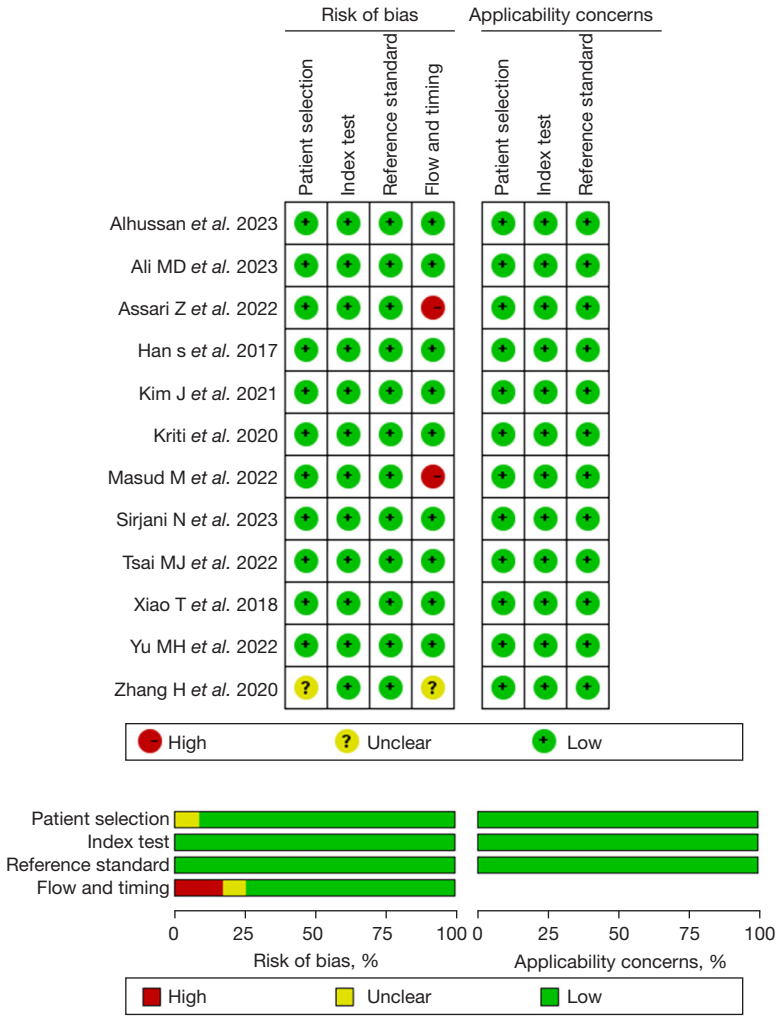

Figure 2 Bias risk of the included studies (based on the QUADAS-2 criteria). Judgements of the review authors about each domain for each included study.

in three articles were augmented with normal images (15,35,36). Seven articles used publicly available datasets (6,15,35,36,38,39,42), four articles used data collected by a single institution (33,34,40,41), and one article used data collected from two centers (37).

Methodology quality assessment

Using RevMan 5.4 software, methodological assessments based on the QUADAS-2 checklist were performed to determine the quality of included studies. As *Figure 2* shows, the majority of the included studies had a low risk of bias in terms of the quality assessment items.

Accuracy of the GoogLeNet model based on US for deep learning in the differential detection of benign and malignant breast masses

Our data analysis revealed that the application of the GoogLeNet deep-learning model, based on US, achieved a PSEN of 0.85 [95% confidence interval (CI): (0.80–0.89)] and a PSPE of 0.86 (95% CI: 0.78–0.92) in distinguishing between benign and malignant breast masses (*Figure 3*). Higgins I^2 statistics revealed significant heterogeneity in terms of sensitivity ($P<0.05$, $I^2=88.36\%$) and specificity ($P<0.05$, $I^2=94.63\%$). Accordingly, the random-effects model was used to analyze the sensitivity and specificity.

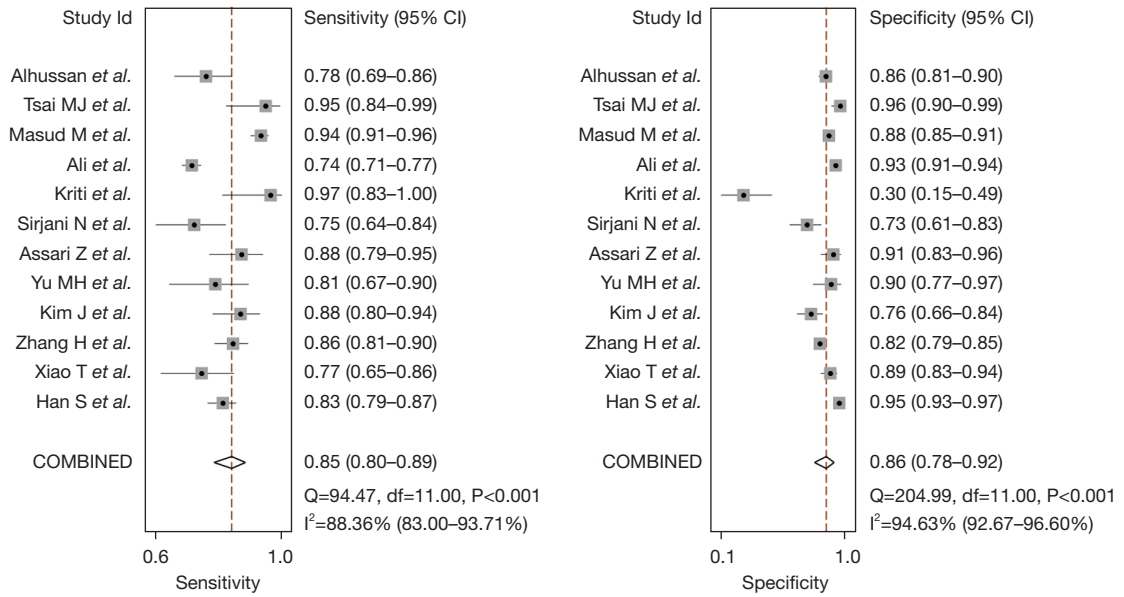

Figure 3 Forest plots showing model sensitivity and specificity for diagnostic breast masses. The horizontal lines illustrate the 95% confidence intervals of the individual studies. CI, confidence interval.

The positive likelihood ratio (PLR) and negative likelihood ratio (NLR) were 6.2 (95% CI: 3.9–9.9) and 0.17 (95% CI: 0.12–0.23), respectively. The diagnostic odds ratio (DOR) was 37.06 (95% CI: 20.78–66.10) (*Figure 4*), and the AUC was 0.92 (95% CI: 0.89–0.94) (*Figure 5*). The Spearman's correlation coefficient (r=–0.31, P=0.10) indicated that extra threshold factors might have contributed to the observed heterogeneity. The results of the above statistical analyses were considered acceptable.

Publication bias

The Deek funnel plot plotted by Stata 17.0 showed a symmetrical distribution of the study, with a P value of 0.70 (P>0.05) (*Figure 6*), which indicated that there was no apparent publication bias in this study.

Heterogeneity detection

In view of the strong heterogeneity among the incorporated studies, this study used meta-regression to analyze the factors related to heterogeneity. The following variables were analyzed: the type of deep-learning architecture (GoogLeNet), the number of breast US image classifications, and the study publication year (≤2021 or >2021). The results of the regression analysis are set out in *Table 2*. Of the variables, the PSEN of Inception V1 was 0.89 (95% CI: 0.84–0.93), and that of Inception V3 was 0.81 (95% CI: 0.75–0.88), P<0.01; the PSPE of Inception V1 was 0.83 (95% CI: 0.73–0.94), and that of Inception V3 was 0.89 (95% CI: 0.81–0.96), P=0.03. Both were statistically significant. The PSEN for studies that included normal, benign, and malignant US images was 0.90 (95% CI: 0.84–0.93), while that for the subgroup that included only benign and malignant images was 0.84 (95% CI: 0.78–0.89) (P<0.01), indicating a statistically significant difference. The PSPE for the studies that included normal, benign, and malignant US images was 0.91 (95% CI: 0.82–1.00), and that for the subgroup that included only benign and malignant images was 0.84 (95% CI: 0.76–0.93) (P=0.06), but no statistically significant difference was found. The PSEN for studies published in and before 2021 was 0.87 (95% CI: 0.80–0.93), and that for studies published after 2021 was 0.85 (95% CI: 0.79–0.91) (P=0.01). The PSPE for studies published in and before 2021 was 0.81 (95% CI: 0.68–0.93), and that for studies published after 2021 was 0.89 (95% CI: 0.83–0.96) (P=0.01). The differences between the PSEN and PSPE before and after 2021 were statistically significant.

Sensitivity analysis

To investigate whether any study affected the stability of the PSEN and PSPE, we eliminated the included studies one

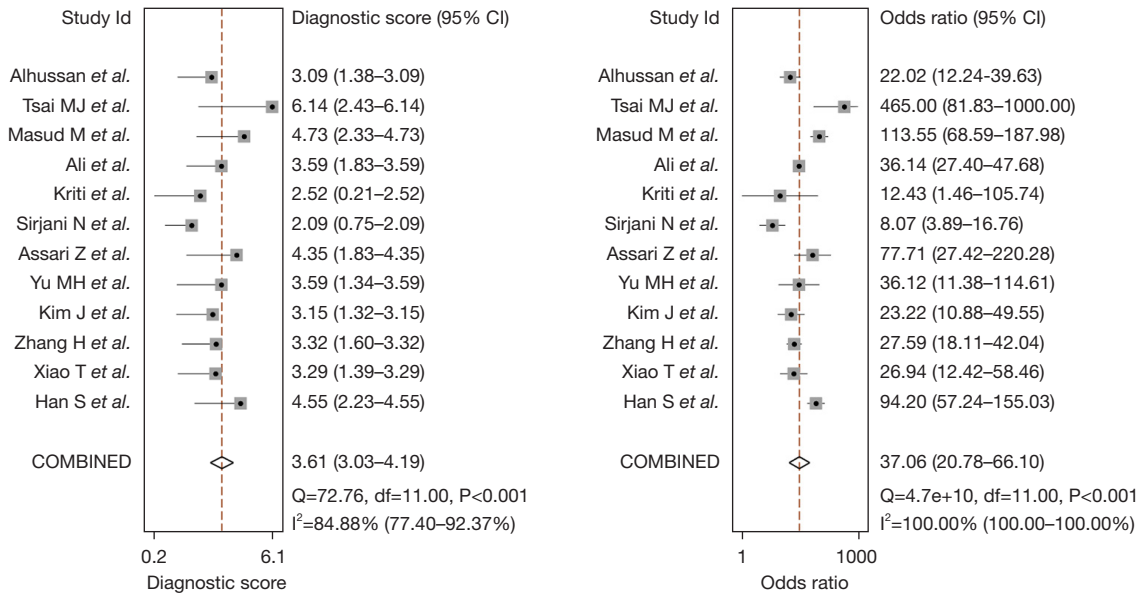

Figure 4 The DORs for the diagnosis of breast masses. The horizontal lines illustrate the 95% confidence intervals of the individual studies. CI, confidence interval; DORs, diagnostic odds ratios.

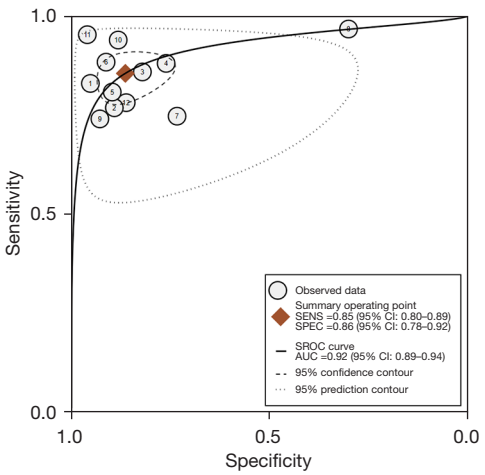

Figure 5 The ROC. 1 represents the study of Han *et al.*; 2 represents the study of Xiao *et al.*; 3 represents the study of Zhang *et al.*; 4 represents the study of Kim *et al.*; 5 represents the study of Yu *et al.*; 6 represents the study of Assari *et al.*; 7 represents the study of Sirjani *et al.*; 8 represents the study of Kriti *et al.*; 9 represents the study of Ali *et al.*; 10 represents the study of Masud *et al.*; 11 represents the study of Tsai *et al.*; and 12 represents the study of Alhussan *et al.* SENS, sensitivity; SPEC, specificity; CI, confidence interval; SROC, summary receiver operating characteristic curve; AUC, area under the curve; ROC, receiver operating characteristic curve.

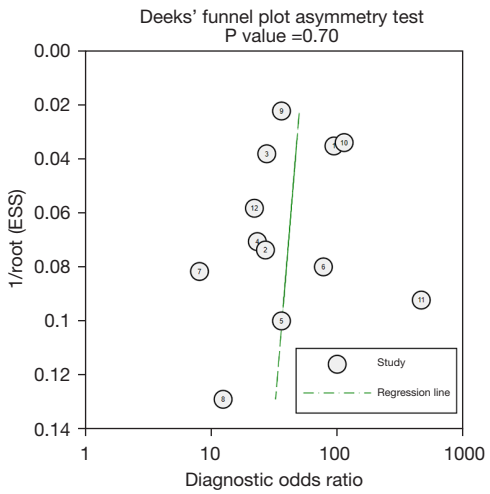

Figure 6 The publication bias of the included studies. No significant publication bias was found in the present meta-analysis. Each circle represents an eligible research study. 1 represents the study of Han *et al.*; 2 represents the study of Xiao *et al.*; 3 represents the study of Zhang *et al.*; 4 represents the study of Kim *et al.*; 5 represents the study of Yu *et al.*; 6 represents the study of Assari *et al.*; 7 represents the study of Sirjani *et al.*; 8 represents the study of Kriti *et al.*; 9 represents the study of Ali *et al.*; 10 represents the study of Masud *et al.*; 11 represents the study of Tsai *et al.*; and 12 represents the study of Alhussan *et al.* 1/root (ESS), square root of the reciprocal of ESS; ESS, effective sample size.

Table 2 Meta-analysis of ultrasound-based deep learning for the differential diagnosis of benign and malignant breast masses

Category	N	PSEN		PSPE	
		SE (95% CI)	P	SP (95% CI)	P
GoogLeNet			<0.01		0.03
Inception V1	6	0.89 (0.84–0.93)		0.83 (0.73–0.94)	
Inception V3	6	0.81 (0.75–0.88)		0.89 (0.81–0.96)	
Number of Classes			<0.01		0.06
2	9	0.84 (0.78–0.89)		0.84 (0.76–0.93)	
3	3	0.90 (0.84–0.93)		0.91 (0.82–1.00)	
Year			0.01		0.01
≤2021	5	0.87 (0.80–0.93)		0.81 (0.68–0.93)	
>2021	7	0.85 (0.79–0.91)		0.89 (0.83–0.96)	
Data set class			0.03		0.03
Extra	8	0.87 (0.82–0.92)		0.82 (0.75–0.93)	
Intra	4	0.82 (0.74–0.91)		0.90 (0.82–0.98)	

N, number of included studies; PSEN, pooled sensitivity; SE, sensitivity; CI, confidence interval; PSPE, pooled specificity; SP, specificity.

by one. The sensitivity and specificity analysis results are set out in *Table 3*. The results revealed no significant change in the PSEN and PSPE or in Higgins I^2, with the exception of non-significant changes noted in individual studies.

Fagan plot analysis

The Fagan plot analysis showed that the GoogLeNet deep-learning model when applied to US images can assist radiologists to differentiate between benign and malignant breast lesions. When the pre-test probability was 25%, a "positive" GoogLeNet result increased the probability of a correct diagnosis to 68%, while a "negative" result decreased the probability of a correct diagnosis to 5% (*Figure 7A*). With pre-test probabilities of 50% and 75%, a "positive" test result changed the post-test probability to 86% and 95%, respectively, while a "negative" test result changed it to 14% and 34%, respectively (*Figure 7B*,*7C*).

Discussion

This study conducted a meta-analysis to evaluate and analyze the value of the GoogLeNet deep-learning model, which is based on US imaging, in the differential diagnosis of benign and malignant breast masses. The PSEN and PSPE of the 12 included studies were 0.85 (95% CI: 0.80–0.89) and 0.86 (95% CI: 0.78–0.92), respectively. The DOR was 37.06 (95% CI: 20.78–66.10), and the AUC was 0.92 (95% CI: 0.89–0.94). These results showed that the GoogLeNet deep-learning model based on US has high diagnostic accuracy in differentiating between benign and malignant breast masses. We prioritized the inclusion of test set data when both test and training sets were available in the study, as test sets can be used to assess the ability of models to generalize over unseen data, help in the selection of the most appropriate model or method, ensure a more accurate assessment of the meta-analysis, and guarantee the scalability of findings.

All the research reports included in our meta-analysis were of relatively good quality, and no obvious publication bias was observed. A lack of uniformity in the gold-standard criteria in a few studies resulted in a slight decline in the quality of some reports. However, there was significant heterogeneity in the sensitivity and specificity of the studies included. This might be partly due to some study data being sourced from a single institution (33,34,40,41), which could have resulted in the high level of heterogeneity. Additionally, it could be related to differences in the study designs, demographic statistics, or imaging acquisition techniques. The meta-regression and subgroup analysis results identified the type of GoogLeNet model, the number of categories in the breast US images, the data set class, and the study publication year as the main sources of heterogeneity. A detailed breakdown of heterogeneity

Table 3 The sensitivity analysis in which articles were eliminated one by one

Eliminated article	PSEN			PSPE			AUC (95% CI)
	SE (95% CI)	I^2 (95% CI), %	P	SP (95% CI)	I^2 (95% CI), %	P	
Han *et al.* (33)	0.86 (0.80–0.90)	89.31 (84.28–94.34)	<0.01	0.85 (0.76–0.91)	93.99 (91.61–96.38)	<0.01	0.92 (0.89–0.94)
Xiao *et al.* (40)	0.86 (0.81–0.90)	89.39 (84.41–94.37)	<0.01	0.86 (0.77–0.92)	95.12 (93.30–96.94)	<0.01	0.92 (0.89–0.94)
Zhang *et al.* (41)	0.86 (0.80–0.90)	88.79 (83.44–94.13)	<0.01	0.87 (0.78–0.92)	94.60 (92.52–96.67)	<0.01	0.92 (0.89–0.94)
Kim *et al.* (37)	0.85 (0.80–0.90)	88.97 (83.73–94.20)	<0.01	0.87 (0.79–0.92)	94.95 (93.04–96.85)	<0.01	0.92 (0.89–0.94)
Yu *et al.* (34)	0.86 (0.80–0.90)	89.40 (84.43–94.37)	<0.01	0.86 (0.77–0.92)	95.10 (93.27–96.93)	<0.01	0.92 (0.89–0.94)
Assari *et al.* (6)	0.85 (0.80–0.90)	88.90 (83.63–94.18)	<0.01	0.86 (0.77–0.92)	94.97 (93.08–96.86)	<0.01	0.92 (0.89–0.94)
Sirjani *et al.* (42)	0.86 (0.81–0.90)	89.46 (84.52–94.40)	<0.01	0.87 (0.79–0.92)	95.03 (93.17–96.89)	<0.01	0.93 (0.90–0.95)
Kriti *et al.* (38)	0.84 (0.79–0.88)	89.28 (84.23–94.32)	<0.01	0.89 (0.84–0.92)	91.64 (87.98–95.30)	<0.01	0.93 (0.90–0.95)
Ali *et al.* (39)	0.86 (0.81–0.88)	78.48 (66.21–90.75)	<0.01	0.85 (0.76–0.91)	93.37 (90.65–96.08)	<0.01	0.92 (0.89–0.94)
Masud *et al.* (35)	0.84 (0.79–0.87)	78.71 (66.60–90.81)	<0.01	0.86 (0.77–0.92)	95.23 (93.46–96.99)	<0.01	0.90 (0.87–0.92)
Tsai *et al.* (15)	0.85 (0.80–0.89)	88.52 (83.01–94.03)	<0.01	0.85 (0.76–0.91)	94.69 (92.66–96.72)	<0.01	0.91 (0.88–0.93)
Alhussan *et al.* (36)	0.86 (0.81–0.90)	89.44 (84.50–94.39)	<0.01	0.86 (0.77–0.92)	95.17 (93.37–96.96)	<0.01	0.92 (0.89–0.94)

PSEN, pooled sensitivity; SE, sensitivity; CI, confidence interval; PSPE, pooled specificity; SP, specificity; AUC, area under the curve.

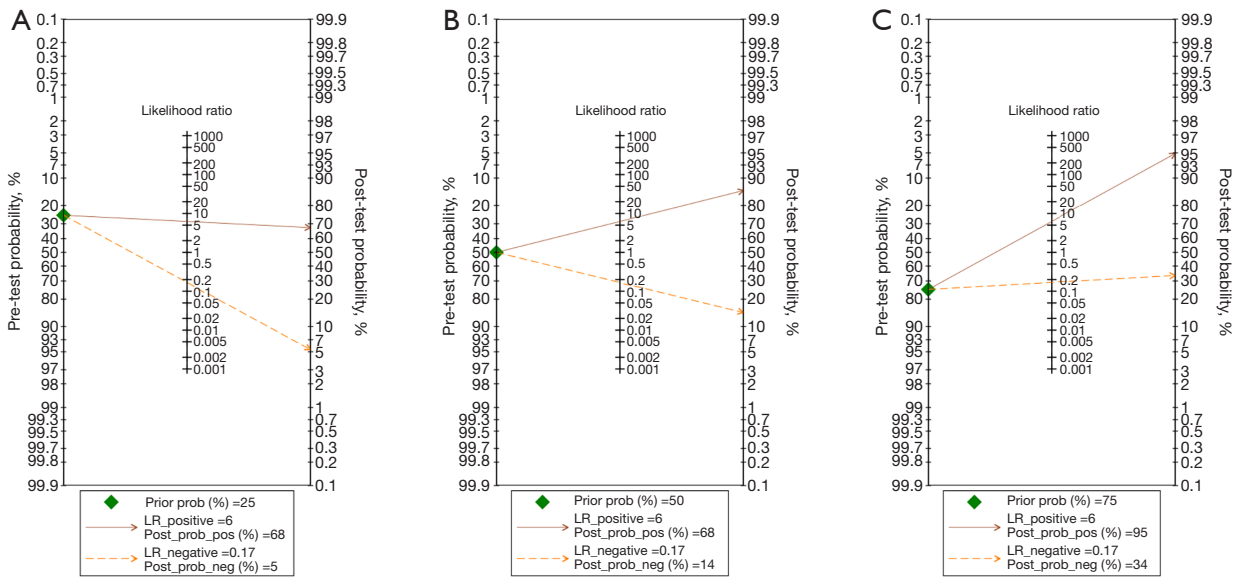

Figure 7 Fagan plot analysis examining the ability of the GoogLeNet model to detect breast masses: (A) pre-test probability at 25%; (B) pre-test probability at 50%; (C) pre-test probability at 75%. The Fagan plot is composed of the left vertical axis representing the pre-test probability, the middle vertical axis representing the likelihood ratio, and the right vertical axis representing the post-test probability. Prob, probability; LR, likelihood ratio; pos, positive; neg, negative.

caused by these factors is presented below.

First, this meta-analysis included five articles published in and before 2021 (33,35,37,38,40,41), and seven articles published after 2021 (6,15,34,36,39,42). The sensitivity of the articles published after 2021 was lower than that of the articles published in and before 2021 (0.85 *vs.* 0.87) (P=0.01). The specificity of the articles published after 2021 was higher than that of the articles published in and before

2021 (0.89 *vs.* 0.81) (P=0.01). The results were statistically significant. To decrease the misdiagnosis rate and improve the accuracy of benign and malignant classification of breast masses, some studies after 2021 introduced new methods and techniques, and while these methods and techniques might have improved the specificity of the models, they also reduced their sensitivity to a certain extent, resulting in an increase in PSPE. Conversely, the gross numbers of benign masses included in the articles after 2021 were more than those included in the articles in and before 2021, and more normal images were also included, which improved the specificity of the models. Most of the studies after 2021 used publicly available datasets that typically contain samples from different sources, different fields, and different characteristics. Due to the broader diversity of data, more malignant images of different pathologic types might have been included, reducing the sensitivity of the models.

Second, in the selection of the ultrasonic images, the different types of images included represented one of the major factors affecting the heterogeneity of this study. Nine articles used images that included both benign and malignant images (6,33,34,37-42), and three articles used images that included normal, benign, and malignant images (15,35,36). The subgroup analysis revealed a statistically significant variation in sensitivity (0.84 *vs.* 0.90) (P<0.05). This might be due to the increased ability of the models to distinguish between categories. The inclusion of normal breast US images enables models to study the characteristics of normal breast tissue and compare them to lumps. This can help the model better understand differences in tumor characteristics, morphology and texture, and improve the sensitivity of classification. The inclusion of normal images in the training data improved the sensitivity of models in some studies; however, this also carries the risk of models overfitting to an unrealistic data distribution compared to clinical populations where most assessed cases have some abnormality. Careful external validation is needed to reduce potential overfitting concerns.

Third, the choice of dataset also affected the results of our meta-analysis. Eight studies used external validation sets, while four used internal validation sets. The subgroup analysis revealed significant differences in sensitivity and specificity (0.87 *vs.* 0.82, P=0.03; 0.82 *vs.* 0.90, P=0.03). Differences across the validation sets might stem from models adequately learning certain features or patterns in the dataset during training, which might lead to overfitting and decreased generalization on unseen data. External validation sets perform better in terms of sensitivity, as they better simulate real-world data distributions, reflecting the models' performance in practical applications. Overall, these findings offer different perspectives on model performance, aiding in comprehensive assessments of accuracy and generalizability. Future research should further investigate and adjust the model to enhance stability.

Finally, Inception V1 and Inception V3 are two different versions of the GoogLeNet model. The Inception V1 model was used in six articles (6,33,35-38), and the Inception V3 model was used in the other six articles (15,34,39-42). Our results showed that the sensitivity of the Inception V1 model was higher than that of the Inception V3 model (0.87 *vs.* 0.81, P<0.05). While the specificity of the Inception V1 model was lower than that of the Inception V3 model (0.89 *vs.* 0.85, P<0.05). This might be because the Inception V3 model is deeper and more complex than the Inception V1 model. The Inception V3 model was introduced with more layers and parameters, allowing the model to better learn the details and features of the image, which improves specificity (43). However, more complex models may be more likely to overfit the training data, resulting in reduced sensitivity on new data (44). Conversely, Inception V1 is relatively simpler and may be easier to generalize to new data, and thus perform better in terms of sensitivity (28,45).

The current meta-analysis had some limitations. First, the inclusion of only two versions of GoogLeNet renders the study less than comprehensive. Inception V4 (46,47) and Xception (40,42) models have been used in some studies, but they were not included in this meta-analysis due to their limited use in studies. Second, this study only included research conducted in Asia and published in English and Chinese languages, which might have introduced language and regional biases. As a result, the generalizability of the findings on a global scale is limited, posing a challenge for healthcare professionals worldwide to trust its relevance and accuracy. To enhance the credibility and applicability of our findings, future research should consider including studies from multiple regions and studies published in multiple languages. Third, only CUS images were included in this study. The GoogLeNet deep-learning model is based on elastic imaging (18,48), and Automated Breast Ultrasound (ABUS) (49) has been shown to perform well in the diagnosis of benign and malignant breast masses. In the future, a wider range of ultrasonic imaging techniques can be incorporated and an attempt can be made to combine different imaging technologies in clinical practice. Fourth, this meta-analysis focused solely on the GoogLeNet model; however, given the swift evolution of deep-learning

architectures, reliance on any single model could create certain limitations. Numerous studies have been conducted using integrated learning with multiple models in the classification of medical images; therefore, comparing or integrating multiple models in future studies could help to improve the broad applicability and validity of the results. Fifth, only 12 articles were included in this meta-analysis, and some of these featured small sample sizes, which might affect the accuracy of our meta-analysis results. Large-scale, prospective, multicenter studies need to be conducted to evaluate the diagnostic efficacy of the GoogLeNet deep-learning model more robustly. This indicates a gap in the existing research and the need for our findings to be more extensively validated before clinical adoption.

This review primarily focused on research that used the GoogLeNet model (based on traditional CNN) to distinguish between benign and malignant breast masses. However, we also acknowledge the limitations of our study. Below, we discuss how to further improve the GoogLeNet deep-learning model. Additionally, we comprehensively review other research achievements in using the GoogLeNet model to characterize breast masses to establish a more solid evidence base for its broader application in medical diagnostics. We aim to provide broader insights and ideas for future improvements in the use of the GoogLeNet model.

The success of an AI model in clinical applications depends not only on its high accuracy and precision but also on the practical integration of the following factors. First, the interpretability of AI decisions is particularly important in the healthcare domain (50-52). While deep-learning models like GoogLeNet excel at image classification tasks, their complex structures make it difficult to explain the decision-making process. To facilitate understanding of diagnostic outcomes by medical professionals, technicians, and patients, interpretability tools and techniques should be introduced (53). For example, Grad-CAM could be used to highlight the image regions that influence the model's decision (54,55) to help healthcare practitioners understand the reasoning behind the predictions. Second, integration with existing healthcare information technology systems cannot be overlooked in practical applications (56). Effective integration should ensure seamless alignment between AI solutions and existing systems, such as electronic health record systems, without disrupting established clinical workflows (57). Finally, the training of healthcare professionals in the use of such technology is crucial. Such training should emphasize that AI solutions are meant to assist rather than replace healthcare professionals in decision making. Through hands-on experiences, like clinical simulations or the use of virtual patient models, medical staff can bolster the efficacy and safety of AI applications in medicine (58).

Aggregation rules are techniques for optimizing feature processing, reducing computational complexity, and enhancing model performance by integrating, compressing, and enhancing features in deep-learning classification models (59). Aggregation rules can significantly reduce the consumption of computational resources in medical image analysis and still maintain efficient and accurate performance when dealing with a large amount of image data, thus improving the overall classification and diagnosis results (60). In addition, when using deep neural networks for classification, the performance of the model can be evaluated based on the statistical information (e.g., the accuracy and F1-score) of the rules (61). Aggregation rules significantly improve the performance and computational efficiency of deep-learning models by integrating features in medical image analysis, while enhancing the generalization and recognition ability of the models (62). For example:

(I) Attention mechanisms significantly improve models' recognition ability by dynamically adjusting the feature weights (63).
(II) Multi-scale feature aggregation techniques (e.g., Feature Pyramid Networks, FPN) enhance the ability of models to capture global and local information (64).
(III) Global average pooling (GAP) retains more global information and improves the generalization ability of models (65).
(IV) Convolutional and pooling layers extract high-dimensional features and can be integrated with aggregation rules to enable classifiers to process this information efficiently (66).
(V) Pooling operation reduces the dimensionality of feature maps, and reduces the computational complexity and the number of parameters, while also preventing overfitting (67).

Thus, optimizing the use of resources and improving the quality of the aggregation rules can greatly enhance the ability of deep-learning models in medical image analysis and improve the efficiency and accuracy of diagnosis.

A great deal of research has been conducted to advance CAD systems to support radiologists (68,69). Due to the difficulty of collecting a huge number of images for the field of medical imaging, the former methods have generally only been able to handle smaller data sets; however, they

have nonetheless shown their potential (70-72). Han *et al.* adopted a deep-learning approach with a substantial dataset (comprising several thousand patients) using the GoogLeNet model (33). Each region of interest sample image was processed, which included margin augmentation, image cropping, and histogram equalization. The images were also scaled to match the input image dimensions of the network. The CNN was trained to distinguish between malignant and benign masses. The AUC of the network was >0.95 with an accuracy of about 0.9 (90%), a sensitivity of 0.83, and a specificity of 0.95. This demonstrates that deep-learning methods have significant potential for clinical application.

Assari *et al.* developed a new GoogLeNet-based dual-mode CAD system that was designed to classify solid breast masses in combination with information from mammograms and CNN images (6). In the proposal framework, each mode is trained initially with two different single-mode models. The dual-mode is then trained using high-level feature maps obtained from each mode. The sensitivity of the dual-mode reached 90.91% and the accuracy reached 90.38%, both of which are higher than those of the single-mode model. The results showed that the dual-mode CAD system improved the accuracy of breast mass classification.

The meta-ensemble learning technique, which merges the outputs of various CNNs, has been shown to enhance the classification accuracy of models. Ali *et al.* applied a meta-learning algorithm to optimize the learning process and combined the output of multiple CNNs using an ensemble learning approach (39). The evaluation results showed that the model using the ensemble learning method had high accuracy and effectiveness. In addition, Zafar *et al.* (73) compared various pre-trained networks and employed a network selection algorithm to determine the best model for breast CNN image classification. They employed an evolutionary optimization (EO) algorithm as a network selection strategy. After thorough testing and analysis, this algorithm was shown to improve classification rates significantly, achieving the highest rates for all the examined pre-trained models. Notably, the Inception-ResNet-v2 model had a classification accuracy rate of 96.15% when the EO algorithm was applied. These results indicate that the integration of multiple models substantially boosts the performance of breast CNN image classification.

A research study proposed a novel intelligent-based, high-performance, low-cost automatic shallow network named the Feature-Preserved Mesh Network for accurately segmenting retinal vessels. This architecture preserves spatial features and employs a series of feature concatenations, contributing to better segmentation performance. This model could inspire improvements in the GoogLeNet model in breast CNN, and in our future research, we intend to expand on this further (74). In addition, techniques such as network selection and information fusion optimization may be used to refine the performance of GoogLeNet models in detecting breast masses. These approaches offer great promise for improving the capabilities of breast CNN classification models (75).

Fukuda *et al.* built a 50-cycle deep-learning model of GoogLeNet architecture based on elastography to predict the probability of malignancy (48). The model was used on the experimental data and compared with the findings of the fat damage ratio assessment and the five-point visual color (elasticity score) assessment. The performance of the model was assessed based on the ROC curve. The model had an AUC of 0.90 and a sensitivity of 0.80. These results indicated that ultrasonic elastography-based GoogLeNet deep-learning had high accuracy in the classification and diagnosis of benign and malignant breast masses. ABUS images can be visualized in horizontal and coronal views. Wang *et al.* adopted an improved Inception V3 architecture to provide an effective method for extracting multi-view features from the two views (49). This method had an AUC of 0.95 with cross-validation performed 50 times. It also had a sensitivity and specificity of 0.89 and 0.88, respectively. It achieved significant improvements in classification performance over conventional machine-learning feature extraction solutions, such as the histogram of oriented gradients and the principal component analysis.

Conclusions

This meta-analysis showed that deep learning based on the CNN GoogLeNet model is an effective method for distinguishing between and diagnosing benign and malignant breast masses. However, due to the limited sample size and the variability in the quality of the studies, additional multicenter or prospective studies need to be conducted in the future to address these limitations.

Acknowledgments

Funding: This work was supported by the Tianshan Young Talent Scientific and Technological Innovation Team: Innovative Team for Research on Prevention and Treatment of High-incidence Diseases in Central Asia (No. 2023TSYCTD0020), the National Natural Science

Foundation of China (No. 82060318), the Corps Science and Technology Key Project (No. 2022CB002-04), The First Affiliated Hospital of Shihezi University School of Medicine Youth Fund Project (No. QN202107), and The First Affiliated Hospital of Shihezi University School of Medicine Youth Fund Project (No. QN202126).

Footnote

Reporting Checklist: The authors have completed the PRISMA-DTA reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-679/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-679/coif). All authors report that this work was supported by the Tianshan Young Talent Scientific and Technological Innovation Team: Innovative Team for Research on Prevention and Treatment of High-incidence Diseases in Central Asia (No. 2023TSYCTD0020), the National Natural Science Foundation of China (No. 82060318), the Corps Science and Technology Key Project (No. 2022CB002-04), the First Affiliated Hospital of Shihezi University School of Medicine Youth Fund Project (No. QN202107), and The First Affiliated Hospital of Shihezi University School of Medicine Youth Fund Project (No. QN202126). The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Kashyap D, Pal D, Sharma R, Garg VK, Goel N, Koundal D, Zaguia A, Koundal S, Belay A. Global Increase in Breast Cancer Incidence: Risk Factors and Preventive Measures. Biomed Res Int 2022;2022:9605439.
2. Yu H, Sun H, Li J, Shi L, Bao N, Li H, Qian W, Zhou S. Effective diagnostic model construction based on discriminative breast ultrasound image regions using deep feature extraction. Med Phys 2021;48:2920-8.
3. Nicosia L, Pesapane F, Bozzini AC, Latronico A, Rotili A, Ferrari F, Signorelli G, Raimondi S, Vignati S, Gaeta A, Bellerba F, Origgi D, De Marco P, Castiglione Minischetti G, Sangalli C, Montesano M, Palma S, Cassano E. Prediction of the Malignancy of a Breast Lesion Detected on Breast Ultrasound: Radiomics Applied to Clinical Practice. Cancers (Basel) 2023.
4. Deb SD, Jha RK. Breast UltraSound Image classification using fuzzy-rank-based ensemble network. Biomedical Signal Processing and Control 2023;85:104871.
5. Aljuaid H, Alturki N, Alsubaie N, Cavallaro L, Liotta A. Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. Comput Methods Programs Biomed 2022;223:106951.
6. Assari Z, Mahloojifar A, Ahmadinejad N. A bimodal BI-RADS-guided GoogLeNet-based CAD system for solid breast masses discrimination using transfer learning. Comput Biol Med 2022;142:105160.
7. Baek J, O'Connell AM, Parker KJ. Improving breast cancer diagnosis by incorporating raw ultrasound parameters into machine learning. Mach Learn Sci Technol 2022;3:045013.
8. Chattopadhyay S, Dey A, Singh PK, Sarkar R. DRDA-Net: Dense residual dual-shuffle attention network for breast cancer classification using histopathological images. Comput Biol Med 2022;145:105437.
9. Kim S-, Choi Y, Kim E-, Han BK, Yoon JH, Choi JS, Chang JM. Deep learning-based computer-aided diagnosis in screening breast ultrasound to reduce false-positive diagnoses. Sci Rep 2021;11:395.
10. Lu YY, Chen YQ, Chen C, Li JL, He KL, Xiao RX. An Intelligent Breast Ultrasound System for Diagnosis and 3D Visualization. Electronics 2022;11:2116.
11. Chabi ML, Borget I, Ardiles R, Aboud G, Boussouar S, Vilar V, Dromain C, Balleyguier C. Evaluation of the accuracy of a computer-aided diagnosis (CAD) system in breast ultrasound according to the radiologist's experience. Acad Radiol 2012;19:311-9.
12. Retson TA, Eghtedari M. Expanding Horizons: The Realities of CAD, the Promise of Artificial Intelligence, and Machine Learning's Role in Breast Imaging beyond Screening Mammography. Diagnostics (Basel)

2023;13:2133.
13. Zhou J, Liu C, Shi Z, Li X, Chang C, Zhi W, Zhou S. Application of ultrasound-based radiomics models of breast masses to predict invasive components of encapsulated papillary carcinoma. Quant Imaging Med Surg 2023;13:6887-98.
14. Dhillon A, Verma GK. Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence 2020;9:85-112.
15. Tsai MJ, Tao YH. Deep Learning Technology Applied to Medical Image Tissue Classification. Diagnostics (Basel) 2022;12:2430.
16. Gao Y, Lin J, Zhou Y, Lin R. The application of traditional machine learning and deep learning techniques in mammography: a review. Front Oncol 2023;13:1213045.
17. Kwon SW, Choi IJ, Kang JY, Jang WI, Lee GH, Lee MC. Ultrasonographic Thyroid Nodule Classification Using a Deep Convolutional Neural Network with Surgical Pathology. J Digit Imaging 2020;33:1202-8.
18. Fujioka T, Kubota K, Mori M, Kikuchi Y, Katsuta L, Kasahara M, Oda G, Ishiba T, Nakagawa T, Tateishi U. 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.
19. Ahmad HM, Khan MJ, Yousaf A, Ghuffar S, Khurshid K. Deep Learning: A Breakthrough in Medical Imaging. Curr Med Imaging 2020;16:946-56.
20. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep Learning in Medical Image Analysis. Adv Exp Med Biol 2020;1213:3-21.
21. Wang Y, Ge XK, Ma H, Qi SL, Zhang GJ, Yao YD. Deep Learning in Medical Ultrasound Image Analysis: A Review. IEEE Access 2021;9:54310-24.
22. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:640-51.
23. Izzuddin TA, Safri NM, Othman MA. Compact convolutional neural network (CNN) based on SincNet for end-to-end motor imagery decoding and analysis. Biocybernetics and Biomedical Engineering 2021;41:1629-45.
24. Abd-Elmoniem KZ, Yassine IA, Metwalli NS, Hamimi A, Ouwerkerk R, Matta JR, Wessel M, Solomon MA, Elinoff JM, Ghanem AM, Gharib AM. Direct pixel to pixel principal strain mapping from tagging MRI using end to end deep convolutional neural network (DeepStrain). Sci Rep 2021;11:23021.
25. Alotaibi B, Alotaibi M. A hybrid deep ResNet and inception model for hyperspectral image classification. PFG-Journal of Photogrammetry, Remote Sensing and Geoinformation Science 2020;88:463-76.
26. Szegedy C, Liu W, Jia YQ, et al. Going Deeper with Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:1-9.
27. Bi N, Chen J, Tan J. The handwritten Chinese character recognition uses convolutional neural networks with the googlenet. International Journal of Pattern Recognition and Artificial Intelligence 2019;33:1940016.
28. Sam SM, Kamardin K, Sjarif NNA, Mohamed N. Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet inception-v1 and inception-v3. Procedia Computer Science 2019;161:475-83.
29. Khosravi P, Kazemi E, Imielinski M, Elemento O, Hajirasouliha I. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. EBioMedicine 2018;27:317-28.
30. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE conference on computer vision and pattern recognition 2016:2818-26.
31. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. BMC Med Imaging 2022;22:69.
32. Morid MA, Borjali A, Del Fiol G. A scoping review of transfer learning research on medical image analysis using ImageNet. Comput Biol Med 2021;128:104115.
33. Han S, Kang HK, Jeong JY, Park MH, Kim W, Bang WC, Seong YK. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol 2017;62:7714-28.
34. Yu MH, Yuan Q, Zeng SE, Cheng H, Li N, Ye HR. The value of transfer learning model based on ultrasound images in differentiating benign and malignant breast masses. Journal of Clinical Ultrasound in Medicine 2022;24:652-6.
35. Masud M, Eldin Rashed AE, Hossain MS. Convolutional neural network-based models for diagnosis of breast cancer. Neural Comput Appl 2022;34:11383-94.
36. Alhussan AA, Eid MM, Towfek SK, Khafaga DS. Breast Cancer Classification Depends on the Dynamic Dipper Throated Optimization Algorithm. Biomimetics (Basel) 2023;8:163.
37. 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. Sci Rep 2021;11:24382.
38. Kriti, Virmani J, Agarwal R. Deep feature extraction and classification of breast ultrasound images. Multimedia Tools and Applications 2020;79:27257-92.
39. Ali MD, Saleem A, Elahi H, Khan MA, Khan MI, Yaqoob MM, Farooq Khattak U, Al-Rasheed A. Breast Cancer Classification through Meta-Learning Ensemble Technique Using Convolution Neural Networks. Diagnostics (Basel) 2023;13:2242.
40. Xiao T, Liu L, Li K, Qin W, Yu S, Li Z. Comparison of Transferred Deep Neural Networks in Ultrasonic Breast Masses Discrimination. Biomed Res Int 2018;2018:4605191.
41. Zhang H, Han L, Chen K, Peng Y, Lin J. Diagnostic Efficiency of the Breast Ultrasound Computer-Aided Prediction Model Based on Convolutional Neural Network in Breast Cancer. J Digit Imaging 2020;33:1218-23.
42. Sirjani N, Ghelich Oghli M, Kazem Tarzamni M, Gity M, Shabanzadeh A, Ghaderi P, Shiri I, Akhavan A, Faraji M, Taghipour M. A novel deep learning model for breast lesion classification using ultrasound Images: A multicenter data evaluation. Phys Med 2023;107:102560.
43. Saini M, Susan S. Data Augmentation of Minority Class with Transfer Learning for Classification of Imbalanced Breast Cancer Dataset Using Inception-V3. In: Morales A, Fierrez J, Sánchez J, Ribeiro B. editors. Pattern Recognition and Image Analysis. IbPRIA 2019. Lecture Notes in Computer Science, Springer, 2019;11867:409-20.
44. Kumar JS, Anuar S, Hassan NH. Transfer Learning based Performance Comparison of the Pre-Trained Deep Neural Networks. International Journal of Advanced Computer Science and Applications 2022;13:797-805.
45. Vijayan N, Kuruvilla J. The impact of transfer learning on lung cancer detection using various deep neural network architectures. 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 2022:1-5.
46. Nazir MS, Khan UG, Mohiyuddin A, Al Reshan MS, Shaikh A, Rizwan M, Davidekova M. A Novel CNN-Inception-V4-Based Hybrid Approach for Classification of Breast Cancer in Mammogram Images. Wireless Communications and Mobile Computing 2022;2022:5089078.
47. Al Husaini MAS, Habaebi MH, Gunawan TS, Islam MR, Elsheikh EAA, Suliman FM. Thermal-based early breast cancer detection using inception V3, inception V4 and modified inception MV4. Neural Comput Appl 2022;34:333-48.
48. Fukuda T, Tsunoda H, Yagishita K, Naganawa S, Hayashi K, Kurihara Y. Deep Learning for Differentiation of Breast Masses Detected by Screening Ultrasound Elastography. Ultrasound Med Biol 2023;49:989-95.
49. Wang Y, Choi EJ, Choi Y, Zhang H, Jin GY, Ko SB. Breast Cancer Classification in Automated Breast Ultrasound Using Multiview Convolutional Neural Network with Transfer Learning. Ultrasound Med Biol 2020;46:1119-32.
50. Chaddad A, Peng J, Xu J, Bouridane A. Survey of Explainable AI Techniques in Healthcare. Sensors (Basel) 2023;23:634.
51. Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20:310.
52. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput Biol Med 2022;140:105111.
53. Rafferty A, Nenutil R, Rajan A. Explainable artificial intelligence for breast tumour classification: Helpful or harmful. International Workshop on Interpretability of Machine Intelligence in Medical Image Computing. Cham: Springer Nature Switzerland, 2022:104-23.
54. Zhang H, Ogasawara K. Grad-CAM-Based Explainable Artificial Intelligence Related to Medical Text Processing. Bioengineering (Basel) 2023;10:1070.
55. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis 2020;128:336-59.
56. Reegu FA, Abas H, Gulzar Y, Xin Q, Alwan AA, Jabbari A, Sonkamble RG, Dziyauddin RA. Blockchain-based framework for interoperable electronic health records for an improved healthcare system. Sustainability 2023;15:6337.
57. Patton MJ, Liu VX. Predictive Modeling Using Artificial Intelligence and Machine Learning Algorithms on Electronic Health Record Data: Advantages and Challenges. Crit Care Clin 2023;39:647-73.
58. Pinto Dos Santos D, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, Maintz D, Baeßler B. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019;29:1640-6.
59. Tiddi I, d'Aquin M, Motta E. Using neural networks to aggregate linked data rules. Using Neural Networks to Aggregate Linked Data Rules. In: Janowicz K, Schlobach S,

Lambrix P, Hyvönen E. editors. Knowledge Engineering and Knowledge Management. EKAW 2014. Lecture Notes in Computer Science, Springer, 2014;8876:547-62.
60. Grisci BI, Krause MJ, Dorn M. Relevance aggregation for neural networks interpretability and knowledge discovery on tabular data. Information Sciences 2021;559:111-29.
61. Rismala R, Maulidevi NU, Surendro K. Personalized neural network-based aggregation function in multi-criteria collaborative filtering. Journal of King Saud University-Computer and Information Sciences 2024;36:101922.
62. Saha A, Tushar FI, Faryna K, D'Anniballe VM, Hou R, Mazurowski MA. Lo JY. Weakly supervised 3D classification of chest CT using aggregated multi-resolution deep segmentation features. Medical Imaging 2020: Computer-Aided Diagnosis. SPIE, 2020;11314:39-44.
63. Fan Z, Gong P, Tang S, Lee CU, Zhang X, Song P, Chen S, Li H. Joint localization and classification of breast masses on ultrasound images using an auxiliary attention-based framework. Med Image Anal 2023;90:102960.
64. Gao WT, Li XJ, Han Y, Liu Y. Multi-scale Vertical Cross-layer Feature Aggregation and Attention Fusion Network for Object Detection. In: Pimenidis E, Angelov P, Jayne C, Papaleonidas A, Aydin M. editors. Artificial Neural Networks and Machine Learning – ICANN 2022. Lecture Notes in Computer Science, Springer, 2022;13532:139-50.
65. Dubey A, Singh SK, Jiang X. Leveraging CNN and Transfer Learning for Classification of Histopathology Images. In: Khare N, Tomar DS, Ahirwal MK, Semwal VB, Soni V. editors. Machine Learning, Image Processing, Network Security and Data Sciences. MIND 2022. Communications in Computer and Information Science, Springer, 2023;1763:3-13.
66. Shallu, Mehra R. Automatic Magnification Independent Classification of Breast Cancer Tissue in Histological Images Using Deep Convolutional Neural Network. In: Luhach A, Singh D, Hsiung PA, Hawari K, Lingras P, Singh P. editors. Advanced Informatics for Computing Research. ICAICR 2018. Communications in Computer and Information Science, Springer, Singapore, 2018;955:772-81.
67. Kassani SH, Kassani PH, Wesolowski MJ, Schneider KA, Deters R. Breast Cancer Diagnosis with Transfer Learning and Global Pooling. 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea (South), 2019:519-24.
68. Renz DM, Baltzer PAT, Kullnig PE, Böttcher J, Vag T, Kaiser WA, Runnebaum IB. Clinical Value of Computer-Assisted Analysis in MR Mammography. A Comparison between two Systems and three Observers with Different Levels of Experience. RofoFortschritte Auf Dem Gebiet Der Rontgenstrahlen Und Der Bildgebenden Verfahren 2008;180:968-76.
69. Chen Z, Ying MTC. Ultrasound-based multiregional radiomics analysis to differentiate breast masses. Quant Imaging Med Surg 2023;13:6353-4.
70. Avendaño-Valencia LD, Yderstræde KB, Nadimi ES, Blanes-Vidal V. Video-based eye tracking performance for computer-assisted diagnostic support of diabetic neuropathy. Artif Intell Med 2021;114:102050.
71. Binol H, Niazi MKK, Plotner A, Sopkovich J, Kaffenberger BH, Gurcan MN. A Multidimensional Scaling and Sample Clustering to Obtain a Representative Subset of Training Data for Transfer Learning-based Rosacea Lesion Identification. Proceedings of the SPIE, 2020. doi: 10.1117/12.2549392.
72. Viriyasaranon T, Chun JW, Koh YH, Cho JH, Jung MK, Kim SH, Kim HJ, Lee WJ, Choi JH, Woo SM. Annotation-Efficient Deep Learning Model for Pancreatic Cancer Diagnosis and Classification Using CT Images: A Retrospective Diagnostic Study. Cancers (Basel) 2023;15:3392.
73. Zafar A, Tanveer J, Ali MU, Lee SW. BU-DLNet: Breast Ultrasonography-Based Cancer Detection Using Deep-Learning Network Selection and Feature Optimization. Bioengineering (Basel) 2023;10:825.
74. Imran SMA, Saleem MW, Hameed MT, Hussain A, Naqvi RA, Lee SW. Feature preserving mesh network for semantic segmentation of retinal vasculature to support ophthalmic disease analysis. Front Med (Lausanne) 2022;9:1040562.
75. Hamza A, Attique Khan M, Wang SH, Alhaisoni M, Alharbi M, Hussein HS, Alshazly H, Kim YJ, Cha J. COVID-19 classification using chest X-ray images based on fusion-assisted deep Bayesian optimization and Grad-CAM visualization. Front Public Health 2022;10:1046296.

Cite this article as: Wang J, Tong J, Li J, Cao C, Wang S, Bi T, Zhu P, Shi L, Deng Y, Ma T, Hou J, Cui X. Using the GoogLeNet deep-learning model to distinguish between benign and malignant breast masses based on conventional ultrasound: a systematic review and meta-analysis. Quant Imaging Med Surg 2024;14(10):7111-7127. doi: 10.21037/qims-24-679

