ORIGINAL PAPER

Endokrynologia Polska
DOI: 10.5603/EP.a2021.0015
Volume/Tom 72; Number/Numer 3/2021
ISSN 0423–104X, e-ISSN 2299–8306

A comparison between deep learning convolutional neural networks and radiologists in the differentiation of benign and malignant thyroid nodules on CT images

Hong-bo Zhao[1], Chang Liu[1], Jing Ye[2], Lu-fan Chang[3], Qing Xu[2], Bo-wen Shi[1], Lu-lu Liu[4], Yi-li Yin[2], Bin-bin Shi[2]

[1]*Department of Radiology, Second Affiliated Hospital of Dalian Medical University, Dalian, China*
[2]*Department of Radiology, Subei People's Hospital of Jiangsu province, Yangzhou, China*
[3]*Beijing Yizhun-ai Technology Co. Ltd., Beijing, China*
[4]*Department of Radiology, Yangzhou University, Yangzhou, China*

Abstract

Introduction: We designed 5 convolutional neural network (CNN) models and ensemble models to differentiate malignant and benign thyroid nodules on CT, and compared the diagnostic performance of CNN models with that of radiologists.

Material and methods: We retrospectively included CT images of 880 patients with 986 thyroid nodules confirmed by surgical pathology between July 2017 and December 2019. Two radiologists retrospectively diagnosed benign and malignant thyroid nodules on CT images in a test set. Five CNNs (ResNet50, DenseNet121, DenseNet169, SE-ResNeXt50, and Xception) were trained-validated and tested using 788 and 198 thyroid nodule CT images, respectively. Then, we selected the 3 models with the best diagnostic performance on the test set for the model ensemble. We then compared the diagnostic performance of 2 radiologists with 5 CNN models and the integrated model.

Results: Of the 986 thyroid nodules, 541 were malignant, and 445 were benign. The area under the curves (AUCs) for diagnosing thyroid malignancy was 0.587–0.754 for 2 radiologists. The AUCs for diagnosing thyroid malignancy for the 5 CNN models and ensemble model was 0.901–0.947. There were significant differences in AUC between the radiologists' models and the CNN models ($p < 0.05$). The ensemble model had the highest AUC value.

Conclusions: Five CNN models and an ensemble model performed better than radiologists in distinguishing malignant thyroid nodules from benign nodules on CT. The diagnostic performance of the ensemble model improved and showed good potential. **(Endokrynol Pol 2021; 72 (3): 217–225)**

Key words: *deep learning; convolutional neural network (CNN); thyroid nodule classification; computed tomography (CT)*

Introduction

Thyroid nodules are a common disease in clinical practice. The incidence is about 65% in the general population, and most are benign [1]. However, in patients with thyroid nodules, approximately 10% of nodules tend toward malignancy [2], and the incidence of thyroid cancer is on the rise worldwide.

Early diagnosis of thyroid nodules is essential for successful treatment. The development of imaging technology and image processing provides an objective basis for diagnosing thyroid nodules [3–5]. Currently, ultrasonography (US) is the first choice for the examination of thyroid nodules. However, US features of benign and malignant nodules show considerable overlap. In previous studies, the sensitivity and specificity of diagnosing thyroid cancer with US have shown some variation, ranging from 27% to 63% and 78.0% to 96.6%, respectively [6–8]. This is probably due to different examiners, different US instruments, and different definitions of US features. US remains highly subjective and depends on clinical experience. Magnetic resonance imaging (MRI) is often used as a supplementary examination to evaluate thyroid disease. Positron emission tomography (PET) plays a role in evaluating thyroid cancers with dedifferentiated tumours [9]. Computed tomography (CT) has unique advantages in the diagnosis of retrosternal goitres, malignant cases with suspicion of extracapsular extension [10, 11], and multiple punctate calcifications [12]. Also, CT scans can help to detect incidental thyroid cancers [13]. In clinical practice, visual examination of many CT images is a tedious and error-prone task for radiologists. The diagnosis of incidental thyroid nodules (ITNs) varies depending on the radiologist's experience, type of practice, and training [14]. Furthermore, some subtle

Yi-li Yin, Department of Radiology, Subei People's Hospital of Jiangsu province, No. 98 of Nantong Road (West), Guangling District, Yangzhou 225001, China, tel: +86 87373620; e-mail: yinli_li15@163.com
Bin-bin Shi, Department of Radiology, Subei People's Hospital of Jiangsu province, No. 98 of Nantong Road (West), Guangling District, Yangzhou 225001, China, tel: +86 87373620; e-mail: shibb154@126.com

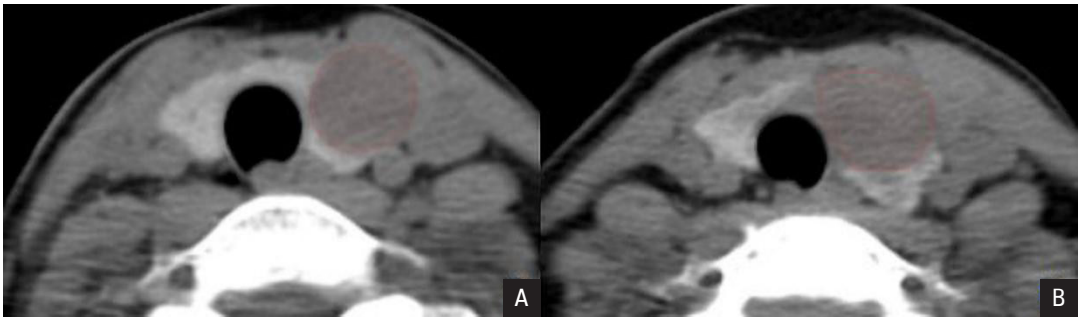

Figure 1. *Regions of interest (ROIs). Figure A shows benign nodules, and Figure B shows malignant nodules. The regions in the red circle are the regions of interest*

CT features, such as calcification, might be neglected in visual inspection.

Fine-needle aspiration (FNA) biopsy has been considered the gold standard for the definitive diagnosis of benign and malignant thyroid nodules. However, the average diagnostic accuracy is approximately 83%, and there is a proportion of false positives. Patients are at risk of secondary biopsy due to limitations in specimen collection and operator experience [15].

Deep convolutional neural networks (CNNs), an emerging form of computer-aided diagnostic (CAD) analysis, are used to form quantitative decisions by automatically extracting features and through the supervised learning of large amounts of data. A growing number of studies have shown that deep learning algorithms have been widely used to solve detection/classification problems, potentially replacing conventional handcrafted methodologies [16–19]. Ko et al. [20] showed that CNNs and experienced radiologists had comparable diagnostic performances to differentiating thyroid malignancy on US. However, to our knowledge, CNN models for the differential diagnosis of benign and malignant thyroid nodules on CT images are infrequent. In this study, we designed 5 CNN models and an ensemble model to differentiate malignant and benign thyroid nodules on CT and compared the diagnostic performance of CNN models with that of radiologists.

Material and methods

Patient data

The institutional review board approved the protocol of this retrospective study. A total of 1527 patients with thyroid nodules were retrospectively enrolled between July 2017 and December 2019 from Northern Jiangsu People's Hospital. The inclusion criteria were as follows: (1) no previous surgical treatment and FNA biopsy, (2) conventional CT examination before the biopsy, and (3) CT image quality meets the diagnostic requirements and calibration analysis. The exclusion criterion was histology with ambiguous diagnostic findings. Demographic information, imaging examination, and clinical baseline characteristics were collected from the hospital PACS (version 4.0.11) workstation. After the screening, we successfully enrolled 880 patients with CT images of 986 nodules. These nodules were then randomly split into training-validation and test sets.

Image acquisition and preprocessing

A GE LightSpeed VCT 64 CT scanner was used for routine thyroid scanning. The tube voltage was 120 kV, and the tube current was 210 mAs. The scanning range was from the skull base to the upper margin of the aortic arch. The layer thickness was 1.25 mm, and the layer spacing was 1.25 mm. The examinations were performed in helical mode, and the helical pitch was 0.984:1 for the CT image. The gantry rotation time was 0.5 seconds for CT scanners. CT images were exported in DICOM format.

CT images of all included patients were exported in DICOM format to the Darwin Intelligent Scientific Research Platform. Two radiologists drew regions of interest (ROIs) (Fig. 1) manually at the edge of the nodule on a layer-by-layer basis on the axial images. They were also involved in the diagnosis of thyroid nodules. Radiologist A (7 years of experience in diagnostic radiology) sketched CT images twice over 2 weeks, while Radiologist B (26 years of experience in diagnostic radiology) performed only 1 feature extraction. Inter- and intra-class correlation coefficients (ICCs) were used to assess the inter- and intra-observer agreement of feature extraction, with an ICC greater than 0.75 indicating good agreement. The ROIs drawn by Radiologist A were entered into the CNN models for subsequent analysis.

A total of 986 cases were marked ROI and were randomly divided into 788 cases of a training-verification set. The test set included 198 cases. For the ROI of the training set, random horizontal flipping and random rotation were performed as image augmentation. All ROI images were adjusted to a window width of 350 and a window level of 40 and scaled to a size of 224 × 224 × 3 (Xception network of 299 × 299 × 3), normalized to pixels between 0 and 1.

Model training-validation and testing, ensemble model

Figure 2 shows the basic architecture of CNN. Five deep learning CNN models were selected to differentiate benign and malignant thyroid nodules based on preoperative CT images. The CNN models used were ResNet50, DenseNet121, DenseNet169, SE-ResNeXt50, and Xception. All networks adopted the pre-trained models on ImageNet. ImageNet is an image database organized according to the WordNet hierarchy, in which each node of the hierarchy is depicted by hundreds or thousands of images. ImageNet is larger in scale and diversity than the other image classification datasets [21]. All models performed 5-fold cross-validation on the training-validation set. The maximum number of iterations in training was 50. The batch size was 4, the optimizer was Adam, the initial learning rate was 5e-5, and the learning rate decayed to the 9th power of the number of iterations. For the 5-fold cross-validation of each model, the model with the highest AUC on the validation set was selected and tested on the test set.

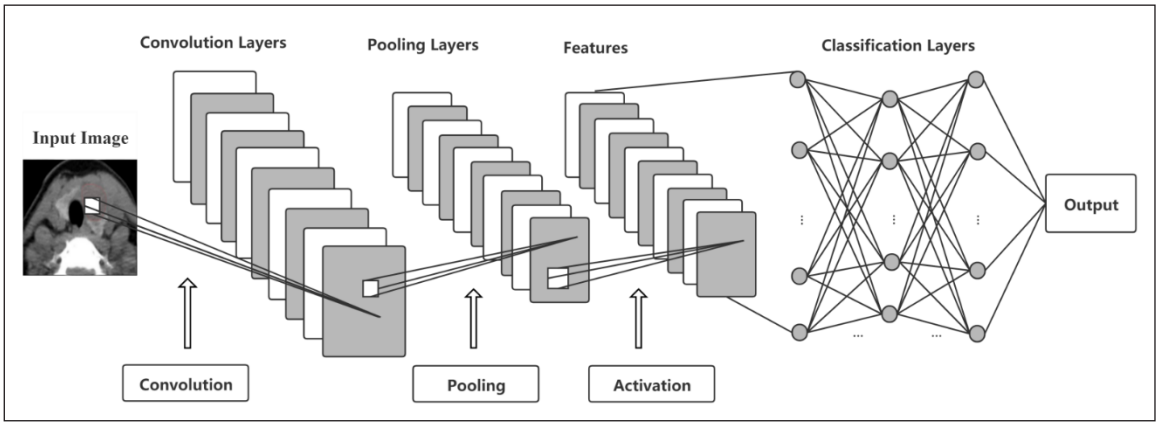

Figure 2. *Basic architecture of convolutional neural network (CNN) for image classification problems*

We selected 3 models with better diagnostic performance, integrated the predicted results of each model's folds on the test set, and finally obtained the ensemble model of the 3 models.

Performance evaluation

The performances of the 5 CNN models and the ensemble model were measured by the area under the receiver operating characteristic curve (ROC), sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) of the test dataset.

Comparison between CNN models and radiologists

Two radiologists, who were blind to the FNA histological results, diagnosed each thyroid nodule as benign or malignant on the CT images of the test set. Their diagnostic performances were compared with the 5 CNN models and the ensemble model.

Attention heat map and lesion detection

In order to understand how CNN interprets CT images for thyroid nodule classification, we extracted the last convolution layer before classification of the fully connected layer of the trained model, used a Class Activation Map (CAM) [22] to calculate the gradients of this layer, and visualized it as a heat map. Then the heat map was overlaid on the original CT image to show the region of interest for the CNN algorithm. In the view of CNN, red and yellow pixel areas correlated more strongly with nodule classification.

Statistical analysis

Based on the prediction results, the sensitivity, specificity, accuracy, PPV, and NPV were calculated to evaluate the diagnostic performances of the different CNN models and radiologists on benign and malignant thyroid nodules. At the same time, the AUC and the 95% confidence interval (CI) were calculated. Additionally, AUCs were compared between each other using DeLong's method. For subject-based comparisons of demographics, the independent 2-sample T-test and chi-square test were used. For nodule-based comparison of nodule characteristics, the generalized estimating equations method was used. $p < 0.05$ was considered statistically significant. All statistical analyses were conducted using SPSS software (version 19.0, IBM Corporation, Armonk, NY) and MedCalc for Windows (version 15.0, MedCalc Software, Ostend, Belgium).

Results

Patient characteristics

There were 541 (53.2%) malignant nodules and 445 (46.8%) benign nodules. These nodules were randomly split into a training-validation set (359 benign and 429 malignant nodules) and a test set (86 benign and 112 malignant nodules). The histopathological results are listed in Table 1. A summary of demographics features can be seen in Table 2. The mean size of the nodules, the female-to-male ratio, and the age of the patients were not significantly different between the benign and malignant thyroid nodules ($p > 0.05$).

Table 1. ***Histopathological results of surgically resected nodules***

	Histopathologic result	Number
Benign (n = 445)	Nodular goitre	247
	Follicular thyroid adenoma	161
	Subacute thyroiditis	25
	Hashimoto's thyroiditis	12
Malignant (n = 541)	Papillary thyroid carcinoma	499
	Medullary thyroid carcinoma	19
	Follicular carcinoma	12
	Anaplastic carcinoma	6
	Primary thyroid lymphoma	5

Diagnostic performances of 5 CNN models for malignant and benign thyroid nodules, and pairwise comparisons between the 5 CNN models

Table 3 presents the diagnostic performances of the 5 CNN models of ResNet50, DenseNet121, DenseNet169, SE-ResNeXt50, and Xception in differentiating malignant and benign thyroid nodules. The AUCs of the 5 models on the test set were 0.945 (95% CI: 0.90–0.97), 0.943 (95% CI: 0.90–0.97), 0.936 (95% CI: 0.89–0.97), 0.920 (95% CI: 0.87–0.95), and 0.901 (95% CI: 0.85–0.94), respectively. ROC curves are shown in Figure 2. The results of pairwise comparisons between all models are shown in Table 4. There were significant AUC differences between ResNet50 and SE-ResNeXt50 (0.945 *vs.* 0.920;

Table 2. *Summary of demographic features*

Features	Benign nodules (n = 445)	Malignant nodules (n = 541)	p value
No. of patients	408	472	
Age, years, x ± s	50.6 ± 12.3	44.3 ± 12.6	< 0.001
Sex (%)			0.358
Male	118 (29.0%)	150 (31.7%)	
Female	290 (71.0%)	322 (68.3%)	
Size [cm]	2.06 ± 1.40	1.81 ± 1.20	< 0.001
≤ 0.5	49 (11.0%)	90 (16.9%)	
0.5-2.0	134 (30.1%)	173 (32.1%)	
≥ 2.0	262 (58.9%)	278 (51.0%)	

Table 3. *The diagnostic performances of 5 convolutional neural network (CNN) models, an ensemble model, and 2 radiologists on the test set*

	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
ResNet50	0.945	0.874	0.837	0.911	0.877	0.872
DenseNet121	0.943	0.869	0.884	0.866	0.833	0.898
DenseNet169	0.936	0.859	0.837	0.884	0.845	0.868
SE-ResNeXt50	0.920	0.859	0.872	0.857	0.822	0.889
Xception	0.901	0.808	0.872	0.768	0.740	0.878
ensemble model	0.947	0.859	0.919	0.821	0.796	0.920
Radiologist A	0.587	0.586	0.593	0.580	0.520	0.644
Radiologist B	0.754	0.748	0.802	0.705	0.677	0.705

PPV — positive predictive value; NPV — negative predictive value, ensemble model: ResNet50, DenseNet121, and DenseNet169; Radiologist A — inexperienced radiologists; Radiologist B — experienced radiologists

Table 4. *Comparisons of diagnostic performances between 5 convolutional neural network (CNN) models and an ensemble model for malignant and benign thyroid nodules*

	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
RN50 *vs*. DN121	0.839	0.881	0.219	0.125	0.425	0.537
RN50 *vs*. DN169	0.234	0.658	0.453	0.453	0.562	0.939
RN50 *vs*. SE-RN50	0.035*	0.685	0.508	0.109	0.323	0.693
RN50 *vs*. Xception	0.005*	0.074	0.581	0.000*	0.022*	0.899
RN50 *vs*. IM	0.680	0.685	0.016*	0.006*	0.151	0.250
DN121 *vs*. DN169	0.365	0.770	0.118	0.727	0.831	0.491
DN121 *vs*. SE-RN50	0.051	0.770	1.000	1.000	0.844	0.825
DN121 *vs*. Xception	0.002*	0.101	0.250	0.007*	0.118	0.693
DN121 *vs*. IM	0.489	0.770	0.250	0.063	0.510	0.585
DN169 *vs*. SE-RN50	0.155	1.000	0.453	0.453	0.684	0.641
DN169 *vs*. Xception	0.015*	0.178	0.453	0.002*	0.082	0.842
DN169 *vs*. IM	0.007*	1.000	0.016*	0.016*	0.389	0.224
SE-RN50 *vs*. Xception	0.173	0.178	1.000	0.021*	0.173	0.800
SE-RN50 *vs*. IM	0.009*	1.000	0.219	0.344	0.674	0.447
Xception *vs*. IM	0.001*	0.178	0.289	0.238	0.352	0.322

PPV — positive predictive value; NPV — negative predictive value; RN50 — ResNet50; DN121 — DenseNet121; DN169 — DenseNet169; SE-RN50 — SE-ResNeXt50, IM — ensemble model; * represent statistically significant ($p < 0.05$)

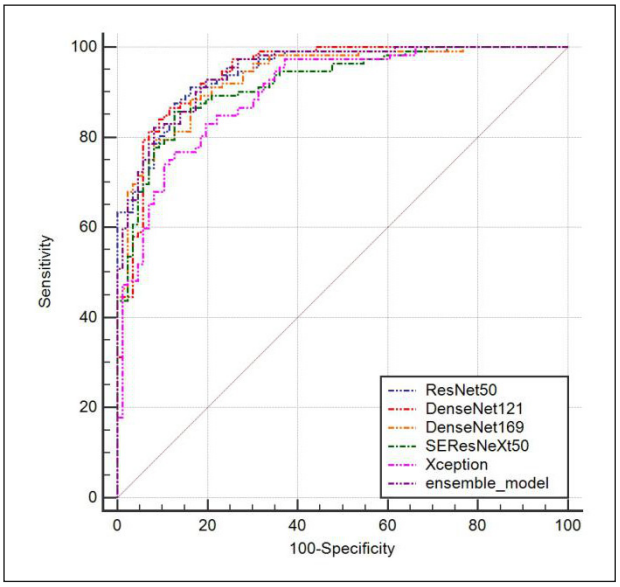

Figure 3. *Receiver operating characteristic curves. The receiver operating characteristic curves of the 5 convolutional neural network (CNN) models and the ensemble model on the test set. Different CNN models are represented by 6 dotted lines, respectively.*

p = 0.035), and there were significant AUC differences among ResNet50, DenseNet121, DenseNet169, and Xception (0.945, 0.943, 0.936, *vs*. 0.901, respectively; p = 0.005, 0.002, and 0.015, respectively). In terms of sensitivity, pairwise comparisons between all models showed no statistically significant differences (p > 0.05), but DenseNet121 had the highest sensitivity. For specificity, there were significant differences among ResNet50, DenseNet121, DenseNet169, SE-ResNeXt50, and Xception (0.911, 0.866, 0.884, 0.857 *vs*. 0.768, respectively; p = 0.000, 0.007, 0.002, and 0.021, respectively). For PPV, there were significant differences between ResNet50 and Xception (0.877 *vs*. 0.740; p = 0.022). Among the 5 CNN models, ResNet50, DenseNet121, and DenseNet169 exhibited better diagnostic performances.

Comparisons of diagnostic performances between the ensemble model and CNN models for malignant and benign thyroid nodules

The prediction of 3 models (Resnet50, Densenet121, and Desnent169) with better effect were further integrated. In the test set, the AUC was 0.947 (95% CI: 0.906–0.974), sensitivity was 0.919, specificity was 0.821, accuracy was 0.859, PPV was 0.798, and NPV was 0.920 (Tab. 3). ROC curves are shown in Figure 3. The comparison results between the ensemble model and the 5 models are shown in Table 4. In terms of AUC, there were significant differences among DenseNet169, SE-ResNeXt50, Xception, and the ensemble model (0.936, 0.920, 0.901, *vs*. 0.947, respectively; p = 0.007, 0.009, and 0.001, respectively). For sensitivity, there were significant differences among ResNet50, DenseNet169, and the ensemble model (0.837, 0.837, *vs*. 0.919, respectively; p = 0.016, and 0.016). For specificity, there were significant differences among Resnet50, DenseNet169, and the ensemble model (0.911, 0.884, *vs*. 0.821, respectively; p = 0.006, and 0.016, respectively). For accuracy, PPV, and NPV, there were no significant differences between the 5 models and the ensemble model (p > 0.05). The ensemble model had the highest AUC value, although it was not statistically significant compared with Resnet50 and Densenet121.

Comparisons of diagnostic performances between the 2 radiologists and CNN models for malignant and benign thyroid nodules

The inter- and intra-class correlation coefficients (ICCs) were 0.878 and 0.961, respectively, indicating good agreement. The diagnosis results of the 2 radiologists for the CT images from the test set are shown in Table 3. Unsurprisingly, the experienced radiologist (Radiologist B) showed significantly better results than the inexperienced radiologist (Radiologist A) (Tab. 5, p > 0.05). The comparison results of the 5 models and ensemble model with Radiologist A and Radiologist B are shown in Table 5. ROC curves are shown in Figure 4. The 5 models and the integrated model showed significantly better results than Radiologist A in the diagnosis of benign and malignant thyroid nodules (p > 0.05). In terms of AUC, there were significant differences among the 5 models, the ensemble model, and Radiologist B (p < 0.05). For specificity, there were significant differences among Resnet50, Densenet121, DenseNet169, SE-ResNeXt50, and Radiologist B (0.911, 0.866, 0.884, 0.857 *vs*. 0.705, respectively; p = 0.000, 0.005, 0.001, and 0.009, respectively). For accuracy, there were significant differences among Resnet50, Densenet121, DenseNet169, SE-ResNeXt50, the ensemble model, and Radiologist B (0.874, 0.869, 0.859, 0.859, 0.859 *vs*. 0.7475, respectively; p = 0.001, 0.002, 0.005, 0.005, and 0.005, respectively). For PPV, there were significant differences between the ensemble model and Radiologist B (0.920 *vs*. 0.705; p = 0.042). For NPV, there were significant differences among Resnet50, Densenet121, DenseNet169, SE-ResNeXt50, and Radiologist B (0.877, 0.833, 0.845, 0.822 *vs*. 0.677, respectively; p = 0.002, 0.012, 0.008, and 0.021, respectively). In conclusion, the 5 CNN models and the ensemble model performed better than the radiologists.

Attention heat map and lesion detection

We generated an attention heat map by a deep learning visualization technique (Fig. 5). By analysing the heat map images, we learned that the CNN model focuses not only on the internal regions of the nodule but also

Table 5. *Comparisons of diagnostic performances between 2 radiologists and convolutional neural network (CNN) models for malignant and benign thyroid nodules*

	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
RN50 *vs*. Radiologist A	< 0.001*	< 0.001*	0.001*	< 0.001*	< 0.001*	< 0.001*
DN121 *vs*. Radiologist A	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*
DN169 *vs*. Radiologist A	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*
SE-RN50 *vs*. Radiologist A	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*
Xception *vs*. Radiologist A	< 0.001*	< 0.001*	< 0.001*	0.002*	< 0.001*	< 0.001*
IM *vs*. Radiologist A	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*
RN50 *vs*. Radiologist B	< 0.001*	0.001*	0.839	< 0.001*	0.002*	0.321
DN121 *vs*. Radiologist B	< 0.001*	0.002*	0.263	0.005*	0.012*	0.119
DN169 *vs*. Radiologist B	< 0.001*	0.005*	0.824	0.001*	0.008*	0.361
SE-RN50 *vs*. Radiologist B	< 0.001*	0.005*	0.359	0.009*	0.021*	0.125
Xception *vs*. Radiologist B	< 0.001*	0.147	0.359	0.360	0.321	0.286
IM *vs*. Radiologist B	< 0.001*	0.005*	0.064	0.053	0.056	0.042*
Radiologist A *vs*. Radiologist B	< 0.001*	0.001*	< 0.001*	0.001*	0.624	0.006*

DN169 — DenseNet169; SE-RN50 — SE-ResNeXt50; IM — ensemble model; Radiologist A — inexperienced radiologists; Radiologist B — experienced radiologists; *represent statistically significant ($p < 0.05$)

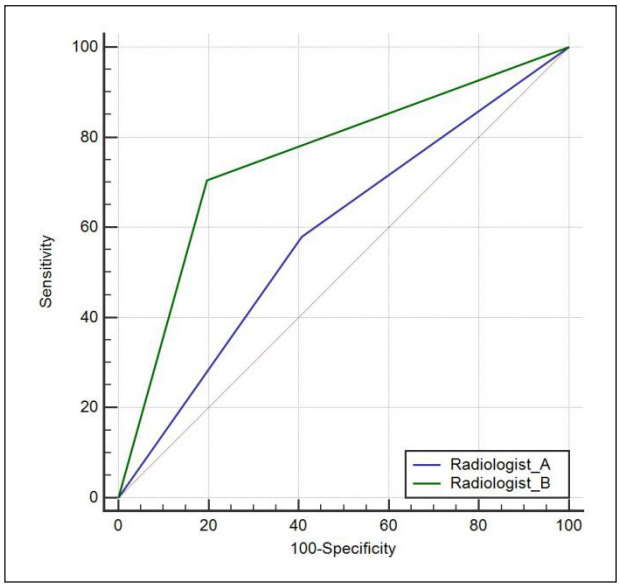

Figure 4. *Receiver operating characteristic curves. The receiver operating characteristic curves of 2 radiologists and the 2 solid lines, respectively, represent the radiologists*

the external parenchyma adjacent to the nodule boundary. Both benign and malignant nodules focus on the external parenchyma adjacent to the nodule boundary. However, unlike benign nodules, malignant nodules also focus on the internal areas.

Discussion

In this study, 5 models and an ensemble model showed favourable diagnostic performances for differentiating malignant and benign thyroid nodules on CT, demonstrating AUCs of 0.901–0.947, sensitivities of 0.837–0.919, specificities of 0.768–0.911, accuracies of 0.808–0.874, PPVs of 0.740–0.877, and NPVs of 0.868–0.920 in the test set. Among the 5 models, the AUC of ResNet50, DenseNet121, and DenseNet169 was significantly better than that of Xception. In this study, we selected 3 models with better AUC for integrating. The AUC of the ensemble model was the best of all the models despite no statistical significance with ResNet50 and DenseNet121. The sensitivity of the ensemble model was noticeably better than ResNet50, but its specificity was not as good as ResNet50.

Compared with the 2 radiologists, all the data of the 5 models and the ensemble model were noticeably better than the inexperienced radiologist (Radiologist A). The AUC of the 5 models and the ensemble model was significantly better than the experienced radiologist (Radiologist B), and the specificity and PPV of ResNet50, DenseNet121, DenseNet169, and SE-ResNeXt50 were significantly better than Radiologist B. In terms of accuracy, the models, except for Xception, all performed better than Radiologist B. The NPV of the ensemble model was significantly better than Radiologist B. The results showed that the diagnostic performances of the 5 models and the ensemble model were noticeably better than that of Radiologist A and somewhat better than that of Radiologist B.

This result is especially important for China. Due to the large gap between eastern and western China and the varying levels of diagnosis and treatment in primary and secondary hospitals, the 5-year survival

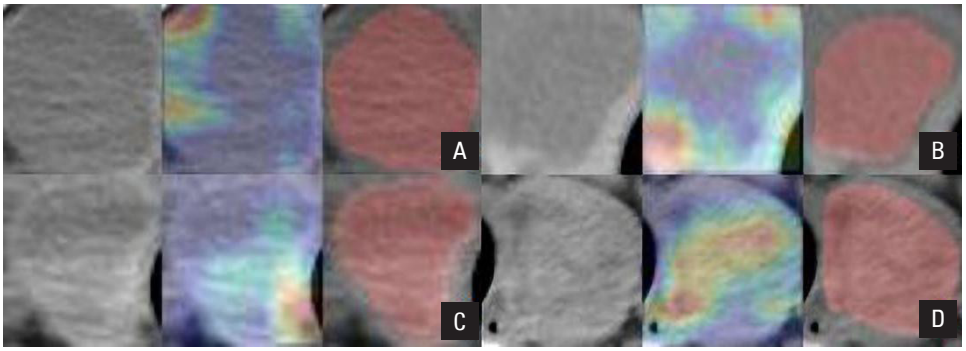

Figure 5. *Attention heatmap.* **A**, **B**. *Heatmaps based on 2 benign thyroid nodule images;* **C**, **D**. *Heatmaps based on 2 malignant thyroid nodule images*

rate of thyroid cancer is only 67.5%, compared with 98.2% and 77.6% in the USA and European countries, respectively [23, 24]. In this study, 5 CNN models and an integrated model performed better diagnostically than 2 radiologists and showed good application value.

Previous researchers have applied manual image feature extraction methods to the classification of thyroid nodules. Chang et al. [25] extracted 78 texture features from US images of thyroid nodules and created a Support Vector Machines (SVMs) model to classify the input images into several categories such as nodules and non-nodules, follicles, and fibrosis. However, handcrafted image feature extractors are designed and selected by the author. They are limited by the author's expertise and can only reflect limited aspects of the problem. Therefore, their classification performance is restricted.

Deep learning, a branch of artificial intelligence, is considered a state of the art image classification technique, which analyses the relationships between existing data points. It has promising applications in clinical diagnosis and risk stratification [26–28]. Unlike handcrafted feature extraction methods, deep learning-based methods, such as CNN, can automatically learn the useful texture features for detection/classification problems, thus yielding better results. With the rapid development of graphic processing units (GPUs), algorithms, and the availability of data, deep learning-based techniques have been widely used to solve image classification problems recently [29, 30].

In a study by Zhu et al. [30], the researchers fine-tuned the residual network based on ResNet18 and obtained good classification results using a public dataset. Similar to the above research, Chi et al. [31] also used the CNN network to classify benign and malignant thyroid nodules on US images. Another study used detection networks such as the multiscale single-shot detection network (multiscale SSD) or Yolo network to differentiate the thyroid nodules by detection-and-classification [32]. The results of the first step of the detection were used to classify the nodules. The method was characterized by the removal of noise and non-nodular regions before performing the classification. However, the method is difficult to use to find small nodules, and the network structure is complex. In our study, we included all sizes of nodules and obtained better results.

Compared with the above research, this research has certain advantages. First, we trained a total of 5 models, while the above research only used a single model. Second, we not only analysed the diagnostic performance of each model for benign and malignant thyroid nodules but also made pairwise comparisons between all models. All 5 models achieved good results, among which ResNet50, DenseNet121, and DenseNet169 had better diagnostic performances. Finally, this study also selected 3 models with better diagnostic performances for the ensemble model. The advantage of an ensemble model is that it can collect each CNN model's architecture and learn characteristics of the input image features, resulting in richer information than can be obtained using individual models. The ensemble model had the highest AUC, sensitivity, and NPV values and improved diagnostic performance, among which the improvement in terms of sensitivity is favourable for clinical screening of malignant nodules. Nguyen et al. [33] integrated the classification results of 2 trained network models, ResNet50 and Inception, to investigate whether the diagnostic performance of the ensemble network for thyroid nodules was better than that of the individual models. This is similar to the present study. After observing its analytical pattern on the heatmaps, we recognized that the internal area was vital for classification. This deep learning visualization technique may help radiologists interpret thyroid CT images more effectively.

This study has several limitations. First, approximately 92.2% of the malignant thyroid nodules in this study were papillary thyroid carcinoma, which may cause the CT presentation of malignant thyroid nodules

to be too homogeneous. Follow-up studies are needed to increase the number of various types of malignant thyroid nodules. Second, this study was based on a single centre and had a small total sample size, requiring an external validation study and an expanded sample size to validate its diagnostic performance and generalizability. Finally, the sketch of regions of interest in this study was a manual sketch, which is not automatic enough and has limitations in clinical application. In the next stage, it will be further improved to carry out an automatic or semi-automatic sketch.

Conclusions

In conclusion, 5 models and the ensemble model performed better than radiologists in distinguishing malignant thyroid nodules from benign nodules on CT. Compared with the single model, the diagnostic performance of the ensemble model improved and showed good potential. Therefore, CNN can be employed as a useful method for distinguishing malignant thyroid nodules from benign ones.

Competing interests

The authors declare that they have no competing interests.

Funding

None.

Ethics approval and consent to participate

I confirm that I have read the Editorial Policy pages. This study was conducted with approval from the Ethics Committee of the Second Affiliated Hospital of Dalian Medical University. This study was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants.

Consent for publication

All patient guardians signed a document of informed consent.

Acknowledgements

We would like to acknowledge the hard and dedicated work of all the staff that implemented the intervention and evaluation components of the study.

References

1. Russ G, Leboulleux S, Leenhardt L, et al. Thyroid incidentalomas: epidemiology, risk stratification with ultrasound and workup. Eur Thyroid J. 2014; 3(3): 154–163, doi: 10.1159/000365289, indexed in Pubmed: 25538897.
2. Angell TE, Maurer R, Wang Z, et al. A Cohort Analysis of Clinical and Ultrasound Variables Predicting Cancer Risk in 20,001 Consecutive Thyroid Nodules. J Clin Endocrinol Metab. 2019; 104(11): 5665–5672, doi: 10.1210/jc.2019-00664, indexed in Pubmed: 31310316.
3. Hoang JK, Branstetter BF, Gafton AR, et al. Imaging of thyroid carcinoma with CT and MRI: approaches to common scenarios. Cancer Imaging. 2013; 13: 128–139, doi: 10.1102/1470-7330.2013.0013, indexed in Pubmed: 23545125.
4. Shie P, Cardarelli R, Sprawls K, et al. Systematic review: prevalence of malignant incidental thyroid nodules identified on fluorine-18 fluorodeoxyglucose positron emission tomography. Nucl Med Commun. 2009; 30(9): 742–748, doi: 10.1097/MNM.0b013e32832ee09d, indexed in Pubmed: 19561553.
5. Brito JP, Gionfriddo MR, Al Nofal A, et al. The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. J Clin Endocrinol Metab. 2014; 99(4): 1253–1263, doi: 10.1210/jc.2013-2928, indexed in Pubmed: 24276450.
6. Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association ManagementGuidelines for Adult Patients with Thyroid Nodulesand Differentiated Thyroid Cancer. Thyroid. 2015; 26(1): 1–133, doi: 10.1089/thy.2015.0020., indexed in Pubmed: 26462967.
7. Remonti LR, Kramer CK, Leitão CB, et al. Thyroid ultrasound features and risk of carcinoma: a systematic review and meta-analysis of observational studies. Thyroid. 2015; 25(5): 538–550, doi: 10.1089/thy.2014.0353, indexed in Pubmed: 25747526.
8. Frates MC, Benson CB, Charboneau JW, et al. Society of Radiologists in Ultrasound. Management of thyroid nodules detected at US: Society of Radiologists in Ultrasound consensus conference statement. Radiology. 2005; 237(3): 794–800, doi: 10.1148/radiol.2373050220, indexed in Pubmed: 16304103.
9. Chaudhary V, Bano S. Imaging of the thyroid: Recent advances. Indian J Endocrinol Metab. 2012; 16(3): 371–376, doi: 10.4103/2230-8210.95674, indexed in Pubmed: 22629501.
10. Moschetta M, Ianora AA, Testini M, et al. Multidetector computed tomography in the preoperative evaluation of retrosternal goiters: a useful procedure for patients for whom magnetic resonance imaging is contraindicated. Thyroid. 2010; 20(2): 181–187, doi: 10.1089/thy.2009.0107, indexed in Pubmed: 20151825.
11. Ishigaki S, Shimamoto K, Satake H, et al. Multi-slice CT of thyroid nodules: comparison with ultrasonography. Radiat Med. 2004; 22(5): 346–353, indexed in Pubmed: 15553016.
12. Wu CW, Dionigi G, Lee KW, et al. Calcifications in thyroid nodules identified on preoperative computed tomography: patterns and clinical significance. Surgery. 2012; 151(3): 464–470, doi: 10.1016/j.surg.2011.07.032, indexed in Pubmed: 21911238.
13. Hoang JK, Choudhury KR, Eastwood JD, et al. An exponential growth in incidence of thyroid cancer: trends and impact of CT imaging. AJNR Am J Neuroradiol. 2014; 35(4): 778–783, doi: 10.3174/ajnr.A3743, indexed in Pubmed: 24113469.
14. Hoang JK, Riofrio A, Bashir MR, et al. High variability in radiologists' reporting practices for incidental thyroid nodules detected on CT and MRI. AJNR Am J Neuroradiol. 2014; 35(6): 1190–1194, doi: 10.3174/ajnr.A3834, indexed in Pubmed: 24407274.
15. Gharib H, Papini E, Garber JR, et al. AACE/ACE/AME Task Force on Thyroid Nodules. American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi medical guidelines for clinical practice for the diagnosis and management of thyroid nodules — 2016 update. Endocr Pract. 2016; 22(5): 622–639, doi: 10.4158/EP161208.GL, indexed in Pubmed: 27167915.
16. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015; 61: 85–117, doi: 10.1016/j.neunet.2014.09.003, indexed in Pubmed: 25462637.
17. Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017; 19: 221–248, doi: 10.1146/annurev-bioeng-071516-044442, indexed in Pubmed: 28301734.
18. Alipanahi B, Delong A, Weirauch MT, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015; 33(8): 831–838, doi: 10.1038/nbt.3300, indexed in Pubmed: 26213851.
19. Aerts HJ, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014; 5: 4006, doi: 10.1038/ncomms5006, indexed in Pubmed: 24892406.
20. Ko SuY, Lee JiH, Yoon JH, et al. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. Head Neck. 2019; 41(4): 885–891, doi: 10.1002/hed.25415, indexed in Pubmed: 30715773.
21. Russako*v*sky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comp Vis. 2015; 115(3): 211–252, doi: 10.1007/s11263-015-0816-y.
22. Bolei Z, Aditya K, Agata L et al. Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Seattle, USA, 2016 June 27–30.
23. Siegel RL, Fedewa SA, Miller KD, et al. Cancer statistics, 2015. CA Cancer J Clin. 2015; 65(1): 5–29, doi: 10.3322/caac.21254, indexed in Pubmed: 25559415.

24. Thomson CS, Forman D. Cancer survival in England and the influence of early diagnosis: what can we learn from recent EUROCARE results? Br J Cancer. 2009; 101 Suppl 2: S102–S109, doi: 10.1038/sj.bjc.6605399, indexed in Pubmed: 19956153.
25. Chang CY, Chen SJ, Tsai MF. Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images. Pattern Rec. 2010; 43(10): 3494–3506, doi: 10.1016/j.patcog.2010.04.023.
26. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553): 436–444, doi: 10.1038/nature14539, indexed in Pubmed: 26017442.
27. Ha EJu, Baek JH, Na DG. Risk Stratification of Thyroid Nodules on Ultrasonography: Current Status and Perspectives. Thyroid. 2017; 27(12): 1463–1468, doi: 10.1089/thy.2016.0654, indexed in Pubmed: 28946821.
28. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542(7639): 115–118, doi: 10.1038/nature21056, indexed in Pubmed: 28117445.
29. 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, doi: 10.1002/jmri.26534, indexed in Pubmed: 30575178.
30. Zhu Y, Fu Z, Fei J. An image augmentation method using convolutional network for thyroid nodule classification by transfer learning. In Proceedings of the 3rd IEEE International Conference on Computer and Communication (pp. 1819–1823) Chengdu, China, 13–16 December 2017.
31. Chi J, Walia E, Babyn P, et al. Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network. J Digit Imaging. 2017; 30(4): 477–486, doi: 10.1007/s10278-017-9997-y, indexed in Pubmed: 28695342.
32. Song W, Li S, Liu Ji, et al. Multitask Cascade Convolution Neural Networks for Automatic Thyroid Nodule Detection and Recognition. IEEE J Biomed Health Inform. 2019; 23(3): 1215–1224, doi: 10.1109/JBHI.2018.2852718, indexed in Pubmed: 29994412.
33. Nguyen DT, Kang JK, Pham TD, et al. Ultrasound Image-Based Diagnosis of Malignant Thyroid Nodule Using Artificial Intelligence. Sensors (Basel). 2020; 20(7), doi: 10.3390/s20071822, indexed in Pubmed: 32218230.