

Effectiveness evaluation of computer-aided diagnosis system for the diagnosis of thyroid nodules on ultrasound

A systematic review and meta-analysis

Wan-Jun Zhao, MD^a, Lin-Ru Fu, MBBS^b, Zhi-Mian Huang, BCom^c, Jing-Qiang Zhu, MD^{a,*}, Bu-Yun Ma, MD^{d,*}

Abstract

Background: More and more automated efficient ultrasound image analysis techniques, such as ultrasound-based computer-aided diagnosis system (CAD), were developed to obtain accurate, reproducible, and more objective diagnosis results for thyroid nodules. So far, whether the diagnostic performance of existing CAD systems can reach the diagnostic level of experienced radiologists is still controversial. The aim of the meta-analysis was to evaluate the accuracy of CAD for thyroid nodules' diagnosis by reviewing current literatures and summarizing the research status.

Methods: A detailed literature search on PubMed, Embase, and Cochrane Libraries for articles published until December 2018 was carried out. The diagnostic performances of CAD systems vs radiologist were evaluated by meta-analysis. We determined the sensitivity and the specificity across studies, calculated positive and negative likelihood ratios and constructed summary receiver-operating characteristic (SROC) curves. Meta-analysis of studies was performed using a mixed-effect, hierarchical logistic regression model.

Results: Five studies with 536 patients and 723 thyroid nodules were included in this meta-analysis. The pooled sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, and diagnostic odds ratio (DOR) for CAD system were 0.87(95% confidence interval [CI], 0.73–0.94), 0.79 (95% CI 0.63–0.89), 4.1 (95% CI 2.5–6.9), 0.17 (95% CI 0.09–0.32), and 25 (95% CI 15–42), respectively. The SROC curve indicated that the area under the curve was 0.90 (95% CI 0.87–0.92). The pooled sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, and DOR for experienced radiologists were 0.82 (95% CI 0.69–0.91), 0.83 (95% CI 0.76–0.89), 4.9 (95% CI 3.4–7.0), 0.22 (95% CI 0.12–0.38), and 23 (95% CI 11–46), respectively. The SROC curve indicated that the area under the curve was 0.96 (95% CI 0.94–0.97).

Conclusion: The sensitivity of the CAD system in the diagnosis of thyroid nodules was similar to that of experienced radiologists. However, the CAD system had lower specificity and DOR than experienced radiologists. The CAD system may play the potential role as a decision-making assistant alongside radiologists in the thyroid nodules' diagnosis. Future technical improvements would be helpful to increase the accuracy as well as diagnostic efficiency.

Abbreviations: AI= artificial intelligence, ATA = American Thyroid Association, AUC = area under the receiver-operating characteristic, CAD = computer-aided diagnosis, CI = confidence interval, DOR = diagnostic odds ratio, FN = false negative, FNA= fine-needle aspiration, FP = false positive, QUADAS-2 = Quality Assessment of Diagnostics Accuracy Studies-2, SROC = summary receiver-operating characteristic, TN = true negative, TP = true positive, TIRADS = Thyroid Imaging Reporting and Data System.

Keywords: computer-aided diagnosis system, meta-analysis, sensitivity and specificity, thyroid nodules, ultrasound

Editor: Giovanni Tarantino.

This study was funded by the Postdoctoral Sustentation Fund of West China Hospital, Sichuan University, China (Grant No. 18HXBH074).

The authors have no conflicts of interest to disclose.

^a Department of Thyroid & Parathyroid Surgery, West China Hospital, ^b West China School of Medicine, Sichuan University, Sichuan, ^c Business College, New York University in Shanghai, Shanghai, ^d Department of Ultrasonography, West China Hospital, Sichuan University, Sichuan, China.

* Correspondence: Jing-qiang Zhu, Department of Thyroid & Parathyroid surgery, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang, Chengdu, Sichuan 610041, China (e-mail: zjq-wkys@163.com), Bu-Yun Ma, Department of Ultrasonography, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang, Chengdu, Sichuan 610041, China (e-mail: Mabu-yun@163.com).

Copyright © 2019 the Author(s). Published by Wolters Kluwer Health, Inc.

This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial License 4.0 (CCBY-NC), where it is permissible to download, share, remix, transform, and buildup the work provided it is properly cited. The work cannot be used commercially without permission from the journal.

Medicine (2019) 98:32(e16379)

Received: 21 February 2019 / Received in final form: 29 May 2019 / Accepted: 14 June 2019

<http://dx.doi.org/10.1097/MD.00000000000016379>

1. Introduction

Thyroid nodules, as the most common endocrine system disease, are detected in up to 19% to 68% of the general population with 3% to 10% of malignancy.^[1,2] In 2018, 567,233 (3.1%) people worldwide were diagnosed as thyroid cancer.^[3] The incidence of thyroid cancer has increased 2.4 times over the last 30 years.^[3,4] Accurately identifying the preoperative distinction between benign vs malignancy has been an ongoing challenge. Ultrasound, as the most accurate imaging modality, is widely used in preoperative evaluation and decision-making for fine-needle aspiration (FNA).^[5-7] It is critical to be able to differentiate benign from malignant process.^[8,9] However, the diagnostic accuracy of ultrasound depends on radiologists' experience, echo perturbation and speckle noise.^[8,10,11] The sensitivity and the specificity of thyroid carcinoma detection by ultrasound were

40.3% to 93% and 54% to 100%, respectively.^[10,12,13] Less experienced radiologists had a higher misdiagnosis rate and increased the rate of unnecessary FNA.

Computer-aided diagnosis (CAD) system with novel artificial intelligence (AI) techniques is widely developed for automated efficient ultrasound image analysis recently.^[14,15] Some studies reported that CAD system had been used to assess breast and lung cancers to provide a suggestion for diagnosis with radiologists.^[16-20] CAD system employs a deep neural network in general, which is a novel computer model automatically judging benignity and malignancy of thyroid nodules in new images after learning lots of images with existing labels.^[21,22] Being a real-time and noninvasive diagnosis suggestion, CAD becomes a widely employed imaging method for the preoperative evaluation and follow-up of thyroid nodules. However, the practicality and

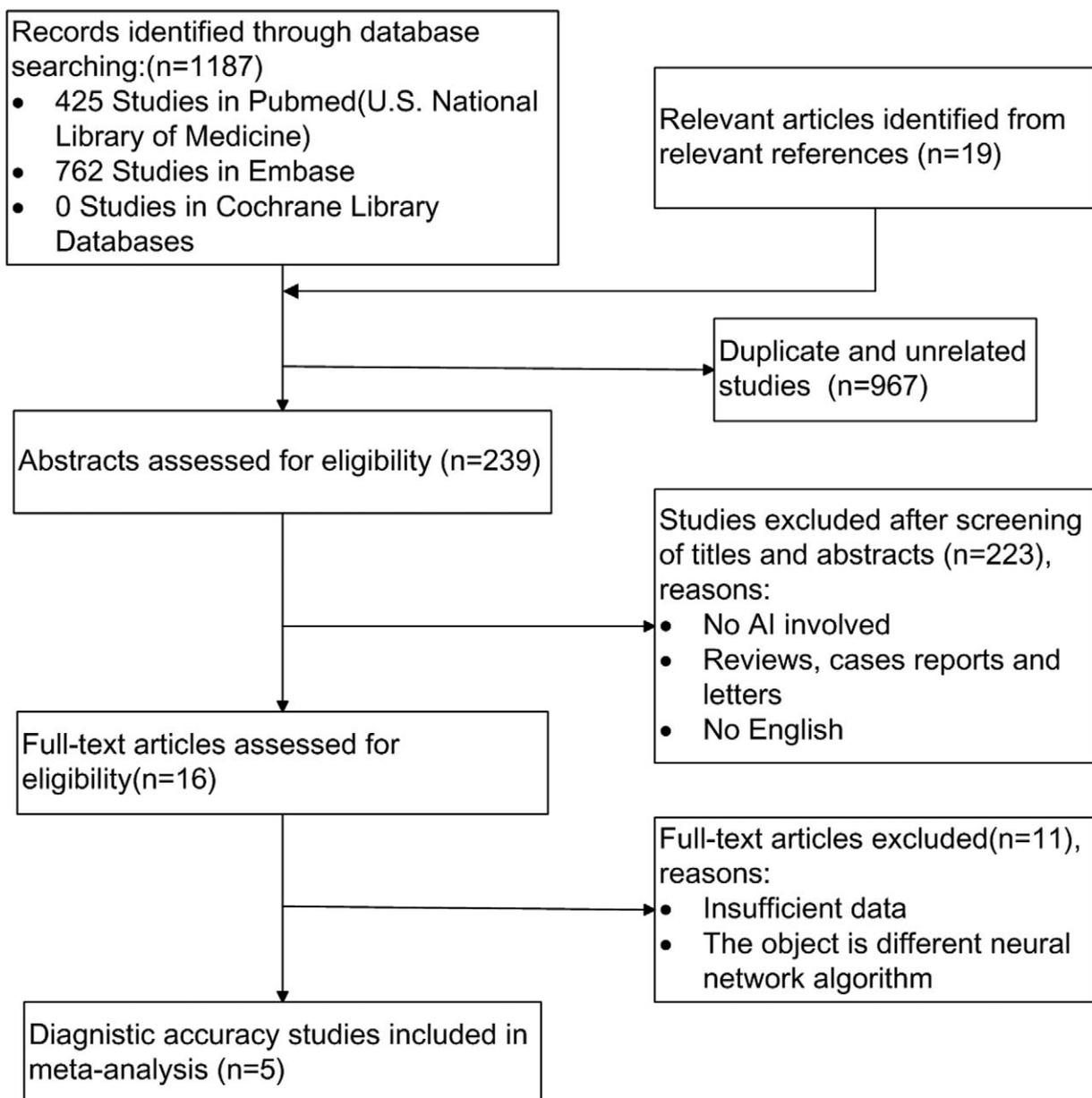


Figure 1. Flow chart of the search for eligible studies on diagnostic performance of computer-aided diagnosis system.

Table 1
Characteristics of the included studies.

Author	Year	Country	Study type	The CAD system	Patients for the test	The number of nodules for test			Criteria for nodules		Size of nodules	
						Total	Benign	Malignant	Benign	Malignant	Benign	Malignant
Jeong et al	2018	Korea	Prospective	Samsung	85	100	56	44	Surgery/FNA	Surgery	1.8±0.8	1.5±0.8
Gitto et al	2018	Italy	Retrospective	Samsung	50	62	14	48	FNA	FNA	1.8±0.7	
Yoo et al	2018	Korea	Prospective	Samsung	50	117	67	50	FNA	Surgery	1.1±0.8	1.2±1.0
Gao et al	2017	China	Retrospective	Independently developed	262	342	103	239	Surgery	Surgery	1.7±1.4	1.0±0.7
Choi et al	2016	Korea	Prospective	Samsung	89	102	59	43	CNB/FNA	CNB/FNA	1.5±0.8	0.9±0.4

CAD = computer-aided diagnosis, CNB = core-needle biopsy, FNA = fine-needle aspiration.

effectiveness of CAD for thyroid nodules were still controversial. The point of contention is time consumption and accuracy of CAD in comparison with the experienced radiologists. The CAD lacks clinical practicality if costing longer time to make the diagnosis of thyroid nodules than radiologists do and failing to achieve a real-time performance. Additionally, regional identification of thyroid nodules^[2,3] and the judgment of thyroid nodules between benignity and malignancy^[2,4,25] are the 2 dimensions of evaluating effectiveness.

To our knowledge, there is no review or meta-analysis of the effectiveness of CAD for thyroid nodules. Therefore, the aim of our study was to conduct the analysis to evaluate the accuracy of CAD for thyroid nodules by reviewing current literatures and summarizing the research status.

2. Materials and methods

2.1. Search strategy and selection criteria

This study is a meta-analysis of previously published studies, thus no ethical approval or patient consent was necessary. Two authors independently searched literatures published in English language until December 2018 by using PubMed (including MEDLINE), Embase, and Cochrane Libraries. The search strategy is listed below: 1# thyroid [Mesh term] 2# ultrasound or ultrasonography or ultrasonic [Title/Abstract] 3# artificial intelligence (AI) or computer-aid diagnosis or neural network or deep learning [All Fields] 4# 1 and 2 and 3. After screening based on title and abstract of studies, the remaining articles were assessed based on full manuscript and excluded with reasons when appropriate. The reference lists of the included articles were examined for any further studies that could be included. We cross-checked the article list in the identified publications and independently extracted the data from each study with full text.

The inclusion criteria for relevant studies were as follows: AI technology or CAD system had been used to diagnose thyroid nodules on ultrasound, clinical studies with sufficient data to reassess sensitivity and specificity of CAD system for diagnosis of thyroid nodules, absolute numbers of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) data had been presented, studies with more than 10 patients, and no data overlap. Studies were excluded as follows: studies on AI algorithms of CAD system, case reports, conference proceeding, editorial comments and letters to the editor, which do not contain the original data, full text not to be accessed online or requested from authors, and studies not in English.

2.2. Quality assessment and data extraction

The overall quality of the included studies in this review was critically appraised by 2 authors independently, based on 15-item modified Quality Assessment of Diagnostics Accuracy Studies-2 (QUADAS-2).^[26] The QUADAS-2 tool primarily assesses 4 domains: risk of bias in patient selection, index test, reference standard, and the timing of reference test. Discrepancies between the researchers were resolved by discussion. To decrease the bias in the selection of studies and in the data extraction, reviewers who were blinded to the journal, author, institution, and date of publication independently selected articles based on the inclusion criteria. Scores were assigned to study design characteristics and examination results by using a standardized form that was based on the QUADAS-2 tool.

A standardized data form was used to extract all data we would evaluate. The information which we collected from each study includes the study basic (author, year of publication and country of origin), study design (prospective or retrospective), patients' characteristics and technical aspects. Each study was analyzed to retrieve the numbers of TP, TN, FP, and FN findings

Table 2
Comparison of diagnostic performance of computer-aided diagnosis system and experienced radiologist.

Author	Year	Country	Sensitivity		Specificity		Positive predictive value		Negative predictive value		Accuracy		AUC	
			Radiologist	CAD	Radiologist	CAD	Radiologist	CAD	Radiologist	CAD	Radiologist	CAD	Radiologist	CAD
Jeong et al	2018	Korea	84.10%	88.60%	96.40%	83.90%	94.90%	81.30%	88.50%	90.40%	91.00%	86.00%	—*	0.718–0.863
Gitto et al	2018	Italy	78.60%	21.40%	66.70%	81.30%	40.70%	25.00%	91.40%	78.00%	69.40%	67.70%	—*	—
Yoo et al	2018	Korea	84.00%	80.00%	95.50%	88.10%	93.30%	83.30%	88.90%	85.50%	90.60%	84.60%	0.898	0.84
Gao et al	2017	China	90.0–96.2%	96.7%	75.7–78.6%	48.5%	90.0–91.2%	81.3%	76.7–89.7%	86.20%	86.0–90.4%	82.2%	0.83–0.87	0.73
Choi et al	2016	Korea	88.40%	90.70%	94.90%	74.60%	92.70%	72.20%	91.80%	91.70%	92.20%	81.40%	0.920	0.83

AUC=area under the receiver-operating characteristic, CAD = computer-aided diagnosis.

*The exact data were not reported in the original articles.

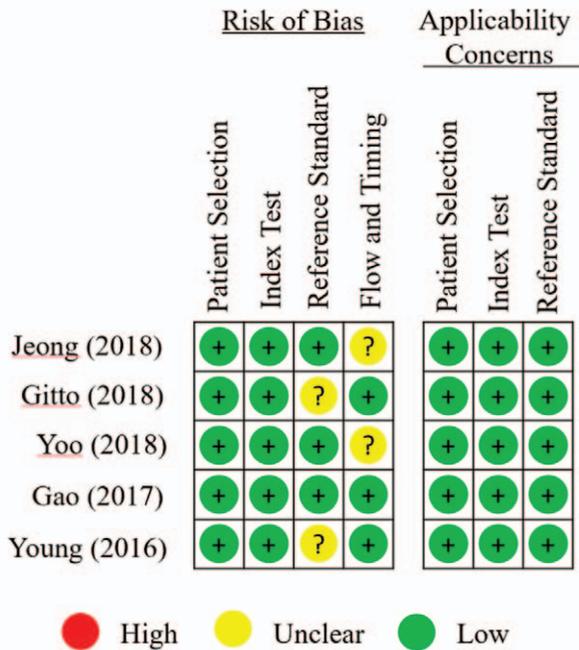


Figure 2. The results of methodologic quality analysis for each study in meta-analysis according to Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2).

of CAD system and radiologists, according to the reference standard. Only studies providing such complete information were finally included in the meta-analysis.

2.3. Statistical analysis

All data from each eligible study were extracted in Excel 2007 (Microsoft Corporation, Redmond, CA) and the statistical analyses in this review were conducted by STATA software, version 14.0 (StataCorp LP, College Station, TX). Summary sensitivity, specificity, positive likelihood ratios, negative likelihood ratios, and diagnostic odds ratio (DOR) with 95% confidence interval (CI) were performed. We used the bivariate random-effects model for analysis and pooling of the diagnostic performance measures across studies.^[27] Each data point of the summary receiver operator characteristic (SROC) graph comes

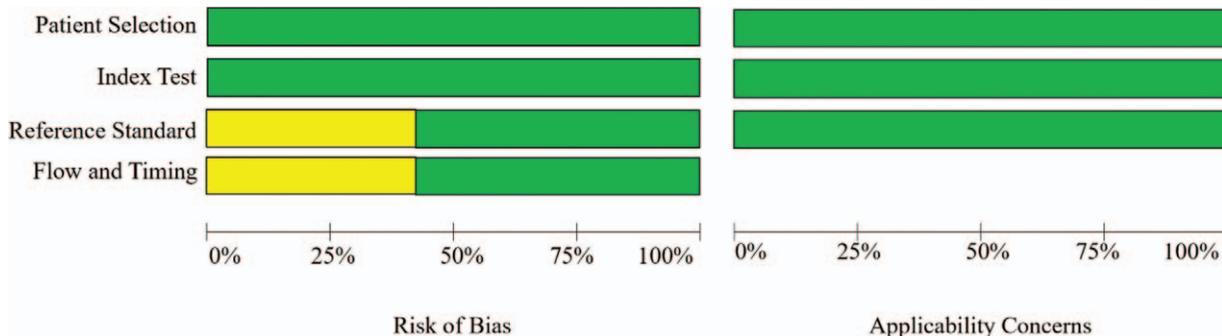


Figure 3. Summary of results of methodologic quality analysis of 5 studies in meta-analysis according to Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2).

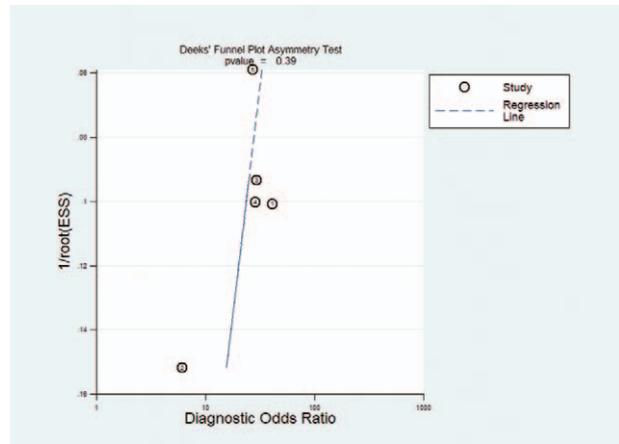


Figure 4. Results of Deeks' funnel plot of asymmetry test for publication bias. Nonsignificant slope indicates that no significant bias was found.

from an individual study.^[28] The SROC curve is formed based on these points to reveal pooled accuracy.^[28] Between-study statistical heterogeneity was assessed using I^2 and the Cochran Q test on the basis of the random-effects analysis. When statistical heterogeneity was substantial, we performed meta-regression to identify potential sources of bias.^[29] Publication bias was examined using the effective sample size funnel plot and associated regression test of asymmetry described by Deeks et al.^[30] Two-sided $P \leq .05$ was considered statistically significant.

3. Results

3.1. Literature search and selection of studies

After the comprehensive computerized search was performed and references lists were extensively cross-checked, our research yielded 1206 records on or before December 15, 2018, including 425 articles from PubMed, 762 articles from Embase, and 19 articles from references. Excluding the duplicates, unrelated topics, unoriginal studies, and non-English articles, 16 full-text articles^[14,24,31-40] were assessed for eligibility. Finally, 5 studies^[31-35] were selected because of the eligibility for the systematic review and meta-analysis. There were 11

Table 3
The meta-analysis results of diagnostic performance of CAD, CAD by Samsung and experienced radiologist.

Groups	Sensitivity		Specificity		Positive likelihood ratio		Negative likelihood ratio		Diagnostic odds ratio	
	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI
CAD	0.87	0.73–0.94	0.79	0.63–0.89	4.1	2.5–6.9	0.17	0.09–0.32	25	15–41
CAD by Samsung	0.82	0.69–0.91	0.83	0.76–0.89	4.9	3.4–7.0	0.22	0.12–0.38	23	11–46
Radiologist	0.88	0.80–0.93	0.92	0.84–0.96	11.1	5.6–21.9	0.13	0.08–0.21	86	47–158

CAD = computer-aided diagnosis, CI=confidence interval.

articles^[14,15,24,36–43] excluded due to various reasons. We excluded 7 full-text articles^[14,24,36–40] because they mainly studied on the design of algorithms of CAD system rather than on clinical practice. In addition, the study conducted by Choi et al^[41] only focused on computerized analysis of calcification of thyroid nodules as visualized by ultrasonography rather than the whole process of diagnosis, the study conducted by Acharya et al^[15] was about Hashimoto Thyroiditis rather than thyroid nodules, the study conducted by Lin et al^[42] focused on CAD-assisted ultrasound diagnosis of [18F] FDG-positive thyroid nodules not simply on CAD systems, and the study conducted by Zhu et al^[43] aimed to build an artificial neural network to discriminate malignant from benign thyroid nodules rather than evaluate an existing diagnostic system, thus we excluded these 4 full-text articles, as well. There were no additional studies found by

screening the references of these articles. The detailed procedure of study selection in the current meta-analysis is shown in Figure 1.

3.2. Study description and characteristics

The basic characteristics of included studies were presented in Table 1. We conducted all analyses based on per-patient data. Among 5 included studies, there were 3 prospective studies^[31,34,35] and 2 retrospective studies.^[32,33] All studies were published during recent 3 years, 3 studies in Korea,^[31,34,35] 1 study^[32] in China, and 1 study^[33] in Italy. The CAD in 4 studies^[31,33–35] was developed by Samsung and the remaining CAD in the Chinese study^[32] was independently developed by laboratory. There was a total of 536 patients and 723 thyroid

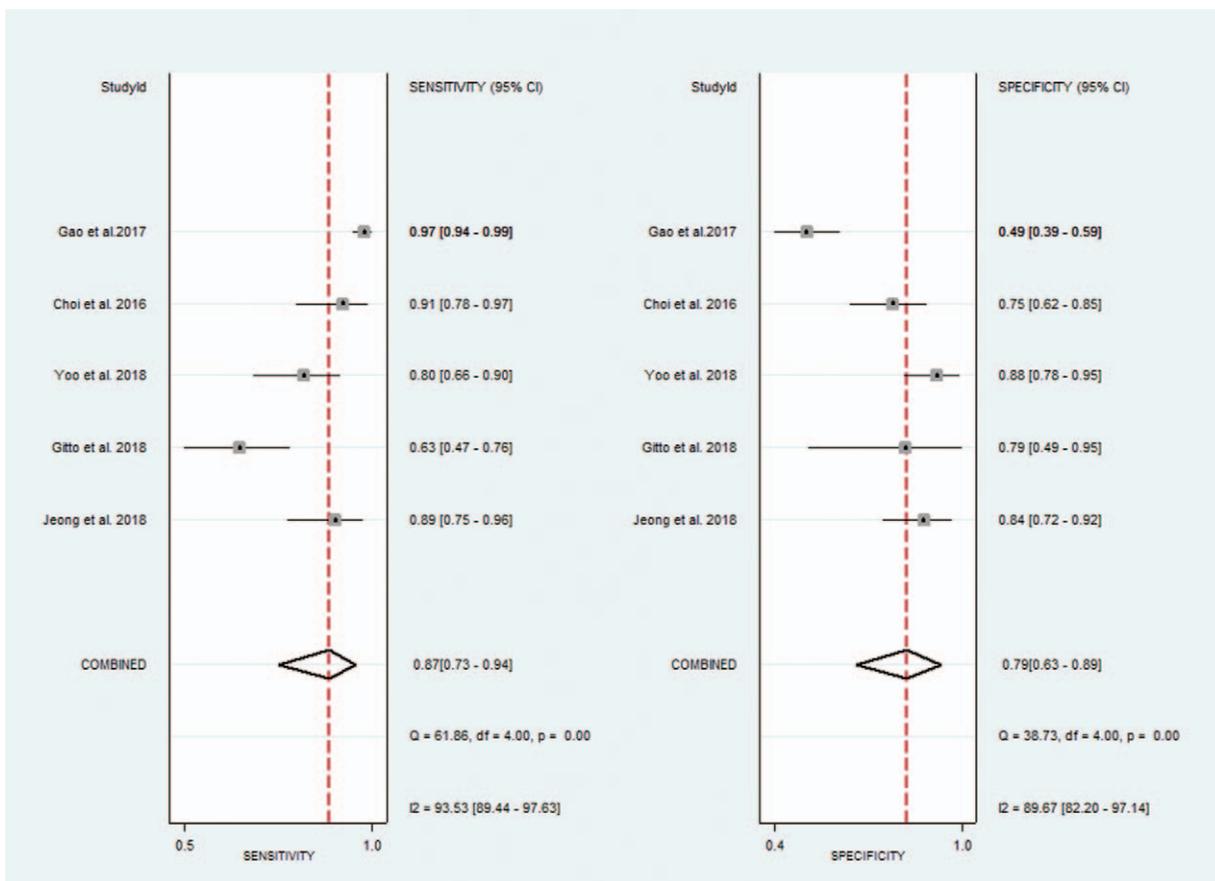


Figure 5. Forest plot of pooled sensitivity and specificity of computer-aided diagnosis system for the diagnosis of thyroid nodules.

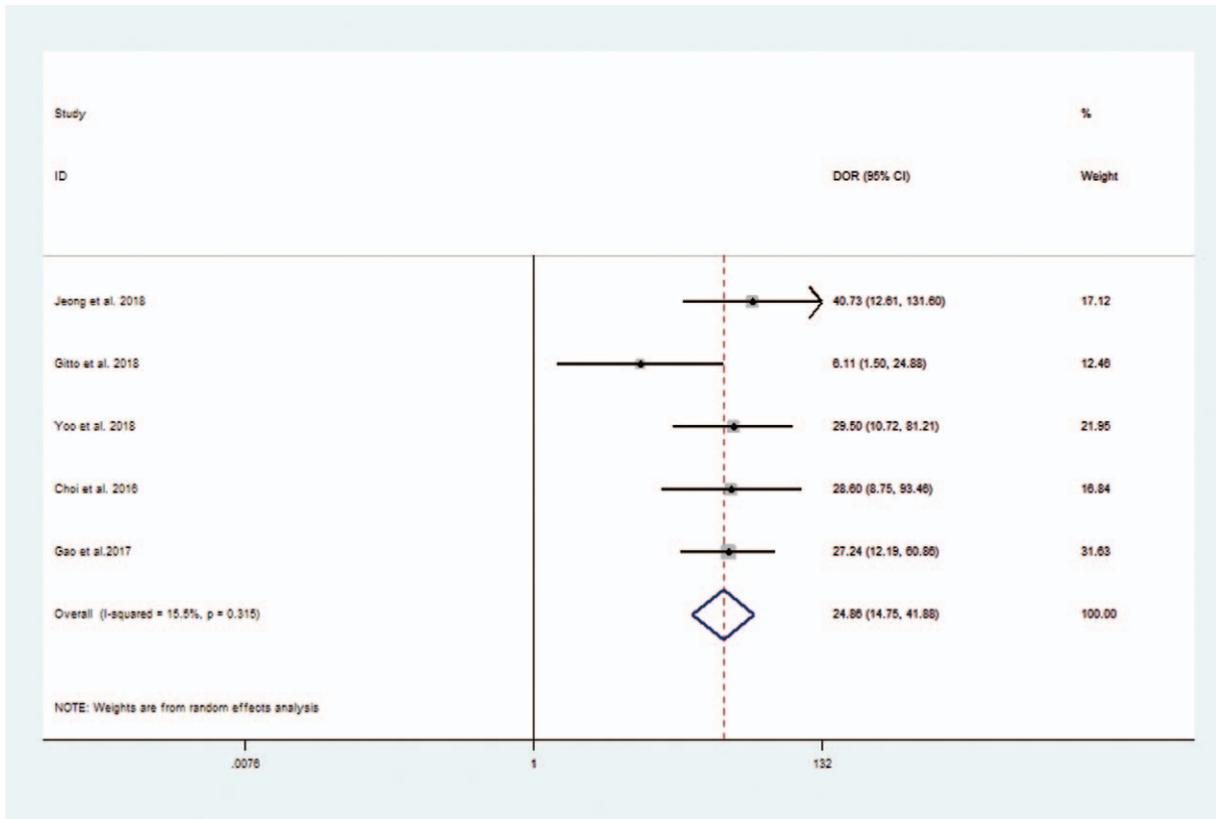


Figure 6. Forest plot of diagnostic odds ratio of computer-aided diagnosis system for the diagnosis of thyroid nodules.

nodules in the included studies. Among thyroid nodules, 299 nodules were benign and 424 nodules were malignant. The average sizes of benign or malignant thyroid nodules were 1.1 to 1.8 or 0.9 to 1.8 cm. The nodules in included studies were

identified by pathologic diagnosis, such as FNA or surgery. Jeong et al^[35] compared the diagnostic performance of CAD system based on the operators' experience (0–5 years) with an experienced radiologist. Gitto et al^[33] compared the diagnostic performance of CAD system with radiologist in the characterization of low-to-high suspicion thyroid nodules based on the Thyroid Imaging Reporting and Data System (TIRADS). Yoo et al^[34] compared the diagnostic performance of the radiologist, the CAD system and the CAD-assisted radiologist. Gao et al^[32] compared the diagnostic efficiency of CAD system with an experience radiologist based on different standards, such as the TIRADS (K-TIRADS and ACR TIRADS) and the American Thyroid Association (ATA). Choi et al^[31] made the subgroup analysis to compare the diagnostic performance of the CAD system and radiologist for thyroid malignancies >1 cm. The diagnostic performance results (sensitivity, specificity, positive predictive value, negative predictive value, accuracy and area under the receiver-operating characteristic [AUC]) of CAD system or experienced radiologists for thyroid nodules in the 5 included studies are presented in Table 2.

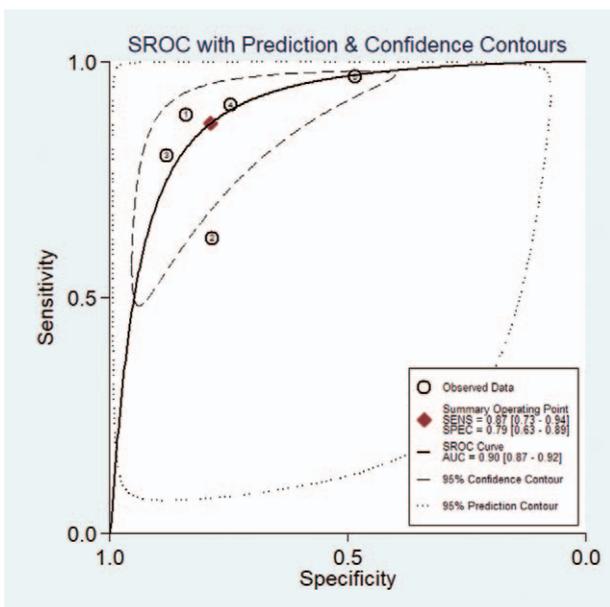


Figure 7. Summary receiver-operating characteristic (SROC) curves of computer-aided diagnosis system for the diagnosis of thyroid nodules.

3.3. Study quality assessment and publication bias

Results of QUADAS-2 assessment are presented in Figures 2 and 3. The quality of studies varied from moderate to high. The main causes of potential bias were attributed to patient selection, index test, and flow and timing. The quality of the included studies was deemed satisfactory. The Deeks funnel plot asymmetry test showed $P = .39$ and indicated no publication bias, as shown in Figure 4.

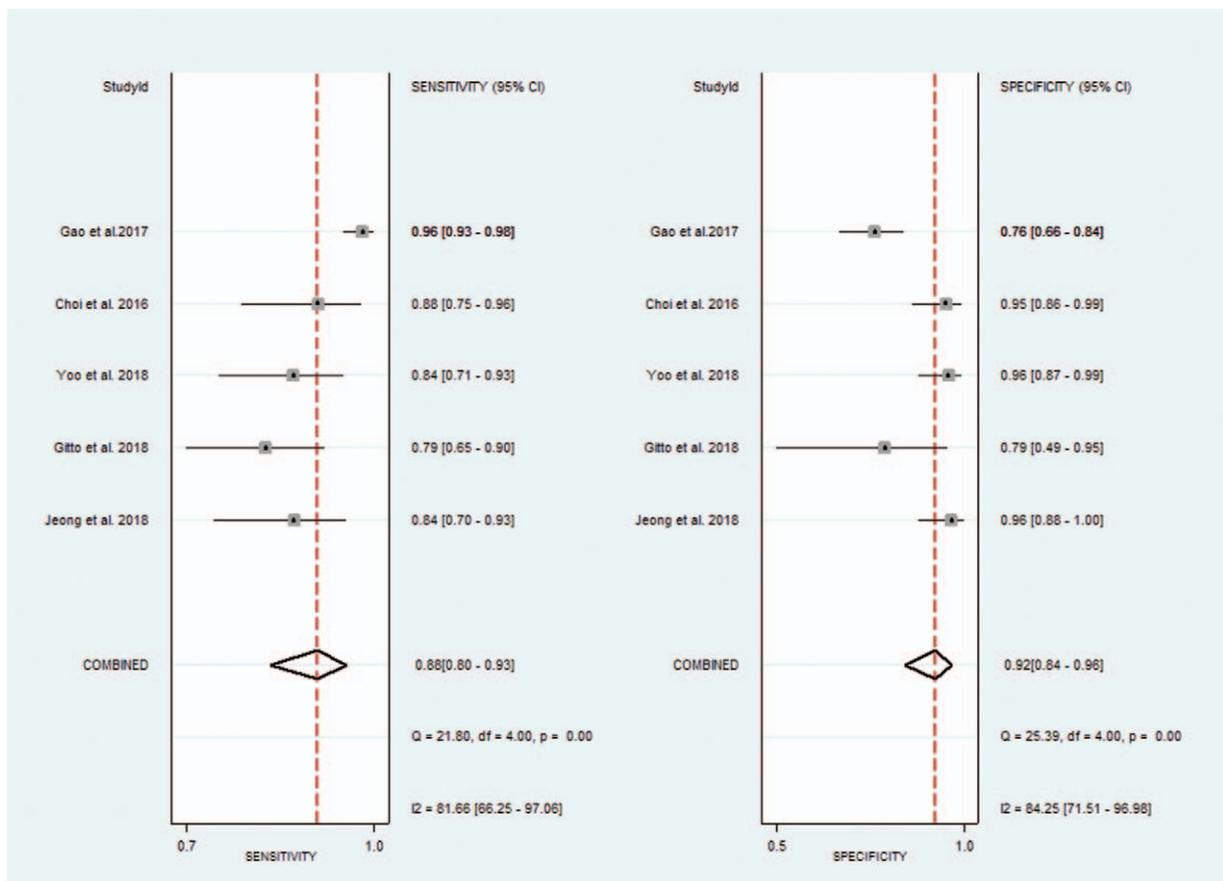


Figure 8. Forest plot of pooled sensitivity and specificity of experienced radiologists for the diagnosis of thyroid nodules.

3.4. Diagnostic accuracy and heterogeneity evaluation

The diagnostic performance results of CAD system or experienced radiologists for thyroid nodules in the 5 included studies in the meta-analysis are presented in Table 3. The pooled sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, and DOR for CAD system were 0.87 (95% CI 0.73–0.94), 0.79 (95% CI 0.63–0.89), 4.1 (95% CI 2.5–6.9), 0.17 (95% CI 0.09–0.32), and 25 (95% CI 15–42), respectively. The Forest figures of sensitivity, specificity, and DOR of CAD system were showed in Figures 5 and 6. The heterogeneity of sensitivity, specificity and DOR of CAD system in the meta-analysis were showed as $I^2 = 93.53\%$ (95% CI 89.44–97.63%), $I^2 = 89.67\%$ (95% CI 82.20–97.14%), and $I^2 = 15.5\%$ ($P = .315$), respectively. Figure 7 shows the SROC curve and indicated that the area under the curve was 0.90 (95% CI 0.87–0.92).

The pooled sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, and DOR for experienced radiologists were 0.88 (95% CI 0.80–0.93), 0.92 (95% CI 0.84–0.96), 11.1 (95% CI 5.6–21.9), 0.13 (95% CI 0.08–0.21), and 86 (95% CI 47–158), respectively. The Forest figures of sensitivity, specificity, and DOR of experienced radiologist are showed in Figures 8 and 9. The heterogeneity of sensitivity, specificity and DOR of experienced radiologist in the meta-analysis were showed as $I^2 = 81.66\%$ (95% CI 66.25–97.06%), $I^2 = 84.25\%$ (95% CI 71.51–96.98%), and $I^2 = 41.1\%$ ($P = .147$), respectively. Figure 10 shows the SROC

curve and indicates that the area under the curve was 0.96 (95% CI 0.94–0.97).

Considering the heterogeneity, subgroup meta-analysis was also performed for CAD by Samsung (shown in Table 3). The pooled sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, and DOR for experienced radiologist were 0.82 (95% CI 0.69–0.91), 0.83 (95% CI 0.76–0.89), 4.9 (95% CI 3.4–7.0), 0.22 (95% CI 0.12–0.38), and 23 (95% CI 11–46), respectively. The Forest figures of sensitivity, specificity and DOR of CAD system by Samsung are showed in Figures 11 and 12. The heterogeneity of sensitivity, specificity, and DOR of CAD system in the meta-analysis were showed as $I^2 = 79.62\%$ (95% CI 59.70–99.54%), $I^2 = 27.52\%$ (95% CI 0.00–98.84%), and $I^2 = 35.9\%$ ($P = .197$), respectively. Figure 13 shows the SROC curve and indicated that the area under the curve was 0.88 (95% CI 0.85–0.91).

Meanwhile, a meta-regression analysis was also performed and showed that there was a lack of heterogeneity based on ethnicity, study design, and so on.

4. Discussion

Ultrasound, as a noninvasive and inexpensive imaging technique, is useful not only for detection but also for discrimination between benign and malignant thyroid nodules, which does not have any radioactive hazards and has a short acquisition

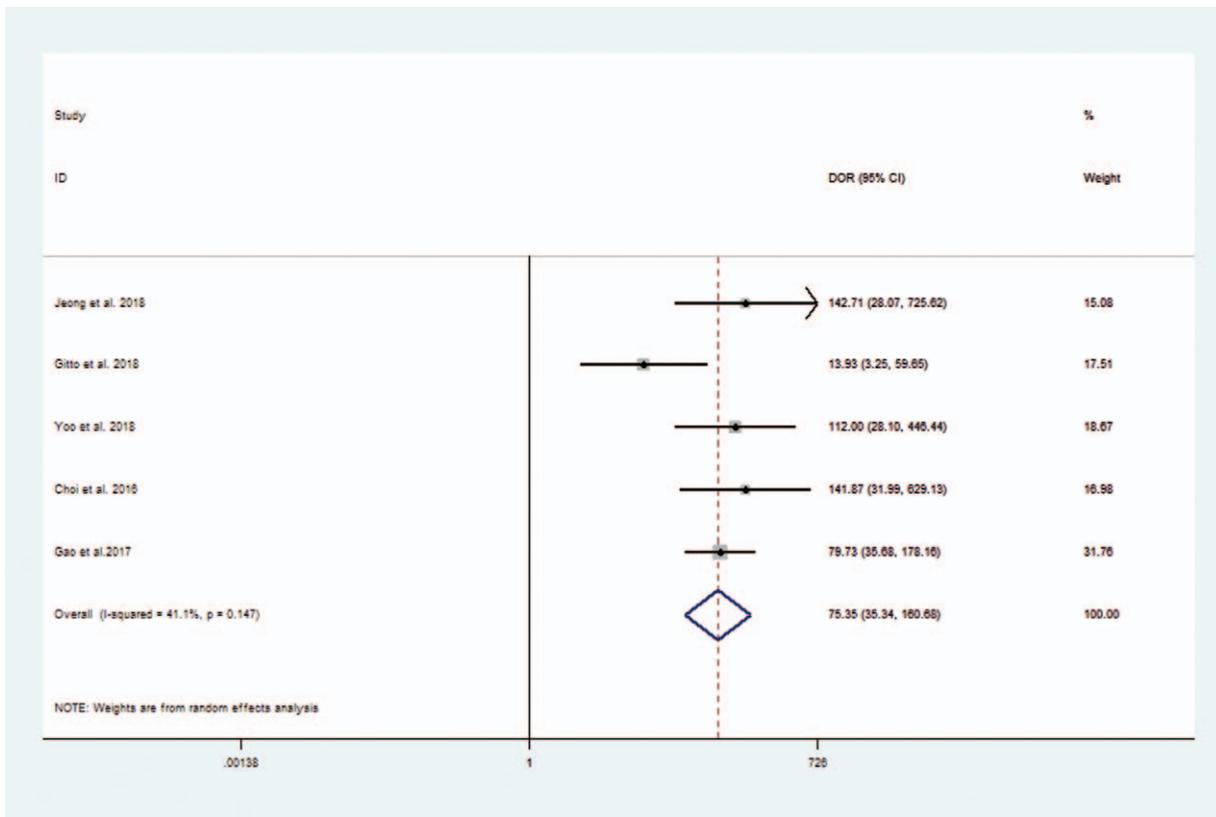


Figure 9. Forest plot of diagnostic odds ratio of experienced radiologists for the diagnosis of thyroid nodules.

time.^[22,44] However, ultrasound has some limitations that the reliability of diagnosis depends on factors such as the quality of images and the expertise of radiologist who interprets the images, which is easily affected by echo perturbation and speckle

noise.^[45,46] Additionally, different types of benign and malignant nodules have different characteristics in an ultrasound image, which adds difficulties to the recognition.^[10,47] Accurate determination of whether thyroid nodules are benign or malignant is necessary, which reduces patients' risk and the significant medical costs with FNA or surgery.^[48] Therefore, interpretation of ultrasound images can only be done by radiologists with lots of experience and training to ensure the accuracy. Meanwhile, more and more automated efficient ultrasound image analysis techniques,^[49–52] such as ultrasound-based CAD, were developed to obtain accurate, reproducible, and objective diagnostic results. The current CAD systems mainly include 2 aspects of technology: the recognition of benign and malignant thyroid nodules and the recognition of thyroid boundary. Table 4 shows characteristics of some studies^[14,24,36–40] on algorithms of CAD system. The concrete implementation flow of most existing CAD systems based on a neural network model is shown in Figure 14. Firstly, the classifiers are trained using the selected features and the ground truth of whether the image is benign or malignant. Secondly, the classifier training parameters are applied on the features selected from the test images to predict the class of the test image. Once, several test images are evaluated in such a way, then the predicted class labels are compared with the ground truth of the test images to calculate the classifier performance measures like accuracy, sensitivity, and specificity. Finally, the classifier resulting in the best accuracy is then chosen as the optimal classifier for future implementations of the CAD system.

So far, whether the diagnostic performance of existing CAD systems can reach the diagnostic level of experienced radiologists

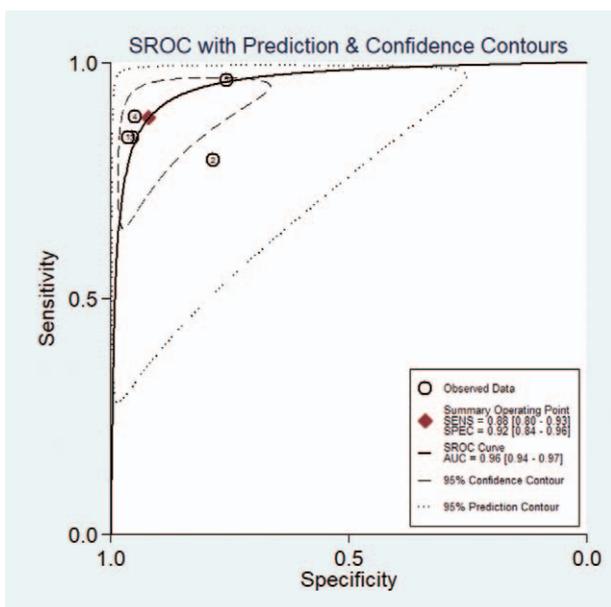


Figure 10. Summary receiver-operating characteristic (SROC) curves of experienced radiologists for the diagnosis of thyroid nodules.

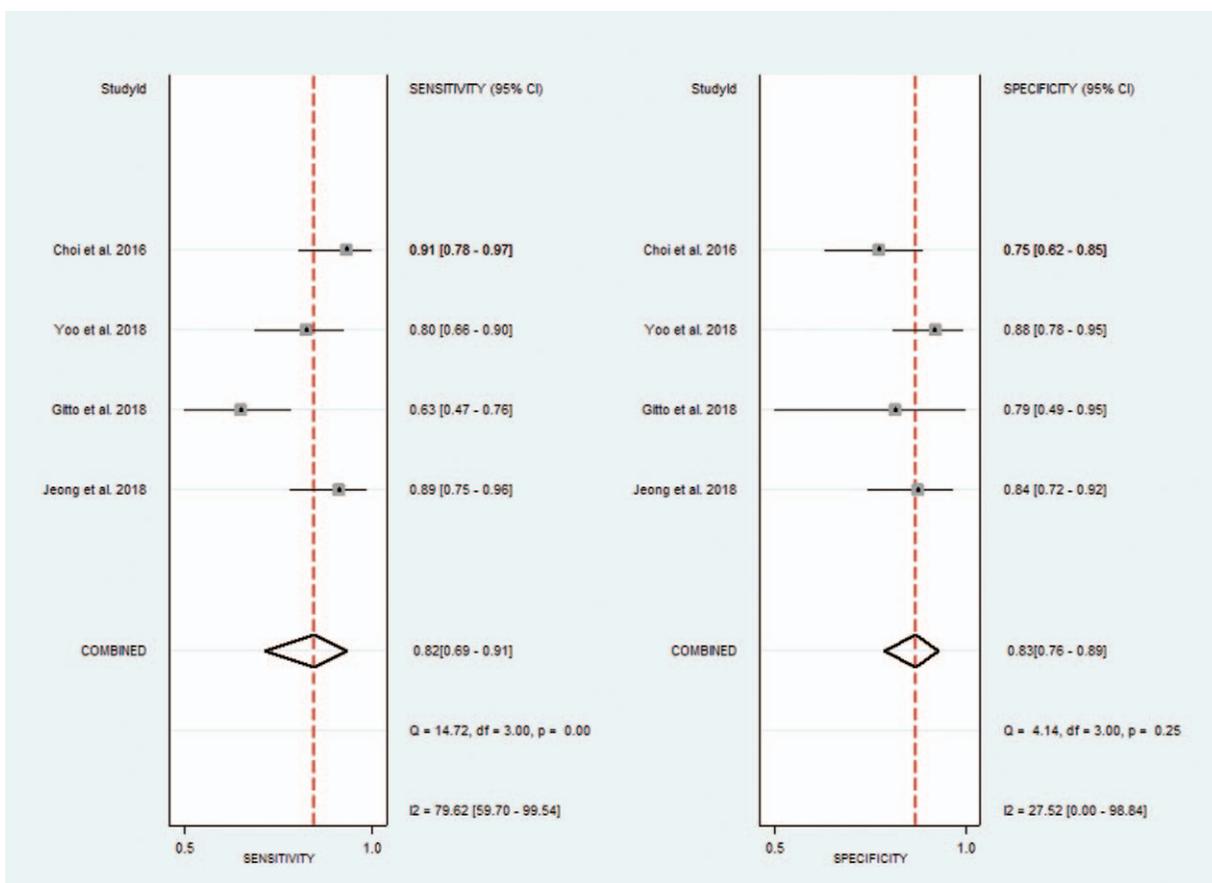


Figure 11. Forest plot of pooled sensitivity and specificity of computer-aided diagnosis system by Samsung for the diagnosis of thyroid nodules.

is still controversial. This study is the 1st meta-analysis study to evaluate the diagnostic value of CAD systems. The finding of this meta-analysis showed that the CAD system had similar sensitivity and negative likelihood ratio compared with experienced radiologists. However, the specificity, the positive likelihood ratio and DOR were relatively low. These results indicated that there was a clear gap between the CAD system and the experienced radiologist in diagnostic performance of thyroid nodules. Besides, the time required for postprocessing needs to be analyzed. Only Gitto et al^[33] reported that the CAD system by Samsung cost four to 5 minutes for per thyroid nodule, which was much longer than radiologists. The unsatisfied time consumption was mainly related to the selection of segmented nodule among a series of candidates provided by the CAD system or to the need of manually modifying the contours of the lesion. It can be seen that segmentation of the existing CAD system also needs manual participation so it belongs to semi-intelligent products. Additionally, successful nodule segmentations were important and influenced the precision of recognition of nodules. Choi et al^[31] reported that the successful nodule segmentations were observed in 87.3% (89/102) of nodules. Poor nodule segmentation occurred more frequently with benign nodules ($n = 11$, 18.6%) than with malignant nodules ($n = 2$, 4.7%) and the difference was statistically significant ($P = .04$). Among those nodules with poor segmentation, 54.6% of the benign nodules (6/11) were also diagnosed as malignant, while all of malignant nodules were diagnosed as malignant. Thus, it can be seen that

poor segmentation on CAD system may increase the FP rate without affection on the FN rate.

The CAD system still has a long way to go to replace experienced radiologists in the process of improving accuracy and reducing time consumption. The existing CAD system's standpoint is to assist radiologists to diagnose the thyroid nodule. Yoo et al^[34] confirmed that the CAD system-assisted radiologist yielded a higher diagnostic sensitivity than the radiologist alone, which implied that the CAD system allowed the radiologist to detect a higher proportion of genuine malignancies.

Regarding the interobserver agreement for the ultrasound characteristics,^[53-55] there was a good agreement for most of the US descriptors between the CAD system and radiologists. However, Jeong et al^[35] found that the CAD system's interpretation of margin and composition failed to match the experienced radiologists. Besides, Choi et al^[31] also reported that fail agreement was found for the margin definitions between the 2. For margin descriptors, it may depend on the semi-automated CAD system's step for the margin descriptors, which requires experience of the operators. Additionally, the current CAD system still has some limitations regarding the interpretation of compositions. Jeong et al^[35] reported that solid nodules with a marked hypoechoic or hypoechoic component were commonly misinterpreted as a partially cystic or partially solid nodule. Therefore, further validation pertaining to this issue and an assessment based on a larger study are required to improve the current CAD system.

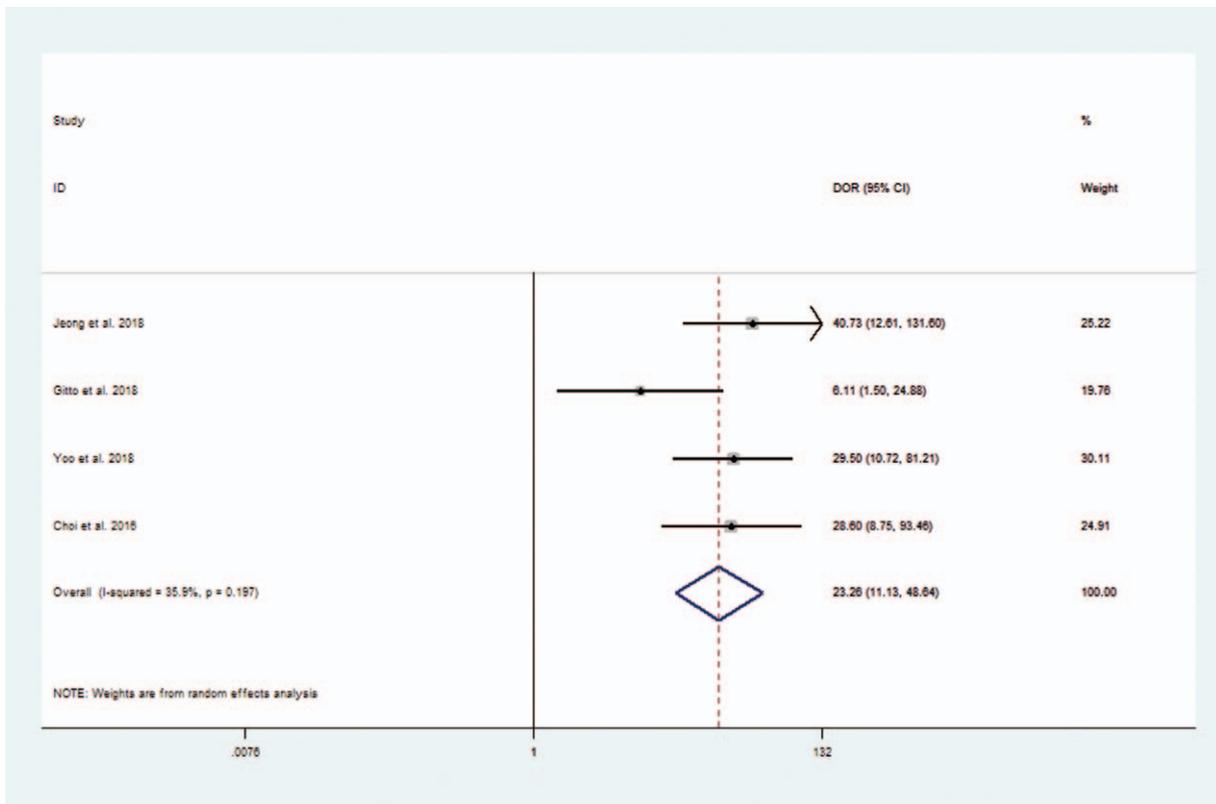


Figure 12. Forest plot of diagnostic odds ratio of computer-aided diagnosis system by Samsung for the diagnosis of thyroid nodules.

The strength of this systematic review is that we included adequate studies in the literature available for meta-analysis. However, our study also had several limitations. Firstly, the

current meta-analysis showed a considerable heterogeneity of sensitivity and specificity among studies. This heterogeneity is likely to arise through diversity in methodologic aspects among different studies, and the basic differences among the patients in the included studies may also have contributed to the observed heterogeneity of the results. However, in sub-meta-analysis and meta-regression analysis of the current review, the interpretation criteria in this study did not possess some sources of heterogeneity. To minimize bias in the selection of studies and the data extraction, reviewers who were blinded to the journal, author, institution, and date of publication independently selected articles based on the inclusion criteria, and scores were assigned to study design characteristics and examination results by using a standardized form that was based on the QUADAS-2. Secondly, the sample size of included studies was relatively small. Therefore, selection bias may have been present. Thirdly, for the small number of thyroid nodules, the gap between benign group and malignant group is large, which would increase measure bias. Additionally, publication bias is a major concern in all meta-analyses as studies reporting significant findings are more likely to be published than those reporting nonsignificant results. However, we assessed the publication bias in our analysis by using Deek funnel plot which showed no definite asymmetry.

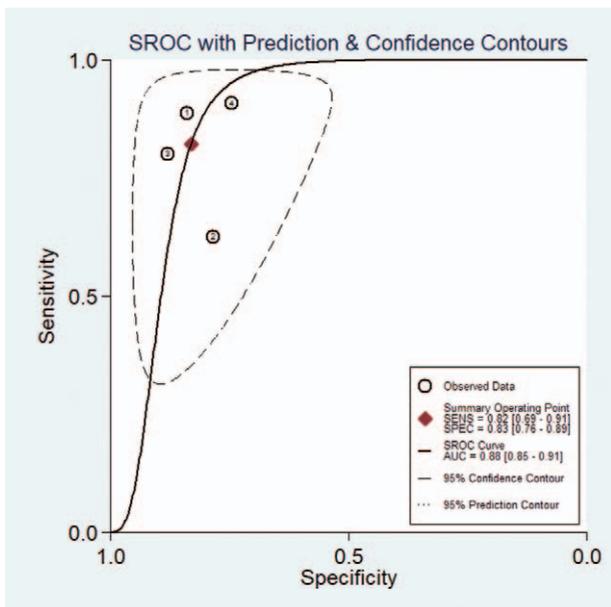


Figure 13. Summary receiver-operating characteristic (SROC) curves of computer-aided diagnosis system by Samsung for the diagnosis of thyroid nodules.

In conclusion, the sensitivity of the CAD system in thyroid nodules was similar to that of experienced radiologists. However, the CAD system had lower specificity and DOR than the experienced radiologist. Larger sample-size prospective studies need be warranted to evaluate the diagnostic performance of CAD system further. The CAD system may play the potential role

Table 4

Characteristics of some studies on algorithms of computer-aided diagnosis system.

Author	Year	Country	Patients for test	The number of nodules for test			The number of nodules for training			Research Content	Details
				Total	Benign	Malignant	Total	Benign	Malignant		
Koundal et al	2018	India	na	na	na	na	na	na	na	NDRLS vs NCM vs NLM	Speckle reduction to reduce speckle noise and preserve the diagnostic features of ultrasound image, automatic generation of ROI that identifies suspicious regions and fully automatic segmentation of nodule in processed ROI image
Lee et al	2018	Korea	na	200	100	100	549	263	286	CAD for metastasis lymph nodes	The accuracy, sensitivity, and specificity for predicting lymph node malignancy were 83.0%, 79.5%, and 87.5%, respectively
Baig et al	2017	China	111	111	84	27	na	na	na	VI vs VI&GSU vs GSU	A novel algorithm for regional subdivision and quantification of thyroid nodular VI in ultrasound images. The optimum cut-off of peripheral, central, and overall VI was 19.7%, 9.1%, and 20.2%, respectively
Ma et al	2017	China	na	na	na	na	10357 (6242 Ps)	na	na	CNN	A deep CNN to automatically segment thyroid nodules from ultrasound images
Qin et al	2017	China	45	50	33	17	610 (543 Ps)	403	207	ANN vs SVM vs ANN&SVM	Improve segmentation performance, an ignition ROI was drawn with landmark points on the edge by a radiologist. The ANN model showed sensitivity, specificity, positive and negative predictive values, Youden index, and accuracy of 88.24%, 90.91%, 83.33%, 93.75%, 79.14%, and 90.00%, respectively, the SVM model 76.47%, 90.91%, 81.25%, 88.24%, 67.38%, and 86.00%, respectively, and in combination, 100.00%, 87.88%, 80.95%, 100.00%, 87.88%, and 92.00%, respectively
Chang et al	2016	Korea	na	59	29	30	na	na	na	SVM vs radiologist	An initial ROI is drawn using two landmark points on the center and edge. The computed areas under curves in the ROC analysis were 0.986 and 0.979 for the proposed CAD system and visual inspection by radiologists, respectively; no significant difference was detected between these groups
Lim et al	2008	Korea	96	109	60	49	na	na	na	ANN vs BLR vs radiologist	The area under the ROC curve (Az) values of ANN and BLR were 0.9492 ± 0.0195 and 0.9046 ± 0.0289 , respectively. The performance of the ANN and the BLR was better than that of the radiologists

ANN = artificial neural network, BLR = binary logistic regression, CNN = convolutional neural network, GSU = gray-scale ultrasound, na = not available, NDRLS = neutrosophic distance regularized level set, NCM = neutrosophic-C means, NLM = neutrosophic-L means, Ps = patients, ROI = region of interest, SVM = support vector machine, VI = vascular index.

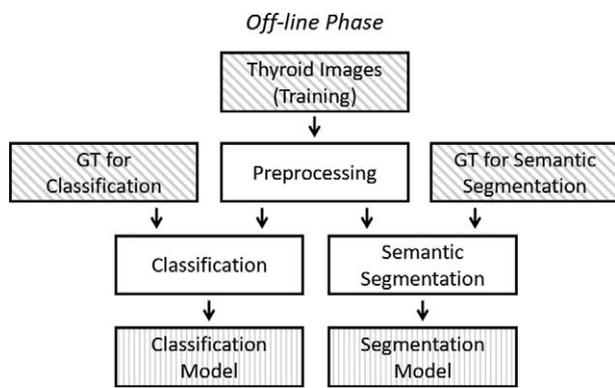


Figure 14. The concrete implementation flow of most existing computer-aided diagnosis systems based on neural network model.

as a decision-making assistant alongside radiologists in the thyroid nodules' diagnosis. Future technical improvements in the image segmentation and recognition, which also includes automatic detection of characteristics such as nodule calcification, would be helpful to increase the accuracy as well as diagnostic efficiency.

Acknowledgments

WJZ acknowledges funding from the Postdoctoral Sustentation Fund of West China Hospital, Sichuan University, China (Grant No. 18HXBH074).

Author contributions

Conceptualization: Wan-Jun Zhao.

Data curation: Lin-Ru Fu, Zhi-Mian Huang.

Formal analysis: Wan-Jun Zhao.

Investigation: Lin-Ru Fu, Zhi-Mian Huang.

Methodology: Wan-Jun Zhao.

Project administration: Wan-Jun Zhao.

Resources: Lin-Ru Fu, Zhi-Mian Huang.

Supervision: Wan-Jun Zhao.

Writing – original draft: Wan-Jun Zhao, Lin-Ru Fu.

Writing – review & editing: Jing-Qiang Zhu, Bu-Yun Ma.

Lin-Ru Fu orcid: 0000-0001-6368-5578.

References

- [1] Guth S, Theune U, Aberle J, et al. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. *Eur J Clin Invest* 2010;39:699–706.
- [2] Wolinski K, Stangierski A, Ruchala M. Comparison of diagnostic yield of core-needle and fine-needle aspiration biopsies of thyroid lesions: systematic review and meta-analysis. *Eur Radiol* 2017;27:431–6.
- [3] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin* 2018;68:7–30.
- [4] Xiaoyi L, General Do. The interpretation of 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: surgery part. *China Oncol* 2017;52:309–15.
- [5] Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid* 2016;26:1–33.
- [6] Gharib H, Papini E, Garber JR, et al. American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi Medical Guidelines for Clinical Practice for the Diagnosis and Management of Thyroid Nodules—2016 Update. *Endocr Pract* 2016;22:622–39.
- [7] Moon WJ, Jung SL, Lee JH, et al. Benign and malignant thyroid nodules: US differentiation—multicenter retrospective study. *Radiology* 2008;247:762–70.
- [8] Mainini AP, Monaco C, Pescatori LC, et al. Image-guided thermal ablation of benign thyroid nodules. *J Ultrasound* 2016;20:11–22.
- [9] Mauri G, Nicosia L, Della Vigna P, et al. Percutaneous laser ablation for benign and malignant thyroid diseases. *Ultrasonography (Seoul, Korea)* 2019;38:25–36.
- [10] Park CS, Kim SH, Jung SL, et al. Observer variability in the sonographic evaluation of thyroid nodules. *J Clin Ultrasound* 2010;38:287–93.
- [11] Park SH, Kim SJ, Kim EK, et al. Interobserver agreement in assessing the sonographic and elastographic features of malignant thyroid nodules. *Am J Roentgenol* 2009;193:W416–23.
- [12] Park SJ, Park SH, Choi YJ, et al. Interobserver variability and diagnostic performance in US assessment of thyroid nodule according to size. *Ultraschall in Der Medizin* 2012;33:E186–90.
- [13] Choi SH, Kim EK, Kwak JY, et al. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. *Thyroid* 2010;20:167–72.
- [14] Chang Y, Paul AK, Kim N, et al. Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: a comparison with radiologist-based assessments. *Med Phys* 2016;43:554.
- [15] Acharya UR, Sree SV, Krishnan MM, et al. Computer-aided diagnostic system for detection of Hashimoto thyroiditis on ultrasound images from a Polish population. *J Ultrasound Med* 2014;33:245–53.
- [16] Zhang P, Xu X, Wang H, et al. Computer-aided lung cancer diagnosis approaches based on deep learning. *J Comput Aided Design Comput Graph* 2018;30:90.
- [17] Paulraj T, Chelliah KSV. Computer-aided diagnosis of lung cancer in computed tomography scans: a review. *Curr Med Imag Rev* 2018;14:374–88.
- [18] Narayanan BN, Hardie RC, Kebede TM. Performance analysis of a computer-aided detection system for lung nodules in CT at different slice thicknesses. *J Med Imaging (Bellingham)* 2018;5:014504.
- [19] Mohebian MR, Marateb HR, Mansourian M, et al. A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning. *Comput Struct Biotechnol J* 2017;15:75–85.
- [20] Ertas G, Doran SJ, Leach MO. A computerized volumetric segmentation method applicable to multi-centre MRI data to support computer-aided breast tissue analysis, density assessment and lesion localization. *Med Biol Eng Comput* 2017;55:57–68.
- [21] Ardakani AA, Gharbali A, Mohammadi A. Classification of benign and malignant thyroid nodules using wavelet texture analysis of sonograms. *J Ultrasound Med* 2015;34:1983–9.
- [22] Acharya UR, Swapna G, Sree SV, et al. A review on ultrasound-based thyroid cancer tissue characterization and automated classification. *Technol Cancer Res Treat* 2014;13:289–301.
- [23] Garg H, Jindal A. Segmentation of thyroid gland in ultrasound image using neural network. Paper presented at: International Conference on Computing, 2013.
- [24] Lim KJ, Choi CS, Yoon DY, et al. Computer-aided diagnosis for the differentiation of malignant from benign thyroid nodules on ultrasonography. *Acad Radiol* 2008;15:853–8.
- [25] Chen D, Niu J, Qiao P, et al. A deep-learning based ultrasound text classifier for predicting benign and malignant thyroid nodules. Paper presented at: International Conference on Green Informatics, 2017.
- [26] Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Int Med* 2011;155:529–36.
- [27] Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56:1129–35.
- [28] Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982–90.
- [29] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2010;20:2865–84.

- [30] Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005;58:882–93.
- [31] Choi YJ, Baek JH, Park HS, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment. *Thyroid* 2017;27:546–52.
- [32] Gao L, Liu R, Jiang Y, et al. Computer-aided system for diagnosing thyroid nodules on ultrasound: a comparison with radiologist-based clinical assessments. *Head Neck* 2017;40:778–83.
- [33] Gitto S, Grassi G, De Angelis C, et al. A computer-aided diagnosis system for the assessment and characterization of low-to-high suspicion thyroid nodules on ultrasound. *Radiol Med* 2019;124:118–25.
- [34] Yoo YJ, Ha EJ, Cho YJ, et al. Computer-aided diagnosis of thyroid nodules via ultrasonography: initial clinical experience. *Korean J Radiol* 2018;19:665–72.
- [35] Jeong EY, Kim HL, Ha EJ, et al. Computer-aided diagnosis system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators. *Eur Radiol* 2019;29:1978–85.
- [36] Baig FN, Lunenburg JTJV, Liu SYW, et al. Computer-aided assessment of regional vascularity of thyroid nodules for prediction of malignancy. *Sci Rep* 2017;7:14350.
- [37] Hoon LJ, Hwan BJ, Han KJ, et al. Deep learning-based computer-aided diagnosis system for localization and diagnosis of metastatic lymph nodes on ultrasound: a pilot study. *Thyroid* 2018;28:1332–8.
- [38] Ma J, Wu F, Jiang TA, et al. Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. *Int J Comput Assist Radiol Surg* 2017;12:1895–910.
- [39] Maroulis DE, Savelonas MA, Karkanis SA, et al. Computer-aided thyroid nodule detection in ultrasound images. *Proceedings of the IEEE Symposium on Computer-Based Medical Systems*, 2005:271–276.
- [40] Yu Q, Jiang T, Zhou A, et al. Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images. *Eur Arch Otorhinolaryngol* 2017;274:2891–7.
- [41] Choi WJ, Park JS, Kim KG, et al. Computerized analysis of calcification of thyroid nodules as visualized by ultrasonography. *Eur J Radiol* 2015;84:1949–53.
- [42] Lin YH, Tsai YC, Lin KJ, et al. Computer-aided diagnostic technique in 2-deoxy-2-[(18)F]fluoro-D-glucose-positive thyroid nodule: clinical experience of 74 non-thyroid cancer patients. *Ultrasound Med Biol* 2019;45:108–21.
- [43] Zhu LC, Ye YL, Luo WH, et al. A model to discriminate malignant from benign thyroid nodules using artificial neural network. *PLoS One* 2013;8:e82211.
- [44] Al NA, Gionfriddo MR, Javed A, et al. Accuracy of thyroid nodule sonography for the detection of thyroid cancer in children: systematic review and meta-analysis. *Clin Endocrinol* 2016;84:423–30.
- [45] Ivanac G, Brkljacic B, Ivanac K, et al. Vascularisation of benign and malignant thyroid nodules: CD US evaluation. *Ultraschall in Der Medizin* 2007;28:502–6.
- [46] Sipos JA. Advances in ultrasound for the diagnosis and management of thyroid cancer. *Thyroid* 2009;19:1363–72.
- [47] Bastin S, Bolland MJ, Croxson MS. Role of ultrasound in the assessment of nodular thyroid disease. *J Med Imaging Radiat Oncol* 2010;53:177–87.
- [48] Gamme G, Parrington T, Wiebe E, et al. The utility of thyroid ultrasonography in the management of thyroid nodules. *Can J Surg* 2017;60:134–9.
- [49] Caresio C, Caballo M, Deandrea M, et al. Quantitative analysis of thyroid tumors vascularity: a comparison between 3-D contrast-enhanced ultrasound and 3-D Power Doppler on benign and malignant thyroid nodules. *Med Phys* 2018;45:3173–84.
- [50] Jin XZ, Lu WW, Zhang HF, et al. Comparative study on the diagnostic values of different ultrasound technologies for malignant thyroid nodules. *Oncol Lett* 2018;16:910–4.
- [51] Lang HH, Woo YC, Chiu WH. Role of second high-intensity focused ultrasound (HIFU) treatment for unsatisfactory benign thyroid nodules after first treatment. *Eur Radiol* 2018;29:1469–78.
- [52] Wang Y, Wei K, Wan P. A method of ultrasonic image recognition for thyroid papillary carcinoma based on deep convolution neural network. *Neuroquantology* 2018;16:757–68.
- [53] Milas M, Mandel SJ, Langer JE, et al. *Advanced Thyroid and Parathyroid Ultrasound [M]*. Switzerland: Springer International Publishing; 2017. 179–194.
- [54] Kabaker AS, Tublin ME, Nikiforov YE, et al. Suspicious ultrasound characteristics predict BRAFV600E-positive papillary thyroid carcinoma. *Thyroid* 2012;22:585–9.
- [55] Smithbindman R, Lebda P, Feldstein VA, et al. Risk of thyroid cancer based on thyroid ultrasound imaging characteristics: results of a population-based study. *JAMA Intern Med* 2013;173:1788–96.