Ultrasound image-based deep learning to assist in diagnosing gross extrathyroidal extension thyroid cancer: a retrospective multicenter study

Qi Qi,[a,i] Xingzhi Huang,[a,i] Yan Zhang,[b] Shuangting Cai,[c] Zhaoyou Liu,[d] Taorong Qiu,[e] Zihan Cui,[e] Aiyun Zhou,[a] Xinchun Yuan,[a] Wan Zhu,[a] Xiang Min,[f] Yue Wu,[a] Weijia Wang,[g] Chunquan Zhang,[h,**] and Pan Xu[a,*]

[a]Department of Ultrasound, The First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China
[b]Department of Ultrasound, Huashan Hospital, Fudan University, Shanghai, Shanghai, China
[c]Ultrasound Department, Jiujiang First People's Hospital, Jiujiang, Jiangxi, China
[d]Ultrasound Department, Affiliated Hospital of Jiangxi University of Chinese Medicine, Nanchang, Jiangxi, China
[e]School of Information Engineering, Nanchang University, Nanchang, Jiangxi, China
[f]Department of Otolaryngology, Head and Neck Surgery, The First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China
[g]Department of Pathology, The First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China
[h]Department of Ultrasound, The Second Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China

eClinicalMedicine
2023;58: 101905

Published Online xxx
https://doi.org/10.1016/j.eclinm.2023.101905

Summary

Background The presence of gross extrathyroidal extension (ETE) in thyroid cancer will affect the prognosis of patients, but imaging examination cannot provide a reliable diagnosis for it. This study was conducted to develop a deep learning (DL) model for localization and evaluation of thyroid cancer nodules in ultrasound images before surgery for the presence of gross ETE.

Methods From January 2016 to December 2021 grayscale ultrasound images of 806 thyroid cancer nodules (4451 images) from 4 medical centers were retrospectively analyzed, including 517 no gross ETE nodules and 289 gross ETE nodules. 283 no gross ETE nodules and 158 gross ETE nodules were randomly selected from the internal dataset to form a training set and validation set (2914 images), and a multitask DL model was constructed for diagnosing gross ETE. In addition, the clinical model and the clinical and DL combined model were constructed. In the internal test set [974 images (139 no gross ETE nodules and 83 gross ETE nodules)] and the external test set [563 images (95 no gross ETE nodules and 48 gross ETE nodules)], the diagnostic performance of DL model was verified based on the pathological results. And then, compared the results with the diagnosis by 2 senior and 2 junior radiologists.

Findings In the internal test set, DL model demonstrated the highest AUC (0.91; 95% CI: 0.87, 0.96), which was significantly higher than that of two senior radiologists [(AUC, 0.78; 95% CI: 0.71, 0.85; $P < 0.001$) and (AUC, 0.76; 95% CI: 0.70, 0.83; $P < 0.001$)] and two juniors radiologists [(AUC, 0.65; 95% CI: 0.58, 0.73; $P < 0.001$) and (AUC, 0.69; 95% CI: 0.62, 0.77; $P < 0.001$)]. DL model was significantly higher than clinical model [(AUC, 0.84; 95% CI: 0.79, 0.89; $P = 0.019$)], but there was no significant difference between DL model and clinical and DL combined model [(AUC, 0.94; 95% CI: 0.91, 0.97; $P = 0.143$)]. In the external test set, DL model also demonstrated the highest AUC (0.88, 95% CI: 0.81, 0.94), which was significantly higher than that of one of senior radiologists [(AUC, 0.75; 95% CI: 0.66, 0.84; $P = 0.008$) and (AUC, 0.81; 95% CI: 0.72, 0.89; $P = 0.152$)] and two junior radiologists [(AUC, 0.72; 95% CI: 0.62, 0.81; $P = 0.002$) and (AUC, 0.67; 95 CI: 0.57, 0.77; $P < 0.001$]. There was no significant difference between DL model and clinical model [(AUC, 0.85; 95% CI: 0.79, 0.91; $P = 0.516$)] and clinical + DL model [(AUC, 0.92; 95% CI: 0.87, 0.96; $P = 0.093$)]. Using DL model, the diagnostic ability of two junior radiologists was significantly improved.

Interpretation The DL model based on ultrasound imaging is a simple and helpful tool for preoperative diagnosis of gross ETE thyroid cancer, and its diagnostic performance is equivalent to or even better than that of senior radiologists.

Abbreviations: DL, deep learning; ETE, extrathyroidal extension; ROC, receiver operating characteristic; AUC, area under the receiver operating characteristic curve; RLN, recurrent laryngeal nerve; US, ultrasound; ResNet, residual network; FPN, feature pyramid network; RPN, region proposal network; ROI, region of interest; Grad-CAM, gradient-weighted class activation mapping; PPV, positive predictive value; NPV, negative predictive value

*Corresponding author.
**Corresponding author.
E-mail addresses: xupan_1989@126.com (P. Xu), rylcq107@126.com (C. Zhang).

[i]These authors contributed equally to this work.

Funding Jiangxi Provincial Natural Science Foundation (20224BAB216079), the Key Research and Development Program of Jiangxi Province (20181BBG70031), and the Interdisciplinary Innovation Fund of Natural Science, Nanchang University (9167-28220007-YB2110).

Keywords: Deep learning; Gross extrathyroidal extension thyroid cancer; Ultrasound

Research in context

Evidence before this study

We searched PubMed and Web of Science with the terms "(extrathyroidal extension thyroid cancer OR gross extrathyroidal extension thyroid cancer) AND (radiomics OR deep learning)" for papers published from database inception to Nov 9, 2022, with no language restrictions. We find that there is no research based on deep learning. Among the 6 studies based on radiology, they were mainly based on MRI or CT images, and only 1 study was based on ultrasound (US) images. The AUC of the model diagnosis of ETE was 0.824. And all articles were single center research.

Added value of this study

As far as we know, no research has tested the feasibility of diagnosing gross extrathyroidal extension (ETE) thyroid cancer based on ultrasonic image deep learning. Our multitask deep learning model can automatically segment and diagnose whether there is gross ETE in the US image. We tested it in internal test set and multi-center set and compared it with senior radiologists and junior radiologists.

Implications of all the available evidence

Our research shows that our multitask deep learning model has excellent ability to diagnose gross ETE, and its diagnostic performance is equivalent to or even better than that of senior radiologists. It provides potential tools to guide individualized treatment strategies.

Introduction

The American Joint Committee on Cancer comprehensively revised thyroid cancer staging in its eighth edition.[1] One of the changes in the T-stage corresponds to the range of extrathyroidal extension (ETE). ETE refers to the invasion of primary tumors into adjacent tissues other than the gland, which contributes to the poor prognosis of patients with thyroid cancer.[2] However, in the eighth edition of staging, minimal ETE was removed from T3 because there was little evidence that it was an independent predictor of persistence, recurrence, or impact on survival in thyroid cancer.[3] The new T3b is defined as a tumor of any size demonstrating the gross ETE invading only the strap muscle. The gross ETE invading major tissue other than the thyroid is classified as T4a, such as invading recurrent laryngeal nerve (RLN), larynx, trachea, esophagus, or subcutaneous soft tissue. The gross ETE invading prevertebral fascia or encasing carotid artery or mediastinal vessels are classified as T4b. This change is hopeful to reduce the overdiagnosis and treatment of low-risk thyroid cancer. Many countries and regions suggested that the observation waiting method should be considered as the preferred choice for patients with low risk thyroid cancer.[4] It was discovered that patients with gross ETE are more likely to have lymph node metastases, distant metastases, higher rates of tumor recurrence, and the worst overall survival.[5] Therefore, gross ETE thyroid cancer is an indication for surgical operation. In contrast, unilateral lobectomy or even dynamic surveillance is an option for low or intermediate-risk thyroid cancer. Accurate diagnosis of gross ETE can reduce the over-treatment rate of thyroid cancer. In that case, the doctors may be able to decide on an effective treatment plan, which also helps in improving the prognosis of patient.

Ultrasound (US) is the most convenient and accurate modality for the preoperative evaluation of superficial organs such as the thyroid gland. In previous studies,[6–9] corresponding US features have been proposed for the invasion of strap muscle, trachea, and RLN, which provide a reference for the preoperative diagnosis of gross ETE. These features had good specificity but low sensitivity (45%–78%). Additionally, the low incidence of gross ETE resulted in percentages of gross ETE in prior studies that were less than 10%, and the diagnostic accuracy of aforementioned indicators requires further validation. In conclusion, there are no ideal US diagnostic criteria for the preoperative diagnosis of gross ETE.

In recent years, various convolutional neural networks based on deep learning (DL) algorithms have been developed for medical image analysis.[10–12] Among the DL networks used in previous studies, there are two major types. The first type directly classifies the whole picture. Commonly used networks include VGGNet, Google Inception Net, ResNet, etc.[13] Some DL models for thyroid US images have used the above networks, and most of their functions focus on the differentiation of benign and malignant thyroid nodules.[14–16] The second type of network can automatically detect and classify objects, and it can extract information from the local part of the image explicitly. The ideal DL model we expect should be able to automatically identify and diagnose nodules in the US interface, and since nodules always have different shapes, we use the target detection model for experiments in the hope of obtaining more accurate and visual diagnosis. Therefore, in this study, we developed a US DL model with automatic localization and identification of gross ETE based on an internal dataset of 663 thyroid cancer nodules and external datasets of 143 thyroid cancer nodules.

Methods

Study population

This study was a retrospective multicenter study using US image sets obtained from four hospitals. Patients that met the inclusion criteria were selected: 1) patients who underwent total or hemisection of the thyroid gland for thyroid cancer; 2) patients who underwent preoperative US evaluation of the thyroid gland. The exclusion criteria include 1) patients who underwent surgery for recurrent thyroid cancer; 2) patients with incomplete operative or pathology reports; 3) patients with low-quality US images (e.g., severe artifacts or low image resolution); 4) patients with evidence of distant metastases. This study followed the Standards for Reporting of Diagnostic Accuracy (STARD) guidelines for diagnostic studies.

We retrospectively reviewed gross ETE patients from January 2016 to December 2021 and no gross ETE patients from January 2018 to December 2019 from the First Affiliated Hospital of Nanchang University as internal datasets (training set, validation set and internal test set) to account for the low prevalence of gross ETE and to reduce the issue of an unbalanced number of categories during training. For the external test set, we retrospectively reviewed gross ETE patients from September 2020 to March 2022 and no gross ETE patients from April 2021 to September 2021 at three hospitals (the Second Affiliated Hospital of Nanchang University, the First People's Hospital of Jiujiang City, and the Affiliated Hospital of Jiangxi University of Traditional Chinese Medicine).

The initial population included 948 patients. We excluded 47 patients who underwent surgery due to recurrence of thyroid cancer, 19 patients with incomplete surgical or pathological reports, and 32 patients with evidence of distant metastasis. After that, a radiologist (XuP) with 9 years of ultrasound experience initially screened the ultrasound images, and a total of 220 nodules in 186 patients were excluded. Finally, 806 nodules from 664 patients were included in this study. In the internal data set, 441 thyroid cancer nodules were randomly selected as the training set and validation set, and the remaining 222 nodules formed the internal test set. There were 143 nodules in the external test set. The details of the screening process of each research queue are shown in Fig. 1.

Ethics statement

The study was approved by the ethics committee of The First Affiliated Hospital of Nanchang University (No. 202112014), the study title is "Study on the diagnosis of benign and malignant thyroid nodules by computer-assisted ultrasound image based on deep learning". An informed consent exemption from the Institutional Review Board was obtained. The principles of the Helsinki Declaration guided the treatment protocols.

Collection of clinical, ultrasound, and pathological data

The medical record system was used to record the clinical and pathological data of patients, including their gender, age, whether they had multiple thyroid cancers, their lymph node status, and the type of ETE. T-staging was performed according to the extent of ETE and the size of thyroid nodules. ETE was divided into gross ETE and no gross ETE (included no ETE and minimal ETE). Gross ETE was defined as gross tumor invasion identified at the time of surgery and confirmed by histopathologic review.[17] Retrospective retrieval of US images from four medical center databases was performed. The thyroid US images were acquired by eight different devices (Appendix Table S1). The US features of all nodules were assessed by two radiologists with 20 years (Yuan X) and 13 years (Liu Z) of clinical experience in thyroid US, respectively. The results served as the clinical and US features for comparing the two types of nodules. In cases of diagnostic disagreement, the two radiologists reached consensus through discussion (The agreement between the two radiologists' assessments was assessed by kappa values, which were 0.859 for contact ratio between nodule and capsule, 0.812 for echo between nodule and capsule, and 0.932 for angle between tumor and Trachea, 0.899 for contact of the nodule with the TEG. This indicated a high agreement between the evaluators.). In addition, none of the radiologists was aware of the clinical history of patient, preoperative US reports, operative records, or pathology results. Table 1, Appendix S1, and Appendix Figure S1 provide more information on the examined ultrasonic features.

Image pre-processing

All thyroid US images were converted to JPEG format. The images in this study had a range of sizes because different US devices were used to gather the images. So we first uniformly adjusted all US images to 1024 × 1024 pixels by clipping or black bar filling. To control the quality of US images, a radiologist (Wu Y) with 3 years of US experience screened all images to remove low-

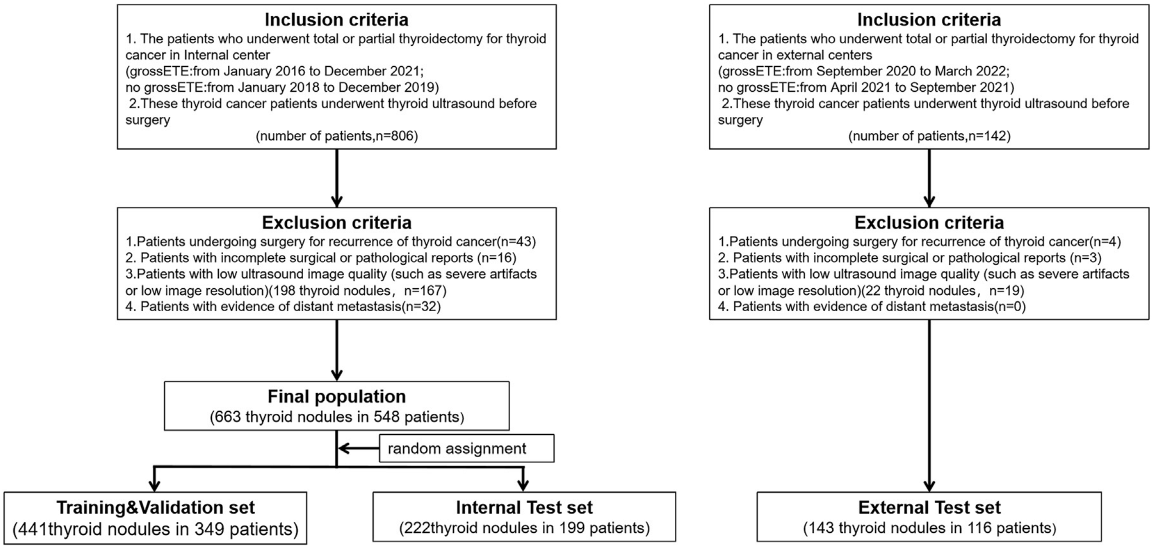

Inclusion criteria
1. The patients who underwent total or partial thyroidectomy for thyroid cancer in Internal center
(grossETE:from January 2016 to December 2021;
no grossETE:from January 2018 to December 2019)
2.These thyroid cancer patients underwent thyroid ultrasound before surgery
(number of patients,n=806)
Inclusion criteria
1. The patients who underwent total or partial thyroidectomy for thyroid cancer in external centers
(grossETE:from September 2020 to March 2022;
no grossETE:from April 2021 to September 2021)
2.These thyroid cancer patients underwent thyroid ultrasound before surgery
(number of patients,n=142)
Exclusion criteria
1.Patients undergoing surgery for recurrence of thyroid cancer(n=43)
2. Patients with incomplete surgical or pathological reports (n=16)
3.Patients with low ultrasound image quality (such as severe artifacts or low image resolution)(198 thyroid nodules，n=167)
4. Patients with evidence of distant metastasis(n=32)
Exclusion criteria
1.Patients undergoing surgery for recurrence of thyroid cancer(n=4)
2. Patients with incomplete surgical or pathological reports (n=3)
3.Patients with low ultrasound image quality (such as severe artifacts or low image resolution)(22 thyroid nodules，n=19)
4. Patients with evidence of distant metastasis(n=0)
Final population
(663 thyroid nodules in 548 patients)
random assignment
Training&Validation set
(441thyroid nodules in 349 patients)
Internal Test set
(222thyroid nodules in 199 patients)
External Test set
(143 thyroid nodules in 116 patients)

Fig. 1: **Inclusion criteria flowchart for the initial population and exclusion.** ETE, extrathyroidal extension.

quality images with severe artifacts or low image resolution. This process excluded 238 images (156 in training set and validation set, 57 in internal test set and 25 in external test set).

Next, the open-source common label tool LabelMe (*http://labelme.csail.mit.edu/Release3.0*) was used to outline all nodules. According to the pathological results, gross ETE nodules were labeled as "advanced", no gross ETE nodules were labeled as "inside", and JSON format files for model training were generated. The work of labeling and sketching ROI in the above pictures was completed by a radiologist (Qi Q) with 6 years of experience in US, which was then reviewed and modified by a radiologist (Xu P) with 9 years of US experience (Fig. 2). [Before formally sketching ROI, we used 200 images (100 no gross ETE, 100 gross ETE) to evaluate the consistency of ROI sketched by two radiologists. The mean dice score between the two radiologists was 0.919 (the mean dice score of no gross ETE was 0.912, and the mean dice score of gross ETE was 0.926).] We further reviewed the image quality during the sketching process. This process excluded 119 images (80 in training set and validation set, 28 in internal test set, and 11 in external test set).

Finally, a total of 4451 images meet the criteria. 2914 images (441 nodules) in the training and validation set that met the criteria were selected. 585 images of 89 nodules (20%) were randomly selected as the validation set, and the rest 2329 images of 352 nodules (80%) were used as the training set. In the internal test set, 974 images of 222 nodules that met the criteria were selected. Additionally, 563 images of 143 nodules that met the criteria were selected in the external test set.

Model architecture

In this study, we used the Mask-RCNN[18] network model based on deep learning to train US images. This model integrates the semantic segmentation task and target detection task. It can locate thyroid nodules from US images, draw the shape of nodules and diagnose them. We first combined the residual network (ResNet) and the feature pyramid network (FPN) as the backbone network to extract features from the input image. Then, the region proposal network (RPN)[19] was used to perform two classifications (foreground or background) and bounding box (BB) regression on the extracted features to generate the region of interest (ROI). Then, the ROI was sent to the ROIAlign layer to reset the feature map size to 7 × 7 pixels and 14 × 14 pixels. Finally, a suggestion frame with classification and probability and a mask for segmenting nodules were generated (Fig. 3). The detailed network structure is displayed in the Appendix Methods. The DL model code is available at *https://github.com/mubik77/DL-Model-to-Assist-in-DiagnosingGrossETE*.

Model training

We tried different combinations of ResNet with FPN as a backbone for training and finally selected ResNet50 + FPN. To reduce overfitting and increase the number and diversity of training sets, we performed image augmentation by random flipping, cropping (cropping 1/5 of image width and height), and adding noise (Gaussian noise processing). The network was fine-tuned to obtain better results and convergence using weights learned in advance on the coco dataset. The training was done in 70 epochs using the Adam optimizer (initialized learning rate was set to 0.0001). The

Characteristic	Levels	Training and validation set	Internal test set	External test set
Patient		349	199	116
Age, mean ± SD		44 ± 12	46 ± 12	45 ± 12
Gender (%)	Female	275 (79)	149 (75)	80 (69)
	Male	74 (21)	50 (25)	36 (31)
Multifocality (%)	No	217 (62)	129 (65)	76 (66)
	Yes	132 (38)	70 (35)	40 (34)
Bilaterality (%)	No	258 (74)	141 (71)	91 (78)
	Yes	91 (27)	58 (29)	25 (22)
Lymph node status (%)	N0/Nx	190 (54)	101 (51)	62 (53)
	N1	159 (46)	98 (49)	54 (47)
Thyroid nodule		441	222	143
Tumor size, mean ± SD, cm		1.39 ± 0.97	1.46 ± 0.92	1.44 ± 0.79
ETE (%)	Absent	248 (56)	121 (55)	71 (50)
	minimal	35 (8)	18 (8)	24 (17)
	Gross	158 (36)	83 (37)	48 (34)
T status (%)	T1	244 (55)	120 (54)	87 (61)
	T2	33 (7)	18 (8)	8 (6)
	T3a	6 (1)	1 (0)	0 (0.00)
	T3b	102 (23)	50 (23)	37 (26)
	T4	56 (13)	33 (15)	11 (8)
Contact of the nodule with the thyroid capsule (%)	No contact	104 (24)	37 (17)	24 (17)
	<25%	31 (7)	17 (8)	19 (13)
	≥25 to <50%	145 (33)	82 (37)	64 (45)
	≥50%	162 (37)	86 (39)	36 (25)
	Capsular disruption	168 (38)	100 (45)	29 (20)
	Contour bulging	76 (17)	50 (23)	22 (15)
	Replacement of strap muscle	76 (17)	50 (23)	29 (20)
Angle between tumor and trachea (%)	No contact	306 (69)	144 (65)	93 (65)
	Acute angle	89 (20)	51 (23)	33 (23)
	Right angle	37 (8)	23 (10)	14 (10)
	Obtuse angle	9 (2)	4 (2)	3 (2)
Contact of the nodule with the TEG (%)	No contact	349 (79)	169 (76)	115 (80)
	Abutting TEG	74 (17)	32 (14)	23 (16)
	Protrusion into TEG	18 (4)	21 (9)	5 (4)

Abbreviations: ETE, Extra thyroidal extension; TEG, tracheoesophageal groove. Note: Contour bulging, Nodules with contour bulging with or without capsular disruption.

***Table 1:* Patient demographics and US features in the training and testing sets.**

specific training and filtering process of DL model is in Appendix S4, Appendix Table S3 and Appendix Figures S3 and S4.

Comparison of diagnostic performance between model and radiologists

The predicted value for a nodule in the DL model was calculated as the average of its predicted probabilities for all images of that nodule. The diagnostic performance of DL model was compared with that of the radiologists, which included two senior radiologists [with professional experience of 25 years (Zhang C) and 20 years (Zhu W), respectively] and two junior radiologists [(with professional experience of six years (Zhang Y) and five years (Huang X), respectively)]. The clinical history of patient, preoperative US examination report, surgical records, and pathological results were unannounced to these radiologists. The radiologists were tasked to diagnose nodules as gross ETE or no gross ETE according to ultrasonic features and diagnostic experience. The diagnosis of internal and external test sets was carried out for two days. The diagnostic performance between the DL model and radiologists was compared by the area under the receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio, negative likelihood ratio, and diagnostic odds ratio (Calculation formulas and definitions are presented in Appendix S5.).

Construction of clinical model, clinical and DL combined model, and compare them with DL model

Based on the available information of clinical features, univariate logistic regression was used to screen out statistically significant features in the training and

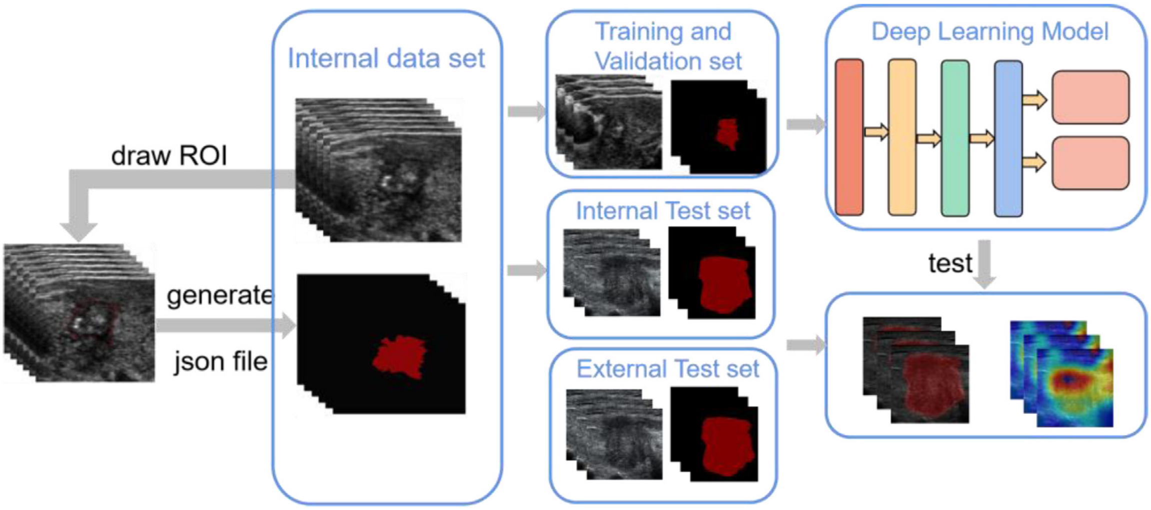


Fig. 2: **Image processing and distribution.** ROI, region of interest.

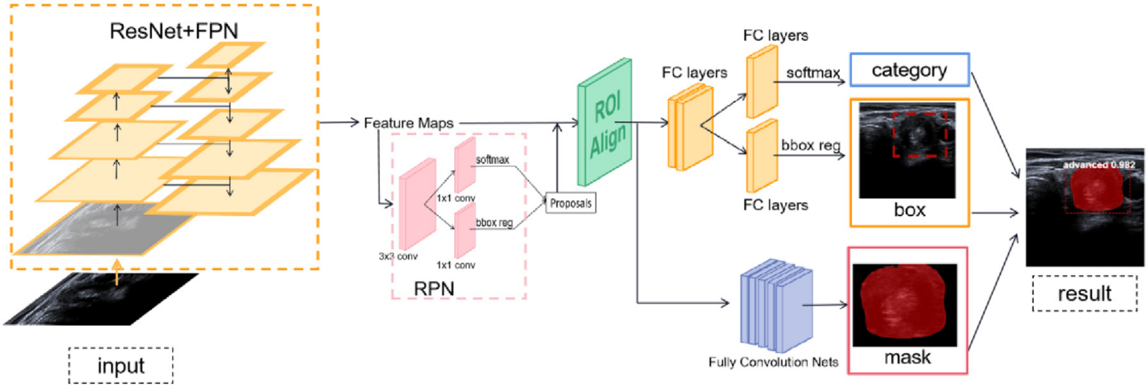


Fig. 3: **Illustration shows the deep learning (DL) neural network architecture.** The data are fed into the backbone network consisting of ResNet and FPN for feature extraction, and the extracted feature maps are fed into the regional proposal network (RPN) to generate regions of interest (ROI). Then all the ROIs are reset by RoIAlign network, and finally the ROI are classified and regressed by the classification branch, and the mask of the object is generated by the mask branch.

validation set. Multivariable logistic regression analysis was used to build a clinical model to diagnose the presence of gross ETE in thyroid cancer. After that, the clinical model and DL model were used to generate a combined model by multivariable logistic regression analysis. The internal test set and external test set were used to compare the diagnostic performance of the clinical model, the DL model, and the clinical + DL combined model.

Visualization and auxiliary diagnosis function of DL model

To make the model useful for assisting radiologists in diagnosis, the diagnostic results of model were visualized. Our DL model can locate the nodule automatically in the US image and depict the extent of nodule with a mask, allowing us to observe the location and shape of nodule. To further interpret the DL model in a human-readable form, we used the Gradient-weighted Class Activation Mapping (Grad-CAM) technique to clarify the focus of model.

Two weeks after the initial diagnosis, four radiologists used the DL model to make another diagnosis on the internal and external test sets while remaining blind to the clinical and pathological information of patients to examine the potential benefits of DL model for treating medical conditions.

Statistical analysis

In this study, the Shapiro–Wilk test was used to assess the normality of data distribution. Measurements conforming to normal distribution were expressed as mean ± standard deviation for comparison by independent samples t-test, and those unconforming were expressed as median (quartiles) for comparison by Mann–Whitney U test. Count data were expressed as frequency (frequencies) for comparison by chi-square test or Fisher exact test. AUC comparisons were carried out via the Delong test. In contrast, the McNemar test was used to compare the differences in sensitivity, specificity, and accuracy between the diagnosis by the DL model and the radiologists. Statistical significance was set at $P < 0.05$. All analyses were performed using R statistical software (version 3.6.3) and IBM SPSS Statistics (version 26). R software packages used in this study included pROC, epiR, STAT.

Role of the funding source

The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. Q. Qi, X. Huang, P. Xu, C. Zhang have accessed the data, and had final responsibility for the decision to submit it for publication.

Results

Demography of the participants

The internal data set was composed of 663 thyroid nodules in 548 patients, of which 422 nodules in 335 patients (268 females and 67 males) were categorized as the no gross ETE group. The average age of patients in this group was 43 ± 11 years, and the average size of nodules was 1.17 ± 0.84 cm. The remaining 241 nodules in 213 patients (156 females and 57 males) were categorized as the gross ETE group. The average age of patients in this group was 48 ± 13 years, and the average size of nodules was 1.87 ± 0.97 cm. Additionally, the external data set was composed of 143 thyroid nodules in 116 patients, of which 95 nodules in 76 patients (50 females and 26 males) were categorized as the no gross ETE group. The average age of patients in this group was 43 ± 14 years, and the average size of nodules was 1.15 ± 0.57. The remaining 48 nodules in 40 patients (30 females and 10 males) were categorized as the gross ETE group. The average age of patients in this group was 49 ± 9 years, and the average size of nodules was 2.03 ± 0.86 cm (Appendix Table S2). According to statistics, the age of patients, whether there were multiple or bilateral thyroid cancer nodules, the status of lymph nodes and the size of nodules were significantly different between the no gross ETE group and the gross ETE group (all $P < 0.05$). The US features used in this investigation, such as the contact between the nodule and the thyroid capsule, the angle between the tumor and the trace, and the contact between the nodule and the TEG, were likewise significantly different between the no gross ETE group and the gross ETE group (all $P < 0.05$) (Appendix Table S2, Appendix Figure S2).

A total of 441 nodules' US images (283 no gross ETE nodules and 158 gross ETE nodules) were randomly selected from the internal data set to be used as the training and validation set. The remaining nodules in the internal data set constitute the internal test set, with 222 nodules (including 139 of no gross ETE nodules and 83 of gross ETE nodules). There were 143 nodules in the external test set (including 95 no gross ETE nodules and 48 gross ETE nodules). Comprehensive data are presented in Table 1.

Comparison of diagnostic performance between DL model and radiologists

In the internal test set (Table 2), the DL model demonstrated an accuracy of 0.87, a sensitivity of 0.80, specificity of 0.92, PPV of 0.86, NPV of 0.88. The AUC was 0.91 (95% CI: 0.87, 0.96) which was the highest in comparison to that of the four radiologists involved in this study (all $P < 0.05$). The AUC of senior radiologist 1 and senior radiologist 2 [0.78 (95% CI: 0.71, 0.85) vs. 0.76 (95% CI: 0.70, 0.83), $P = 0.59$] were significantly higher ($P < 0.05$) than that of junior radiologist 1 and junior radiologist 2 [(0.65 (95% CI: 0.58, 0.73) vs. (0.69 (95% CI: 0.62, 0.77), $P = 0.044$] (Fig. 4). The sensitivity of DL model (0.80) was non-significantly different from that of two seniors radiologists (0.65 and 0.65, respectively) ($P > 0.05$), but significantly higher than that of two junior radiologists (0.46 and 0.45, respectively) ($P < 0.05$). Additionally, the specificity of DL model (0.92) was non-significantly different from that of two seniors radiologists (0.91 and 0.86, respectively) and two junior radiologists (0.85 and 0.94, respectively) ($P > 0.05$). Appendix Tables S4–S7 show the contrast P-value of AUC, accuracy, sensitivity, and specificity.

In the external test set (Table 2), the DL model obtained the highest AUC [0.88 (95% CI: 0.81, 0.94)] in the external test set (Table 2), with an accuracy of 0.85, sensitivity of 0.92, specificity of 0.81, PPV of 0.71, and NPV of 0.95. While there was a non-significant difference between the AUC of DL model and that of senior radiologists 1 [0.81 (95% CI: 0.72, 0.89); $P = 0.152$], a significant difference in the AUC between the DL model and senior radiologist 2 [0.75 (95% CI: 0.66, 0.84), $P = 0.008$)] and two junior radiologists [0.72 (95% CI: 0.62, 0.81), $P = 0.002$; 0.67 (95 CI: 0.57, 0.77), $P < 0.001$] (Fig. 4). Appendix Tables S8–S11 show the contrast P-value of AUC, accuracy, sensitivity, and specificity.

Since the sensitivity of each radiologist to the DL model was more variable than the specificity, we compared the sensitivity of DL model with that of four radiologists to diagnose the no ETE, minimal ETE, T3B, and T4A nodules. Two test sets were combined to create the data set (Table 3). The contrast demonstrated that in non-ETE nodules, the sensitivity of DL model was highest (0.95), and significantly higher than that of senior radiologist 2 (0.89, $P = 0.042$) and junior radiologist 1 (0.87, $P = 0013$). In minimal ETE nodules, the

	AUC (95% CI)†	P value	ACC†	P value	SEN†	P value	SPE†	P value	PPV	NPV	PL ratio	NL ratio	DOR
Internal testing set													
DL Model	0.91 (0.87–0.96)		0.87		0.80		0.92		0.86	0.88	10.05	0.22	45.68
Senior radiologist 1	0.78 (0.71–0.85)*	<0.001	0.82	0.116	0.65	0.06	0.91	1.000	0.82	0.81	7.54	0.38	19.84
Senior radiologist 2	0.76 (0.70–0.83)*	<0.001	0.78*	0.019	0.65	0.071	0.86	0.176	0.74	0.81	4.76	0.40	11.90
Junior radiologist 1	0.65 (0.58–0.73)*	<0.001	0.70*	<0.001	0.46*	<0.001	0.85	0.091	0.64	0.72	3.03	0.64	4.73
Junior radiologist 2	0.69 (0.62–0.77)*	<0.001	0.76*	0.002	0.45*	<0.001	0.94	0.634	0.82	0.74	7.75	0.59	13.14
Senior radiologist 1 + DL	0.82 (0.76–0.89)*	0.001	0.85	0.494	0.73	0.464	0.91	1.000	0.84	0.85	8.51	0.29	29.34
Senior radiologist 2 + DL	0.83 (0.77–0.89)*	0.001	0.84	0.344	0.78	1.000	0.87	0.239	0.78	0.87	6.05	0.25	24.20
Junior radiologist 1 + DL	0.83 (0.76–0.89)*	<0.001	0.84	0.414	0.76	0.709	0.89	0.537	0.81	0.86	7.03	0.27	26.04
Junior radiologist 2 + DL	0.85 (0.79–0.91)*	0.003	0.86	0.888	0.78	1.000	0.91	1.000	0.84	0.88	9.07	0.24	37.79
External testing set													
DL model	0.88 (0.81–0.94)		0.85		0.92		0.81		0.71	0.95	4.84	0.11	44.00
Senior radiologist 1	0.81 (0.72–0.89)	0.152	0.84	1.000	0.71*	0.019	0.91	0.097	0.79	0.86	7.48	0.32	23.38
Senior radiologist 2	0.75 (0.66–0.84)*	0.008	0.78	0.223	0.65*	0.003	0.85	0.561	0.69	0.83	4.38	0.42	10.43
Junior radiologist 1	0.72 (0.62–0.81)*	0.002	0.76	0.101	0.58*	<0.001	0.85	0.561	0.67	0.80	3.96	0.49	8.08
Junior radiologist 2	0.67 (0.57–0.77)*	<0.001	0.72*	0.015	0.52*	<0.001	0.82	1.000	0.60	0.77	2.91	0.58	5.02
Senior radiologist 1 + DL	0.86 (0.79–0.93)	0.682	0.87	0.736	0.83	0.354	0.88	0.226	0.78	0.91	7.20	0.19	37.89
Senior radiologist 2 + DL	0.82 (0.74–0.90)	0.163	0.85	1.000	0.75	0.055	0.89	0.152	0.78	0.88	7.12	0.28	25.43
Junior radiologist 1 + DL	0.81 (0.73–0.89)	0.076	0.83	0.872	0.75	0.055	0.87	0.320	0.75	0.87	5.94	0.29	20.48
Junior radiologist 2 + DL	0.83 (0.75–0.91)	0.125	0.85	1.000	0.77	0.092	0.88	0.226	0.77	0.88	6.66	0.26	25.62

Abbreviations: DL, deep learning; AUC, area under the curve; ACC, accuracy; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; PL Ratio, Positive Likelihood Ratio; NL Ratio, Negative Likelihood Ratio; DOR, diagnostic odds ratio. Note: † The differences between radiologists and DL model, and the differences among radiologists were compared, *P* values were calculated. Detailed results are presented in Appendix Tables S4–S11; *$P < 0.05$, Significant difference with DL model.

Table 2: **Performance comparison among DL model and radiologists.**

sensitivity of DL model was 0.60, which was non-significantly different from that of the four radiologists (all $P > 0.05$). In T3B nodules, the sensitivity of DL model was highest (0.89), significantly higher than that of senior radiologist 2 (0.67, $P = 0.001$) and two junior radiologists (0.52 and 0.52, respectively, with $P < 0.05$). In T4 nodules, the sensitivity of DL model was highest (0.75), which was significantly higher than that of two

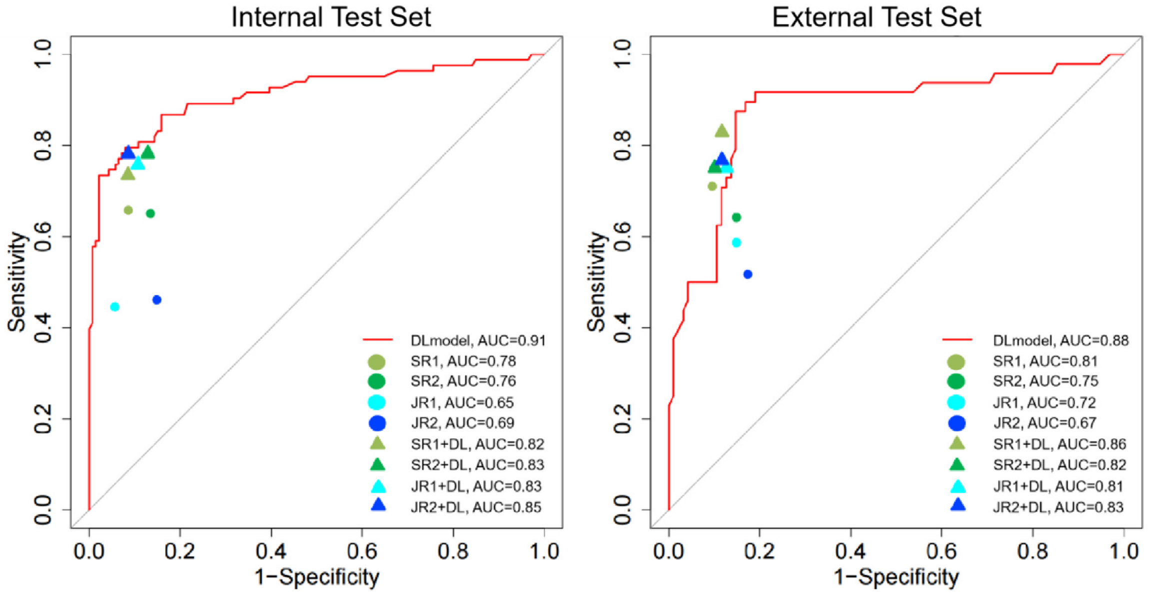


Fig. 4: **Diagnostic performance comparison among DL models and radiologists.** SR, senior radiologist; JR, junior radiologist; SR + DL, DL model assistant senior radiologist; JR + DL, DL model assistant junior radiologist.

	DL model	Senior radiologist 1	Senior radiologist 2	Junior radiologist 1	Junior radiologist 2
No ETE	0.95 (182/192)	0.93 (179/192)	0.89 (170/192)*	0.87 (167/192)*	0.92 (176/192)
Min ETE	0.60 (25/42)	0.81 (34/42)	0.71 (/30/42)	0.76 (32/42)	0.79 (33/42)
T3b	0.89 (77/87)	0.70 (61/87)	0.67 (58/87)*	0.52 (45/87)*	0.52 (45/87)*
T4	0.75 (33/44)	0.61 (27/44)	0.61 (27/44)	0.48 (21/44)*	0.39 (17/44)*

Abbreviations: DL, deep learning. Note: *$P < 0.05$, significant difference with DL model. Detailed P values were presented in Appendix Table S12.

Table 3: **Sensitivities of DL models and radiologists in diagnosis of thyroid carcinoma at different sites of invasion.**

juniors radiologists (0.48 and 0.39, respectively, with $P < 0.05$). (Detailed P values were presented in Appendix Table S12).

In order to verify the performance of DL model in various subgroups of clinical characteristics, we calculated various diagnostic indicators of DL model in different age, sex, multiple, bilateral, tumor size and lymph node status groups (Appendix Table S13). The results showed that the AUCs of DL model in each group had no statistical difference (all $P > 0.05$).

Comparison of diagnostic performance between DL model and clinical model and clinical + DL combined model

Univariate analysis showed that gender, age, lymph node status, tumor size, multiple, bilateral, capsule contact ratio, capsule echo state, angle between tumor and trachea, contact of the node with the TEG were significantly correlated with gross ETE ($P < 0.05$). In multivariable logistic regression analysis, age, capsule contact ratio, capsule echo state, angle between tumor and trachea, and lymph node status were considered as independent predictors of gross ETE. The results of multivariable logistic regression analysis are shown in Table 4.

In the internal test set, the AUC [0.84 (95% CI: 0.79, 0.89), $P = 0.019$], accuracy (0.76, $P = 0.002$) and specificity (0.68, $P < 0.001$) of the clinical model (Fig. 5) were lower than those of the DL model, and the sensitivity (0.89, $P = 0.135$) was not statistically different from the DL model. In the external test set, the AUC [0.85 (95% CI: 0.79, 0.91), $P = 0.516$], accuracy (0.75, $P = 0.055$), sensitivity (0.92, $P = 1$) of the clinical model were not significantly different from the DL model, and the specificity (0.66, $P = 0.032$) was significantly lower than the DL model (Table 5).

Risk factors	OR (95% CI)	P value
Age	1.055 (1.025–1.086)	< 0.001
Envelope contact ratio	1 (reference)	0.013
25%–50%	1.673 (0.361–7.754)	0.510
>50%	3.712 (1.654–8.332)	0.001
Envelope echo state	1 (reference)	< 0.001
Capsular disruption	0.102 (0.037–0.280)	< 0.001
Replacement of strap muscle	0.672 (0.255–1.772)	0.422
Contour bulging	1.826 (0.716–4.658)	0.208
Angle between tumor and trachea	2.581 (1.558–4.277)	< 0.001
Lymph node status	0.439 (0.237–0.811)	0.009

Table 4: **Clinical risk factors for gross ETE.**

In the internal test set, the AUC [0.94 (95% CI: 0.91, 0.97), $P = 0.143$] and accuracy (0.84, $P = 0.414$) of the clinical + DL model (Fig. 5) were not significantly different from the DL model. The sensitivity (0.94, $P = 0.012$) was higher than DL model, and the specificity (0.78, $P = 0.002$) was lower than DL model. In the external test set, the AUC [0.92 (95% CI: 0.87, 0.96), $P = 0.093$], accuracy (0.87, $P = 0.736$), sensitivity (0.92, $P = 1$), specificity (0.84, $P = 0.702$) of the clinical + DL model were not significantly different from the DL model (Table 5, Appendix Figure S6).

Visualization and auxiliary diagnosis function of DL model

We have selected four representative cases to demonstrate the visualization capabilities of DL model (Fig. 6). The DL model can automatically locate the nodule in the US image and depict the extent of nodule with a mask. With the heat map drawn by Grad-CAM, we can observe that the DL model focuses on the area where the nodule is in contact with the adjacent structures of thyroid gland.

The detailed changes in each diagnostic index of four radiologists with the aid of DL model are shown in Table 2. In the internal test set, except for senior radiologist 1 ($P = 0.061$), the AUCs of remaining three radiologists were significantly higher than the corresponding previous values ($P < 0.05$). Additionally, the AUC of all four radiologists in the external test set improved significantly ($P < 0.05$). The senior radiologists improved by 0.06 on average and junior radiologists by 0.15 on average. The findings demonstrate that the DL model helps radiologists positively increase their capacity for diagnosis.

Discussion

In this study, a DL model was constructed. The DL model was used to diagnose the presence of gross ETE by mining and learning the ultrasonic features of thyroid cancer from ultrasonic images. The AUC of DL model in the internal test set was 0.91, while the AUC in

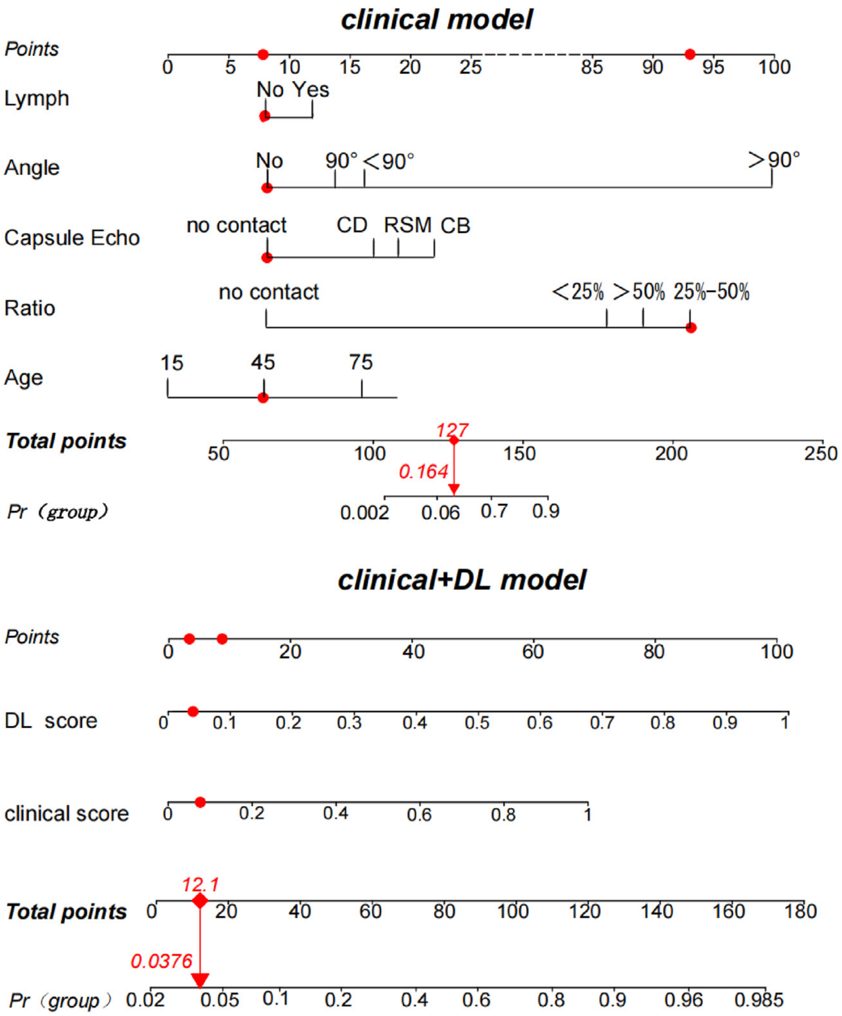


Fig. 5: **Clinical model nomogram and clinical deep learning combined nomogram.** Lymph, lymph node status; Angle, angle between tumor and trachea; Capsule echo, capsule echo state; Ratio, capsule contact ratio; Age, patient's age; DL, deep learning.

the external test set was 0.88, indicating that the model had an excellent ability to identify gross ETE. Furthermore, the mAP of this model was 0.78, indicating that it had a strong capacity to locate and segment thyroid cancer nodules automatically. In addition to diagnosis, the model can depict the location and range of nodules, thereby improving clinical interpretability. Our study is the first to report on the deep-learning study to identify the gross ETE in thyroid cancer based on US images following the 8th edition of AJCC staging system.

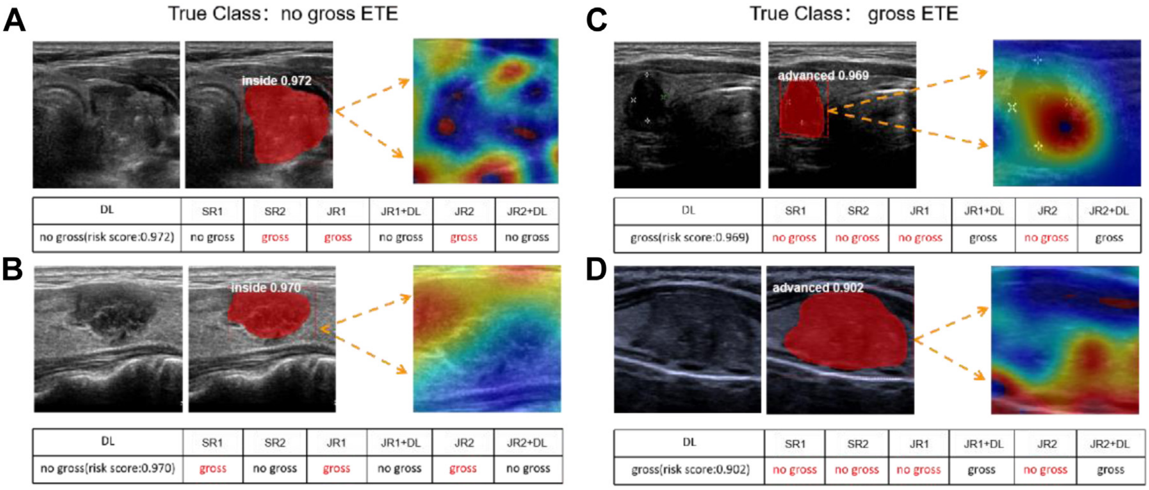
Thyroid cancer has a cancer-related mortality rate of less than 10%,[20] with most thyroid cancer patients having a good prognosis. Consequently, treatment recommendations frequently promote individualized treatment. Questions such as “should active surveillance or surgery be performed?” or “how is the extent of surgery determined?” need to be evaluated specifically in the context of preoperative imaging, clinical information, and intraoperative conditions. One of the key factors to evaluate is the ETE range. Several studies have shown that age, tumor size, multifocality, lymph node metastasis, distant metastasis, recurrence rate, and recurrence-free survival are non-significantly differ between patients with minimal ETE and no ETE, but differ significantly between patients with T3b and T4 and those with no ETE.[3,20–24] Similar findings could be made from the baseline patient information in this investigation. These findings imply that thyroid cancer with and without gross ETE may have different clinical implications for patients.

The commonly used target detection frameworks include two methods. The first type is single-stage detector such as U-NET, SSD.[25] The second type is two-stage detector such as Fast R–CNN[26] and FPN. Among them, FPN can take into account both deep and shallow features by using top-down paths and horizontal

	AUC (95%CI)	*P* value	ACC	*P* value	SEN	*P* value	SPE	*P* value	PPV	NPV	PL ratio	NL ratio	DOR
Internal testing set													
DL model	0.91 (0.87–0.96)		0.87		0.80		0.92		0.86	0.88	10.05	0.22	45.68
Clinical model	0.84 (0.79–0.89)*	0.019	0.76*	0.002	0.89	0.135	0.68*	<0.001	0.37	0.62	2.75	0.16	17.19
Clinical + DL model	0.94 (0.91–0.97)	0.143	0.84	0.414	0.94*	0.012	0.78*	0.002	0.72	0.96	4.35	0.08	54.38
External testing set													
DL model	0.88 (0.81–0.94)		0.85		0.92		0.81		0.71	0.95	4.84	0.11	44.00
Clinical model	0.85 (0.79–0.91)	0.516	0.75	0.055	0.92	1	0.66*	0.032	0.58	0.94	2.72	0.13	20.92
Clinical + DL model	0.92 (0.87–0.96)	0.093	0.87	0.736	0.92	1	0.84	0.702	0.75	0.95	5.81	0.10	58.1

Abbreviations: DL, deep learning; AUC, area under the curve; ACC, accuracy; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; PL Ratio, Positive Likelihood Ratio; NL Ratio, Negative Likelihood Ratio; DOR, diagnostic odds ratio. Note: *P < 0.05, significant difference with DL model.

Table 5: **Performance comparison among DL model, clinical model, clinical + DL model.**



A

DL	SR1	SR2	JR1	JR1+DL	JR2	JR2+DL
no gross(risk score:0.972)	no gross	gross	gross	no gross	gross	no gross

B

DL	SR1	SR2	JR1	JR1+DL	JR2	JR2+DL
no gross(risk score:0.970)	gross	no gross	gross	no gross	gross	no gross

C

DL	SR1	SR2	JR1	JR1+DL	JR2	JR2+DL
gross(risk score:0.969)	no gross	no gross	no gross	gross	no gross	gross

D

DL	SR1	SR2	JR1	JR1+DL	JR2	JR2+DL
gross(risk score:0.902)	no gross	no gross	no gross	gross	no gross	gross

Fig. 6: **Visualization and focus of the DL model and the guidance to radiologists.** ETE, extrathyroidal extension; DL, deep learning model; SR, senior radiologist; JR, junior radiologist. Inside, no gross ETE nodule; in advance, gross ETE nodule. In four sets of representative cases, the first picture of each set is the US picture of thyroid nodule (TN), the second is the diagnostic effect picture of DL model, and the third is the heat map of the nodule and peri-tumor. A: the heat map highlights the area where the TN is in contact with the surrounding capsule; B: the heat map highlights the area where the TN is in contact with the anterior capsule; C: the heat map highlights the area where the TN is in contact with the trachea and nerve; D: the heat map highlights the area where the TN is in contact with the trachea and tracheoesophageal groove.

connections, and generate a feature pyramid with strong semantic information at all scales. In the detection task, it is very important to use the multi-scale features to improve the performance. Mask R–CNN is a successful instance segmentation framework based on FPN. Since nodule areas always have different shapes, we use Mask R–CNN to visualize them to obtain more accurate and visual diagnosis. Through comparison, we finally chose ResNet50 combined with FPN as the backbone of Mask R–CNN. As a classic and excellent deep learning network, ResNet50 has been applied to many medical image analysis scenarios. For example, in the studies of Peng et al.[14] and Wang et al.,[9] ResNet50 was used to design excellent models to identify benign and malignant thyroid nodules. In addition to enhancing the diagnostic performance of model for the internal testing set, ResNet50 enables the model to continue to function consistently in the external testing set.

All four radiologists in this study demonstrated high specificity and low sensitivity in their diagnosis of gross ETE. When sensitivity was compared between the four radiologists and the DL Model, we discovered that the DL Model was more sensitive than most radiologists in no ETE, T3b, and T4a, particularly in T3b, and T4. Gross ETE occurs most often in the strap muscles and RLN, which is difficult for image-based diagnosis.[27–29] In the study by Chung et al.,[5] the sensitivity of radiologists to T3b was 0.45, while in the study by Newman et al.,[30] imaging made correct cues in only a quarter of the patients who invaded the RLN. In the internal dataset of this study, 71.8% of the gross ETE had an invasion of the strap muscle, of which only 46.8% of the nodules had an US presentation typical of the replacement of strap muscle. Additionally, 34.4% of the gross ETE invaded the RLN, but only 40.9% of the nodules had an US image akin to protrusion into tracheoesophageal groove. The differences between the two are particularly challenging to distinguish because of the proximity of the strap muscles to the soft tissues of thyroid gland. Similarly, the proximity of thyroid to the tracheoesophageal groove and trachea makes it difficult for the doctor to determine whether the nodule has broken through the thyroid envelope and invaded the nerve or trachea. Moreover, cases of variation in the location of recurrent laryngeal nerve are common. In a study by Newman et al.,[30] 3.3% of nodules that invaded the recurrent laryngeal nerve were located in areas far from the trachea.

In the internal test set of this study, the diagnostic performance of DL model (AUC of 0.91) was significantly higher than that of the two senior radiologists (AUC of 0.78 and 0.76, respectively). In the external test set, the diagnostic performance of DL model (AUC of 0.88) was also higher than that of the two senior radiologists (AUC of 0.81 and 0.75). However, there was no significant difference compared with senior radiologist 1. The DL model outperformed the two junior radiologists in both test sets, and their diagnostic performance significantly improved to as good as that of the senior radiologists when they applied the DL model. The AUC of the clinical + DL model in the internal test set was 0.94, and that in the external test set was 0.92. Although its diagnostic performance was improved compared with DL model, the improvement was not significant (all $P > 0.05$). Therefore, on the premise that the diagnosis performance is equivalent, only using DL model can simplify the diagnosis process. In terms of visualization capacity, the DL model proposed in this study can automatically locate the nodules and outline the boundaries of nodules, allowing for visualization of the nodule size. Additionally, Grad-CAM further reveals the visualized response area for the model decision. The visualized examples demonstrate that our DL model not only focuses on the thyroid nodule but also on the contact site between the nodule and the surrounding thyroid tissues. The demonstration also indicates that the model is powerful for the flexible recognition of image features, suggesting the potential application of DL model proposed in this study as a tool for junior training radiologists.

However, this study still has several shortcomings that will need to be addressed in the future. First, although we captured several photos from different US sections for each node, missing image features could still occur as this study employed static US images and was conducted retrospectively. At the same time, due to the low incidence rate of gross ETE, we extended the recruitment time of gross ETE patients in consideration of the need for data balance during model training, which may cause selection bias. Secondly, although this study includes external test sets from three centers, they are from one province. Therefore, the popularization of DL model needs to be tested in more centers and large data sets and further improved. In this study, the number of radiologists participating in the comparison is small, which will also lead to selection bias and cannot represent the average level of radiologists. So we also anticipate applying our model to actual clinical circumstances to validate our approach clinically. However, this approach requires special approval from the relevant authorities, which is difficult to obtain in a short period. We will actively pursue this in our future work.

In conclusion, we constructed an US image-based DL model that can determine the presence of gross ETE in thyroid cancer with a diagnostic performance equal or even exceeding that of senior radiologists. The model offers an efficient method for the preoperative diagnosis of gross ETE thyroid cancer.

Contributors

Literature research: Q. Qi, X. Huang, Y. Zhang, Y. Wu, C. Zhang, P. Xu; clinical studies: Q. Qi, X. Huang, Y. Zhang, S. Cai, Z. Liu, A. Zhou, X. Yuan, W. Zhu, X. Min, Y. Wu, C. Zhang, P. Xu; experimental studies: Q. Qi, X. Huang, T. Qiu, Z. Cui; statistical analysis: Q. Qi, X. Huang; and manuscript Writing: Q. Qi, X. Huang, P. Xu, C. Zhang; Q. Qi, X. Huang, P. Xu, C. Zhang have accessed the data, and had final responsibility for the decision to submit it for publication.

Data sharing statement

Due to the privacy of patients, the data related to patients cannot be available for public access but can be obtained from the corresponding author on reasonable request approved by the institutional review board of all enrolled centres.

Declaration of interests

All authors declare no competing interests.

Acknowledgements

Supported by Jiangxi Provincial Natural Science Foundation (20224BAB216079), the Key Research and Development Program of Jiangxi Province (20181BBG70031), and the Interdisciplinary Innovation Fund of Natural Science, Nanchang University (9167-28220007-YB2110).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2023.101905.

References

1 Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin*. 2017;67(2):93–99. https://doi.org/10.1089/thy.2021.0310.

2 Andersen PE, Kinsella J, Loree TR, et al. Differentiated carcinoma of the thyroid with extrathyroidal extension. *Am J Surg*. 1995;170(5):467–470. https://doi.org/10.1016/s0002-9610(99)80331-6.

3 Radowsky JS, Howard RS, Burch HB, et al. Impact of degree of extrathyroidal extension of disease on papillary thyroid cancer outcome. *Thyroid*. 2014;24(2):241–244. https://doi.org/10.1089/thy.2012.0567.

4 Vaccarella S, Franceschi S, Bray F, et al. Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. *N Engl J Med*. 2016;375:614–617. https://doi.org/10.1056/NEJMp1604412.

5 Tran B, Roshan D, Abraham E, et al. An analysis of the American Joint Committee on Cancer 8th edition T staging system for papillary thyroid carcinoma. *J Clin Endocrinol Metab*. 2018;103(6):2199–2206. https://doi.org/10.1210/jc.2017-02551.

6 Chung SR, Baek JH, Choi YJ, et al. Sonographic assessment of the extent of extrathyroidal extension in thyroid cancer. *Korean J Radiol*. 2020;21(10):1187–1195. https://doi.org/10.3348/kjr.2019.0983.

7 Lamartina L, Bidault S, Hadoux J, et al. Can preoperative ultrasound predict extrathyroidal extension of differentiated thyroid cancer? *Eur J Endocrinol*. 2021;185(1):13–22. https://doi.org/10.1530/EJE-21-0091.

8 Kwak JY, Kim EK, Youk JH, et al. Extrathyroid extension of well-differentiated papillary thyroid microcarcinoma on US. *Thyroid*. 2008;18(6):609–614. https://doi.org/10.1089/thy.2007.0345.

9 Moon SJ, Kim DW, Kim SJ, et al. Ultrasound assessment of degrees of extrathyroidal extension in papillary thyroid microcarcinoma. *Endocr Pract*. 2014;20(10):1037–1043. https://doi.org/10.4158/EP14016.OR.

10 Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

11 Cao C, Liu F, Tan H, et al. Deep learning and its applications in biomedicine. *Dev Reprod Biol*. 2018;16(1):17–32. https://doi.org/10.1016/j.gpb.2017.07.003.

12 Mazurowski MA, Buda M, Saha A, et al. Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging*. 2019;49(4):939–954. https://doi.org/10.1002/jmri.26534.

13 Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. 2016;35(5):1285–1298. https://doi.org/10.1109/TMI.2016.2528162.

14 Wang J, Jiang J, Zhang D, et al. An integrated AI model to improve diagnostic accuracy of ultrasound and output known risk features in suspicious thyroid nodules. *Eur Radiol*. 2022;32(3):2120–2129. https://doi.org/10.1007/s00330-021-08298-7.

15 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. https://doi.org/10.1016/S2589-7500(21)00041-8.

16 Hoang JK, Middleton WD, Farjat AE, et al. Reduction in thyroid nodule biopsies and improved accuracy with American College of Radiology Thyroid Imaging Reporting and Data System. *Radiology*. 2018;287(1):185–193. https://doi.org/10.1148/radiol.2018172572.

17 Li G, Li R, Song L, et al. Implications of extrathyroidal extension invading only the strap muscles in papillary thyroid carcinomas. *Thyroid*. 2020;30(1):57–64. https://doi.org/10.1089/thy.2018.0801.

18 He K, Gkioxari G, Dollar P, et al. Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell*. 2020;42(2):386–397. https://doi.org/10.1109/TPAMI.2018.2844175.

19 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(6):1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.

20 Hay ID, Johnson TR, Thompson GB, et al. Minimal extrathyroid extension in papillary thyroid carcinoma does not result in increased rates of either cause-specific mortality or postoperative tumor recurrence. *Surgery*. 2016;159(1):11–19. https://doi.org/10.1016/j.surg.2015.05.046.

21 Amit M, Boonsripitayanon M, Goepfert RP, et al. Extrathyroidal extension: does strap muscle invasion alone influence recurrence and survival in patients with differentiated thyroid cancer? *Ann Surg Oncol*. 2018;25(11):3380–3388. https://doi.org/10.1245/s10434-018-6563-x.

22 Park SY, Kim HI, Kim JH, et al. Prognostic significance of gross extrathyroidal extension invading only strap muscles in differentiated thyroid carcinoma. *Br J Surg*. 2018;105(9):1155–1162. https://doi.org/10.1002/bjs.10830.

23 Moon HJ, Kim EK, Chung WY, et al. Minimal extrathyroidal extension in patients with papillary thyroid microcarcinoma: is it a real prognostic factor? *Ann Surg Oncol*. 2011;18(7):1916–1923. https://doi.org/10.1245/s10434-011-1556-z.

24 Ito Y, Kakudo K, Hirokawa M, et al. Clinical significance of extrathyroid extension to the parathyroid gland of papillary thyroid carcinoma. *Endocr J*. 2009;56(2):251–255. https://doi.org/10.1507/endocrj.k08e-297.

25 Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector. *Comput Vis (ECCV)*. 2016;9905:21–37. https://doi.org/10.1007/978-3-319-46448-0_2.

26 Lee YS, Park WH. Diagnosis of depressive disorder model on facial expression based on fast R-CNN. *Diagnostics*. 2022;12(2). https://doi.org/10.3390/diagnostics12020317.

27 Dralle H, Sekulla C, Haerting J, et al. Risk factors of paralysis and functional outcome after recurrent laryngeal nerve monitoring in thyroid surgery. *Surgery*. 2004;136(6):1310–1322. https://doi.org/10.1016/j.surg.2004.07.018.

28 Randolph GW, Kamani D. The importance of preoperative laryngoscopy in patients undergoing thyroidectomy: voice, vocal cord function, and the preoperative detection of invasive thyroid malignancy. *Surgery*. 2006;139(3):357–362. https://doi.org/10.1016/j.surg.2005.08.009.

29 Shindo ML, Caruana SM, Kandil E, et al. Management of invasive well-differentiated thyroid cancer: an American Head and Neck Society consensus statement. AHNS consensus statement. *Head Neck*. 2014;36(10):1379–1390. https://doi.org/10.1002/hed.23619.

30 Newman SK, Harries V, Wang L, et al. Invasion of a recurrent laryngeal nerve from small well-differentiated papillary thyroid cancers: patient selection implications for active surveillance. *Thyroid*. 2022;32(2):164–169. https://doi.org/10.1089/thy.2021.0310.