Radiology

ORIGINAL RESEARCH • HEAD AND NECK IMAGING

Generalizability and Diagnostic Performance of AI Models for Thyroid US

WenWen Xu, MD • XiaoHong Jia, MD* • ZiHan Mei, MD • XiaoLin Gu, PhD • Yang Lu, MS • Chi-Cheng Fu, PhD • RuiFang Zhang, MD • Ying Gu, MD • Xia Chen, MD • XiaoMao Luo, MD • Ning Li, MD • BaoYan Bai, MD • QiaoYing Li, MD • JiPing Yan, MD • Hong Zhai, MD • Ling Guan, MD • Bing Gong, MD • KeYang Zhao, MS • Qu Fang, MS • Chuan He, BS • WeiWei Zhan, MD • Ting Luo, MD • HuiTing Zhang, MD • YiJie Dong, MD • JianQiao Zhou, MD • for The Chinese Artificial Intelligence Alliance for Thyroid and Breast Ultrasound*

From the Department of Ultrasound, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, 197 Ruijin Er Road, 200025, Shanghai, China (W.W.X., X.H.J., Z.H.M., W.W.Z., T.L., H.T.Z., Y.J.D., J.Q.Z.); Department of Scientific Research, Shanghai Aitrox Technology Corporation Limited, Shanghai, China (X.L.G., Y.L., C.C.F., K.Y.Z., Q.F., C.H.); Department of Ultrasound, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China (R.F.Z.); Department of Medical Ultrasound, Affiliated Hospital of Guizhou Medical University, Guiyang, China (Y.G., X.C.); Department of Medical Ultrasound, Yunnan Cancer Hospital & The Third Affiliated Hospital of Kunming Medical University, Kunming, China (X.M.L.); Department of Ultrasound, Yunnan Kungang Hospital, The Seventh Affiliated Hospital of Dali University, Anning, China (N.L.); Department of Ultrasound, Affiliated Hospital of Yan'an University, Yan'an, China (B.Y.B.); Department of Ultrasound, Tangdu Hospital, Fourth Military Medical University, Xi'an, China (Q.Y.L.); Department of Ultrasound, Shanxi Provincial People's Hospital, Taiyuan, China (J.P.Y.); Department of Ultrasound, Traditional Chinese Medical Hospital of Xinjiang Uygur Autonomous Region, Urumqi, Xinjiang Uygur Autonomous Region, China (H.Z.); Department of Ultrasound, Gansu Provincial Cancer Hospital, Lanzhou, China (L.G.); Department of Ultrasound, Jilin Central General Hospital, Jilin, China (B.G.); and College of Health Science and Technology, Shanghai Jiaotong University School of Medicine, Shanghai, China (J.Q.Z.). Received May 9, 2022; revision requested July 19; revision received April 17, 2023; accepted April 25. **Address correspondence to** J.Q.Z. (email: *zhousu30@126.com*).

* W.W.X. and X.H.J. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

Radiology 2023; 307(5):e221157 • https://doi.org/10.1148/radiol.221157 • Content codes: HN US AI

牛魔的208家医院：10000多病人视频直接随机分，和多中心有个屁关系

Background: Artificial intelligence (AI) models have improved US assessment of thyroid nodules; however, the lack of generalizability limits the application of these models.

Purpose: To develop AI models for segmentation and classification of thyroid nodules in US using diverse data sets from nationwide hospitals and multiple vendors, and to measure the impact of the AI models on diagnostic performance.

Materials and Methods: This retrospective study included consecutive patients with pathologically confirmed thyroid nodules who underwent US using equipment from 12 vendors at 208 hospitals across China from November 2017 to January 2019. The detection, segmentation, and classification models were developed based on the subset or complete set of images. Model performance was evaluated by precision and recall, Dice coefficient, and area under the receiver operating characteristic curve (AUC) analyses. Three scenarios (diagnosis without AI assistance, with freestyle AI assistance, and with rule-based AI assistance) were compared with three senior and three junior radiologists to optimize incorporation of AI into clinical practice.

Results: A total of 10 023 patients (median age, 46 years [IQR 37–55 years]; 7669 female) were included. The detection, segmentation, and classification models had an average precision, Dice coefficient, and AUC of 0.98 (95% CI: 0.96, 0.99), 0.86 (95% CI: 0.86, 0.87), and 0.90 (95% CI: 0.88, 0.92), respectively. The segmentation model trained on the nationwide data and classification model trained on the mixed vendor data exhibited the best performance, with a Dice coefficient of 0.91 (95% CI: 0.90, 0.91) and AUC of 0.98 (95% CI: 0.97, 1.00), respectively. The AI model outperformed all senior and junior radiologists ($P < .05$ for all comparisons), and the diagnostic accuracies of all radiologists were improved ($P < .05$ for all comparisons) with rule-based AI assistance.

Conclusion: Thyroid US AI models developed from diverse data sets had high diagnostic performance among the Chinese population. Rule-based AI assistance improved the performance of radiologists in thyroid cancer diagnosis.

Supplemental material is available for this article.

US is essential for thyroid nodule evaluation. Thyroid nodules are very common, with prevalence ranging 19%–67% (1). Most nodules are benign, with 4.5%–6% being malignant (2). To increase the accuracy and specificity of imaging characteristics associated with malignancy, radiologists interpret US imaging findings in terms of the size, shape, margin type, and content features of nodules according to the Thyroid Imaging Reporting and Data System (TI-RADS) (3,4). For further management, patients with suspicious nodules that meet the size criteria are advised to undergo fine-needle aspiration (FNA) biopsy. However, because of the heterogeneity of thyroid nodules, critical diagnostic decisions about whether to perform FNA biopsy on the basis of US findings involve labor-intensive tasks that require radiologists to have substantial experience and expertise.

With recent developments in artificial intelligence (AI) applications in medical fields, deep learning has continued to achieve meaningful improvements in the management of thyroid nodules at US (5). However, the poor performance of well-trained deep learning models from a single institution or single vendor when applied to external validation data sets raises a question of the broad clinical application of thyroid US AI. US imaging

Abbreviations

AI = artificial intelligence, AUC = area under the ROC curve, ROC = receiver operating characteristic, TI-RADS = Thyroid Imaging Reporting and Data System

Summary

Artificial intelligence models for thyroid US developed on diverse data sets from mixed vendors and broad regions resulted in high diagnostic performance and can potentially enhance generalizability.

Key Results

- In this retrospective study of 10 023 patients with pathologically confirmed thyroid nodules, US artificial intelligence (AI) models based on diverse data were developed.
- The highest Dice coefficient (0.91) and area under the receiver operating characteristic curve (0.98) were found for the segmentation model trained on nationwide data and the classification model trained on mixed vendor data, respectively.
- The diagnostic performances of all radiologists were improved (*P* < .05 for all comparisons) with rule-based AI assistance.

has unique characteristics when compared with other modalities, including operator dependence. Other factors of imaging variabilities include medical equipment, scan modes, settings, imaging protocols, and interpretation techniques. Thus, a deep learning model must be developed based on large numbers of diverse data sets from various geographic regions, multiple vendors, and various clinical settings to reflect broad generalizability and applicability. It is known that a model with an appropriate training set (ie, a data set representing the complexity of the data) can be applied to diverse sample sets because the model can learn unbiased characteristic features from such a training set (6). However, investigators usually face difficulty in accessing diverse data sets to develop models due to the uniform characteristics of images caused by limited data sources from any single institution (7).

The purpose of the current study was to develop generalizable US AI models based on large real-world US data sets from nationwide hospitals across regions and multiple vendors in the segmentation and classification tasks of thyroid nodules, and to measure the diagnostic improvement of radiologists by integrating the AI models.

Materials and Methods

X.L.G., Y.L., C.C.F., K.Y.Z., Q.F., and C.H. are employees of Shanghai Aitrox, which provided technical support in the construction of the AI models. The authors who are not employees or consultants of Shanghai Aitrox had full control of any data and information that might present a conflict of interest.

Patient Data Collection

This retrospective study was approved by the institutional review board of Ruijin Hospital, and written informed consent to undergo US was acquired from patients before examinations. Data sets were obtained from 208 hospitals that are members of the Chinese Artificial Intelligence Alliance for Thyroid and Breast Ultrasound, or CAAU, across 31 provinces and municipalities. Consecutive patients who underwent diagnostic thyroid US examination and had corresponding surgical pathologic assessment results for the determination of benignity or malignancy from November 2017 to January 2019 were included. Refer to Appendix S1 for detailed inclusion and exclusion criteria.

In the current study, thyroid US examinations were performed by radiologists rather than sonographers, and the US systems were equipped with linear probes ranging 5–18 MHz across different vendors or grades (Appendix S1). Only one of the most suspicious nodules of a patient was selected for the study. Typically, at least one longitudinal US image and one cross-sectional US image of the target nodule were stored and used for analysis. All US imaging data and related clinical information were gathered and sent to Ruijin Hospital, the primary study center, for subsequent analysis. Adhering to the institutional protocols, patient data were deidentified and encrypted before they were assessed. The data included *(a)* static images, on which the AI models were built, validated, and tested; and *(b)* cine clips, on which the potential and feasibility of the developed models in processing real-time cine loops was tested.

A portion of the US images (10 512 images in 5478 patients from 175 hospitals) were manually annotated for the nodule boundary by two radiologists (X.H.J. and W.W.X., with 11 and 5 years of experience) in a retrospective manner in preparation for the subsequent segmentation task. For the classification task, 24 944 images from all 10 023 patients were used. Part of the data (1068 patients) has been used in a study with different design and purpose (8). The sample allocation flowchart is shown in Figure 1.

Detection, Segmentation, and Classification Models for US

All US images were resized and normalized before training and testing the models. YOLOv5 *(https://pytorch.org/hub/ultralytics_yolov5/)* (9), DenseNet121_FPN *(https://keras.io/api/applications/densenet/)* (10), and DenseNet 121 *(https://pytorch.org/hub/pytorch_vision_densenet/)* were used as backbones for the detection, segmentation, and classification tasks, respectively. See Appendix S1 for algorithm details and GitHub for the source code of the developed models *(https://github.com/personnoicon/Generalizability-and-diagnostic-efficacy-of-AI-models-for-thyroid-ultrasound.git)*.

Generalization on Subgroups according to Hospitals, Vendors, and Regions

The US images were divided into subgroups according to hospitals, vendors, and regions based on their source. To test the generalizability of the segmentation and classification models, models with data from one subset or a combination of multiple subsets were trained and then tested internally and externally on separate samples. See also Appendix S1.

Evaluation of the Performance and Generalizability of Models

Performance of the detection model was evaluated with precision and recall analysis, sensitivity, and F1 score. Feasibility of applying the detection model trained with static images to US cine clips of 30 frames per second was assessed by *(a)* disassembling the cine clips into an image series, *(b)* feeding the generated

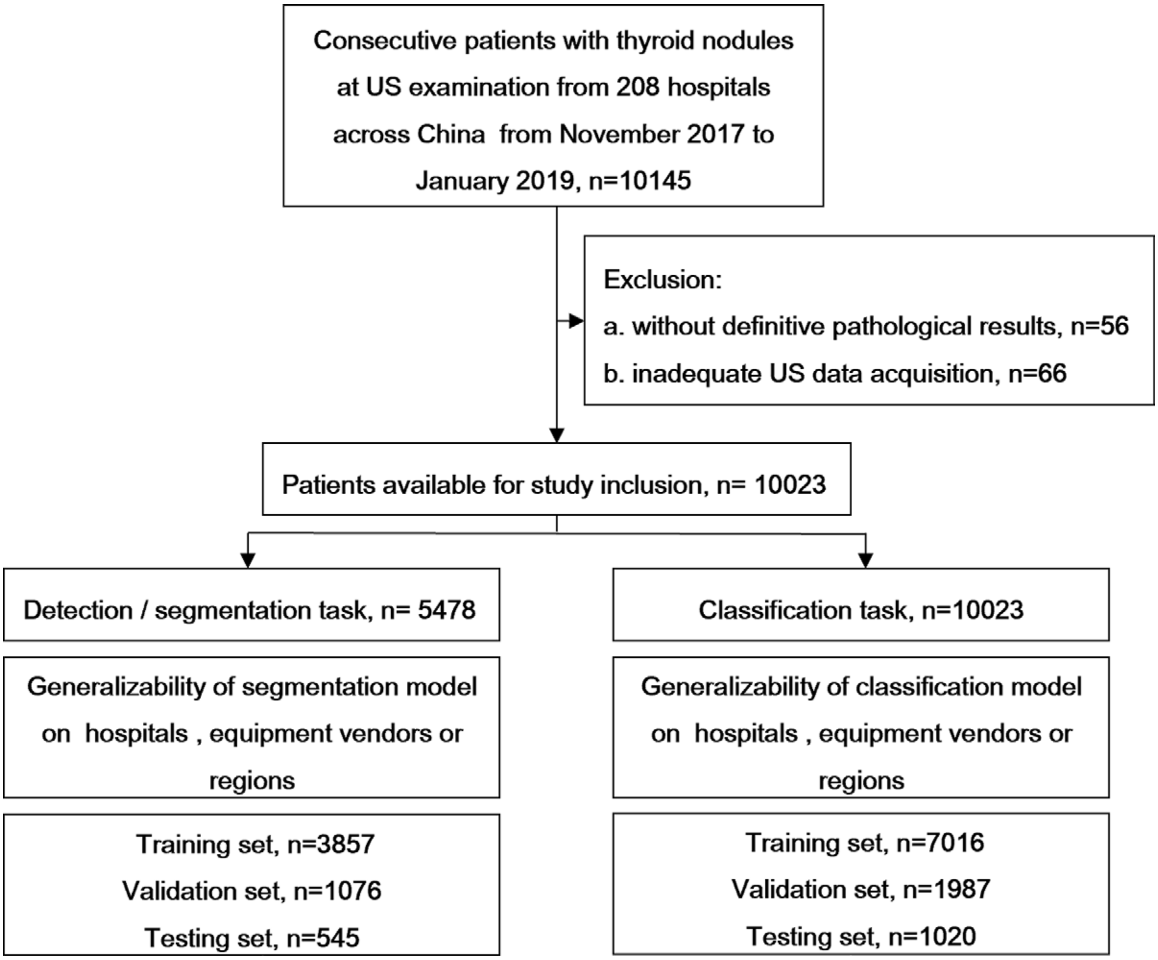
Consecutive patients with thyroid nodules at US examination from 208 hospitals across China from November 2017 to January 2019, n=10145
Exclusion:
a. without definitive pathological results, n=56
b. inadequate US data acquisition, n=66
Patients available for study inclusion, n= 10023
Detection / segmentation task, n= 5478
Classification task, n=10023
Generalizability of segmentation model on hospitals , equipment vendors or regions
Generalizability of classification model on hospitals , equipment vendors or regions
Training set, n=3857
Validation set, n=1076
Testing set, n=545
Training set, n=7016
Validation set, n=1987
Testing set, n=1020

Figure 1: Flowchart shows the development and evaluation of the deep learning models for computer-aided diagnosis of thyroid nodules.

Table 1: Baseline Patient Characteristics

		Detection and Segmentation Tasks			Classification Task		
Characteristic	Total Group (*n* = 10023)	Training Set (*n* = 3857)	Validation Set (*n* = 1076)	Test Set (*n* = 545)	Training Set (*n* = 7016)	Validation Set (*n* = 1987)	Test Set (*n* = 1020)
Overall							
Age (y)*	46 (37–55)	47 (37–55)	48 (38–55)	47 (36–55)	47 (37–55)	48 (39–56)	45 (34–55)
Patients with malignant tumors	6188 (61.74)	2228 (57.77)	603 (56.04)	321 (58.90)	4332 (61.74)	1217 (61.25)	639 (62.65)
No. of images	15809	4344	1180	637	11652	2920	1237
Patients with benign tumors	3835 (38.26)	1629 (42.23)	473 (44.08)	224 (41.10)	2684 (38.26)	770 (38.75)	381 (37.35)
No. of images	9135	3046	885	420	6825	1643	667
Male sex	2354 (23.49)	893 (23.15)	283 (26.30)	119 (21.83)	1609 (22.93)	467 (23.50)	278 (27.25)
Age (y)*	46 (37–55)	…	…	…	…	…	…
No. of images	5922	1703	546	230	4302	1099	521
Female sex	7669 (76.51)	2964 (76.85)	793 (73.70)	426 (78.17)	5407 (77.07)	1520 (76.50)	742 (72.75)
Age (y)*	47 (37–55)	…	…	…	…	…	…
No. of images	19022	5687	1519	827	14,175	3464	1383

Note.—Except where indicated, data are numbers of patients, with percentages in parentheses. Data with detection annotations and data with segmentation labels (*n* = 5478 for both) were randomly divided into training, validation, and test sets with the ratio of 7:2:1. No statistically significant difference was found in age between male and female sex (*P* = .439).

* Data are medians, with IQRs in parentheses.

image series to the model for inference, and *(c)* testing if the model can detect thyroid nodules on each frame in real time.

To evaluate performance of the segmentation model, Dice coefficients were calculated. For the classification model, performance was assessed by the receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), and sensitivity. Heat maps generated by the gradient-weighted class activation mapping (Grad-CAM; *https://github.com/jacobgil/pytorch-grad-cam*) approach (11) were used to reveal the key regions where the classification model paid attention for predictions.

Image Analysis

A total of 1904 images in 1020 patients with Chinese TI-RADS grading reports were used as the test set (4). Diagnostic performance of the classification model was assessed for patients older than or equal to 45 years and younger than 45 years of age, respectively (12,13). Furthermore, the diagnostic performance of the classification model was compared with the results of Chinese TI-RADS grading by radiologists. Probability of the malignancy of thyroid nodules on each image was determined by the classification model. Performance of the model was assessed by ROC curve analysis, with the pathologic assessment results considered the ground truth.

Table 2: Vendor Distribution

Vendor	Patients (*n* = 10 023)	Images (*n* = 24 944)
Mindray	2592 (25.86)	8696 (34.86)
Philips	2517 (25.11)	6028 (24.17)
Esaote	1945 (19.41)	3674 (14.73)
GE Healthcare	1344 (13.41)	3044 (12.20)
Siemens Healthineers	541 (5.40)	1157 (4.64)
Toshiba	492 (4.91)	1016 (4.07)
Samsung	145 (1.45)	348 (1.40)
Supersonic Imagine	132 (1.32)	319 (1.28)
Hitachi	114 (1.14)	249 (1.00)
Other	201 (2.01)	413 (1.66)

Note.—Data are numbers of patients or images, with percentages in parentheses. “Other” vendors include Zonare, SonoScape, and VINNO.

An experiment was designed to compare the diagnostic performance of six US radiologists, including three senior radiologists (J.Q.Z., B.Y.B., and Y.J.D., with 25, 20, and 14 years of experience in thyroid US, respectively) and three junior radiologists (T.L., Z.H.M., and H.T.Z., with 4, 4, and 3 years of experience, respectively). All six radiologists independently assessed the static US images of 1020 target nodules from 1020 patients in the test set and were blinded to clinical history, original clinical records, and pathologic assessment findings. They were asked to identify malignant nodules in the following three scenarios. *(a)* For diagnosis without AI assistance, radiologists were required to perform the first evaluation to rate the US images in the 1020 patients comprising the test set according to the Chinese TI-RADS (images were scored as 3, 4a, 4b, 4c, or 5 with an increasing malignant potency). *(b)* In the freestyle AI–assisted mode, radiologists were required to perform the second evaluation of the same US images from the patients in the test set after a 2-week washout period, and to flexibly determine the Chinese TI-RADS category of the nodules with reference to the malignancy prediction (malignant vs benign) of the AI classification model. *(c)* In the rule-based AI–assisted mode, if the AI classification model predicted a malignant thyroid nodule, the original Chinese TI-RADS category in the first evaluation was upgraded by one level (eg, 4a upgraded to 4b); otherwise, the original Chinese TI-RADS category was downgraded by one level (eg, 4a downgraded to 3). Radiologists were blinded to the

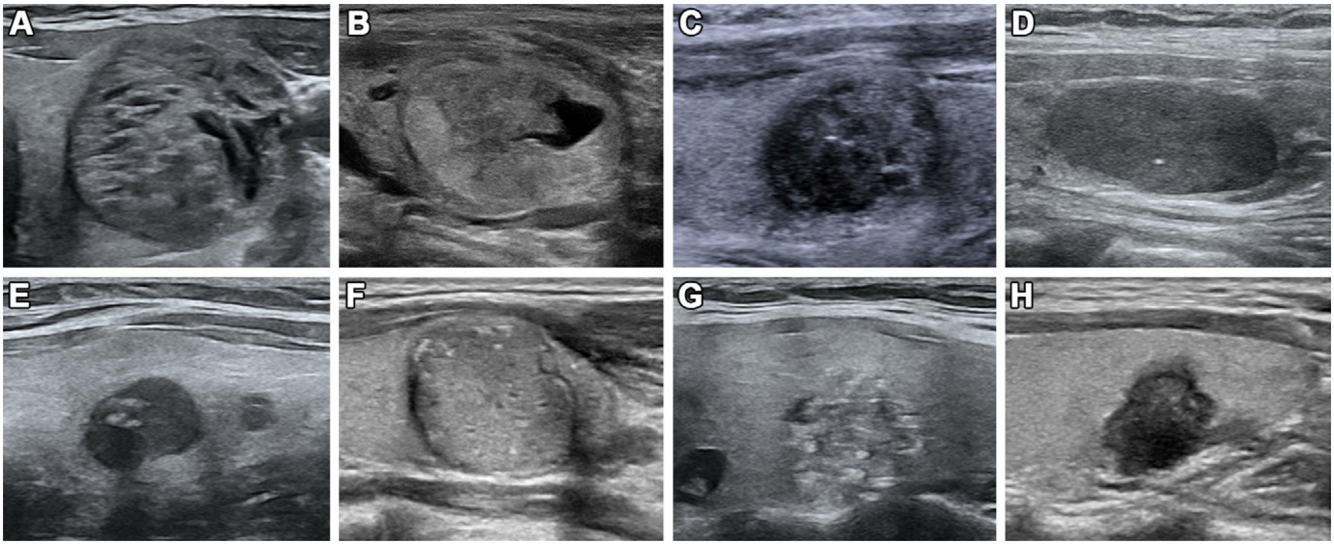

Figure 2: Typical US images of malignant and benign thyroid nodules show **(A)** nodular goiter in a 32-year-old female patient, **(B)** nodular goiter in a 24-year-old male patient, **(C)** nodular goiter in a 43-year-old female patient, **(D)** follicular adenoma in a 56-year-old female patient, **(E)** medullary carcinoma in a 44-year-old female patient, **(F)** papillary thyroid carcinoma in a 62-year-old female patient, **(G)** papillary thyroid carcinoma in a 26-year-old male patient, and **(H)** papillary thyroid carcinoma in a 30-year-old female patient.

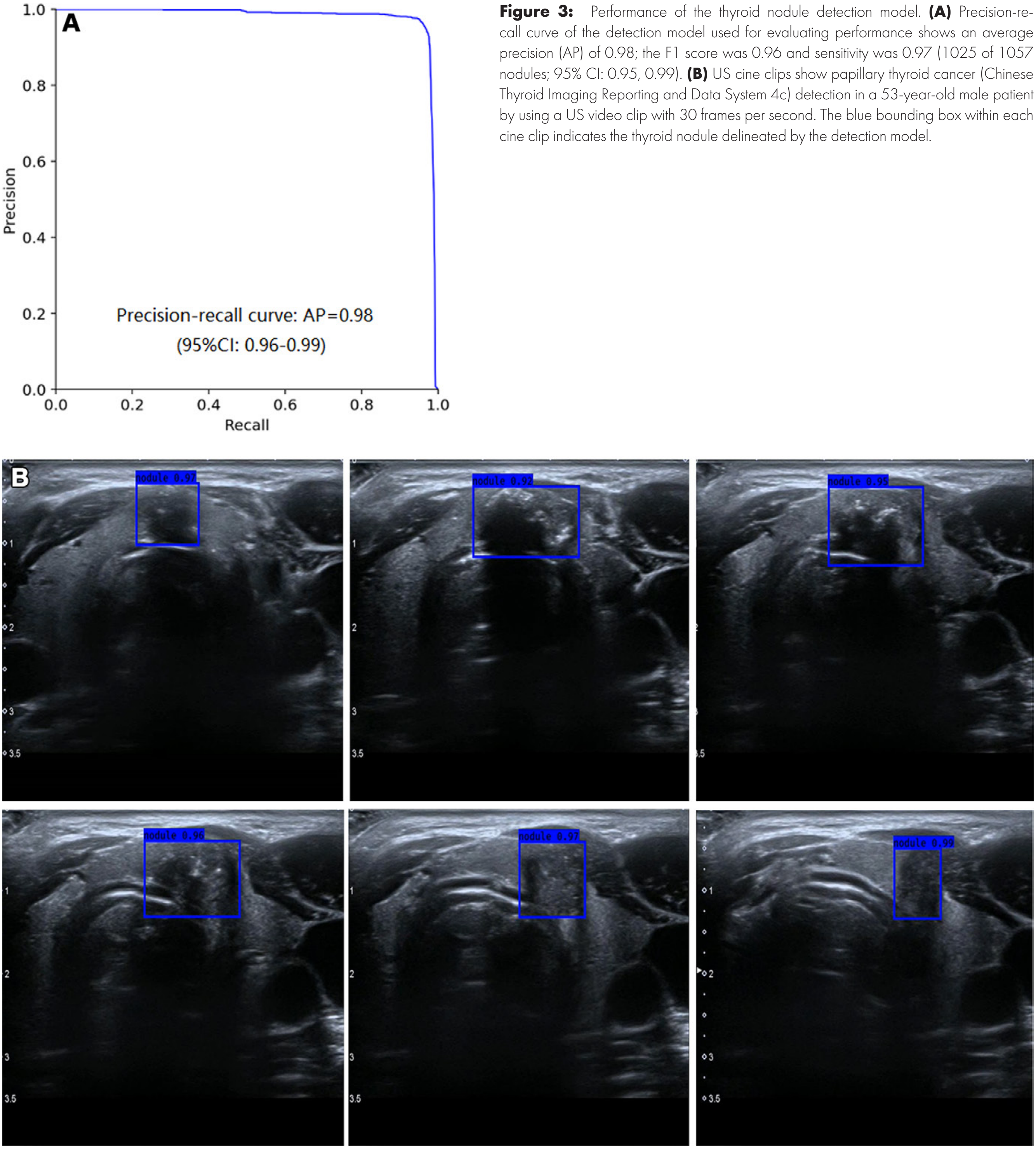

Figure 3: Performance of the thyroid nodule detection model. **(A)** Precision-recall curve of the detection model used for evaluating performance shows an average precision (AP) of 0.98; the F1 score was 0.96 and sensitivity was 0.97 (1025 of 1057 nodules; 95% CI: 0.95, 0.99). **(B)** US cine clips show papillary thyroid cancer (Chinese Thyroid Imaging Reporting and Data System 4c) detection in a 53-year-old male patient by using a US video clip with 30 frames per second. The blue bounding box within each cine clip indicates the thyroid nodule delineated by the detection model.

results predicted by the AI classification model in the rule-based AI–assisted mode. ROC curves were plotted for all scenarios, and AUC values were calculated for performance evaluation.

Statistical Analysis

The pROC package *(https://cran.r-project.org/web/packages/pROC/pROC.pdf)* (14) on the R (version 4.0.0; The R Foundation) language platform was used for drawing ROC curves. AUCs of the classification model were compared with those of radiologists by using the DeLong test. The optimal threshold was determined by maximizing the Youden index. To measure interobserver variability, the Kendall coefficient of concordance (W) was used. The observers' agreement was interpreted as follows: 0–0.20, poor; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.80–1, excellent (15). The sample size calculation for the comparison of diagnostic performance between AI and

Table 3: Segmentation of Dice Similarity Coefficients Based on Vendor

Training Data Set	Testing Data Set						
	GE	Philips	Esaote	Mindray	Siemens	Other	Mixed
GE	0.75 (0.73, 0.78)	0.78 (0.74, 0.82)	0.74 (0.70, 0.78)	0.80 (0.77, 0.83)	0.76 (0.70, 0.82)	0.83 (0.79, 0.87)	0.77 (0.76, 0.79)
Philips	0.72 (0.70, 0.75)	0.79 (0.75, 0.83)	0.74 (0.70, 0.78)	0.77 (0.75, 0.80)	0.73 (0.67, 0.78)	0.79 (0.75, 0.83)	0.76 (0.74, 0.77)
Esaote	0.72 (0.70, 0.74)	0.79 (0.75, 0.83)	0.84 (0.81, 0.88)	0.77 (0.74, 0.80)	0.81 (0.75, 0.86)	0.81 (0.77, 0.85)	0.80 (0.78, 0.81)
Mindray	0.63 (0.60, 0.65)	0.78 (0.74, 0.82)	0.66 (0.62, 0.69)	0.79 (0.76, 0.82)	0.76 (0.70, 0.81)	0.81 (0.77, 0.85)	0.72 (0.71, 0.74)
Siemens	0.71 (0.69, 0.74)	0.78 (0.74, 0.82)	0.74 (0.70, 0.78)	0.77 (0.75, 0.80)	0.85 (0.79, 0.90)	0.79 (0.75, 0.83)	0.76 (0.75, 0.78)
Other	0.75 (0.73, 0.78)	0.82 (0.78, 0.86)	0.78 (0.75, 0.82)	0.83 (0.80, 0.86)	0.83 (0.78, 0.89)	0.84 (0.80, 0.88)	0.81 (0.79, 0.82)
Mixed	0.83 (0.80, 0.85)	0.88 (0.84, 0.92)	0.87 (0.83, 0.91)	0.85 (0.82, 0.88)	0.87 (0.81, 0.92)	0.90 (0.86, 0.94)	0.87 (0.85, 0.88)

Note.—Data in parentheses are 95% CIs. "Other" vendors include Zonare, SonoScape, and VINNO.

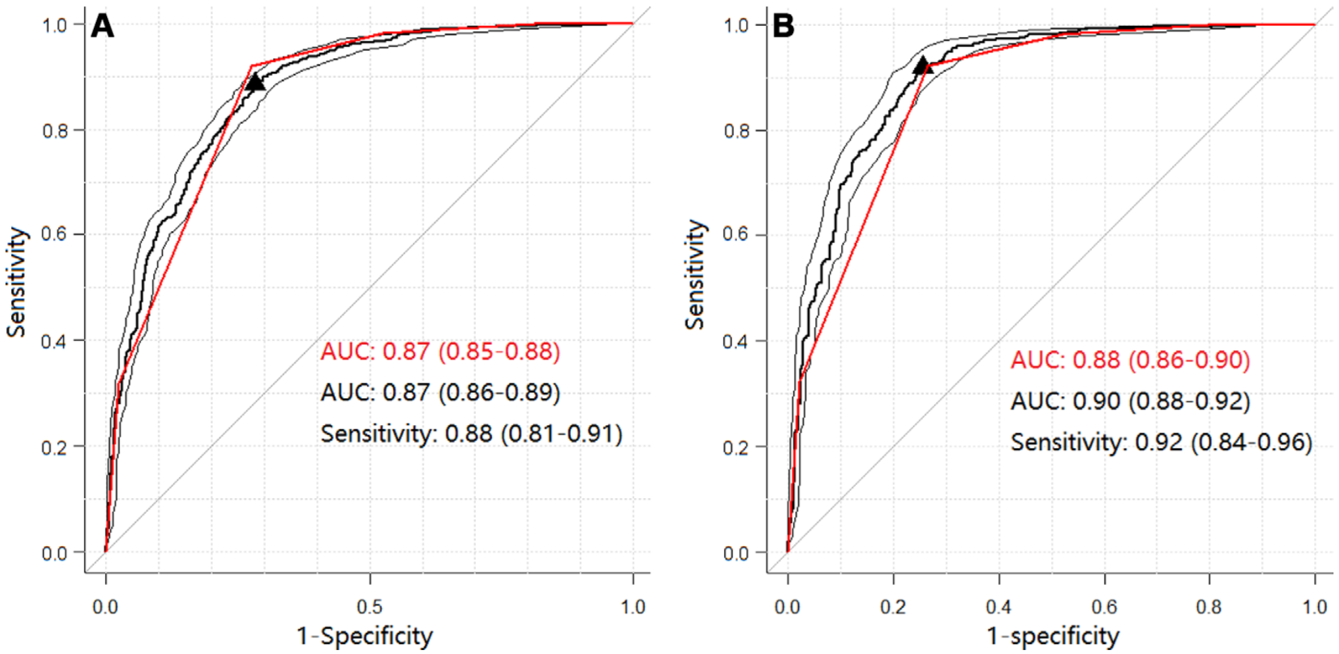

Figure 4: Receiver operating characteristic (ROC) curves show the classification model (black) and Chinese Thyroid Imaging Reporting and Data System (TI-RADS)–based radiologist diagnosis (red) at **(A)** the image level and **(B)** the patient level. The ROC curves, areas under the ROC curve (AUCs), and sensitivities in black correspond to the performance of the classification model, while the red curves and AUCs correspond to the performance of the radiologists using TI-RADS. The black triangles indicate the operation points for the classification model at the maximum Youden Index.

radiologists is shown in Appendix S1. *P* < .05 was considered indicative of a statistically significant difference.

Results

Patient Characteristics

A total of 24944 images in 10023 patients (median age, 46 years [IQR 37–55 years]; 7669 female, 2354 male) were collected from 208 hospitals across 31 provinces and municipalities of China. Table 1 lists the patient characteristics for each task and data set, including 7390 thyroid nodule images in 3857 patients, 2065 images in 1076 patients, and 1057 images in 545 patients used for training, validating, and testing, respectively, the detection and segmentation models; and 18477 thyroid nodule images in 7016 patients, 4563 images in 1987 patients, and 1904 images in 1020 patients used for training, validating, and testing, respectively, the classification model. Images were acquired with the US equipment of 12 vendors,

Table 4: Classification of AUC Values Based on Vendor

Training Data Set	Testing Data Set GE	Philips	Esaote	Mindray	Siemens	Other	Mixed
GE	0.71 (0.64, 0.79)	0.65 (0.60, 0.70)	0.63 (0.56, 0.69)	0.73 (0.68, 0.78)	0.60 (0.42, 0.78)	0.69 (0.61, 0.77)	0.58 (0.55, 0.61)
Philips	0.91 (0.87, 0.96)	0.82 (0.78, 0.86)	0.89 (0.85, 0.93)	0.85 (0.82, 0.89)	0.93 (0.85, 1.00)	0.84 (0.78, 0.90)	0.72 (0.69, 0.75)
Esaote	0.66 (0.58, 0.74)	0.76 (0.71, 0.80)	0.85 (0.81, 0.89)	0.78 (0.74, 0.82)	0.83 (0.70, 0.95)	0.55 (0.47, 0.62)	0.82 (0.79, 0.84)
Mindray	0.90 (0.86, 0.95)	0.68 (0.63, 0.73)	0.79 (0.74, 0.85)	0.83 (0.79, 0.86)	0.77 (0.62, 0.91)	0.85 (0.80, 0.90)	0.79 (0.76, 0.81)
Siemens	0.48 (0.39, 0.56)	0.74 (0.70, 0.79)	0.81 (0.76, 0.86)	0.83 (0.80, 0.87)	0.89 (0.80, 0.98)	0.85 (0.80, 0.91)	0.61 (0.58, 0.64)
Others	0.80 (0.73, 0.87)	0.73 (0.68, 0.78)	0.86 (0.81, 0.90)	0.84 (0.80, 0.88)	0.87 (0.75, 0.98)	0.83 (0.78, 0.89)	0.73 (0.70, 0.75)
Mixed	0.98 (0.97, 1.00)	0.88 (0.85, 0.91)	0.93 (0.90, 0.96)	0.92 (0.90, 0.94)	0.90 (0.79, 1.00)	0.92 (0.88, 0.97)	0.85 (0.83, 0.87)

Note.—Data in parentheses are 95% CIs. AUC = area under the receiver operating characteristic curve. “Other” vendors include Zonare, SonoScape, and VINNO.

the top five of which were Mindray (2592 of 10 023 patients, 26%), Philips (2517 of 10 023 patients, 25%), Esaote (1945 of 10 023 patients, 19%), GE Healthcare (1344 of 10 023 patients, 14%), and Siemens Healthineers (541 of 10 023 patients, 5%) (Table 2). Figure 2 illustrates typical US images of malignant and benign thyroid nodules.

Model Performance and Generalization

Figure 3A shows the performance of the detection model. The average precision of the precision-recall curve was 0.98 (95% CI: 0.96, 0.99), with an F1 score of 0.96 and sensitivity of 0.97 (1025 of 1057 nodules; 95% CI: 0.95, 0.99). When applied to US cine clips of 30 frames per second, the detection model inferenced fast enough to localize the thyroid nodule on each frame in real time (Fig 3B).

The performance of the segmentation models was evaluated according to hospitals, vendors, or regions. For the performance of models based on hospitals, the performance differences of the hospital #001 model and mixed hospital model are not statistically significant for six of the eight test sets ($P > .05$ for all comparisons) (Appendix S1, Fig S1). For the performance of models evaluated based on vendors, the mixed vendor model had the highest Dice coefficients among all test sets, ranging from 0.83 (GE test set; 95% CI: 0.80, 0.85) to 0.90 (“other” vendor test set; 95% CI: 0.86, 0.94). The lowest Dice coefficient of the evaluation of vendor-based generalizability was 0.63 (95% CI: 0.60, 0.65), which was for the model trained with the Mindray data set and tested on the GE test data set (Table 3). For the generalizability of models based on regions, the nationwide model had the best performance, with the highest Dice coefficient of 0.91 (95% CI: 0.90, 0.91) (Appendix S1, Table S1). The mean Dice coefficient of the general segmentation model, which was trained with the training set irrespective of hospital, vendor, or region, was 0.86 (95% CI: 0.86, 0.87) in the test set.

Table 5: Classification of AUC Values Based on Region

Training Data Set	Testing Data Set East Coast	Inland	Nationwide
East coast	0.84 (0.81, 0.87)	0.72 (0.69, 0.76)	0.81 (0.79, 0.83)
Inland	0.64 (0.60, 0.68)	0.82 (0.78, 0.85)	0.64 (0.61, 0.67)
Nationwide	0.86 (0.83, 0.88)	0.84 (0.81, 0.87)	0.84 (0.82, 0.86)

Note.—Data in parentheses are 95% CIs. Data sets were obtained from 208 hospitals that are members of the Chinese Artificial Intelligence Alliance for Thyroid and Breast Ultrasound across 31 provinces and municipalities. AUC = area under the receiver operating characteristic curve.

The results of evaluating the diagnostic performance of the classification models are presented in Figure 4. The diagnostic performance of the classification models at the image level was evaluated according to hospitals, vendors, or regions. For the generalizability of models trained on hospitals, the hospital #001 and mixed hospital models have significant difference in performance in five of the seven test sets ($P < .001$ for all comparisons) (Appendix S1, Fig S2). For the generalizability of models trained on various vendors, the model trained on the mixed vendor data set showed the best performance, as indicated by the highest AUC in all test sets except the Siemens test set trained using the Philips data set, from 0.85 (mixed vendor test set; 95% CI: 0.83, 0.87) to 0.98 (GE test set; 95% CI: 0.97, 1.00) (Table 4). For the generalizability of models trained on regions, the highest AUC values were found in all test sets when the nationwide model was evaluated; the AUC values of the east coast, inland, and nationwide test sets were 0.86 (95% CI: 0.83, 0.88), 0.84 (95% CI:

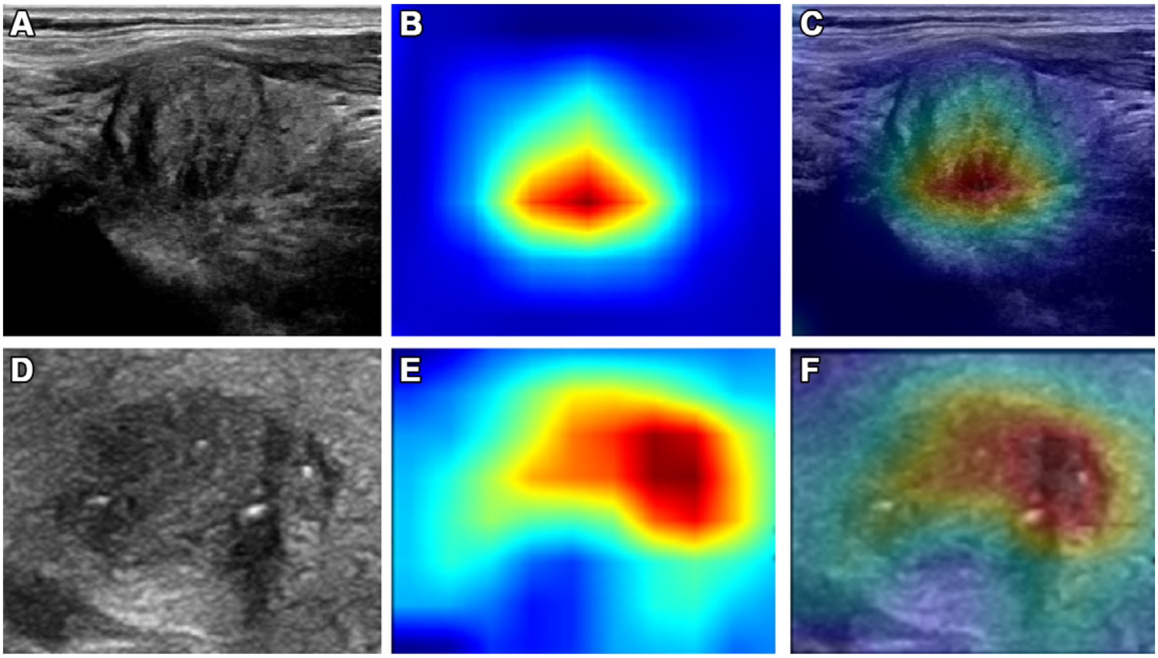

Figure 5: Visualization of benign and malignant thyroid nodules. **(A)** US image in a 38-year-old female patient with no relevant symptoms shows a benign thyroid nodule (nodular goiter). **(B)** Heat map generated by the classification model shows the corresponding thyroid nodule in **(A)**. The classification model determines whether the nodule is benign or malignant based on the image features from the highlighted region. **(C)** Overlapped US–heat map image shows that the classification model focuses on the nodule region consisting of isoechoic and markedly hypoechoic areas. **(D)** US image in a 44-year-old female patient with no relevant symptoms shows a malignant thyroid nodule (papillary thyroid carcinoma). **(E)** Heat map generated by the classification model shows the corresponding thyroid nodule in **(D)**. **(F)** Overlapped US–heat map image shows that the classification model focuses on the nodule region consisting of echogenic foci and hypoechoic area.

0.81, 0.87), and 0.84 (95% CI: 0.82, 0.86), respectively. The lowest AUCs were found in the nationwide test set when the inland training set was used (0.64; 95% CI: 0.61, 0.67) and the east coast test set when the inland training set was used (0.64; 95% CI: 0.60, 0.68) (Table 5). The heat map results indicate that the classification model could cover the majority of thyroid nodules in the US images (Fig 5).

For the test set consisting of 1020 patients, the classification model trained on a set of images from multiple hospitals, vendors, and regions obtained an AUC of 0.87 (95% CI: 0.86, 0.89) and sensitivity of 0.88 (1088 of 1237 images; 95% CI: 0.81, 0.91) at the image level (Fig 4A). Considering that diagnosis is not established on the basis of a single US image in real-world clinical settings, the highest predicted malignancy potential from multiple images of each patient was selected to reevaluate the model performance. At the patient level, the classification model had an AUC of 0.90 (95% CI: 0.88, 0.92) and sensitivity of 0.92 (588 of 639 patients; 95% CI: 0.84, 0.96) (Fig 4B). Moreover, the classification model at the patient level exhibited no statistically significant difference in diagnostic performance between the age groups of younger than 45 years (n = 490) and older than or equal to 45 years (n = 530), with AUCs of 0.90 (95% CI: 0.87, 0.94) and 0.90 (95% CI: 0.88, 0.93), respectively.

Comparison and Combination of AI Model and Radiologists

For the test set, the AUCs of radiologists' staging based on the Chinese TI-RADS ranged from 0.79 (95% CI: 0.77, 0.81) to 0.85 (95% CI: 0.83, 0.87) at the image level, and from 0.82 (95% CI: 0.79, 0.84) to 0.88 (95% CI: 0.86, 0.90) at the patient level. The diagnostic performance of all six radiologists was inferior to that of the AI classification model at both the image level and the patient level (Fig 4, Table 6). The same thyroid AI classification model showed different performances when used by radiologists in different ways in a virtual clinical scenario. In the freestyle AI–assisted mode (Fig 6), all three junior radiologists and one senior radiologist did not show any improvement in diagnostic performance compared with that of AI, regardless of whether at the image or patient level ($P \geq .07$ for all). The performance of the other two senior radiologists improved at both the image level and the patient level ($P \leq .01$ for all). In the rule-based AI–assisted mode, which artificially increases or decreases a point of Chinese TI-RADS classification based on malignant or benign diagnosis by the AI model, all six radiologists achieved an improvement in their diagnostic accuracies at both the image level and the patient level ($P \leq .02$ for all). At the patient level, the Kendall coefficients of concordance without AI assistance were 0.69, 0.74, and 0.77 for all radiologists, senior radiologists, and junior radiologists, respectively; and in the rule-based AI assistance mode, the corresponding coefficients were 0.86, 0.88, and 0.89, respectively.

Discussion

The generalizability of an artificial intelligence (AI) model indicates its feasibility to be applied universally in other medical

Table 6: Comparison of Diagnostic Performance of the AI Classification Model and Radiologists

	AUC						
Scoring and Experience Level	AI Model	Diagnosis without AI Assistance	Diagnosis with Freestyle AI Assistance	Diagnosis with Rule-based AI Assistance	*P* Value 1*	*P* Value 2†	*P* Value 3‡
Image-level scoring							
AI model	0.87 (0.86, 0.89)	NA	NA	NA	NA	NA	NA
Junior radiologists							
Reader 1		0.80 (0.78, 0.82)	0.79 (0.78, 0.81)	0.85 (0.84, 0.87)	.81	<.001	<.001
Reader 2		0.81 (0.79, 0.83)	0.82 (0.80, 0.84)	0.85 (0.84, 0.87)	.07	<.001	<.001
Reader 3		0.81 (0.79, 0.83)	0.82 (0.80, 0.84)	0.85 (0.83, 0.87)	.08	<.001	<.001
Senior radiologists							
Reader 4		0.84 (0.82, 0.86)	0.83 (0.81, 0.85)	0.86 (0.84, 0.88)	.38	<.001	<.001
Reader 5		0.79 (0.77, 0.81)	0.83 (0.82, 0.85)	0.86 (0.84, 0.88)	<.001	<.001	<.001
Reader 6		0.85 (0.83, 0.87)	0.87 (0.85, 0.88)	0.88 (0.86, 0.90)	.005	<.001	.004
Patient-level scoring							
AI model	0.90 (0.88, 0.92)	NA	NA	NA	NA	NA	NA
Junior radiologists							
Reader 1		0.84 (0.82, 0.87)	0.83 (0.81, 0.86)	0.88 (0.86, 0.91)	.34	<.001	<.001
Reader 2		0.83 (0.80, 0.85)	0.84 (0.82, 0.87)	0.87 (0.85, 0.90)	.12	<.001	<.001
Reader 3		0.82 (0.79, 0.84)	0.83 (0.80, 0.86)	0.87 (0.85, 0.89)	.19	<.001	<.001
Senior radiologists							
Reader 4		0.86 (0.84, 0.89)	0.85 (0.82, 0.87)	0.88 (0.86, 0.90)	.34	.02	<.001
Reader 5		0.84 (0.81, 0.86)	0.86 (0.84, 0.88)	0.89 (0.87, 0.91)	.01	<.001	<.001
Reader 6		0.88 (0.86, 0.90)	0.90 (0.88, 0.92)	0.90 (0.88, 0.92)	.01	<.001	.02

Note.—Data in parentheses are 95% CIs. The general AI classification model used a training set of multiple hospitals, vendors, and regions. Radiologist diagnosis without AI assistance was based on the Chinese Thyroid Imaging Reporting and Data System. AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, NA = not applicable.

* Comparison for diagnosis without AI assistance versus freestyle AI–assisted mode.

† Comparison for diagnosis without AI assistance versus rule-based AI–assisted mode.

‡ Comparison for diagnosis without AI assistance versus the result of automatic discrimination between benign and malignant nodules suggested by the AI classification model.

centers, assuring the clinical usability of the model. In principle, training the model with a larger data set consisting of images of a greater diversity would improve the generalizability. Our study has suggested that the segmentation and classification models based on mixed hospital, mixed vendor, and nationwide data sets showed the best performance in model testing.

However, similarity was found between the Dice coefficients of hospital #001 and those of the mixed hospital segmentation models in 75% (six of eight) of the test set. The segmentation model identified only the boundary of thyroid nodules, while the classification

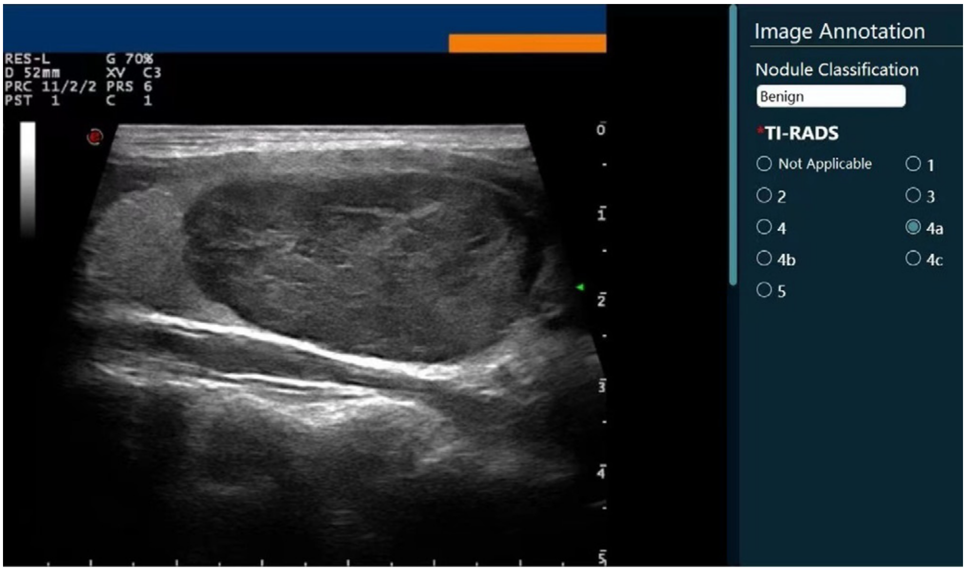

Image Annotation

Nodule Classification: Benign

*TI-RADS

○ Not Applicable ○ 1
○ 2 ○ 3
○ 4 ● 4a
○ 4b ○ 4c
○ 5

Figure 6: Screenshot shows the software interface of the freestyle artificial intelligence (AI)–assisted mode with which radiologists rate US images. The AI classification result is given in the white box in the upper right corner.

model required a composite analysis of the nodule boundary, internal structure, echogenicity, and calcification before determining the pathologic characteristics of the nodule. Therefore, a high-quality segmentation model can be built based on larger sample data from mixed vendors provided by a single hospital. In contrast, the classification model has limited robustness and would require a larger amount of sample data from mixed hospitals.

For the generalizability of segmentation and classification models based on vendors, the mixed vendor model demonstrated the best performance among all test sets. However, some vendor results were not intuitive. The most typical example was seen in the model trained on the Philips data set, which performed better on the test data sets of the other top four vendors (AUC, 0.85–0.93) than the Philips test data set (AUC, 0.82). A possible explanation is that the same vendor has different US systems ranging from high- to low-end machines, resulting in a wide variation in the quality and style of US images, and thus the best performance does not necessarily occur in the test data set of the same vendor.

A larger number of tertiary care hospitals are in the east coast region of China and thus benefit from broad referral bases. Hospitals in the east coast region are often equipped with high-end US equipment with higher performance, which may explain why the segmentation and classification models trained on the east coast region data set had better performance than those trained on the inland region data set.

Our models showed performance comparable with that of many proposed US AI models for thyroid (16–20). Our detection model reached an average precision of 0.98, while the cascade convolutional neural network developed by Ma et al (17) had an AUC of 0.99. Our segmentation model had a mean Dice coefficient of 0.86, comparable with those studies with Dice coefficients between 0.760 and 0.909 (18,19). The classification model in our study had an AUC of 0.90, similar to that reported in other studies to date (16,20). It has been reported that the US features of malignant thyroid nodules differ between younger and older individuals. The developed AI model exhibited no significant difference in the diagnostic performance between patients older than or equal to 45 years and those younger than 45 years of age (P = .925).

A study by Peng et al (5) showed that the AUC of thyroid AI-assisted US (0.922) was significantly higher than that of radiologists (0.839) and our results confirmed this finding, with AI-assisted US outperforming both junior and senior radiologists ($P \leq .02$ for all). Typically, junior radiologists obtain more help from AI than senior radiologists (5,21,22). However, in clinical practice there is no standard pattern for how radiologists and AI-assisted thyroid US should be combined. These may include the freestyle AI–assisted mode (5), the rule-based AI-assisted mode (23), and the combination of both modes (22). In our study, when using the rule-based AI–assisted mode, the diagnostic performance of all six radiologists improved ($P < .05$ for all), although the AUC improvements were relatively modest. As a comparison, 33% (one of three) of senior radiologists and all junior radiologists did not benefit when using the freestyle AI–assisted mode. Mai et al (22) found that all senior radiologists did not benefit from both AI-assisted modes. These findings highlight the potential challenge of implementing AI models into clinical practice. The freestyle AI-assisted mode allows radiologists to weigh in when they need to "consult AI," as opposed to the rule-based AI model that automatically upgrades or downgrades the Chinese TI-RADS score, which artificially increases the confidence of radiologists. Both our study and the study by Mai et al indicated that the rule-based AI–assisted mode may be more suitable for junior radiologists. For senior radiologists, the use of the freestyle AI–assisted mode could lead to better diagnostic performance. However, we should acknowledge that radiologists' performance in classifying thyroid nodules at US is already quite high without AI, with AUCs ranging 0.82–0.88 for patient-level scoring; thus, the AI-assisted model can only modestly improve diagnostic performance. The practical benefits of clinical application of an AI model are yet to be determined.

Our study had several limitations. First, only B-mode US images of thyroid nodules were studied, and other imaging modalities, such as color Doppler US and elastography, were not included. Second, the established models were not compared with those established by other research institutions. Third, only nodules with pathologic assessment results were included, which likely led to an underrepresentation of low-suspicion or benign nodules. Fourth, our study did not address clinical information, including body habitus, which may have potential implications for generalization of the model. Lastly, although the incorporation of data from many hospitals across China has improved the generalizability, it remains a single-country data set. For example, the model performance still has geographic variations, and thus its generalizability has room for improvement.

In conclusion, our study findings indicate the importance of including diverse data sets in terms of hospitals, vendors, and regions for the development of a deep learning model. Our study provides a useful reference for the development of generalizable thyroid US artificial intelligence models in the future.

Acknowledgment: We thank Dr Mingzi Zhang, PhD, of Shanghai Aitrox Co Ltd for his valuable advice and assistance throughout the revision of this manuscript.

Author contributions: Guarantors of integrity of entire study, **W.W.X., Z.H.M., R.F.Z., Y.G., X.C., B.Y.B., Q.Y.L., H.Z., B.G., H.T.Z., J.Q.Z.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **W.W.X., X.H.J., Z.H.M., X.L.G., Y.L., C.C.F., R.F.Z., Y.G., X.M.L., N.L., B.Y.B., Q.Y.L., L.G., Q.F., C.H., W.W.Z., T.L., H.T.Z., J.Q.Z.**; clinical studies, **W.W.X., Z.H.M., X.L.G., R.F.Z., Y.G., X.C., X.M.L., N.L., B.Y.B., Q.Y.L., J.P.Y., H.Z., L.G., B.G., T.L., H.T.Z., Y.J.D., J.Q.Z.**; experimental studies, **Z.H.M., X.L.G., Y.L., R.F.Z., Y.G., N.L., B.Y.B., Q.Y.L., L.G., K.Y.Z., Q.F., H.T.Z.**; statistical analysis, **Z.H.M., Y.L., R.F.Z., Y.G., N.L., B.Y.B., Q.Y.L., L.G., T.L., H.T.Z., J.Q.Z.**; and manuscript editing, **W.W.X., X.H.J., Z.H.M., X.L.G., Y.L., C.C.F., R.F.Z., Y.G., X.M.L., N.L., B.Y.B., Q.Y.L., L.G., T.L., H.T.Z., J.Q.Z.**

Disclosures of conflicts of interest: **W.W.X.** No relevant relationships. **X.H.J.** No relevant relationships. **Z.H.M.** No relevant relationships. **X.L.G.** No relevant relationships. **Y.L.** No relevant relationships. **C.C.F.** No relevant relationships. **R.F.Z.** No relevant relationships. **Y.G.** No relevant relationships. **X.C.** No relevant relationships. **X.M.L.** No relevant relationships. **N.L.** No relevant relationships. **B.Y.B.** No relevant relationships. **Q.Y.L.** No relevant relationships. **J.P.Y.** No relevant relationships. **H.Z.** No relevant relationships. **L.G.** No relevant relationships.

B.G. No relevant relationships. **K.Y.Z.** No relevant relationships. **Q.F.** No relevant relationships. **C.H.** No relevant relationships. **W.W.Z.** No relevant relationships. **T.L.** No relevant relationships. **H.T.Z.** No relevant relationships. **Y.J.D.** No relevant relationships. **J.Q.Z.** No relevant relationships.

References

1. 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):1–133.
2. Lin JD, Chao TC, Huang BY, Chen ST, Chang HY, Hsueh C. Thyroid cancer in the thyroid nodules evaluated by ultrasonography and fine-needle aspiration cytology. Thyroid 2005;15(7):708–717.
3. Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. J Am Coll Radiol 2017;14(5):587–595.
4. Zhou J, Yin L, Wei X, et al. 2020 Chinese guidelines for ultrasound malignancy risk stratification of thyroid nodules: the C-TIRADS. Endocrine 2020;70(2):256–279.
5. Peng S, Liu Y, Lv W, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. Lancet Digit Health 2021;3(4):e250–e259.
6. Yi J, Kang HK, Kwon JH, et al. Technology trends and applications of deep learning in ultrasonography: image quality enhancement, diagnostic support, and improving workflow efficiency. Ultrasonography 2021;40(1):7–22.
7. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin 2019;69(2):127–157.
8. Jia X, Ma Z, Kong D, et al. Novel Human Artificial Intelligence Hybrid Framework Pinpoints Thyroid Nodule Malignancy and Identifies Overlooked Second-Order Ultrasonographic Features. Cancers (Basel) 2022;14(18):4440.
9. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016; 779–788.
10. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015; 234–241.
11. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision, 2017; 618–626.
12. Wang Z, Zhang H, Zhang P, He L, Dong W. Diagnostic value of ultrasound-detected calcification in thyroid nodules. Ann Acad Med Singap 2014;43(2):102–106.
13. Shi C, Li S, Shi T, Liu B, Ding C, Qin H. Correlation between thyroid nodule calcification morphology on ultrasound and thyroid carcinoma. J Int Med Res 2012;40(1):350–357.
14. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(1):77.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174.
16. Li X, Zhang S, Zhang Q, 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(2):193–201.
17. Ma J, Wu F, Jiang T, Zhu J, Kong D. Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images. Med Phys 2017;44(5):1678–1691.
18. Kumar V, Webb J, Gregory A, et al. Automated Segmentation of Thyroid Nodule, Gland, and Cystic Components From Ultrasound Images Using Deep Learning. IEEE Access 2020;8:63482–63496.
19. Koundal D, Sharma B, Guo Y. Intuitionistic based segmentation of thyroid nodules in ultrasound images. Comput Biol Med 2020;121:103776.
20. Shi G, Wang J, Qiang Y, et al. Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification. Comput Methods Programs Biomed 2020;196:105611.
21. 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.
22. Mai W, Zhou M, Li J, et al. 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(10):7936–7944.
23. He LT, Chen FJ, Zhou DZ, et al. A Comparison of the Performances of Artificial Intelligence System and Radiologists in the Ultrasound Diagnosis of Thyroid Nodules. Curr Med Imaging Rev 2022;18(13):1369–1377.