Original Article

# Assessment of the statistical optimization strategies and clinical evaluation of an artificial intelligence-based automated diagnostic system for thyroid nodule screening

**Fangqi Guo[1,2#], Wanru Chang[3#], Jiaqi Zhao[2]^, Lei Xu[4], Xuan Zheng[5], Jia Guo[6]**

[1]Department of Ultrasound, Second Affiliated Hospital (Changzheng Hospital) of Naval Medical University, Shanghai, China; [2]Department of Ultrasound, Shanghai Fourth People's Hospital, School of Medicine, Tongji University, Shanghai, China; [3]School of Mathematical Sciences, Zhejiang University, Hangzhou, China; [4]Zhejiang Qiushi Institute for Mathematical Medicine, Hangzhou, China; [5]Demetics Medical Technology, Hangzhou, China; [6]Shuguang Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai, China

*Contributions:* (I) Conception and design: J Zhao, L Xu; (II) Administrative support: J Zhao, J Guo; (III) Provision of study materials or patients: F Guo, W Chang; (IV) Collection and assembly of data: F Guo, W Chang, L Xu; (V) Data analysis and interpretation: W Chang, L Xu, X Zheng; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

*Correspondence to:* Jiaqi Zhao. Department of Ultrasound, Shanghai Fourth People's Hospital, School of Medicine, Tongji University, Shanghai 200434, China. Email: ultrasoundczzjq@163.com.

**Background:** Thyroid cancer is the most common endocrine cancer in the world. Accurately distinguishing between benign and malignant thyroid nodules is particularly important for the early diagnosis and treatment of thyroid cancer. This study aimed to investigate the best possible optimization strategies for an already-trained artificial intelligence (AI)-based automated diagnostic system for thyroid nodule screening and, in addition, to scrutinize the clinically relevant limitations using stratified analysis to better standardize the application in clinical workflows.

**Methods:** We retrospectively reviewed a total of 1,092 ultrasound images associated with 397 thyroid nodules collected from 287 patients between April 2019 and January 2021, applying postoperative pathology as the gold standard. We applied different statistical approaches, including averages, maximums, and percentiles, to estimate per-nodule-based malignancy scores from the malignancy scores per image predicted by AI-SONIC Thyroid v. 5.3.0.2 (Demetics Medical Technology Ltd., Hangzhou, China) system, and we assessed its diagnostic efficacy on nodules of different sizes or tumor types with per-nodule analysis using performance metrics.

**Results:** Of the 397 thyroid nodules, 272 thyroid nodules were overrepresented by malignant nodules according to the results of the surgical pathological examinations. Taking the median of the malignancy scores per image to estimate the nodule-based score with a cutoff value of 0.56 optimized for the means of sensitivity and specificity, the AI-based automated detection system demonstrated slightly lower sensitivity, significantly higher specificity (almost independent of nodule size), and similar accuracy to that of the senior radiologist. Both the AI system and the senior radiologist demonstrated higher sensitivity in diagnosing smaller nodules (≤25 mm) and comparable diagnostic performances for larger nodules. The mean diagnostic time per nodule of the AI system was 0.146 s, which was in sharp contrast to the 2.8 to 4.5 min of the radiologists.

**Conclusions:** Using our optimization strategy to achieve nodule-based diagnosis, the AI-SONIC Thyroid automated diagnostic system demonstrated an overall diagnostic accuracy equivalent to that of senior

^ ORCID: 0000-0003-1212-9525.

radiologists. Thus, it is expected that it can be used as a reliable auxiliary diagnostic method by radiologists for the screening and preoperative evaluation of malignant thyroid nodules.

**Keywords:** Thyroid nodule; radiologists; artificial intelligence (AI); AI automated diagnostic system

Submitted Jan 27, 2022. Accepted for publication Nov 24, 2022. Published online Jan 02, 2023.
doi: 10.21037/qims-22-85
**View this article at:** https://dx.doi.org/10.21037/qims-22-85

## Introduction

Thyroid cancer is the most common endocrine cancer in the world (1). Over the past few decades, published studies have shown that the incidence of thyroid cancer has been on the rise (2). The incidence of thyroid cancer in the United States has nearly doubled since 2000, and it is now the most common cancer among women (3-5). Accurately distinguishing benign and malignant thyroid nodules is particularly important for the early diagnosis and treatment of thyroid cancer. High-frequency ultrasound has been widely used in thyroid nodule examinations due to its noninvasiveness, convenience, and accuracy. However, due to the overlap of some features of benign and malignant thyroid nodules, the interpretation of image features is susceptible to the subjective biases of radiologists with different experiences (6). These factors may lead to misdiagnosis or even overdiagnosis, resulting in irreversible negative consequences (7). In May 2017, the American College of Radiology (ACR) released a standard guide for the diagnostic classification of thyroid nodules on ultrasound, the Thyroid Imaging Reporting and Data System (TI-RADS) (8). In this system, 5 categories of thyroid nodular features—composition, echogenicity, shape, margin, and echogenic foci—are assessed, and risk classification is determined according to a sum of scores (9). Despite offering a more standardized approach to diagnosis, the TI-RADS is still not able to objectify subjective human assessments.

As emerging and innovative technology, artificial intelligence (AI)-based automated diagnostic systems have great potential for overcoming the subjective limitations of human interpretation and have attracted the significant attention of researchers (10). Several studies have shown that, in terms of sensitivity, the AI-based automated diagnostic systems for thyroid nodules provide comparable diagnoses to those of senior radiologists (11-16). However, while in clinical practice, diagnoses are made for individual nodules, an AI system outputs a malignancy probability for each ultrasound image. In the studies where AI systems were used to discriminate benign from malignant nodules, most of the authors (14-17) made no mention of the concept of per-nodule–based diagnosis or of any postprocessing of returned malignancy probabilities for multiple images associated with the same nodule by an AI system. One exception was a study (11) in which the authors simply took the maximum probability value and used a naïve cutoff value of 0.5. To our knowledge, there has been no study that investigates the best possible way to convert per-image–based to per-nodule–based malignancy diagnosis. Additionally, tests of nodule-size stratification are often ignored during evaluation of AI systems (16-19). However, there have been numerous studies showing size-dependent diagnostic difficulty for thyroid nodules by radiologists using conventional B-mode ultrasound-based TI-RADS criteria (20-22). It is important for radiologists to know the limitation of an AI system for certain categories of cases, such that they need to exercise more caution if an AI system shows considerably worse diagnostic accuracy than that achieved by the radiologists themselves. Furthermore, the assessment of the efficiency of such a system to aid clinical examinations has been overlooked, especially for busy consultation periods when radiologists experience pressure to examine an overabundance of patients.

In this study, we evaluated an AI-based automated diagnostic system, the AI-SONIC Thyroid v. 5.3.0.2 (Demetics Medical Technology Ltd., Hangzhou, China) by inspecting not only typical cases but also nodules of different sizes, thyroid tumors of uncertain malignant potential, and actively growing nodules at a quantitatively optimized malignancy probability cutoff value. We then compared the diagnostic results reported by radiologists of different levels of experience. In total, ultrasound imaging data of 397 thyroid nodules were acquired from the 287 surgical patients recruited for this study, with postoperative pathological diagnosis being taken as the gold standard. We present the following article in accordance with the STARD

 

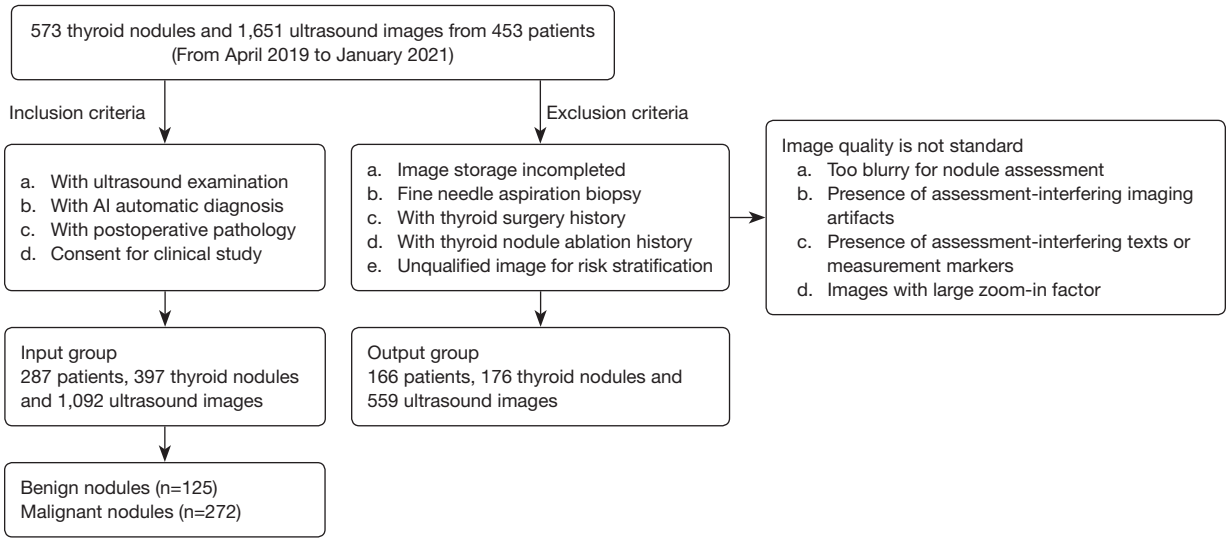


**Figure 1** The flowchart of the inclusion criteria and exclusion criteria for data collection. AI, artificial intelligence.

reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-22-85/rc).

## Methods

### *Ethics and consent*

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by the Ethics Committee of the Changzheng Hospital of Naval Medical University, and the requirement for informed consent was waived due to the retrospective nature of this study.

### *Database*

A total of 287 patients with thyroid nodules (397 nodules in total) who underwent surgeries at the Department of General Surgery, Second Affiliated Hospital (Changzheng Hospital) of Naval Medical University, between April 2019 and January 2021 were selected. Patients with complete clinical information who underwent preoperative ultrasonography and complete examinations were consecutively included in the study. Patients with ultrasound images not of standard quality were excluded. The histopathological diagnoses of all thyroid nodules were determined surgically. Our study fulfilled the statistical requirements for sample size in diagnostic test evaluations conducted in the field of biomedical informatics (23). In summary, our study included 272 malignant nodules and 125 benign nodules. The flow of the inclusion and exclusion criteria in data collection is shown in *Figure 1*.

### *Instruments and methods*

The thyroid (bilateral lobes, isthmus, and conical lobes) of preoperative patients was scanned by 2 radiologists: a senior radiologist who had been engaged in thyroid ultrasonography for more than 10 years and a junior radiologist who had been engaged in thyroid ultrasonography for fewer than 5 years. All patients were placed in the supine position, the neck was fully exposed, the high-frequency linear array probe was placed in the anterior cervical area, and the scans were performed in the order of right first and then left. The vertical section along the long diameter of the thyroid gland from outside to inside or from inside to outside was scanned in a series of vertical slide scans. The transverse section was scanned from top to bottom until the lower pole of the thyroid disappeared. If nodules were found during the examination, it was necessary to focus on repeated scanning, and the acoustic characteristics of the detected thyroid nodules were evaluated according to the following ACR TI-RADS classification criteria: solid, hypoechoic or extremely hypoechoic, and irregular borders; aspect ratio >1; small margin lobulation; and 5 features of microcalcification defined as suspicious malignant features. The TI-RADS 1–3 categories were suggestive of benign nodules, and the TI-RADS 4–5 categories were suggestive

of malignant nodules.

The AI-SONIC Thyroid is an AI-based automated diagnostic system for thyroid nodules. The procedure using the AI-SONIC Thyroid is the same as that of conventional ultrasonography, and during the examination, ultrasound imaging features, such as size, boundary, shape, internal echo, and nodule calcification, are observed. Using its AI algorithm, the AI-SONIC Thyroid system carries out automatic labeling, processing, and analysis, and automatically quantifies and identifies the following 5 types of characteristics of thyroid nodules: boundary, internal echo, morphology, calcification, and morphology. The AI-SONIC Thyroid records the highest risk probability based on its automatic interpretation of the nodule using the TI-RADS classification and the probability values of benign and malignant nodules in the probability value range of 0–1 (0–0.4, benign; 0.4–0.6, requires further examination; 0.6–1, malignant).

### *Acquisition of ultrasound images and radiology analysis*

Routine ultrasound examination was performed using an ultrasound imaging device (Hi Vision Preirus; Hitachi, Tokyo, Japan) with a 7.5 MHz linear-array probe (EUP-L74M). The following system parameters were maintained: gain, 30 dB; gain compensation, intermediate (zero compensation); depth of focus, 2.0–3.75 cm; tissue heat index, <0.4; and mechanical index, 1.2. One senior radiologist (with 13 years of working experience) and one junior radiologist (with 4 years of working experience) performed the clinical ultrasound examinations on patients with preoperative thyroid glands (on both sides of the lobes and isthmus) according to the ACR TI-RADS without being exposed to patients' pathological data. Nodule sizes were manually recorded in terms of lengths in the major and minor axes of the thyroid nodules in the ultrasound images with the largest cross-sectional areas. To evaluate the diagnostic performances on different nodule sizes, original lengths in the major and minor axes were converted into single values by simply computing their Euclidean norms.

The ultrasound images of each nodule were acquired at different cross-sectional planes including the vertical and transverse sections, stored in real time, and transferred to the AI-SONIC Thyroid system. It generated a malignancy score from each individually detected nodule in an ultrasound image ranging from 0 to 1, indicating the malignancy risk probability. The higher the score was, the higher the probability of the detected nodule being malignant. The score for the TI-RADS classification calculated by the radiologists and the malignancy probability value returned by the system were recorded. The system was configured with an Intel Core i5-7500T (4×2.7–3.3 GHz) CPU (Intel Corp., Santa Clara, CA, USA), with 16 GB RAM, 256 GB SSD, and 2 TB hard disk drive (HDD) as well as a NVIDIA GeForce GTX1060/6GB GPU (NVIDIA Corp., Santa Clara, CA, USA). The runtime of the AI system for analyzing each inputted ultrasound image was recorded using the Software Development Kit v. 2.3.1.5 (Demetics Medical Technology Ltd.). The grouping of ultrasound images to their corresponding thyroid nodules was performed on annotation software provided by the same supplier. The runtime per nodule by the AI system was estimated by multiplying the average image processing time with the average number of images per nodule. The retrospective diagnostic time per nodule by the radiologists was estimated based on diagnoses performed on 117 nodules acquired within a half-year interval by recording the total diagnostic time divided by the number of assessed nodules. All images included during this time of diagnostic assessment went through the same preprocessing procedures, including the inclusion and exclusion criteria, and were provided in the same way to the AI system.

### *Cut-off value optimization strategy*

It has been common to maximize the Youden index = sensitivity + specificity – 1 (24) to optimize the cutoff value that defines the sensitivity and specificity of diagnostic systems. By definition, the Youden index is always lower than both the sensitivity and specificity at the corresponding threshold because subtracting 1 from either of them will lead to a negative number, making its absolute value less interpretable. Instead, the average value of sensitivity and specificity allows for quick estimation of the approximate ranges of sensitivity and specificity around the optimal cutoff value, while following the same trend as the Youden index with its maximum value locating at an identical threshold position. We therefore proposed to use the average of sensitivity and specificity for optimizing the cut-off value to differentiate malignant from benign nodules.

From each nodule detected in an image, a malignancy score per image was computed by the AI system. Methodologically, the malignancy of a nodule was evaluated on a per-nodule basis. To compute the nodule-specific malignancy score from the malignancy scores of individual images associated with the same nodule, arbitrary percentiles including their median values and the

 

**Table 1** Characteristics of study participants

| Parameters | Value |
|---|---|
| Mean age (years), mean ± SD | 46.5±12.5 |
| Gender, n (%) | |
| Male | 64 (22.3) |
| Female | 223 (77.7) |
| No. of nodules, n (%) | |
| All | 397 |
| Benign | 125 (31.5) |
| Malignant | 272 (68.5) |
| Nodule sizes (mm), mean ± SD, range [IQR 25, 50, 75] | |
| All | 16.34±14.8 [7.64, 11.82, 20.01] |
| Benign | 25.82±20.6 [12.21, 20.62, 35.23] |
| Malignant | 11.98±8.0 [6.79, 9.90, 14.16] |

SD, standard deviation; IQR, interquartile range.

**Table 2** The number of different nodules and their respective histological diagnoses

| Histological diagnoses | No. of nodules |
|---|---|
| Benign | 125 |
| Thyroid follicular nodular disease | 75 |
| Follicular adenoma | 47 |
| Oncocytic adenoma of the thyroid | 3 |
| Malignant | 272 |
| Papillary thyroid carcinoma | 254 |
| Follicular thyroid carcinoma | 10 |
| Thyroid tumors of uncertain malignant potential | 6 |
| Medullary thyroid carcinoma | 2 |

average values were evaluated. To choose a threshold as close to 0.5 as possible while maintaining a relatively high diagnostic efficacy, we took the median value as the nodule-specific malignancy score. In such a case, the cutoff value for differentiating malignant from benign nodules was set to 0.56.

### *Statistical analysis*

The results were analyzed and plotted using Python v. 3.8 (Python Software Foundation, Wilmington, DE, USA). The areas under the receiver operating characteristic (ROC) curves (25) were compared using the DeLong test via the Python rpy2 package. A P value less than 0.05 was considered statistically significant. Taking into consideration the pathological results, the sensitivity, specificity, and accuracy of the AI-based diagnostic system were calculated. The diagnostic performance of nodule size was compared to the pathology as defined by the 2022 World Health Organization (WHO) Classification of Thyroid Neoplasms (26), using the Kappa (κ) test. The final P value was calculated according to the 2-tail pairwise *t*-test based on 5 independent tests on randomly subdivided datasets.

## Results

### *Patient data*

A total of 287 patients with 397 thyroid nodules were included in this study (*Table 1*). There were 272 cases (68.5%) of malignant tumors, including 254 cases of papillary carcinoma, 10 cases of follicular carcinoma, 6 cases of thyroid tumors of uncertain malignant potential, and 2 cases of medullary thyroid carcinoma. There were 125 benign cases (31.5%), including 75 cases of thyroid follicular nodular disease (of which 22 cases were accompanied by other lesions) and 50 cases of thyroid adenoma (of which 47 cases were follicular adenoma and 3 cases were oncocytic adenoma of the thyroid). Determined pathological types of fewer than 5 cases are not specified here (*Table 2*).

### *AI-based automated diagnostic system malignancy score threshold for predicting the nature of thyroid nodules*

We performed per-image analysis of the AI model using the Youden index and our own recommended averages of the sensitivity and specificity scores as the efficacy score, as shown in *Figure 2*. These 2 measures followed an identical trend while the conventional Youden index failed to show the values of sensitivity and specificity at their balanced optimum for gauging the diagnostic efficacy of the model.

### *Per-nodule evaluation of diagnostic performance*

The optimal threshold for each nodule (referred to as the conditionally optimized threshold) was chosen by optimizing the average sensitivity and specificity (*Figure 3*). It can be seen from *Figure 3* that the conditionally optimized threshold increased with the percentile, while the efficacy score (average of the sensitivity and specificity scores)

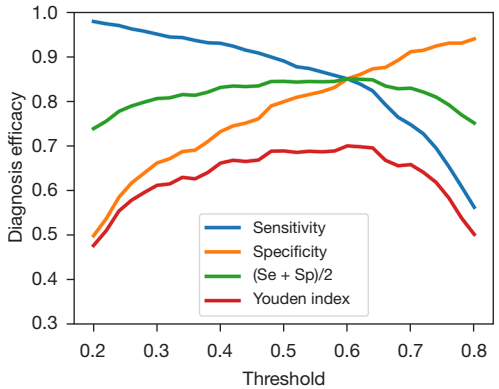


**Figure 2** Diagnostic efficacy of the AI system measured by sensitivity, specificity, average sensitivity and specificity, and the Youden index per image as a function of threshold for the computed malignancy score. Se, sensitivity; Sp, specificity; AI, artificial intelligence.

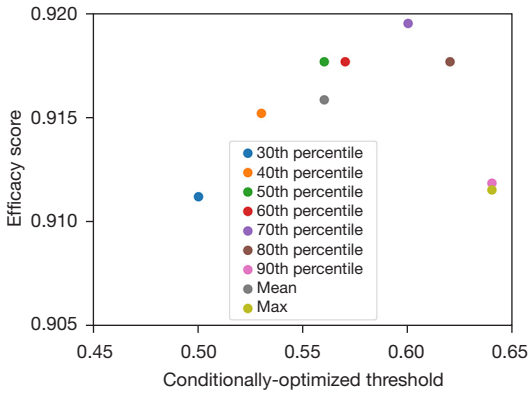


**Figure 3** The calculated efficacy scores with respect to the conditionally optimized for each case of defining the nodule-specific malignancy score

first increased and then decreased, taking a negatively skewed form. The peak performance was reached at the 70$^{th}$ percentile with a threshold of 0.6, and the conventional averaging or maximum method to define the nodule-specific malignancy score underestimated the diagnostic performance by clear margins in comparison with the best case.

Having defined the method to calculate the nodule-specific malignancy score, we compared the diagnostic efficacy for thyroid nodules on the per-image and nodule basis using the ROC curves and the associated area under the curve (AUC) values, as shown in *Figure 4*. There is a substantial gap between the 2 types of analyses, as can be

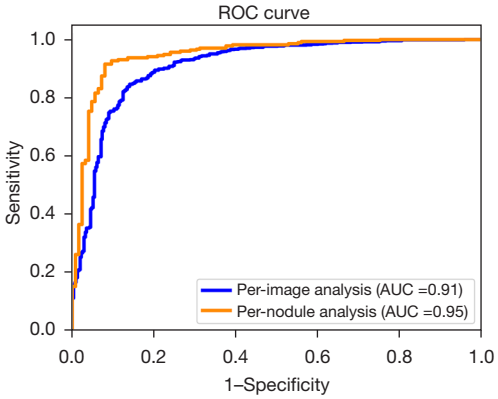


**Figure 4** The ROC curves and associated AUC values of diagnoses of the AI model on a per-image and nodular basis. ROC, receiver operator characteristic; AUC, area under the curve; AI, artificial intelligence.

seen from the figure and the P value of 0.0198 calculated from the Delong test, confirming that the observed difference was statistically significant.

### *Diagnostic performance of the AI system and radiologists in different groups*

Of the 397 cases of thyroid nodules, 125 cases were benign thyroid nodules and 272 cases were malignant thyroid nodules. Of these, 87 benign nodules (69.6%) and 209 (76.8%) malignant ones were correctly diagnosed by junior radiologists [sensitivity: 95% confidence interval (CI): 71.3–81.6%; specificity: 95% CI: 60.6–77.3%]. In comparison, 99 benign cases (79.2%) and 263 malignant cases (96.7%) were correctly diagnosed by senior radiologists (sensitivity: 95% CI: 93.6–98.4%; specificity: 95% CI: 70.8–85.7%). A total of 115 benign cases (92.0%) and 249 malignant cases (91.5%) were correctly diagnosed by the AI system (sensitivity: 95% CI: 87.4–94.5%; specificity: 95% CI: 85.4–95.9%). The complete set of statistics summarizing the performance metrics of sensitivity, specificity, and accuracy are provided in *Table 3*. The sensitivity of the AI system in diagnosing benign and malignant thyroid nodules was lower than that of senior radiologists (91.5% *vs.* 96.7%, respectively) but higher than that of junior radiologists (91.5% *vs.* 76.8%, respectively), with statistical significance. The specificity of the AI system was higher than that of both the junior and senior radiologists (92.0% *vs.* 69.6% and 79.2%, respectively), with statistically significant P values of (P=0.0023 and P=0.0095, respectively). The

**Table 3** Diagnostic performance of the AI system with radiologists (according to ACR TI-RADS) in different groups

| Ultrasonography examiner | Pathology results (n) | | Sensitivity (%) | Specificity (%) | Accuracy (%) | κ value |
|---|---|---|---|---|---|---|
| | Benign | Malignant | | | | |
| Junior radiologist | | | | | | |
| ACR TI-RADS 1–3 | 87 | 63 | 76.8*** (209/272) | 69.6** (87/125) | 74.6**** (296/397) | 0.441 |
| ACR TI-RADS 4–5 | 38 | 209 | | | | |
| 95% CI | | | (71.3, 81.6) | (60.6, 77.3) | | |
| Senior radiologist | | | | | | |
| ACR TI-RADS 1–3 | 99 | 9 | 96.7* (263/272) | 79.2** (99/125) | $91.2^{ns}$ (362/397) | 0.788 |
| ACR TI-RADS 4–5 | 26 | 263 | | | | |
| 95% CI | | | (93.6, 98.4) | (70.8, 85.7) | | |
| Diagnostic AI System | | | | | | |
| Benign (predicted) | 115 | 23 | 91.5 (249/272) | 92.0 (115/125) | 91.7 (364/397) | 0.813 |
| Malignant (predicted) | 10 | 249 | | | | |
| 95% CI | | | (87.4, 94.5) | (85.4, 95.9) | | |

Statistical 2-tail pairwise *t*-tests for sensitivity, specificity, and accuracy were conducted to assess the radiologists and the AI system by randomly subdividing the entire dataset into 5 partitions. *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001; $^{ns}$, not significant. AI, artificial intelligence; ACR, American College of Radiology; TI-RADS, Thyroid Imaging Reporting and Data System.

overall diagnostic accuracy of the AI system was equivalent to that of the senior radiologists (91.7% *vs.* 91.2%, respectively) with a P value of 0.72 as evaluated by randomly subdividing the entire dataset. The results showed that the diagnoses of benign and malignant thyroid nodules made by the junior radiologists generally had moderate consistency with the postoperative pathological diagnoses (κ=0.441), the ultrasound diagnoses of thyroid nodules made by the senior radiologists were generally highly consistent with the postoperative pathological diagnoses (κ=0.788), and the AI system for per-nodule diagnoses had very good consistency with the postoperative pathological diagnoses (κ=0.813; AUC =0.949).

### *Comparison of the diagnostic performance of AI system to senior radiologists for the diagnosis of thyroid nodules of different sizes*

*Table 4* summarizes the sensitivity, specificity, accuracy, κ value, and 95% CI of the AI system and the radiologists in the diagnosis of thyroid nodules of different sizes. We split the dataset into 2 groups using a cutoff size of 25 mm. The senior radiologists and the AI system both had excellent sensitivity (senior radiologists: 98.4%, 95% CI: 95.7–99.5%; AI system: 92.9%, 95% CI: 88.8–95.6%) and overall accuracy (91.5% *vs.* 91.5%) for smaller thyroid nodule diagnosis as well as excellent specificity for larger nodule diagnosis (senior radiologists: 95.9%, 95% CI: 84.9–99.3%; AI system: 100%, 95% CI: 90.9%–100%). The AI system surpassed the senior radiologists in diagnostic specificity for smaller nodules by a large margin (AI system: 86.8%, 95% CI: 76.7–93.2%; senior radiologists: 68.4%, 95% CI: 56.6–78.3%). Both the senior radiologists and the AI system had mediocre sensitivity in diagnosing larger thyroid nodules (senior radiologists: 73.7%, 95% CI: 48.6–89.9%; AI system: 73.7%, 95% CI: 48.6–89.9%). The AI system showed a consistent tendency of higher κ values than compared to the senior radiologists for different nodule sizes.

### *Comparison of the diagnostic efficiency of AI system with radiologists in different groups*

The AI system had an average processing time for diagnosing each inputted ultrasound image of 52.96 ms, recorded by running the software development kit on a standard hardware package (details provided in "Instruments and methods"), and the average runtime per nodule was estimated at 0.146 s. This is in sharp contrast with the mean diagnostic time of the radiologists, as shown in *Table 5*. Junior radiologists required on average 273.5 s for an ultrasound-based thyroid nodule diagnosis per nodule while the senior

 

**Table 4** Comparison of the diagnostic performance of the AI system with that of the radiologists in differentiating the thyroid nodules of different sizes

| Ultrasonography examiner | Pathology results (n) | | Sensitivity (%) | Specificity (%) | Accuracy (%) | κ value |
|---|---|---|---|---|---|---|
| | Benign | Malignant | | | | |
| d≤25 mm (junior radiologist) | | | | | | |
| ACR TI-RADS 1–3 | 42 | 59 | 76.7 (194/253) | 55.3 (42/76) | 71.7 (236/329) | 0.29 |
| ACR TI-RADS 4–5 | 34 | 194 | | | | |
| 95% CI | | | (70.9, 81.6) | (43.5, 66.5) | | |
| d>25 mm (junior radiologist) | | | | | | |
| ACR TI-RADS 1–3 | 45 | 4 | 78.9 (15/19) | 91.8 (45/49) | 88.2 (60/68) | 0.71 |
| ACR TI-RADS 4–5 | 4 | 15 | | | | |
| 95% confidence interval | | | (53.9, 93.0) | (79.5, 97.3) | | |
| d≤25 mm (senior radiologist) | | | | | | |
| ACR TI-RADS 1–3 | 52 | 4 | 98.4 (249/253) | 68.4 (52/76) | 91.5 (301/329) | 0.74 |
| ACR TI-RADS 4–5 | 24 | 249 | | | | |
| 95% CI | | | (95.7, 99.5) | (56.6, 78.3) | | |
| d>25 mm (senior radiologist) | | | | | | |
| ACR TI-RADS 1–3 | 47 | 5 | 73.7 (14/19) | 95.9 (47/49) | 89.7 (61/68) | 0.73 |
| ACR TI-RADS 4–5 | 2 | 14 | | | | |
| 95% CI | | | (48.6, 89.9) | (84.9, 99.3) | | |
| d≤25 mm (AI System) | | | | | | |
| Benign (predicted) | 66 | 18 | 92.9 (235/253) | 86.8 (66/76) | 91.5 (301/329) | 0.77 |
| Malignant (predicted) | 10 | 235 | | | | |
| 95% CI | | | (88.8, 95.6) | (76.7, 93.2) | | |
| d>25 mm (AI System) | | | | | | |
| Benign (predicted) | 49 | 5 | 73.7 (14/19) | 100 (49/49) | 92.6 (63/68) | 0.80 |
| Malignant (predicted) | 0 | 14 | | | | |
| 95% CI | | | (48.6, 89.9) | (90.9, 100) | | |

AI, artificial intelligence; ACR, American College of Radiology; TI-RADS, Thyroid Imaging Reporting and Data System.

**Table 5** Comparison of mean diagnostic times of the AI system with those of radiologists in different groups

| Ultrasonography examiner | Time (s) per nodule |
|---|---|
| Junior radiologist | 273.5 |
| Senior radiologist | 172.4 |
| AI system | 0.146 |

AI, artificial intelligence.

radiologists required 172.4 s on average.

## Discussion

In this study, we showed that when the cutoff value was optimized to predict thyroid nodule malignancy with an AI system, the means of sensitivity and specificity had exactly

the same tendency as the commonly used Youden index but with the advantage of having better interpretable absolute values for estimating the approximate ranges of sensitivity and specificity around the optimal cutoff value. We then evaluated different statistical strategies to estimate the per-nodule-based malignancy score from the per-image-based malignancy scores outputted by the AI system on multiple images associated with each individual thyroid nodule. We showed that taking the average or the maximum values was suboptimal for per-nodule analysis and recommend the use of the median value, although further optimization using other percentiles was possible. In case sensitivity is compromised for an exchange of higher specificity, one can take the maximum method when calculating the nodule-specific scores. Nevertheless, our results support the practice of per-nodule analysis for improving the diagnostic performance of an AI system, despite the fact that additional preprocessing efforts are required (17).

We compared the diagnostic performances of an AI system, AI-SONIC Thyroid v. 5.3.0.2, to that of junior and senior radiologists on ultrasound images of 397 collected thyroid nodules collected from 287 patients, all with definite diagnoses as determined by postoperative pathology. Our results showed that although the sensitivity of the AI automated detection system was slightly lower than that of the senior radiologist (91.5% *vs.* 96.7%), the specificity of the AI automated detection system was significantly higher than that of the senior radiologist (92.0% *vs.* 79.2%, P=0.0095). The accuracy of the AI system was similar to that of the senior radiologist (91.7% *vs.* 91.2%). The junior radiologists, in contrast, had significantly lower performances than did he AI system in terms of sensitivity (76.8%), specificity (69.6%), and accuracy (74.6%), suggesting that with the aid of the AI system, an improvement in diagnostic performance can be presumably expected, although this was not tested in this study.

Meanwhile, among the 33 cases misdiagnosed by the AI system, 14 cases of papillary thyroid carcinoma were predicted to be benign. For these 14 misdiagnosed papillary thyroid carcinoma cases, capsule infiltration by nodules was commonly observed, resulting in blurred boundaries between the nodules and the thyroid glands. In contrast, among the 35 cases misdiagnosed by senior radiologists, 7 cases of nodular goiter and 6 cases of thyroid adenoma were considered suspicious for thyroid cancer. The cases misdiagnosed by both the AI system and the senior radiologists were diverse in terms of their subcategorization on pathological examinations. In the clinical practice of thyroid cancer screening, it may be helpful to further boost the sensitivity of the AI system by setting a slightly lower threshold with only a moderate compromise for the specificity. Nevertheless, concerning the overall diagnostic accuracy in this study where malignant samples were overrepresented (68.5%), the AI system demonstrated a slightly better performance than did the senior radiologists (91.7% *vs.* 91.2%), although this was not statistically significant. This implies that in scenarios where benign and malignant samples are more balanced or benign cases are more frequently occurring, the AI system could have a statistically measurable higher accuracy.

We further analyzed how a diagnosis of malignancy by the AI system was affected by the sizes of the thyroid nodules compared to the diagnosis made by senior radiologists. The results showed that the sensitivities of both the AI system and senior radiologists for diagnosing medium-sized and large-sized nodules (>25 mm) were greatly reduced compared to those for smaller nodules. Large thyroid nodules may protrude from fibrous capsules toward the sternum, causing degradation of the contrast between the nodules and glands. For large nodules, tumor heterogeneity may also be present, necessitating the acquisition of more ultrasound images at different cross-sections to capture adequate evidence of malignancy in the form of relevant ultrasound features to improve the sensitivity during the per-nodule analysis. It may also simply be a sampling-bias effect, as only a small fraction of our dataset was composed of thyroid nodules larger than 25 mm (68/397, 17.1%). All raters showed a reduction in specificity for diagnosing small nodules, especially the radiologists. The ultrasound features of small nodules are typically less distinctive compared with the large nodules, which seemed to influence all raters, including the AI system. Nevertheless, a dramatic reduction in the specificity of radiologists' diagnoses of small nodules might reflect the psychological tendency of not overlooking malignant nodules. AI systems are not affected by these psychological effects. However, the trade-off between sensitivity and specificity is not avoidable. With our proposed cutoff value optimization strategy, satisfactory results on the AI system have been achieved. In terms of κ values, the AI system was consistently better compared to the senior radiologists in diagnosing thyroid nodules of different sizes.

In addition to the evaluations of the diagnostic performance metrics of the AI system, we also compared the throughput of the AI system with that of radiologists performing diagnoses. The mean diagnostic time per image

 

in the AI system was about 53 ms, resulting in a subsecond mean diagnostic time of 0.146 s on a per-nodule basis. This is in sharp contrast to the mean diagnostic time of about 2.8 to 4.5 minutes taken by the radiologists. In real practice, however, extra time is needed to assist radiologists with applying the AI system, as the images need to be transferred from an ultrasound instrument to the AI system, and the association of each image to its corresponding nodule has to be performed manually in its current form.

Our study found that the sensitivity of the AI-based automated diagnostic system was slightly lower than that of senior radiologists (91.5% *vs.* 96.7%), but the specificity of the AI-based automated detection system was higher than that of senior radiologists (92.0% *vs.* 79.2%). Furthermore, the accuracy of AI was similar to that of senior doctors (91.7% *vs.* 91.2%), but we found that this result was different from other research results (18,19), in which the computer-aided designs (CADs) presented a higher sensitivity (90.5% *vs.* 81.1%) and lower specificity (41.2% *vs.* 83.5%) than did the radiologists. We analyzed the possible reasons for this and included nodules subjected to ultrasound-guided fine-needle aspirations (FNAs) or ultrasound examinations prior to scheduled surgery. In fact, many patients with benign nodules chose to continue follow-up observation and those with malignant nodules chose surgical treatment, so that the proportion of benign and malignant nodules in this experimental sample might have been biased. The proportion of malignancies was rather high, which might have influenced the diagnostic performance of the CADs due to the increased sensitivity and accuracy of the malignant nodules. In the future, we will collect more benign cases to achieve a balanced proportion of benign and malignant thyroid nodules, so as to improve diagnostic accuracy. The AI-based automated diagnostic system is expected to provide reliable assistance with the preoperative evaluation of thyroid malignant nodules.

This study was not without limitations. First, a larger sample size for the evaluation of the diagnostic performance would allow a more statistically reliable assessment of the AI system for diagnosing indeterminate thyroid nodules of Bethesda categories III and IV (23), which pose a significant challenge even for FNA cytology. It would be interesting to see whether the generally trained AI-SONIC Thyroid system could be helpful in assessing different follicular types of nodules. Follicular thyroid cancer, which accounts for 5% to 10% of all thyroid cancers (27), is less prone to lymph node metastases than is papillary thyroid cancer, the most common type, but is more likely to relapse and metastasize to the lung and bone. Papillary thyroid carcinoma consists of a papillary structure within the tumor cells. The most confusing nodules are follicular variant papillary thyroid carcinoma (FVPTC). At present, an AI-based diagnosis has the advantages of rapid and noninvasive diagnosis, but the current data volume may not be sufficiently large to support further study on the differentiation of such nodules. However, in the future, we will continue to accumulate data and further identify and study such follicular papillary thyroid carcinoma (PTC) nodules through the combination of ultrasound imaging features identified by AI and FNA results.

Ideally, it would also be better to evaluate the diagnostic performance of the AI system in a prospective setting, in which we could additionally assess how effective it can be in helping to reduce the need for fine-needle aspiration cytology (FNAC) in making malignant diagnoses in real-world clinical scenarios. However, a possible challenge that we might encounter is that for certain cases, FNAC might be taken as the endpoint diagnostic method. The drawback of this is that the FNAC itself suffers from the aforementioned disadvantages in diagnosing Bethesda categories of indeterminate thyroid nodules and, thus, is not 100% accurate, making the assessment of AI diagnostic performance less reliable.

In addition, most AI systems only process and analyze static, gray-scale, ultrasound images, and cannot effectively analyze multimode, ultrasonic, AI-based diagnoses using dynamic video, color Doppler ultrasound, and ultrasonic elastography (28-30). It would be beneficial if in the future, the AI system in could process multimode ultrasound images, as mentioned above, and facilitate 3D visualization of the nodule as it is examined in real time.

To relieve the burden of radiologists, we look forward to the future research and development of effective analyses using ultrasound videos, color Doppler ultrasound, and ultrasonic elastography, as well as diagnostic tools that can accurately display the positions of thyroid nodules, characterize them in real time, and associate individual slices of nodules automatically (31). Ideally, it will also be beneficial for an AI system to offer the selection and formulation of surgical plans for thyroid cancer cases depending on the characterizations of the nodules (32).

## Acknowledgments

*Funding:* This work was funded by the National Natural Science Foundation of China (No. 81501492), the Natural Science Foundation of Shanghai of China (No. 20ZR1457900), and the Three-year Action Plan of Talent Construction of Changzheng Hospital, Military Medical Talent Project of Pyramid Talent Project.

## Footnote

*Reporting Checklist:* The authors have completed the STARD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-22-85/rc

*Conflicts of Interest:* All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-85/coif). XZ is a clinical application doctor at Demetics Medical Technology. The other authors have no 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. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by the Ethics Committee of the Changzheng Hospital of Naval Medical University, and the requirement for informed consent was waived due to the retrospective nature of this study.



## References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Kitahara CM, Sosa JA. The changing incidence of thyroid cancer. Nat Rev Endocrinol 2016;12:646-53.
3. Lim H, Devesa SS, Sosa JA, Check D, Kitahara CM. Trends in Thyroid Cancer Incidence and Mortality in the United States, 1974-2013. JAMA 2017;317:1338-48.
4. Morris LG, Sikora AG, Tosteson TD, Davies L. The increasing incidence of thyroid cancer: the influence of access to care. Thyroid 2013;23:885-91.
5. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. CA Cancer J Clin 2017;67:7-30.
6. Lee HJ, Yoon DY, Seo YL, Kim JH, Baek S, Lim KJ, Cho YK, Yun EJ. Intraobserver and Interobserver Variability in Ultrasound Measurements of Thyroid Nodules. J Ultrasound Med 2018;37:173-8.
7. La Vecchia C, Malvezzi M, Bosetti C, Garavello W, Bertuccio P, Levi F, Negri E. Thyroid cancer mortality and incidence: a global overview. Int J Cancer 2015;136:2187-95.
8. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC, Hammers LW, Hamper UM, Langer JE, Reading CC, Scoutt LM, Stavros AT. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol 2017;14:587-95.
9. Wang Y, Lei KR, He YP, Li XL, Ren WW, Zhao CK, Bo XW, Wang D, Sun CY, Xu HX. Malignancy risk stratification of thyroid nodules: comparisons of four ultrasound Thyroid Imaging Reporting and Data Systems in surgically resected nodules. Sci Rep 2017;7:11560.
10. Zhang Y, Wu Q, Chen Y, Wang Y. A Clinical Assessment of an Ultrasound Computer-Aided Diagnosis System in Differentiating Thyroid Nodules With Radiologists of Different Diagnostic Experience. Front Oncol 2020;10:557169.
11. Mai W, Zhou M, Li J, Yi W, Li S, Hu Y, Ji J, Zeng W, Gao B, Liu H. The value of the Demetics ultrasound-assisted diagnosis system in the differential diagnosis of benign from malignant thyroid nodules and analysis of the influencing factors. Eur Radiol 2021;31:7936-44.
12. Yoo YJ, Ha EJ, Cho YJ, Kim HL, Han M, Kang SY. Computer-Aided Diagnosis of Thyroid Nodules via Ultrasonography: Initial Clinical Experience. Korean J Radiol 2018;19:665-72.
13. Choi YJ, Baek JH, Park HS, Shim WH, Kim TY, Shong YK, Lee JH. 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.
14. Wang L, Yang S, Yang S, Zhao C, Tian G, Gao Y, Chen Y, Lu Y. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. World J Surg Oncol 2019;17:12.
15. Wu H, Deng Z, Zhang B, Liu Q, Chen J. Classifier Model Based on Machine Learning Algorithms: Application to Differential Diagnosis of Suspicious Thyroid Nodules via Sonography. AJR Am J Roentgenol 2016;207:859-64.
16. Peng S, Liu Y, Lv W, Liu L, Zhou Q, Yang H, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. Lancet Digit Health 2021;3:e250-9.
17. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2019;20:193-201.
18. Jeong EY, Kim HL, Ha EJ, Park SY, Cho YJ, Han M. 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.
19. Xia S, Yao J, Zhou W, Dong Y, Xu S, Zhou J, Zhan W. A computer-aided diagnosing system in the evaluation of thyroid nodules-experience in a specialized thyroid center. World J Surg Oncol 2019;17:210.
20. Du YR, Ji CL, Wu Y, Gu XG. Combination of ultrasound elastography with TI-RADS in the diagnosis of small thyroid nodules (≤10 mm): A new method to increase the diagnostic performance. Eur J Radiol 2018;109:33-40.
21. Li X, Gao F, Li F, Han XX, Shao SH, Yao MH, Li CX, Zheng J, Wu R, Du LF. Qualitative analysis of contrast-enhanced ultrasound in the diagnosis of small, TR3-5 benign and malignant thyroid nodules measuring ≤1 cm. Br J Radiol 2020;93:20190923.
22. Xi X, Wang Y, Gao L, Jiang Y, Liang Z, Ren X, Gao Q, Lai X, Yang X, Zhu S, Zhao R, Zhang X, Zhang B. Establishment of an Ultrasound Malignancy Risk Stratification Model for Thyroid Nodules Larger Than 4 cm. Front Oncol 2021;11:592927.
23. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform 2014;48:193-204.
24. Zhu J, Zhang S, Yu R, Liu Z, Gao H, Yue B, Liu X, Zheng X, Gao M, Wei X. An efficient deep convolutional neural network model for visual localization and automatic diagnosis of thyroid nodules on ultrasound images. Quant Imaging Med Surg 2021;11:1368-80.
25. Niu S, Huang J, Li J, Liu X, Wang D, Wang Y, Shen H, Qi M, Xiao Y, Guan M, Li D, Liu F, Wang X, Xiong Y, Gao S, Wang X, Yu P, Zhu J. Differential diagnosis between small breast phyllodes tumors and fibroadenomas using artificial intelligence and ultrasound data. Quant Imaging Med Surg 2021;11:2052-61.
26. Baloch ZW, Asa SL, Barletta JA, Ghossein RA, Juhlin CC, Jung CK, LiVolsi VA, Papotti MG, Sobrinho-Simões M, Tallini G, Mete O. Overview of the 2022 WHO Classification of Thyroid Neoplasms. Endocr Pathol 2022;33:27-63.
27. Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. Thyroid 2017;27:1341-6.
28. Sobrinho-Simões M, Eloy C, Magalhães J, Lobo C, Amaro T. Follicular thyroid carcinoma. Mod Pathol 2011;24 Suppl 2:S10-8.
29. Liu T, Ge X, Yu J, Guo Y, Wang Y, Wang W, Cui L. Comparison of the application of B-mode and strain elastography ultrasound in the estimation of lymph node metastasis of papillary thyroid carcinoma based on a radiomics approach. Int J Comput Assist Radiol Surg 2018;13:1617-27.
30. Yeon EK, Sohn YM, Seo M, Kim EJ, Eun YG, Park WS, Yun SJ. Diagnostic Performance of a Combination of Shear Wave Elastography and B-Mode Ultrasonography in Differentiating Benign From Malignant Thyroid Nodules. Clin Exp Otorhinolaryngol 2020;13:186-93.
31. Görges R, Eising EG, Fotescu D, Renzing-Köhler K, Frilling A, Schmid KW, Bockisch A, Dirsch O. Diagnostic value of high-resolution B-mode and power-mode sonography in the follow-up of thyroid cancer. Eur J Ultrasound 2003;16:191-206.
32. Maddaloni E, Briganti SI, Crescenzi A, Beretta Anguissola G, Perrella E, Taffon C, Palermo A, Manfrini S, Pozzilli P, Lauria Pantano A. Usefulness of Color Doppler Ultrasonography in the Risk Stratification of Thyroid Nodules. Eur Thyroid J 2021;10:339-44.

**Cite this article as:** Guo F, Chang W, Zhao J, Xu L, Zheng X, Guo J. Assessment of the statistical optimization strategies and clinical evaluation of an artificial intelligence-based automated diagnostic system for thyroid nodule screening. Quant Imaging Med Surg 2023;13(2):695-706. doi: 10.21037/qims-22-85