INTERNATIONAL JOURNAL OF SURGERY

Automated diagnosis and management of follicular thyroid nodules based on the devised small-dataset interpretable foreground optimization network deep learning: a multicenter diagnostic study

Zheyu Yang, MD[a], Siqiong Yao, PhD[b], Yu Heng, MD[c], Pengcheng Shen, PhD[b], Tian Lv, MD[h], Siqi Feng, MD[g], Lei Tao, PhD[c], Weituo Zhang, PhD[d,f], Weihua Qiu, PhD[a,e], Hui Lu, PhD[b], Wei Cai, PhD[a]

Background: Currently, follicular thyroid carcinoma (FTC) has a relatively low incidence with a lack of effective preoperative diagnostic means. To reduce the need for invasive diagnostic procedures and to address information deficiencies inherent in a small dataset, we utilized interpretable foreground optimization network deep learning to develop a reliable preoperative FTC detection system.

Methods: In this study, a deep learning model (FThyNet) was established using preoperative ultrasound images. Data on patients in the training and internal validation cohort ($n = 432$) were obtained from Ruijin Hospital, China. Data on patients in the external validation cohort ($n = 71$) were obtained from four other clinical centers. We evaluated the predictive performance of FThyNet and its ability to generalize across multiple external centers and compared the results yielded with assessments from physicians directly predicting FTC outcomes. In addition, the influence of texture information around the nodule edge on the prediction results was evaluated.

Results: FThyNet had a consistently high accuracy in predicting FTC with an area under the receiver operating characteristic curve (AUC) of 89.0% [95% CI 87.0–90.9]. Particularly, the AUC for grossly invasive FTC reached 90.3%, which was significantly higher than that of the radiologists (56.1% [95% CI 51.8–60.3]). The parametric visualization study found that those nodules with blurred edges and relatively distorted surrounding textures were more likely to have FTC. Furthermore, edge texture information played an important role in FTC prediction with an AUC of 68.3% [95% CI 61.5–75.5], and highly invasive malignancies had the highest texture complexity.

Conclusion: FThyNet could effectively predict FTC, provide explanations consistent with pathological knowledge, and improve clinical understanding of the disease.

Keywords: automated diagnosis, deep learning, follicular thyroid nodules, interpretable foreground optimization network

[a]Department of General Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, [b]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, [c]Department of Otolaryngology, Eye, Ear, Nose and Throat Hospital, Fudan University, [d]Shanghai Tong Ren Hospital and Clinical Research Institute, [e]Department of General Surgery, Ruijin Hospital Gubei Campus, Shanghai Jiao Tong University School of Medicine, [f]Hong Qiao International Institute of Medicine, Shanghai, [g]Department of General Surgery, Liaoning Cancer Hospital & Institute, Shenyang and [h]Department of Head, Neck and Thyroid Surgery, Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou, People's Republic of China

Z.Y., S.Y. and Y.H. are contributed equally to this article.

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

*Corresponding author. Address: Address: Shanghai Tong Ren Hospital and Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, 1111 Xianxia Road, Shanghai 200050, China. Tel.: +86 21 62909911; fax: + 86 21 62776196. E-mail: zhangweituo@sjtu.edu.cn (W. Zhang); Department of General Surgery, Ruijin Hospital, Shanghai Jiaotong University School of medicine, 197 Ruijin 2nd Road, Shanghai 200025, PR China. Department of General Surgery, Civil Aviation Shanghai Hospital, 398 Hongbaoshi Road, Shanghai 201103, PR China. Tel.: +86 21 64370045; fax: +86 21 64454473. E-mail: qwh11072@rjh.com.cn (W. Qiu); School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, No.800, Dongchuan Road, Minhang District, Shanghai 200240, PR China. Tel.: +86 21 63846590; fax: +86 21 342068. E-mail: huilu@sjtu.edu.cn (H. Lu); Department of General Surgery, Ruijin Hospital, Shanghai Jiaotong University School of medicine, 197 Ruijin 2nd Road, Shanghai 200025, PR China. Tel: +86-18917762028. Tel.: + 86 21 64370045; fax: +86 21 64454473. E-mail: caiwei@shsmu.edu.cn (W. Cai).

International Journal of Surgery (2023) 109:2732–2741

Received 6 January 2023; Accepted 10 May 2023

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.lww.com/international-journal-of-surgery.

Published online 18 May 2023

http://dx.doi.org/10.1097/JS9.0000000000000506

Introduction

Follicular thyroid carcinoma (FTC) is the second most common well-differentiated thyroid carcinoma derived from thyroid follicular epithelium, exhibiting clinical and biological characteristics that are distinct from those of papillary thyroid carcinoma (PTC)[1-3]. To date, the diagnosis of FTC remains challenging, especially in the context of separating FTC from benign follicular nodules (FNs) (follicular adenoma, FA) through preoperative detection using cytologic, sonographic, or clinical features. Research findings reported that ultrasound, which can efficiently diagnose PTC following characteristics of malignancy, cannot reliably distinguish between FA and FTC[4]. The only reliable way of differentiating them is through diagnostic surgery to investigate definite capsular invasion (CI) and/or vascular invasion (VI) by pathological examinations[1,4-6]. Although differentiated thyroid carcinoma (DTC) presents a perfect prognosis with a 5-year disease-free survival rate of over 98%, FTC is, in contrast, generally considered to be more aggressive and carries risks of distant metastasis[1,7,8]. Thus, patients with suspicious FNs are recommended to undergo diagnostic surgeries in clinical guidelines, leading to a large portion of patients receiving surgery for FN that were later pathologically confirmed to be benign tumors[9,10]. Therefore, an effective differentiating method targeting benign and malignant FNs, especially a preoperative one using noninvasive means, may carry welcomed significance for the current diagnosis and treatment of thyroid follicular tumors.

Artificial intelligence (AI), typified by deep learning, has been reported to assist clinicians in providing precise strategies for diagnosis, staging, medication, and other aspects of diagnostic and treatment processes[11,12]. According to the latest list in U.S. Food and Drug Administration (FDA), more than 178 AI and machine learning-enabled medical devices were added to the clinical approval. AI systems also show great potential in the head and neck, especially thyroid oncology, which has been reported to meet or exceed human experts in medical imaging based on its outstanding ability of feature extraction. With predictive efficiency and repeatability, these systems presented the possibility of fewer or more invasive diagnostic procedures[12,13]. However, challenges remain where the poor characterization capabilities of models arising from the high parameter complexity of the deep learning algorithm result in the inability to provide a clear explanation for the predicted results. Researchers have also pointed out that the application of AI still has many limitations[14,15]. Yet, the value of pursuing AI-assisted diagnosis should not be overlooked, for it addresses issues of limited predictive efficiency and repeatability in traditional statistical models or radiomics models.

At present, several research teams have established reliable AI diagnostic systems for thyroid nodules[12,16,17]. However, for specific pathological subtypes of thyroid cancer, especially FTC, there is still a lack of relevant research. The aim of this multicenter, cross-machine, multi-operator study was to establish a specific deep learning model (FThyNet) for differentiating the benign and malignant FNs with well-matched external validation and provide a proper clinical explanation for this predicting system. The present research fills the gap in providing efficient preoperative prediction models for differentiating FTC and FA, and also provides a new possibility to alter the traditional therapeutic pattern of follicular thyroid tumors.

HIGHLIGHTS

- This newly established model was the first to differentiate follicular thyroid carcinoma (FTC) and follicular adenoma (FA) with a preoperative approach.
- The present study provided the largest sample of FTC patients and ultrasound images in multi/single-center clinical research on FTC so far.
- The model is more efficient in identifying higher clinical stage FTC (grossly invasive or distant metastasis).
- The FThyNet provided outstanding efficiency in training, testing, and valid external sets. The excellent generalization is shown not only in the external validation results from other clinical centers, but also in the regrouping result of high-stage FTC.
- The FThyNet provided reasonable interpretability consistent with pathological features.

Methods

Study design and patients

A retrospectively collected multicenter database over a 5-year period (2015–2020) from China was analyzed. All patients underwent a first-time thyroidectomy with complete thyroid examinations for FNs. Clinical centers involved in the study include the Department of General Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine (Center A), Eye and ENT Hospital of Fudan University (Center B), Ruijin Hospital Gubei Campus (Center C), Liaoning Carcinoma Hospital & Institute (Center D), Zhejiang Provincial People’s Hospital (Center E). A total of 2735 patients diagnosed with FNs were enrolled in the study, among which 298 cases were histologically proven FTC.

The exclusion criteria were as follows: no accessible preoperative ultrasound examinations ($n=21$); histologically proven non-invasion follicular thyroid neoplasm with papillary-like nuclear features (NIFTP) or no definite histologically proven FTC ($n=28$); history of thyroidectomy or coexistence of other head and neck carcinoma ($n=11$). After exclusion, 248 patients with pathologically proven FTC were studied, together with 1:1 matched 255 cases randomly sampled from patients histopathologically diagnosed with FA. Two datasets were included in the present study. Dataset 1 enrolled 432 patients (212 FTC and 220 FA) from clinical center A and was used as the main cohort. Dataset 2 enrolled 71 patients (36 FTC and 35 FA) from clinical centers B, C, D, and E were used as an external test set. This work has been reported in line with the STROCSS (strengthening the reporting of cohort, cross-sectional and case–control studies in surgery) criteria[18], Supplemental Digital Content 1, http://links.lww.com/JS9/A506.

Data sources, surgical approach, and pathological diagnosis

Data sources, including demographics, serum index, and fine-needle aspiration (FNA), were collected from Electronic Medical Records System (EMRS) for further analysis. Original ultrasound images were collected from EMRS and the medical image management system. All ultrasound data were provided by ultrasound radiologists with more than 5 years of experience in thyroid ultrasound. All of the radiologists had received rigorous training to standardize the imaging parameter adjustment method and the ultrasound scanning procedure of the thyroid according to the AIUM (American Institute of Ultrasound in Medicine) practice guideline for performing thyroid ultrasound. Longitudinal and transverse image sets contained at least the target nodules to be used for subsequent analysis. The image covering the complete outline of the smaller part of the nodule can be obtained by adjusting the position of the scanning section when the nodule is beyond the display range of the probe. Data from each subcenter were gathered and reviewed by two ultrasound radiologists, and only data that passed the quality control examination were enrolled.

Classification of the preoperative ultrasound examinations was based on Kwak Thyroid Imaging Reporting and Data System (TI-RADS)[19]. Surgical treatment was decided following the recommendations in the 2015 American Thyroid Association (ATA) Management Guidelines[20]. Surgery was performed by general surgeons or thyroid specialists at the main center, and by thyroid specialists or head and neck surgeons at the other subcenters. Surgical procedures included total thyroidectomy and thyroid lobectomy with/without neck lymph node dissection. Patients with distant metastasis only underwent surgical biopsy for metastatic lesions. Pathological examination of all patients included intraoperative cryopathology together with postoperative immunohistochemistry and molecular examination. All acquired specimens were examined by two or more board-certified pathologists from each center. Diagnosis of FTC followed the WHO definition and standard diagnostic criteria for FTC[21,22]. All patients enrolled were classified based on postoperative histopathological results, according to the 2015 Tumor Node Metastasis (TNM) staging system of the American Joint Committee on Carcinoma (AJCC), 8th edition[23].

Development of deep learning model

Sufficient and frequent communication was done between researchers and doctors to ensure the clinical significance of the study, as shown in Figure 1. First, the doctors highlighted the need for recognition improvement of follicular carcinoma and its significance. The researchers then obtained and visualized the information corresponding to the subject, confirming with the doctor that the subject's information was correct. Thereafter, the researchers organized the data, designed the method, integrated the model, and cemented the model to address the problem. The doctors then tested whether it was effective. The data collected in the experiment included ultrasound images and pathological reports. Clinical Center A provided the training set and the validation set. The training set included 1392 images, including 705 FA images and 687 FTC images, and the validation set included 349 images, including 177 FA images and 172 FTC images. The external test set was provided by four hospitals, which included 309 images containing 150 FA images and 159 FTC images.

Based on known pathology, we modeled the thyroid FN based on the morphological information, which returned a benign or malignant classification model. The network architecture is shown in Figure 2. In the image preprocessing stage, we used the median filter operator to denoise the image and then used the gray threshold binarization to obtain the gray image, which was three-channel. We needed to segment the nodules effectively in advance, which could be used as the input to classify benign or malignant. However, compared with natural images or other medical images, thyroid ultrasound images had the characteristics of large-scale change, low contrast between organs, fuzzy background, relatively small detection targets, and small numbers. Due to the large intraclass variance of the background, serious misjudgment was an issue, which brought great challenges to the segmentation

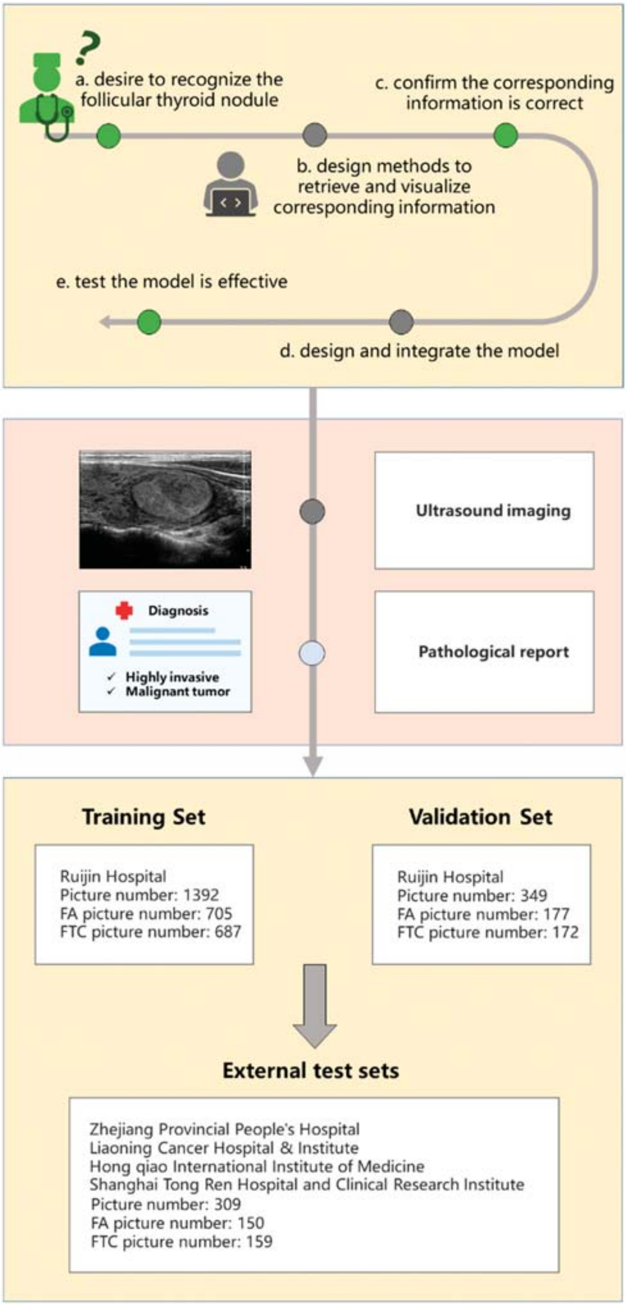
?
a. desire to recognize the follicular thyroid nodule
c. confirm the corresponding information is correct
b. design methods to retrieve and visualize corresponding information
e. test the model is effective
d. design and integrate the model
Ultrasound imaging
Diagnosis
✓ Highly invasive
✓ Malignant tumor
Pathological report
Training Set
Validation Set
Ruijin Hospital
Picture number: 1392
FA picture number: 705
FTC picture number: 687
Ruijin Hospital
Picture number: 349
FA picture number: 177
FTC picture number: 172
External test sets
Zhejiang Provincial People's Hospital
Liaoning Cancer Hospital & Institute
Hong qiao International Institute of Medicine
Shanghai Tong Ren Hospital and Clinical Research Institute
Picture number: 309
FA picture number: 150
FTC picture number: 159

Figure 1. Study design for the development and validation of a deep learning model to predict FTC. Doctors and researchers conducted multiple rounds of discussions and exchanges on the follicular topic. Data includes ultrasound imaging and pathological reports, where the training set and validation set are provided by Ruijin Hospital, and the data of the external validation set are provided by four other hospitals. FA, follicular adenoma; FTC, follicular thyroid carcinoma.

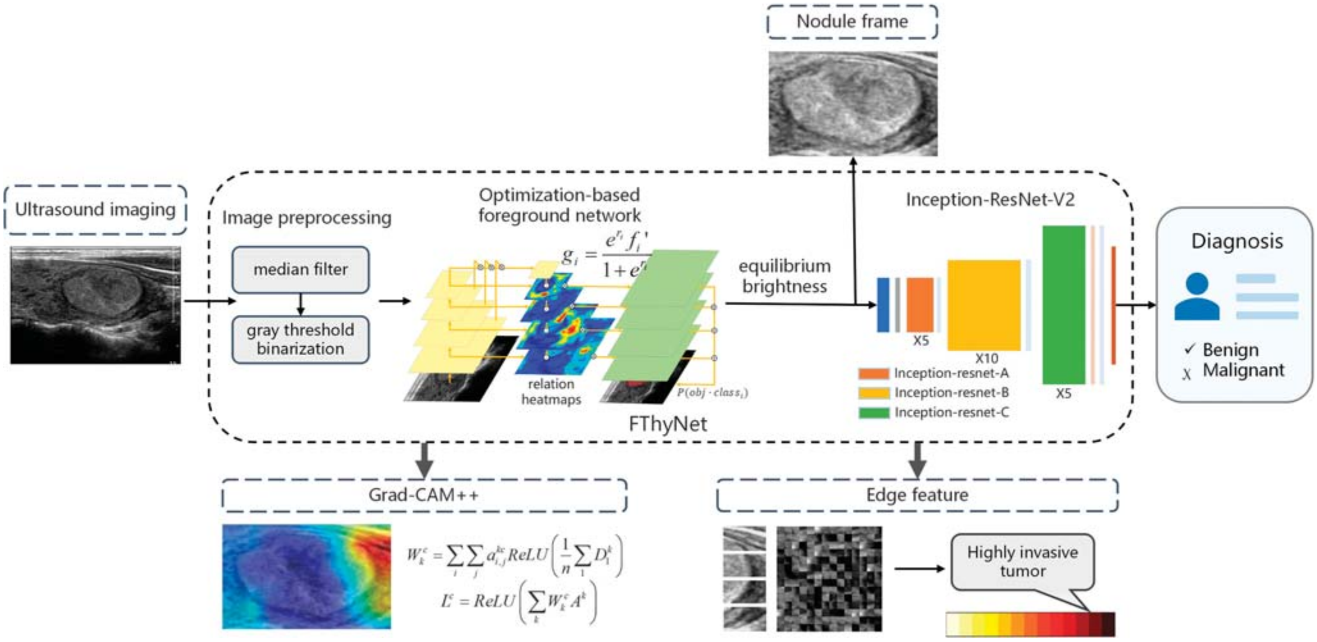

Figure 2. Technical flowchart of the model. The nodule was obtained after the ultrasound imaging was segmented by the optimization-based foreground network. Then the nodule detection frame was obtained with operations such as brightness equalization. The Inception-ResNet-V2 model was used to classify the frame data. Then the Grad-CAM + + visualization module and edge feature module provided clinical interpretability of the classification.

task[24-27]. This study used the foreground optimization method to realize the accurate segmentation of thyroid nodules[28,29]. The feature pyramid network (FPN) in this method could solve the multiscale input problem. The foreground–background relationship module could improve the discrimination of the nodule itself by correlating the surrounding semantic information related to the nodule morphological pixels, so as to enhance the difference between the nodule and the surrounding background. The foreground optimization module was embedded in the model to focus on the classification error samples misjudged as nodules in the background and to adjust the descending direction of the gradient to improve the recognition rate.

Among them, the specific solution method of the foreground–background relation network was as follows.

We started with definitions (1)–(3):

$$f_i = \begin{cases} \Gamma(C_i) + I(f_{i+1}),\ i = 2, 3, 4 \\ \Gamma(C_i),\ i = 5 \end{cases} \tag{1}$$

C_i denotes the ith feature map extracted from ResNets, where the feature map C_i has an output stride of 2^i pixels with respect to the input image. The top–down pathway and lateral connections are used to generate pyramidal feature maps. $\{f_i,\text{l}, i = 2, 3,4,5\}$, where Γ denotes the lateral connection implemented by a learnable 1×1 convolutional layer and I denotes the nearest neighbor upsampling with a scale factor of 2.

$$P:\quad \mathbb{R}^{d\times H\times W} \rightarrow \mathbb{R}^{d^u\times H\times W} \tag{2}$$

$$f_i^u = P_i(\theta_i)\circ f_i \tag{3}$$

f_i^u is the feature map and f_i transformed by the scale-aware projection function $P_i(\theta_i)$, where θ_i denotes the learnable parameters of $P_i(\theta_i)$. We adopt a simple form of $P_i(\theta_i)$, which is just implemented by 1×1 convolutional layer followed by batch normalization and ReLU in order.

Then we obtain the scene embedding vector v:

$$v = l\left(C_2, \ldots, C_i, \ldots, C_6\right),\ i = 3, 4, 5, \tag{4}$$

where l denotes a learnable projection function for scene representation with output space $\mathbb{R}^{d^u}$.

The variable r_i denotes the similar estimation implemented by pointwise inner product, it is got by calculation of v and f_i^u:

$$r_i = sim\left(v, f_i^u\right) = v\cdot f_i^u. \tag{5}$$

Feature map f_i will be recoded to get re-encoding feature map.f_i'

$$f_i' = re - encoding\left(f_i, \tau_i\right), \tag{6}$$

where the re-encoder has a learnable parameter τ_i. Re-encoding is the operation in the re-encoder module. We designed the re-encoder to introduce an extra non-linear unit to avoid feature degradation since the weighting operation is a linear function. There was a simple form of this re-encoder, which was implemented by a 1×1 convolutional layer followed by batch normalization and ReLU for high efficiency of parameters and computation. The item including r_i of Eqn. (5) is used to weight the re-encoded feature maps, which is the normalized relation

Table 1

Characteristics of two patient cohorts.

	Main cohort, $N=432$								
	Training set, $N=352$			Validation set, $N=80$			External test set, $N=71$		
Characteristics	**FA**	**FTC**	***P***	**FA**	**FTC**	***P***	**FA**	**FTC**	***P***
Age, years, mean	41.2 (23–71)	47.5 (27–78)	< 0.01	43.3 (30–73)	48.0 (33–75)	< 0.01	40.8 (26–66)	49.2 (39–68)	< 0.01
Sex			0.259			0.950			0.269
Male	65	53	–	15	14	–	13	9	–
Female	114	120	–	26	25	–	22	27	–
Primary tumor size, cm[a]									
Mean	3.17 (0.5–7)	3.36 (0.6–7)	0.422	3.06 (0.8–5)	3.31 (0.4–5.5)	0.167	2.76 (1.1–6)	3.06 (0.9–6.5)	0.347
> 4 cm	28	20	0.265	5	4	0.784	8	12	0.327
Minimally capsular invasion	0	131 (75.7%)	–	0	31 (79.5%)	–	0	22 (61.10%)	–
Grossly capsular/vessel invasion	0	42 (24.3%)	–	0	8 (20.5%)	–	0	14 (38.90%)	–
Distant metastasis	0	9 (5.2%)	–	0	3 (7.7%)	–	0	3 (8.30%)	–
Kwak TI-RADS			0.076			0.052			0.359
3	33.0%	31.8%	–	36.6%	30.8%	–	37.1%	25%	–
4A	67.0%	63.6%	–	63.4%	48.7%	–	62.9%	72.2%	–
4B	0	2.9%	–	0	7.7%	–	0	2.8%	–
4C	0	1.2%	–	0	10.3%	–	0	0	–
5	0	0.6%	–	0	2.6%	–	0	0	–
Total	179	173		41	39		35	36	

[a]The longest diameter of the largest lesion.

FA, follicular adenoma; FTC, follicular thyroid carcinoma; Kwak TI-RADS, Kwak Thyroid Imaging Reporting and Data System.

map using the sigmoid gate function based on a simple self-gating mechanism.

Finally, we obtain the relation-enhanced foreground feature map:g_i

$$g_i = \frac{e^{r_i} f_i'}{1 + e^{r_i}} \tag{7}$$

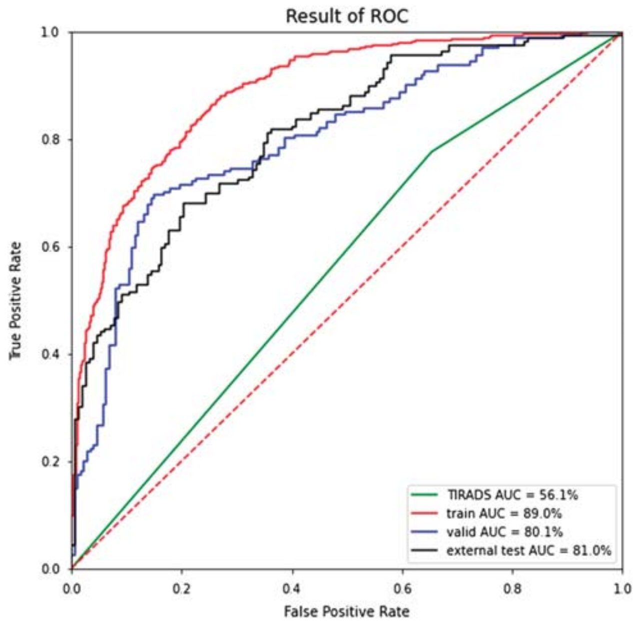

Figure 3. ROC curves of results. The curves show that the AUC of TIRADS is low, and the results of FThyNet are significantly better than those of TIRADS. AUC, area under the receiver operating characteristic curve; ROC, receiver operating characteristic; TIRADS, Thyroid Imaging Reporting and Data System.

This was used to model the data obtained from the segmentation, and Inception-ResNet-V2[30] was used to realize the binary classification prediction of benign or malignant thyroid nodules based on morphology. We verified the effectiveness of the model through model comparison experiments (STable 1, Supplemental Digital Content 2, http://links.lww.com/JS9/A507 and STable 2, Supplemental Digital Content 2, http://links.lww.com/JS9/A507). The thermodynamic feature map and nodule edge information analysis were made based on the results (Refer to the following content and supplement for details).

Segmentation network training parameters

Classification network training parameters

Comparative experiments of segmentation and classification network

Model interpretability approach

Comparison with human readers

Model evaluation and statistical analysis

Human-test Verifying the Understandability

See details in supplementary files, Supplemental Digital Content 2, http://links.lww.com/JS9/A507.

Ethical statement

All clinical data, including demographics, operative procedures, pathology, and complications, were retrospectively collected. The research was approved by the local Ethics Committee and the Institutional Review Board of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine Hospital, and was also approved by Chinese Clinical Trial (ChiCTR2200060823). Written informed consent was obtained from the patient for publication in this study and accompanying images. A copy of the written consent was available for review by the Editor-in-Chief of this journal on request.

Table 2

Classification results of TIRADS and FThyNet.

		FThyNet		
	TIRADS (*n* = 2050)	**Training set (*n* = 1392)**	**Validation set (*n* = 349)**	**Test set (*n* = 309)**
Accuracy	57.9% [52.8–62.3%]	79.7% [77.0–82.1%]	77.5% [73.1–81.7%]	73.0% [68.2–8.1%]
Specificity	34.7% [28.0–40.9%]	85.9% [83.0–88.8%]	85.6% [80.0–90.5%]	79.6% [73.2–88%]
Sensitivity	77.6% [72.6–82.7%]	73.6% [69.8–77.4%]	68.9% [61.5–75.9%]	66.7% [59.4–73.5%]
AUC	56.1% [51.8–60.3%]	89.0% [87.0–90.9%]	80.1% [75.0–84.6%]	81.0% [76.1–85.5%]

AUC, area under the receiver operating characteristic curve; TIRADS, Thyroid Imaging Reporting and Data System

Results

Patients characteristics of study cohorts

The present study enrolled 503 patients diagnosed as FA or FTC, of which 432 cases from center A were used as the main cohort and 71 from other multicenters as the external test set. The main cohort was separated into the training set ($N = 352$) and the validation set ($N = 80$) randomly. The clinical characteristics of patients are summarized in Table 1. In the main cohort, the mean age of FTC patients was significantly older than that of FA ($P < 0.01$) in both the training and validation sets, whereas gender showed no statistical correlation between FA and FTC patients. The mean size of the primary tumor in FA and FTC patients was 3.17 cm, 3.36 cm in the training set, and 3.06 and 3.31 cm in the validation sets. Furthermore, tumors greater than 4 cm also showed no statistical correlation between FA and FTC patients in both sets. Taken together with both sets of 212 FTC patients, postoperative histopathological results showed that 162 cases (76.4%) were diagnosed with minimally CI, while 50 cases (23.6%) with grossly capsular or vessel invasion, and 12 cases (5.7%) with distant metastasis at initial diagnosis. Preoperative ultrasound results using Kwak TI-RADS showed that no statistically significant differences were found between the staging classification of FA and FTC patients in either training or validation sets. Comparable patient characteristics were observed in the external test set similarly.

Predicting system and external testing

Our segmentation model achieved good accuracy. The training set had a total of 1392 images, and the mIOU reached 0.973. The test set had a total of 349 images, and the mIOU reached 0.917. This demonstrated that the optimization-based foreground network was effective for thyroid follicular tumor images.

We showed the classification results of TIRADS and the classification results of FThyNet on training, validation, and test sets in Figure 3 and Table 2. It was found that the AUC of TIRADS was relatively low, only 56.1%, and the specificity was very low at 34.7%, indicating that they were more inclined to classify tumors as malignant. The AUC of FThyNet was above 80% on the three sets, which was far better than the classification results of TIRADS. The AUC of the validation set is the same as the test set, which showed that the classification effect of FThyNet is stable. Among them, the FThyNet model showed higher classification accuracy for highly invasive tumors. In the training set and validation set, the recognition accuracy of the model for high-

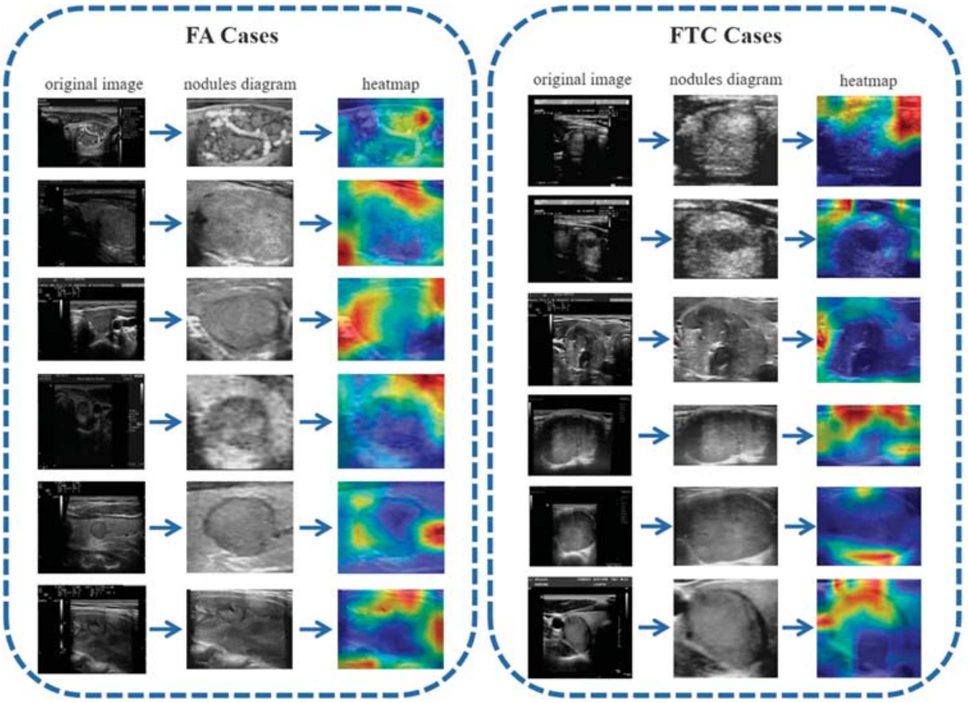

Figure 4. Characteristic heat maps. FA cases are shown on the left, and FTC cases are shown on the right. Each displayed image is divided into three columns, followed by the original image, nodules diagram, and the heat map. FA, follicular adenoma; FTC, follicular thyroid carcinoma.

Table 3

Classification results of nodule edge texture.

	Edge feature		
	Training set	Validation set	Test set
	(*n*=1392)	(*n*=349)	(*n*=309)
Accuracy	70.2% [67.4–73.3%]	69.1% [63.0–75.2%]	64.9% [58.8–71.0%]
Specificity	72.7% [69.0–76.2%]	73.4% [65.4–80.9%]	68.6% [59.8–77.3%]
Sensitivity	66.5% [61.7–71.1%]	63.7% [54.9–72.8%]	61.3% [53.3–69.5%]
AUC	71.4% [68.6–74.7%]	68.3% [61.5–75.5%]	66.2% [59.0–73.3%]

AUC, area under the receiver operating characteristic curve.

invasive tumors reached 88.2 and 90.3%, respectively. This suggests that the model has a stronger ability to discriminate highly invasive tumors.

Interpretability analysis

Characteristic heat maps of benign and malignant FNs were presented in Figure 4 by Grad-CAM. In the classification model, the network paid more attention to the features of nodules' edge regions. The benign nodular edge was clear, and the attention scope in the heat map focused across the whole edge, whereas the edge features of malignant nodules were relatively fuzzy, and FThyNet focused on only specific regions of the edge. The result was consistent with the surrounding tissue features extracted from the symbiotic relationship heat map and pathological characteristics between FA and FTC (SFigure 1, Supplemental Digital Content 2, http://links.lww.com/JS9/A507).

We generated symbiotic relationship heat maps between benign and malignant FNs and surrounding tissues from the system to investigate how FThyNet differentiated FTC from FA. The results showed that the texture features around benign nodules were regular and clear, whereas the texture features around malignant nodules were relatively irregular and unclear. Typical feature images and heat maps are shown in Figure 4. This is consistent with the diagnostic criteria of benign or malignant tumors in pathological knowledge.

Table 3 illustrates the impact of nodule margin features on the prediction results of follicular carcinoma. The results indicate that the accuracy rate of benign and malignant classification based on the edge features of nodules falls between 66 and 72%, emphasizing that edge features are an important evaluation factor for canceration, but other influencing factors exist. Furthermore, Figure 5 evaluates the impact of nodule margin features on the prediction of high-invasive and low-invasive tumors, and benign nodules. The findings indicate that 72% of highly invasive tumors fell within the 0.8–1 interval of the feature classification results, 48% of low-invasive tumors fell within the 0.5–0.8 interval, and 39% fell within the 0–0.5 interval. Moreover, 73% of benign tumors fell within the 0.5–0.8 interval, while nodules were distributed between 0 and 0.5, indicating that edge features can significantly distinguish highly aggressive tumors from benign nodules.

Discussion

In contrast to other types of DTC, FTC was reported to have a higher rate of distant metastatic disease (DM) and disease-specific

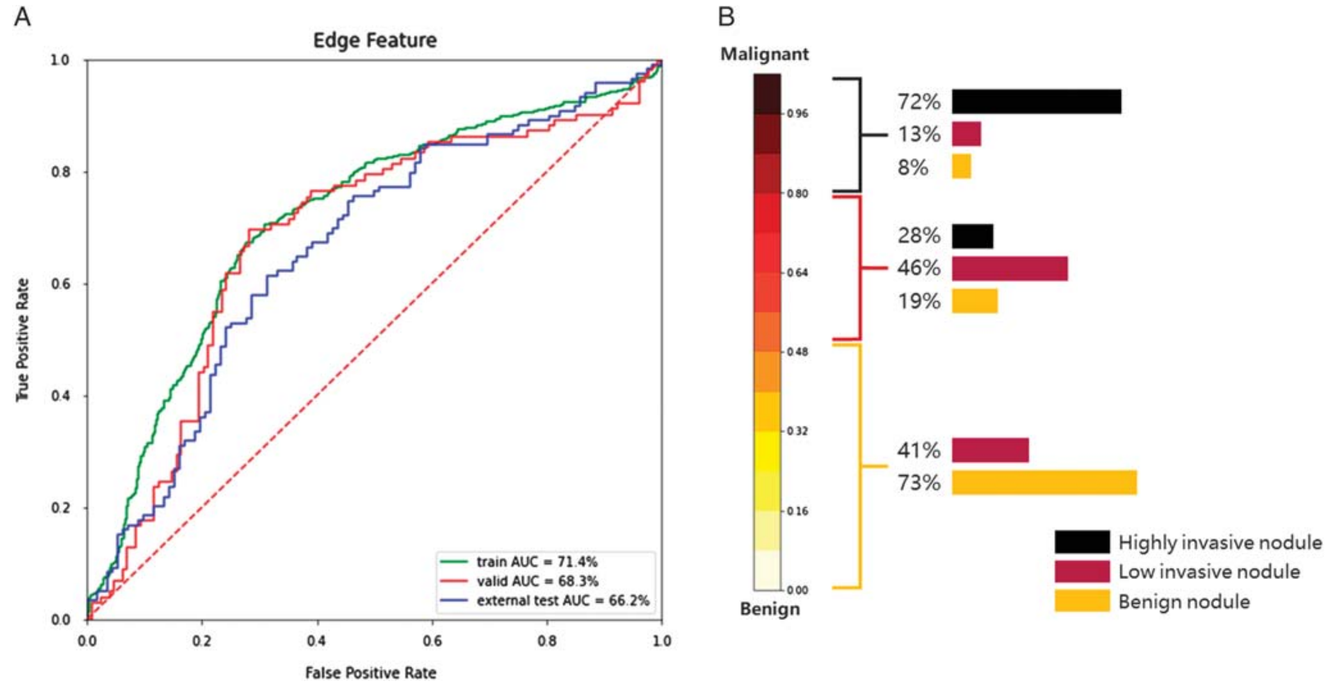

Figure 5. ROC curves and malignant tumor distribution. (A) ROC curves of edge texture results. (B) Distribution of highly invasive nodules, low-invasive nodules, and benign nodules in the edge texture classification results. AUC, area under the receiver operating characteristic curve; ROC, receiver operating characteristic

mortality (DSM), which occurs mainly in grossly invasive FTC[4,31]. In the present study, 15 (6.0%) patients were diagnosed with DM in 248 FTC patients, whereas the DM ratio was reported in 1–1.4% PTC patients in several studies based on the SEER database[4,32,33]. Despite the general acceptance of ultrasound's inefficiency in distinguishing between FTC and FA[4], we defined TD-4A as a cutoff to objectively quantify the prediction ability of preoperative ultrasound on FNs. The results showed that the AUC was only 56.1%, while the specificity was only 34.7%. Given the low detection rates of ultrasonography and cytology preoperatively, a large number of patients with FTC suspected underwent diagnostic surgeries considering the potential risks and poor prognosis. Statistics from the main center showed that more than 60% of suspected FNs were diagnosed with pathological benign postoperatively. These unnecessary diagnostic procedures not only increased the unknown surgical risk and potential surgical injury for each patient but also increased the social and medical burden on thyroid diseases. Thus, the key to overcoming the dilemma of FTC was to develop a trustworthy and efficient diagnostic mode preoperatively.

The present study established the FThyNet predicting system based on foreground optimization recognition to differentiate FTC from FA, overcoming the limitations of a small sample size, which was the first attempt of its kind for this topic. FThyNet provided outstanding predicting ability with all AUC over 0.8 in training, valid and external cohorts. This model was particularly good at grossly invasive FTC diagnosis, with a detection accuracy of over 90%. The diagnostic ability of FThyNet on grossly invasive FTC greatly improved its clinical application potential, which also indicated that the focus area recognized by deep learning was the key difference between FTC and FA. With its excellent generalization ability, FThyNet is expected to become an effective diagnostic tool to assist clinicians in decision-making with suspected FNs.

Given that few FTC displayed cytological characteristics of atypia of cellular/nuclear atypia or follicular lesion of undetermined significance, the distinction between FTC and FA required pathological examination after tumor excision[4,31]. The pathological diagnosis of FTC required evidence of capsular invasion (CI), vascular invasion (VI), or extrathyroidal extension (ETE)[4,34]. The subtle pathological distinctions between FTC and FA were hard to extract by preoperative means such as ultrasound or computed tomography (CT) merely through clinicians. The present study showed no significant difference between classifications of TIRADS in FTC and FA patients; hence, ultrasonography alone is not recommended to be used to precisely diagnose FTC by physicians. However, images provided by these medical examinations contained multidimensional information about nodular features that could be identified by AI, such as deep learning with the help of multilayer network architecture, thus bestowing these images with greater clinical significance[35].

Based on the mentioned theoretical feasibility, together with ultrasound's advantages of noninvasive, painless, real-time, and convenient acquisition, the diagnostic system modeling on ultrasound images may have more potential for clinical application. Due to the pathological features of FTC, that is, invasive tumor cells may also affect the biological characteristics of the surrounding tissues, image information of the tumor together with a small region of surrounding tissues are modeled. Subsequently, the FThyNet showed good interpretability by focusing on the nodules' edge regions, which was consistent with pathological features between FTC and FA.

Few previous studies were carried out on differential diagnosis of FTC and FA, most of which focused on improvements in diagnosis methods in cytology or pathology[6,7,36]. Our study exhibited a noninvasive, high-precision preoperative predicting system on follicular thyroid neoplasms for the first time. The present FThyNet showed good prediction efficiency as well as excellent generalization ability, with AUC in the training set, validation set, and test set of 89.0%, 80.1%, and 81.0%, respectively. Compared with the prediction efficiency, the excellent generalization ability of FThyNet validated by data from five independent clinical centers reflected more of its potential in clinical application. The key technical challenges were the fuzzy texture features of ultrasonic images and the serious recognition errors caused by the large intra-class variance of background information, which would be magnified in FTC and FA images. Thus, the prospect optimization algorithm was used to identify the target nodules, and features from nodules and surrounding tissue were extracted to generate the symbiotic relationship heat maps. Surprisingly, the heat maps reflected that FThyNet focused on only specific regions of the edge, and the texture features around malignant nodules were relatively distorted and unclear, which was consistent with the main histopathological focus. The mentioned results provided good interpretability for FThyNet, as well as a reason for the excellent prediction and generalization ability of the present system. Furthermore, the clinical interpretation provided by FThyNet could greatly improve the understanding of the disease of clinical experts, for the AI yielded a predictive ability increasing by more than 30% over human counterparts. This may also provide auxiliary help for subsequent prospective studies.

The FThyNet may also play a role in the active surveillance of follicular thyroid neoplasms. For FN patients with no surgical preference and low malignant risk, the FThyNet system will provide a clear direction in the course of active surveillance. At present, active surveillance was suggested as a viable option, mostly for papillary thyroid microcarcinomas (< 1 cm)[37,38]. The characteristics of lacking preoperative diagnostic means and a higher risk of distant metastasis in FTC were the biggest obstacles to proceed active surveillance on suspected FNs. The FThyNet provides more options for patients with suspected FNs than merely undergoing diagnostic surgeries, and manages the risk of potential malignancy in patients proceeding with active surveillance instead of diagnostic surgeries. Less is more is a safe rule in individualized therapy and precision medicine that is increasingly being relied on in thyroid surgery[38–40], in which AI will play an indispensable role. Thus, the FThyNet proposed for the first time has the great potential of revising the disease management model of FNs.

The present study has limitations. Firstly, the low incidence of FTC limits the sample size, which potentially affects the generalization and robustness of the prediction system. In the following studies, we will include a larger sample size of different regions and different populations to achieve better generalization. Secondly, the design of this retrospective study and only static images involved limits the clinical evidence quality; however, the establishment of FThyNet provides an important basis for future prospective clinical trials. Thirdly, whereas the controversy about the biological behavior of minimal invasive FTC, this study enrolls these patients as a malignancy cohort according to the ATA guidelines, which may affect the efficacy of the model. However, considering the clear malignant biological features of minimal invasive FTC, the high sensitivity at the expense of partial specificity ensures that no high-risk patients will be missed. Fourthly, as for the common problem of overconfidence in AI systems, we will solve it by calculating the confidence interval of each case through algorithm implementation in the following research.

Ethical approval

This research was supported by Shanghai Jiao Tong University (YG2019ZDA15), the National Natural Science Foundation of China (NSFC, 81772558, 82072948). This work was also supported by Neil Shen's SJTU Medical Research Fund, Clinical Research Plan of SHDC (SHDC2020CR6028), SJTU-Yale Collaborative Research Seed Fund, NSFC81903417.

Sources of funding

This research was supported by Shanghai Jiao Tong University (YG2019ZDA15) and the National Natural Science Foundation of China (NSFC, 81772558, 82072948). This work was also supported by Neil Shen's SJTU Medical Research Fund, Clinical Research Plan of SHDC (SHDC2020CR6028), SJTU-Yale Collaborative Research Seed Fund, NSFC81903417.

Author contribution

Z.Y., S.Y., W.C., and H.L.: study conception and design; Z.Y., Y.H., T.L. and S.F.: material preparation and data curation. Formal analysis and methodology were performed by S.Y., P.S., W.Z., and H.L. W.C., W.Q., and H.L. provided funding acquisition. The first draft of the manuscript was written by Z.Y., S.Y., P.S. and Y.H., W.C., W.Q., H.L., W.Z., and L.T. edited and revised the article, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of interest disclosure

The authors report no conflicts of interest. The authors are responsible for the content and writing of the paper.

Research registration unique identifying number (UIN)

1. Name of the registry: Chinese Clinical Trial (http://www.chictr.org.cn/).
2. Unique identifying number or registration ID: ChiCTR2200060823.
3. Hyperlink to your specific registration (must be publicly accessible and will be checked): https://www.chictr.org.cn/showproj.html?proj=164286.

Guarantors

Wei Cai and Zheyu Yang.

Data availability statement

Data are available on request due to privacy or other restrictions. The data that support the findings of this study are available on request from the corresponding author Wei Cai. The data are not publicly available due to them containing information that could compromise research participant privacy.

Provenance and peer review

Not commissioned, externally peer-reviewed.

Acknowledgements

This research was supported by Shanghai Jiao Tong University (YG2019ZDA15) and National Natural Science Foundation of China (NSFC, 81772558, 82072948). This work was also supported by Neil Shen's SJTU Medical Research Fund, Clinical Research Plan of SHDC (SHDC2020CR6028) and SJTU-Yale Collaborative Research Seed Fund, NSFC81903417.

References

[1] Carty SE, Ohori NP, Hilko DA, *et al*. The clinical utility of molecular testing in the management of thyroid follicular neoplasms (Bethesda IV nodules). Ann Surg 2020;272:621–7.

[2] Burman KD, Wartofsky L. Clinical practice. Thyroid nodules. N Engl J Med 2015;373:2347–56.

[3] Cipriani NA, Nagar S, Kaplan SP, *et al*. Follicular thyroid carcinoma: how have histologic diagnoses changed in the last half-century and what are the prognostic implications? Thyroid 2015;25:1209–16.

[4] Daniels GH. Follicular thyroid carcinoma: a perspective. Thyroid 2018; 28:1229–42.

[5] Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. Thyroid 2017;27:1341–6.

[6] Sun Y, Li L, Zhou Y, *et al*. Stratification of follicular thyroid tumours using data-independent acquisition proteomics and a comprehensive thyroid tissue spectral library. Mol Oncol 2022;16:1611–24.

[7] Saburi S, Tsujikawa T, Miyagawa-Hayashino A, *et al*. Spatially resolved immune microenvironmental profiling for follicular thyroid carcinoma with minimal capsular invasion. Mod Pathol 2022;35:721–7.

[8] Simões-Pereira J, Mourinho N, Ferreira TC, *et al*. Avidity and outcomes of radioiodine therapy for distant metastasis of distinct types of differentiated thyroid cancer. J Clin Endocrinol Metab 2021;106:e3911–22.

[9] Conzo G, Avenia N, Ansaldo GL, *et al*. Surgical treatment of thyroid follicular neoplasms: results of a retrospective analysis of a large clinical series. Endocrine 2017;55:530–8.

[10] Hu MI, Waguespack SG, Dosiou C, *et al*. Afirma Genomic Sequencing Classifier and Xpression Atlas Molecular Findings in Consecutive Bethesda III–VI Thyroid Nodules. J Clin Endocrinol Metab 2021;106: 2198–207.

[11] Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.

[12] DeJohn CR, Grant SR, Seshadri M. Application of machine learning methods to improve the performance of ultrasound in head and neck oncology: a literature review. Cancers (Basel) 2022;14:665.

[13] 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:e250–9.

[14] Saxe A, Nelli S, Summerfield C. If deep learning is the answer, what is the question? Nat Rev Neurosci 2021;22:55–67.

[15] Potnis KC, Ross JS, Aneja S, *et al*. Artificial intelligence in breast cancer screening: evaluation of FDA device regulation and future recommendations. JAMA Intern Med 2022;182:1306–12.

[16] Wang L, Zhang L, Zhu M, *et al*. Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. Med Image Anal 2020;61: 101665.

[17] Buda M, Wildman-Tobriner B, Hoang JK, *et al.* Management of thyroid nodules seen on US images: deep learning may match performance of radiologists. Radiology 2019;292:695–701.

[18] Mathew G, Agha R. for the STROCSS Group. STROCSS 2021: strengthening the reporting of cohort, cross-sectional and case–control studies in surgery. Int J Surg 2021;96:106165.

[19] Yoon JH, Han K, Kim EK, *et al.* Diagnosis and management of small thyroid nodules: a comparative study with six guidelines for thyroid nodules. Radiology 2017;283:560–9.

[20] 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;26:1–133.

[21] WHO. WHO Classification of Tumours of Endocrine Organs, 4th ed. World Health Organization; 2017.

[22] Grani G, Lamartina L, Durante C, *et al.* Follicular thyroid cancer and Hürthle cell carcinoma: challenges in diagnosis, treatment, and clinical management. Lancet Diabetes Endocrinol 2018;6:500–14.

[23] Tuttle RM, Haugen B, Perrier ND. Updated American Joint Committee on Cancer/Tumor-Node-Metastasis Staging System for Differentiated and Anaplastic Thyroid Cancer (Eighth Edition): What Changed and Why? Thyroid 2017;27:751–6.

[24] Lin G, Liu F, Milan A, *et al.* RefineNet: Multi-Path Refinement Networks for Dense Prediction. IEEE Trans Pattern Anal Mach Intell 2020;42:1228–42.

[25] Chollet F. Xception: deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. pp. 1251–8.

[26] Zhang L, Wang X, Yang D, *et al.* Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans Med Imaging 2020;39:2531–40.

[27] Jinlian M, Dexing K. Deep learning models for segmentation of lesion based on ultrasound images. AUDT 2018;2:83–93.

[28] Zheng Z, Zhong Y, Wang J, *et al.* Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. pp. 4096–105.

[29] Jang WD, Lee C, Kim CS. Primary object segmentation in videos via alternate convex optimization of foreground and background distributions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. pp. 696–704.

[30] Szegedy C, Ioffe S, Vanhoucke V, , *et al.* Inception-v4, Inception-ResNet and the impact of residual connections on learning. Thirty-first AAAI Conference on Artificial Intelligence, 2017.

[31] Sugino K, Ito K, Nagahama M, *et al.* Prognosis and prognostic factors for distant metastases and tumor mortality in follicular thyroid carcinoma. Thyroid 2011;21:751–7.

[32] Zaydfudim V, Feurer ID, Griffin MR, *et al.* The impact of lymph node involvement on survival in patients with papillary and follicular thyroid carcinoma. Surgery 2008;144:1070–8.

[33] Goffredo P, Cheung K, Roman SA, *et al.* Can minimally invasive follicular thyroid cancer be approached as a benign lesion?: a population-level analysis of survival among 1,200 patients. Ann Surg Oncol 2013;20:767–72.

[34] Asa SL. The evolution of differentiated thyroid cancer. Pathology 2017; 49:229–37.

[35] Harvey H, Topol EJ. More than meets the AI: refining image acquisition and resolution. Lancet 2020;396:1479.

[36] Macerola E, Poma AM, Proietti A, *et al.* Digital gene expression analysis on cytology smears can rule out malignancy in follicular-patterned thyroid tumors. J Mol Diagn 2020;22:179–87.

[37] Chou R, Dana T, Haymart M, *et al.* Active surveillance versus thyroid surgery for differentiated thyroid cancer: a systematic review. Thyroid 2022;32:351–67.

[38] Molinaro E, Campopiano MC, Elisei R. Management of Endocrine Disease: papillary thyroid microcarcinoma: toward an active surveillance strategy. Eur J Endocrinol 2021;185:R23–34.

[39] Wang TS, Sosa JA. Thyroid surgery for differentiated thyroid cancer – recent advances and future directions. Nat Rev Endocrinol 2018;14: 670–83.

[40] Kim BW, Yousman W, Wong WX, *et al.* Less is more: comparing the 2015 and 2009 American Thyroid Association guidelines for thyroid nodules and cancer. Thyroid 2016;26:759–64.