European Journal of Radiology 127 (2020) 108992

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Differential Diagnosis of Benign and Malignant Thyroid Nodules Using Deep Learning Radiomics of Thyroid Ultrasound Images

Hui Zhou[a,b,c,1], Yinhua Jin[a,1], Lei Dai[a,1], Meiwu Zhang[a], Yuqin Qiu[a], Kun wang[b,c,**], Jie Tian[b,c,d,**], Jianjun Zheng[a,*]

[a] HwaMei Hospital, University of Chinese Academy of Sciences, 41 Xibei Street, Ningbo, 315010, China
[b] CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing, 100190, China
[c] School of Artificial Intelligence, University of Chinese Academy of Sciences, No.19 (A) Yuquan Road, Shijingshan District, Beijing, 100049, China
[d] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, 100191, China

ARTICLE INFO

Keywords:
Thyroid nodules
Thyroid ultrasound
Deep learning
Ultrasound Radiomics
Diagnosis

ABSTRACT

Purpose: We aimed to propose a highly automatic and objective model named deep learning Radiomics of thyroid (DLRT) for the differential diagnosis of benign and malignant thyroid nodules from ultrasound (US) images.

Methods: We retrospectively enrolled and finally include US images and fine-needle aspiration biopsies from 1734 patients with 1750 thyroid nodules. A basic convolutional neural network (CNN) model, a transfer learning (TL) model, and a newly designed model named deep learning Radiomics of thyroid (DLRT) were used for the investigation. Their diagnostic accuracy was further compared with human observers (one senior and one junior US radiologist). Moreover, the robustness of DLRT over different US instruments was also validated. Analysis of receiver operating characteristic (ROC) curves were performed to calculate optimal area under it (AUC) for benign and malignant nodules. One observer helped to delineate the nodules.

Results: AUCs of DLRT were 0.96 (95% confidence interval [CI]: 0.94-0.98), 0.95 (95% confidence interval [CI]: 0.93-0.97) and 0.97 (95% confidence interval [CI]: 0.95-0.99) in the training, internal and external validation cohort, respectively, which were significantly better than other deep learning models ($P < 0.01$) and human observers ($P < 0.001$). No significant difference was found when applying DLRT on thyroid US images acquired from different US instruments.

Conclusions: DLRT shows the best overall performance comparing with other deep learning models and human observers. It holds great promise for improving the differential diagnosis of benign and malignant thyroid nodules.

1. Introduction

Thyroid nodules are defined as discrete lesions within the thyroid gland, radiologically distinct from surrounding thyroid parenchyma [1]. They are becoming increasingly common in clinical practice, being detected in up to 65% of the general population [2]. Among the large number of detected nodules, most of them are benign, clinically insignificant, and safely managed by the surveillance program. However, approximately 10% of patients presenting thyroid nodules are at risk of malignancy [3], and the incidence of thyroid cancer has continuously increased worldwide [4]. Therefore, the accurate identification of benign and malignant thyroid nodules is vital in clinical decision-making and management.

Fine-needle aspiration (FNA) biopsy has gained worldwide

Abbreviations: DLRT, deep learning Radiomics of thyroid; US, ultrasound; CNN, convolutional neural network; TL, transfer learning; ROC, receiver operating characteristic curve; AUC, area under curve; CI, confidence interval; FNA, Fine-needle aspiration; CT, computed tomography; MRI, magnetic resonance imaging; ROI, region-of-interest; CAM, class activation map; SD, standard deviation; PPV, positive predictive values; NPV, negative predictive values; LR+, positive diagnostic likelihood ratio; LR-, negative diagnostic likelihood ratio

* Corresponding author at: HwaMei Hospital, University of Chinese Academy of Sciences, 41 Xibei Street, Ningbo, 315010, China.
** Corresponding authors at: CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing, 100190, China.
E-mail addresses: kun.wang@ia.ac.cn (K. wang), jie.tian@ia.ac.cn (J. Tian), zhengjianjun@ucas.ac.cn (J. Zheng).
[1] These authers contributed equally to this work.

https://doi.org/10.1016/j.ejrad.2020.108992
Received 19 November 2019; Received in revised form 27 March 2020; Accepted 5 April 2020

acceptance as the golden standard for the definitive diagnosis of benign and malignant thyroid nodules [1,5]. However, it is invasive and limited by specimen collection and operator experience [6]. With the continuous improvement of ultrasonic instruments, the application of high-frequency ultrasound (US) to small organs has become an important part of the non-invasive ultrasound diagnosis, particularly in the field of thyroid imaging. Currently, US is the first clinical choice of thyroid nodules screening, because of its high sensitivity, non-radioactivity, easy-to-operate, and rapid diagnostic work-up. The American Thyroid Association (ATA) 2015 guidelines emphasize the significance of ultrasonography in detecting thyroid nodules [1], and it is also recommended by the 2012 European Society of Oncology (ESMO) thyroid cancer guidelines as the first-line diagnostic method.

US features can be utilized to differentiate malignancies from benign thyroid nodules. For example, a cystic or spongiform appearance usually suggest a benign nodule only needed a long-term follow-up, whereas the solid composition, hypoechogenicity, infiltrative or irregular margins, and micro-calcifications are generally considered to be risk factors of malignancy [7,8] which may need further treatment, such as resection. Some studies demonstrated that a combination of US features provided certain diagnostic accuracy [9]. However, many other studies indicated a considerable overlap of US features appearing in both benign and malignant nodules [10,11]. The sensitivity and specificity of using US for thyroid cancer diagnosis varied from 27% to 63% and 78.0% to 96.6% in various studies [1,8,12]. This is likely due to interobserver variability in assigning sonographic features to nodules and that US is highly operator dependent. Different examiners, different US instruments, and different definitions of US features will eventually affect the diagnostic accuracy. As a result, US remains highly subjective and depends on clinical experience.

At present, an emerging technology named Radiomics based on machine learning can extract and analyze thousands of quantitatively calculated image features (also called Radiomics features) from medical images, which has the potential to reveal disease characteristics that is impossible for human to recognize by naked eyes in daily practice [13]. Radiomics has been widely used for analyzing CT and MR images with impressive effectiveness [14–17], and these study enhanced the clinical practice or assist the radiologist. But its applications in US are still rarely reported [9,18–21]. Therefore, it is worthy of investigating whether a Radiomics approach can make better use of thyroid ultrasound images and achieve more accurate diagnosis of differentiating malignant from benign thyroid nodules.

A few Radiomics studies have been conducted on ultrasound images for classifying benign and malignant thyroid nodules [9,18,21]. However, they were limited either by the relatively small number of patient population [9,18], or lacking cytology results as gold standard [18], or too much labor work for operators [9,21]. Most of them still utilized human defined image features to establish the diagnostic model to classify benign and malignant nodules [9,21], which inevitably brought subjective and experience dependent bias.

Here, we developed a convolutional neural network (CNN) based transfer learning method tailed for the quantitative analysis of thyroid ultrasound images. It is a deep learning approach, named as DLRT (deep learning Radiomics of thyroid), that does not require complicated manual segmentations of thyroid nodule boundaries [22].

2. Materials and Methods

2.1. Design and overview

This was a retrospective study. A new diagnostic approach named DLRT was used for the differential diagnosis of benign and malignant thyroid nodules. FNA biopsy was used as the golden standard, and DLRT was compared with two other deep learning models as well as two radiologists. This retrospective study was approved by the Ethics Committee of Ningbo No.2 hospital in China. The requirement for informed patient consent was obtained.

2.2. Patient enrollments

From January 2017 to March 2018, 2284 consecutive thyroid patients who underwent US examination and US-guided FNA biopsy were recruited. The inclusion criteria were as follows: (1) no previous fine-needle aspiration biopsy, (2) no previous surgical treatment, and (3) conventional US examination before the biopsy, with thyroid nodule indication in recorded US images. The exclusion criteria were: (1) nodule diameter < 10 mm, (2) unqualified histology with ambiguous diagnostic findings (too few cells or atypical pathology). Demographic information, imaging examination, and clinical baseline characteristics were collected from the hospital PACS (eWorldUIS, version 3.2) workstation.

2.3. US examination

All thyroid nodules were assessed using either My Lab90 (Esaote, Genoa, Italy) or IU22 (Philips, Eindhoven, Netherlands) ultrasound instruments with their default modes of thyroid examination. They both equipped with 5-13 MHz linear probes. All patients were examined in supine position with extended neck and good exposure of the lower thyroid margins. Both thyroid lobes and isthmus were scanned in longitudinal and transverse planes. Longitudinal and transverse images of the thyroid were acquired by following the American College of Radiology accreditation standards [1]. All images were saved as DICOM in the PACS workstation. Two senior thyroid radiologists with at least seven years of clinical experience performed all these US examinations.

2.4. US-guided fine-needle aspiration

US-guided FNA was performed for highly suspicious nodules according to ATA guideline by using a PTC needle (22 G × 70 mm, Hakko Co., Ltd., Tokyo, Japan) right after the US examination. All biopsy specimens were examined by cytopathologists with more than six years of work experience.

2.5. DLRT development and validation

To train and validate DLRT for the differential diagnosis of benign and malignant thyroid nodules, we successfully enrolled 1629 patients from the Ningbo No.2 hospital in China, with ultrasound images of 1003 benign nodules and 642 malignant nodules. We also enrolled an external validation of 105 thyroid nodules with 75 benign nodules and 30 malignant nodules to further test the performance of the model from HwaMei Hospital. Moreover, cytology results of all nodules were obtained by US-guided FNA biopsies and used as the golden reference. The performance of DLRT was compared with a basic CNN model, a TL model, as well as two ultrasound radiologists.

To develop the DLRT model, 1629 enrolled patients from the Ningbo No.2 hospital were randomly divided into the training cohort (1097, two-thirds of patients) and internal validation cohort (532, one-third of patients). US images and FNA biopsies of the training cohort were used to optimize a large number of parameters in the DLRT model, whereas the validation cohort was to evaluate the performance of the trained model. In the training cohort, to reduce the potential bias caused by the unbalanced data and the limited size of population, we applied the data augmentation strategy before the training procedure [23]. Thyroid images in the training cohort were augmented through a number of random transformations, which increased the training data pool and decreased the overfitting of the generated radiomics model. After the model was set, we test the performance with an external validation cohort.

For applying DLRT, we designed a simple manual initiation by defining multiple region-of-interests (ROIs). After manual indication of

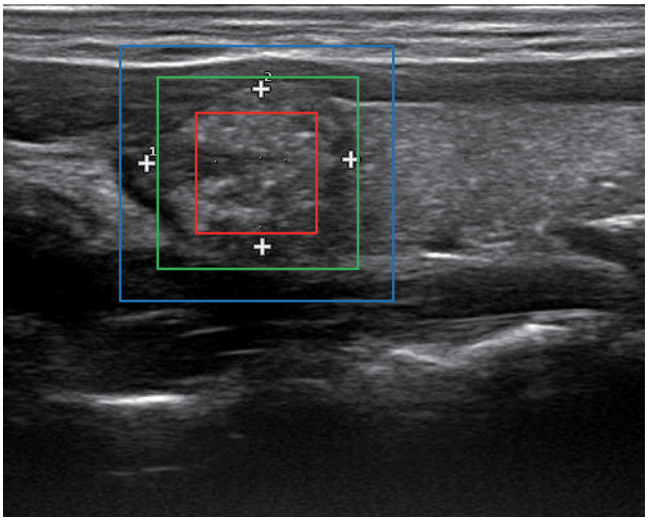

Fig. 1. Illustration of the region-of-interests (ROIs). Three different size of ROIs (sizes: 150 × 150 pixels for the red, 200 × 200 pixels for the green and 250 × 250 pixels for the blue) were automatically generated by a simple designed manual initiation.

the center point of a node, a square ROI will be cropped. For each thyroid nodule, three square ROIs (sizes: 150 × 150 pixels, 200 × 200 pixels and 250 × 250 pixels) whose sizes were based on statistics, were automatically generated after one mouse click on the nodule centre area (Fig. 1). Then, the corresponding three cropped images were used as input layers to trigger the DLRT model (Fig. 2). DLRT adopted the CNN architecture and transfer learning strategy [24]. It consisted of four hidden layers. The first three layers were transferred from one of our previous studies without any modification (Fig. 2) [22], whereas the last hidden layer was fine-tuned using enrolled thyroid US images. This layer contained 32 feature maps, and the size of the convolution filter and the max pooling was 3 × 3 pixels and 2 × 2 pixels, respectively. Finally, a fully-connected layer with 32 nodes was connected to every neuron in the last three pooling layers, and the probability (a malignancy score) of the binary classification (benign or malignant) can be calculated in the output layer (Fig. 2). The DLRT architecture was based on Keras library, and we used ReLU activation function in all the convolutional layers, a drop out of 0.5 was adopted in the last fully connected layer, the sum of squared error was used as loss function. The detailed introduction and the mathematical description of DLRT are demonstrated in the Supplementary Materials. The datasets and some codes generated during and/or analysed during the current study are available on reasonable request.

2.6. Comparison between different Radiomics models

As DLRT adopted both CNN architecture and transfer learning strategy, we compared the performances of the basic CNN model, the TL model, and DLRT. The Basic CNN model had exactly the same network architecture with DLRT (four hidden layers followed with a fully connected layer), but all parameters of every layer were trained by US images and FNA histological results of the training cohort. Differently, the TL model employed transferred parameters for the first three layers from another study [22] without using any data from our training cohort. Only the parameters of the last hidden layer were trained by the training cohort, which was the same as DLRT. That means the model need to be adapted or refined on the input-output pair data available for

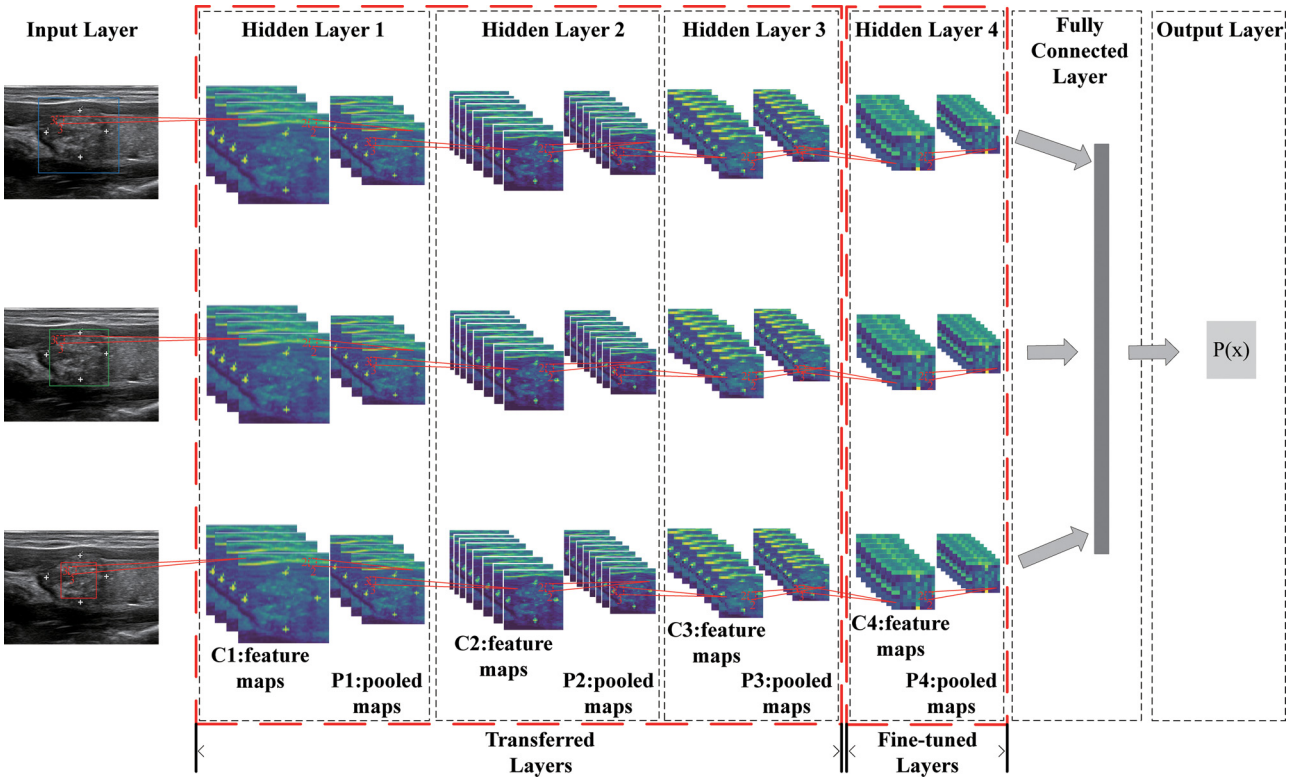

Fig. 2. Illustration of the deep learning radiomics of thyroid (DLRT) flowchart. Each row (channel) of the figure stands for a basic convolutional neural network (CNN) model, which consists of four convolutional layers. Transfer learning (TL) model is based on the CNN model, which consists of three transferred layers and one fine-tuning layer. The DLRT model consists of three channels, and each channel adopted the TL strategy. As for DLRT model, three different size of region-of-interests (ROIs) were sent into the input layer, followed by three transferred layers, a fine-tuned layer was connected to the transferred layers, and then a fully connected layer was connected to the fine-tuned layer to combine different features extracted by the previous layers, at last an output layer was used to calculate the probability for the classification. The parameters of the DLRT model were automatically optimized by using all ultrasound thyroid images in the training cohort.

the task of DLRT. This part involved freezing the first three hidden layers and fine-tuning the rest part of the model. The biggest difference between these two models and DLRT was that they only can use one ROI as the input layer for each US image, whereas DLRT was designed to take three different ROIs from a single image as the input. Therefore, we chose the middle size ROI (Fig. 1, green box) as the input for the basic CNN and TL models.

All Radiomics models were developed and validated on an Ubuntu 16.04 operating system (Canonical Group Limited, London, United Kingdom) with a graphics processing unit of GeForce 980 Ti (NVidia Corporation, Santa Clara, California, United States), whose graphics memory is 6 G Bits. The deep learning framework is Keras (version 1.4, François Chollet, California, United States), whose backend is Tensorflow (version 1.3, Google, Inc., California, United States).

2.7. Comparison between Radiomics and human observers

Thyroid nodule images from internal and external validation cohort was given to two ultrasound radiologists who were blind to the FNA histological results and did not review any of the images that were acquired during the original ultrasound examination. One has more than 12 years of experience in thyroid diagnosis, the other has only three years of experience. Their diagnostic performances were compared with DLRT, the basic CNN model, and the TL model.

2.8. Comparison between different ultrasound instruments

As all enrolled US images were acquired by two ultrasound instruments (Esaote and Philips), we compared the diagnostic performance of DLRT over different systems, in order to evaluate its generalization ability.

2.9. Visualization of DLRT

To understand how DLRT interpret US images for thyroid nodule classification, we applied a deep learning visualization technique called Class Activation Map (CAM) [25], which produced a heat map of class activation over input images. Therefore, based on the quantitative analysis of DLRT, original grey scale US images were transferred into pseudo-colored images using this approach. In the view of DLRT, pixels with warmer colors (e.g., red and yellow) indicated stronger correlation with the nodule classification than pixels with colder colors (e.g., blue and green).

2.10. Statistical analysis

Descriptive statistics were summarized as mean ± standard deviation (SD) or with 95% confidence interval (CI). Analysis of receiver operating characteristic (ROC) curves were performed to calculate optimal area under it (AUC) for benign and malignant nodules. Differences between various AUCs were compared by using a Delong test [26]. Sensitivity, specificity, positive and negative predictive values (PPV, NPV), positive and negative diagnostic likelihood ratios (LR+, LR-) were also calculated. *P* values less than 0.05 indicated statistical significance. The statistical analyses were performed using SPSS software for Windows, version 20.0 (SPSS, Chicago, IL).

3. Results

3.1. Baseline characteristics

A total of 2317 nodules from 2284 potentially eligible patients were retrospectively enrolled in this study (Fig. 3). Among them, 423 patients were excluded due to too few cells from US-guided FNA for pathology, atypical pathology. Another 127 patients were excluded, because their nodule diameter was less than 10 mm. Finally, 1734 patients with 1750 thyroid nodules were enrolled, and FNA was performed for each enrolled nodules.

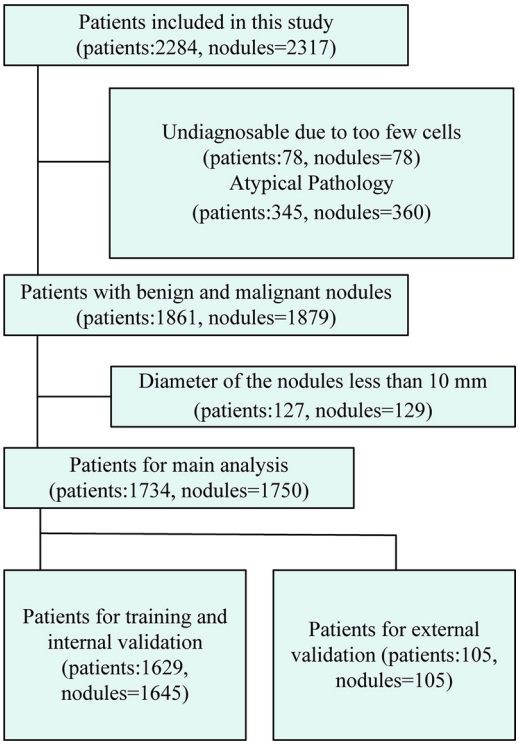

Fig. 3. The results of the patient enrolments. In total, 1734 out of 2284 patients (1750 out of 2317 nodules) from two hospitals were enrolled in this study. 1629 out of 1734 patients (1645 out of 1750 nodules) were for training and internal validation from one hospital, 105 out of 1734 patients (105 out of 1750 nodules) were for external validation from another hospital.

After randomization of enrolled 1734 patients (1750 nodules), 1097 nodules (428 malignant, 39.0%) were assigned to the training cohort, 548 nodules (214 malignant, 39.1%) composed the internal validation cohort and the other 105 nodules (30 malignant, 28.6%) composed the external validation cohort. Their characteristics are summarized in Table 1. Features of these nodules are summarized in Table 2.

3.2. Diagnostic accuracy of DLRT, the basic CNN, and the transfer learning model

Both in training and validation (internal and external) cohorts, DLRT demonstrated the highest diagnostic accuracy comparing with the other two models for the differential diagnosis of benign and malignant thyroid nodules (Fig. 4A, B and C). Differences of AUCs were all statistically significant ($P < 0.01$, Table 3). AUCs of DLRT reached 0.96,

Table 1
Baseline Characters of Patients

Variables	All patients	Training cohort	Internal Validation	External Validation
Number of patients (%)	1734	1097 (63.3%)	532 (30.6%)	105(6.1%)
Age (y)	47.3 ± 12.9	48.6 ± 12.4	46.8 ± 12.8	47.9 ± 12.5
Gender (%)				
Male	421	264(62.7%)	132(31.3%)	25(6.0%)
Female	1313	833(63.4%)	400(30.5%)	80(6.1%)
Nodule type (%)				
Benign	1078	669 (62.1%)	334 (31.0%)	75(6.9%)
Malignant	672	428(63.7%)	214(31.8%)	30(4.5%)

Qualitative variables are in n (%), and quantitative variables are in mean ± SD, when appropriate.

Table 2
Comparison of features of benign and malignant thyroid nodules

Features	Benign nodules (n = 1078)	Malignant nodules (n = 672)	P value
Size (cm)			
1.0-2.0	672 (62.3)	438 (65.2)	< 0.001
≥ 2.0	406 (37.7)	234 (34.8)	
Echogenicity			
anechoic	34 (3.2)	16 (2.4)	< 0.001
isoechoic	498 (46.2)	122 (18.2)	
hypoechoic	437 (40.5)	516 (76.7)	
hyperechoic	109 (10.1)	18 (2.7)	
Margins			
Well-defined	735 (68.2)	238 (35.4)	< 0.001
Ill-defined	343 (31.8)	434 (64.6)	
Internal composition			
solid	45 (4.2)	566 (84.3)	< 0.001
cystic	811 (75.2)	24 (3.5)	
mixed	222 (20.6)	82 (12.2)	
Shape			
regular	770 (71.4)	136 (20.3)	< 0.001
irregular	308 (28.6)	536 (79.7)	
Calcifications			
absent	736 (68.3)	165 (24.5)	
micro	126 (11.7)	385 (57.3)	< 0.001
macro	134 (12.4)	69 (10.2)	
micro + macro	82 (7.6)	53 (8.0)	

Qualitative variables are in n (%), when appropriate.

0.95 and 0.97 in training internal and external validation cohorts, respectively, which were 0.09, 0.10 and 0.10 higher than these of the TL model who offered the second highest AUCs. The basic CNN model offered the worst AUCs in both training, internal and external validation cohorts, which were 0.82, 0.81 and 0.82, respectively. Sensitivities of DLRT reached 90.1%, 89.3% and 89.5% in training internal and external validation cohorts, respectively, which were much higher than these of the TL model and the basic CNN model ($P < 0.01$, Table 3).

3.3. Comparison between Radiomics and human observers

A senior and a junior US Radiologist who were blind to cytology data performed differential diagnosis using US images from the internal and external validation cohort. Without surprise, the senior observer (Doctor A) outperformed the junior one (Doctor B) with a significant specificity improvement of 12.9% and 13.0% in internal and external validation cohorts, respectively (Table 3, $P < 0.05$). However, both human observers provided lower sensitivity and specificity than all three Radiomics models (Table 3).

3.4. Comparison between different ultrasound instruments

As DLRT showed the best performance over other approaches, we further investigated whether its diagnostic accuracy was influenced by different US instruments. In the validation cohort, 284 and 264 thyroid nodules were scanned by two instruments, respectively. DLRT offered almost the same AUC (0.96 with 95%CI: 0.94-0.98 vs. 0.95 with 95%CI: 0.93-0.97) for the two subgroups. Their ROC curves overlapped each other (Fig. 5A), revealing similar performance for using different instruments. Statistical comparisons of AUC, sensitivity, and specificity also confirmed that no significant difference (all $P > 0.05$) was found, if DLRT was applied to US images acquired by different scanners (Fig.5B).

3.5. Visualization of DLRT

The Radiomics features automatically extracted and learned by DLRT were mapped and visualized by pseudo-color on corresponding pixels (Fig. 6). The obtained heat map images revealed which parts of a US image were strongly associated with the decision-making of DLRT. Then, we learned two common patterns from these images. First, for both benign (Fig. 6A and B) and malignant (Fig. 6C and D) cases, DLRT did not only pay attentions on nodule internal areas, but also analyzed external parenchyma adjacent to the nodule boundary. Second, for the

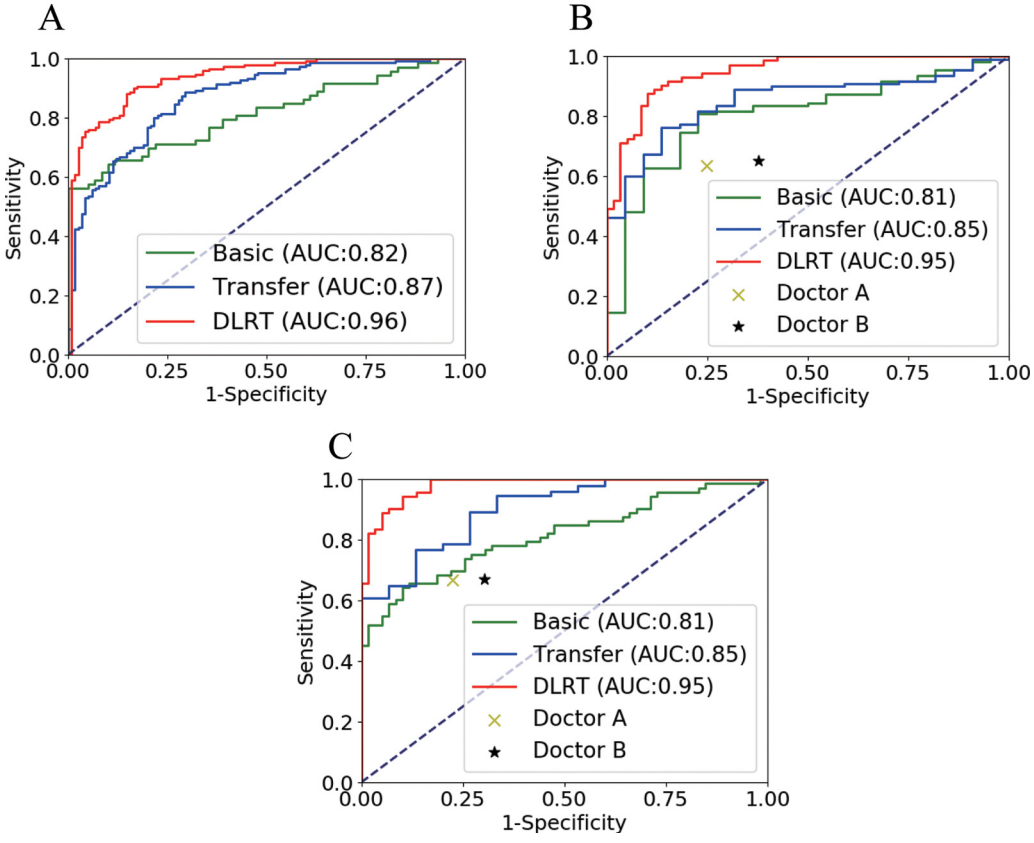

Fig. 4. Comparison of receiver operating characteristic (ROC) curves, area under the curve (AUC), sensitivity and specificity between radiomics models (DLRT, the basic CNN, and the transfer learning model) and human observers (a senior and a junior US Radiologist) for the differential diagnosis of benign and malignant thyroid nodules in training, internal and external validation cohorts, respectively. (A) ROC curves for radiomics models (DLRT, the basic CNN, and the transfer learning model) in training cohorts, (B) (C) ROC curves for radiomics models (DLRT, the basic CNN, the transfer learning model) versus human observers (a senior and a junior US Radiologist) in internal and external validation cohorts.

Table 3
Diagnostic Performance of DLRT, the transfer learning, the basic CNN and two Radiologist for the differential diagnosis of benign and malignant thyroid nodules in training internal and external validation cohorts.

		AUC	Sensitivity %	Specificity %	PPV %	NPV %	LR+	LR-
DLRT	T	0.96 (0.94-0.98)	90.1 (86.6-93.6)	82.7 (79.5-85.9)	87.7 (83.4-92.0)	86.5 (82.1-90.9)	5.2 (4.6-5.8)	0.12 (0.07-0.17)
	IV	0.95 (0.93-0.97)	89.3 (86.1-92.5)	83.5 (80.1-86.9)	87.4 (83.1-91.7)	87.2 (82.5-91.9)	5.4 (4.8-6.0)	0.13 (0.09-0.17)
	EV	0.97 (0.95-0.99)	89.5 (86.3-92.7)	84.1 (80.7-87.5)	87.5 (83.2-91.2)	87.5 (82.8-92.2)	5.5 (4.9-6.1)	0.14 (0.10-0.18)
Transfer Learning	T	0.87** (0.85-0.89)	78.4 (75.2-81.6)	80.2 (76.8-83.6)	80.8 (76.5-85.1)	82.7 (78.2-87.1)	4.0 (3.5-4.5)	0.27 (0.22 – 0.32)
	IV	0.85** (0.83-0.87)	78.6 (75.4-81.8)	81.4 (78.2-84.6)	80.1 (75.8-84.4)	81.4 (76.9-85.9)	4.2 (3.5-4.9)	0.26 (0.21- 0.31)
	EV	0.87** (0.85-0.89)	78.3 (75.1-81.5)	81.2 (78.0-84.4)	80.0 (75.7-84.3)	81.1 (76.6-85.6)	4.3 (3.6-5.0)	0.25 (0.20- 0.30)
Basic CNN	T	0.82** (0.79-0.85)	67.3 (63.8-70.8)	82.4 (78.7-86.1)	78.7 (74.4.8-83.0)	79.8 (75.5-84.1)	3.8 (3.3-4.3)	0.40 (0.34-0.46)
	IV	0.81** (0.78-0.84)	64.7 (61.1-68.3)	88.9 (85.4-92.4)	78.2 (73.8-82.6)	79.3 (74.7-83.9)	5.8 (5.2-6.4)	0.41 (0.38-0.45)
	EV	0.82** (0.79-0.85)	65.1 (61.5-68.7)	88.2 (84.7-91.7)	78.0 (73.6-82.4)	79.1 (74.5-83.7)	5.9 (5.3-6.5)	0.42 (0.39-0.46)
Doctor A	IV	-	63.6 (60.1-67.1)	75.1 (71.4-78.8)	79.3 (74.7-83.9)	63.6 (59.6-67.6)	2.1 (1.7-2.5)	0.39 (0.35-0.43)
	EV	-	64.2 (60.7-67.7)	75.5 (71.8-79.2)	78.1 (73.5-82.7)	64.2 (60.2-68.2)	1.9 (1.5-2.3)	0.38 (0.34-0.42)
Doctor B	IV	-	65.2 (61.5-68.9)	62.2 (58.8-65.6)	75.4 (70.8-80.0)	62.5 (58.4-66.6)	1.8 (1.4-2.2)	0.58 (0.54-0.62)
	EV	-	65.0 (61.3-68.7)	62.5 (59.1-65.9)	75.1 (70.5-79.7)	62.3 (58.2-66.4)	1.9 (1.5-2.3)	0.57 (0.53-0.61)

Statistical quantifications were demonstrated with 95% confidence interval.
Abbreviations: T, training cohort; IV, internal validation cohort; EV, external validation cohort; AUC, area under the receiver-operator-characteristic curve; PPV, positive predictive value; NPV, negative predictive value; LR+, positive diagnostic likelihood ratio; LR-, negative diagnostic likelihood ratio; T, training cohort; V, validation cohort.
AUC of DLRT was statistically compared to AUC of the transfer learning and the basic CNN, respectively (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$).

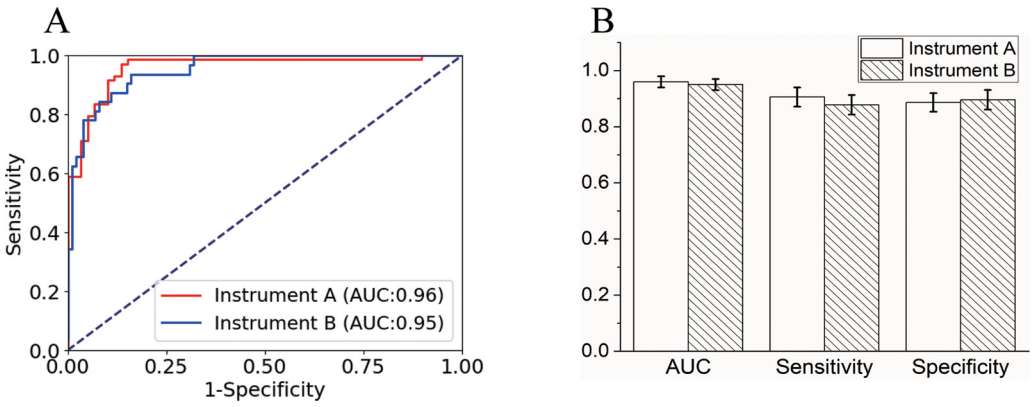

Fig. 5. Comparison of receiver operating characteristic (ROC) curves, area under the curve (AUC), sensitivity and specificity between two different ultrasound instruments for the differential diagnosis of benign and malignant thyroid nodules. (A) ROC curves for two different ultrasound instruments in validation cohorts, (B) AUC, sensitivity and specificity for two different ultrasound instruments in validation cohorts.

"easy" case, when the US image exhibited typical malignant characteristics (Fig. 6C), DLRT showed similar analytical patterns on benign and malignant images (Fig. 6A to C). However, for the "difficult" case, when a malignant nodule showed similar appearance with benign, DLRT focused more on nodule adjacent parenchyma than its internal area (Fig. 6D).

4. Discussion

In this study, we developed and validated three deep learning based Radiomics models, the basic CNN model, the TL model, and DLRT, for the differential diagnosis of benign and malignant thyroid nodules by automatic and quantitative analysis of thyroid US images. We retrospectively enrolled US images and FNA cytology data from 1750 thyroid nodules to compare their performances. Their diagnostic accuracy was further compared with human observers. Moreover, the robustness of DLRT over different ultrasound imaging instruments was also investigated.

In both training, internal and external validation cohorts, DLRT demonstrated the highest diagnostic accuracy comparing with the basic CNN and TL model. AUCs of DLRT researched 0.96 (95%CI: 0.94-0.98), 0.95 (95%CI: 0.93-0.97) and 0.97 (95%CI: 0.95-0.99) in training internal and external validation, respectively, which were significantly better than the other two methods (both $P < 0.01$). The TL model showed the second highest diagnostic accuracy, and the basic CNN model was the worst. Besides these, another unique characteristic of DLRT was that its sensitivity was better than other two models, which is favorable for clinical screening of malignant nodules. These results indicated that the multiple ROIs strategy made critical contribution for accuracy improvement, because it enabled independent analysis targeting regions inside and outside each thyroid nodule. Furthermore, with partial adjustment by transfer learning, a pre-trained deep learning Radiomics model designed for US images [22] can be effectively applied for another US diagnosis scenario, which was even better than re-training the entire model from scratch.

After the comparison with human observers, DLRT offered significant better sensitivity and specificity than both senior and junior US radiologists did ($P < 0.001$). The comparison between different US instruments also revealed that DLRT had a consistent performance, regardless the input US images were acquired by which scanner. All these findings further proved that the DLRT effectively analyzed thyroid US images and achieved accurate and reliable differential diagnosis of benign and malignant thyroid nodules. After observing its analytical pattern on transferred heat maps, we recognized that the nodule surrounding adjacent parenchyma was vital for classification, especially for these challenging cases in human eyes. This deep learning visualization technique is likely to assist radiologists for more efficient interpretation of thyroid US images.

Our work demonstrated several advantages over other studies attempted to differentiate malignant and benign nodules using computer-aided analysis on thyroid US images [9,18,21]. The majority of those studies used human-defined US features and machine learning based classifier, which inevitably brought intensive labor work to extract features from US images. Therefore, the study with the largest array of features only involved 12 features for classification [21]. DLRT was a highly automatic end-to-end approach. It only required one mouse click

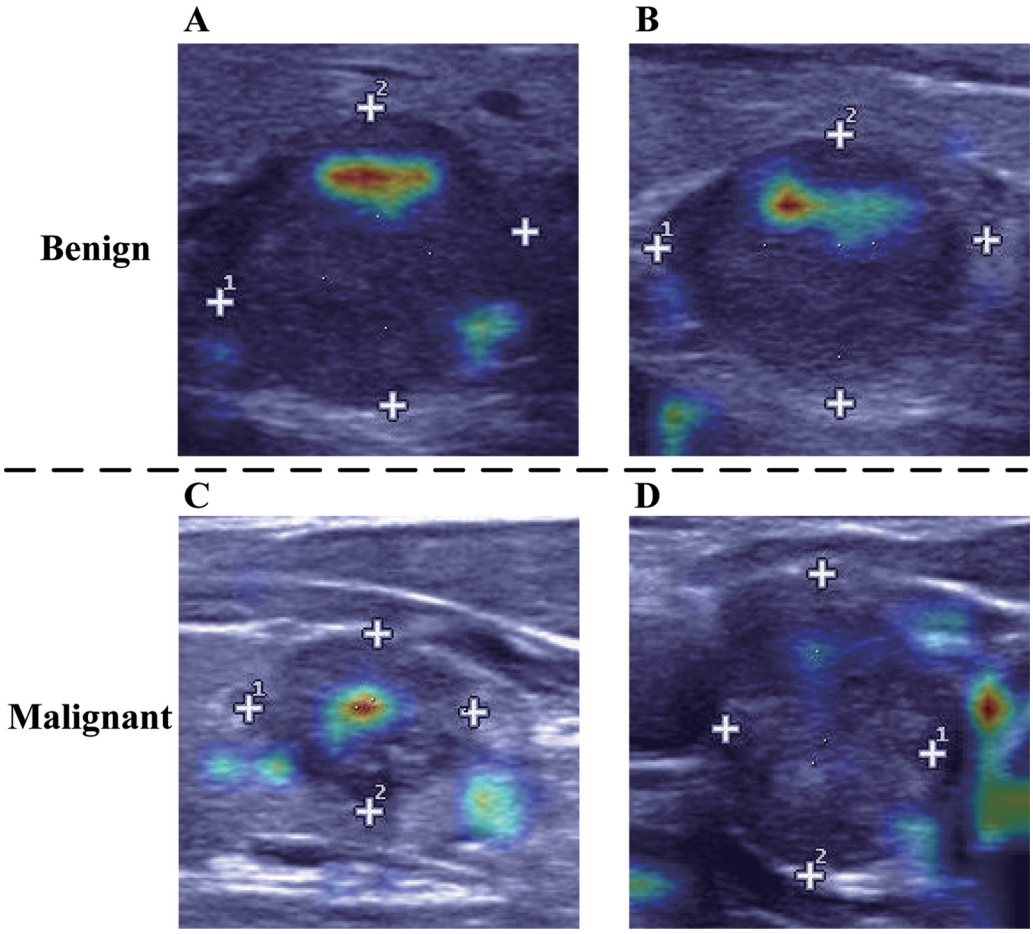

Fig. 6. Visualization of deep learning radiomics of thyroid (DLRT). (A) and (B) heat maps of the DLRT model based on two benign thyroid nodule images, (C) and (D) heat maps of the DLRT model based on two malignant thyroid nodule images.

on the nodule center as the manual trigger. Then, it automatically extracted thousands of computer-defined features and adopted deep learning based classifier to optimize the diagnosis model. Thus, it can be seamlessly integrated into the conventional work-flow of thyroid US examinations without extra time and labor cost. This DLRT method could be incorporated into ultrasound devices to provide auxiliary diagnosis results for clinical observers. Besides that, there was only one study also adopted deep learning based transfer learning for differentiating thyroid nodules [18]. However, it had a much smaller population size, with only 428 and 164 thyroid US images in training and validation cohorts. It did not use FNA biopsy as gold reference.

The major limitation in our study was that the data came from a single center retrospectively. The performance of DLRT needs to be further validated in a multicenter perspective study. A larger dataset acquired from different hospitals with more types of US instruments is necessary for consisting a more comprehensive training cohort, so that the accuracy and reliability of DLRT can be continuously improved for each US scanner, as well as the question of whether it will have worse performance on certain US scanners can be properly addressed.

5. Conclusions

In conclusion, DLRT achieved the most accurate differential diagnosis of benign and malignant thyroid nodules comparing with other deep learning models and human observers. Its performance was not affected by different US instruments. It holds a good potential for improving the overall diagnostic efficacy in routine thyroid US examinations.

Contributors

Study conception: JJZ, JT, and KW; data collection: MWZ, YQQ, HZ; data analysis: HZ, YHJ, LD, MWZ and YQQ; administrative support: JJZ, JT; manuscript drafting: HZ, KW, LD, and MWZ; All authors read and approved the final version of the manuscript.

Declaration of Competing Interest

None.

Acknowledgements

The work is supported by Ministry of Science and Technology of China under Grant No. 2017YFA0205200, National Natural Science Foundation of China under Grant No. 81227901, 81527805, and 61671449, Chinese Academy of Sciences under Grant No. GJJSTD20170004, KFJ-STS-ZDTP-059, YJKYYQ20180048, and XDB32030200, Ningbo Technology and Public Welfare Foundation of China under Grant No. 2017C50070, Ningbo Natural Science Foundation of China under Grant No. 2017A610207, Reasearch Foundation of Hwa Mei Hospital, University of Chinese Academy of Sciences, China under Grant No. 2020HMKY50. We would like to thank Doctor Baowen Zheng and Congde Chen for their help with the pathological diagnoses.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the

online version, at doi:https://doi.org/10.1016/j.ejrad.2020.108992.

References

[1] B.R. Haugen, E.K. Alexander, K.C. Bible, G.M. Doherty, S.J. Mandel, Y.E. Nikiforov, F. Pacini, G.W. Randolph, A.M. Sawka, M. SchlumbergeR, K.G. Schuff, S.I. Sherman, J.A. Sosa, D.L. Steward, R.M. Tuttle, L. Wartofsky, 2015 American Thyroid Association ManagementGuidelines for Adult Patients with Thyroid Nodulesand Differentiated Thyroid Cancer, Thyroid 26 (1) (2015) 1–133.

[2] G. Russ, S. Leboulleux, L. Leenhardt, L. Hegedüs, Thyroid incidentalomas: epidemiology, risk stratification with ultrasound and workup, Eur. Thyroid J. 3 (3) (2014) 154–163.

[3] T.E. Angell, R. Maurer, Z. Wang, M.I. Kim, C.A. Alexander, J.A. Barletta, C.B. Benson, E.S. Cibas, N.L. Cho, G.M. Doherty, P.M. Doubilet, M.C. Frates, A.A. Gawande, J.F. Krane, E. Marqusee, F.D. Moore, M.A. Nehs, P.R. Larsen, E.K. Alexander, A Cohort Analysis of Clinical and Ultrasound Variables Predicting Cancer Risk in 20,001 Consecutive Thyroid Nodules, J. Clin. Endocrinol. Metab. 104 (11) (2019) 5665–5672.

[4] American Thyroid Association (ATA) Guidelines Taskforce on Thyroid Nodules and Differentiated Thyroid Cancer, D.S. Cooper, G.M. Doherty, B.R. Haugen, R.T. Kloos, S.L. Lee, S.J. Mandel, E.L. Mazzaferri, B. McIver, F. Pacini, M. Schlumberger, S.I. Sherman, D.L. Steward, R.M. Tuttle, Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer, Thyroid 19 (11) (2009) 1167–1214.

[5] S.D. Daniel, P.K. Joshua, C.D. James, F. Lyssa, C.K. Giulia, B.L. Richard, M. Bryan, The Impact of Benign Gene Expression Classifier Test Results on the Endocrinologist–Patient Decision to Operate on Patients with Thyroid Nodules with Indeterminate Fine-Needle Aspiration Cytopathology, Thyroid 22 (10) (2012) 996–1001.

[6] W.J. Moon, S.L. Jung, J.H. Lee, D.G. Na, J.H. Baek, Y.H. Lee, J. Kim, H.S. Kim, J.S. Byun, D.H. Lee, Thyroid Study Group, Korean Society of Neuro- and Head and Neck Radiology, Benign and malignant thyroid nodules: US differentiation–multicenter retrospective study, Radiology 247 (3) (2008) 762–770.

[7] H.J. Moon, J.M. Sung, E.K. Kim, J.H. Yoon, J.H. Youk, J.Y. Kwak, Diagnostic performance of gray-scale US and elastography in solid thyroid nodules, Radiology 262 (3) (2012) 1002–1013.

[8] L.R. Remonti, C.K. Kramer, C.B. Leitão, L.C. Pinto, J.L. Gross, Thyroid ultrasound features and risk of carcinoma: a systematic review and meta-analysis of observational studies, Thyroid 25 (5) (2015) 538–550.

[9] S.Y. Kim, E.K. Kim, H.J. Moon, J.H. Yoon, J.Y. Kwak, Application of Texture Analysis in the Differential Diagnosis of Benign and Malignant Thyroid Nodules: Comparison With Gray-Scale Ultrasound and Elastography, AJR, Am. J. Roentgenol. 205 (3) (2015) W343–351.

[10] J.D. Iannuccilli, J.J. Cronan, J.M. Monchik, Risk for malignancy of thyroid nodules as assessed by sonographic criteria: the need for biopsy, J. Ultrasound Med. 23 (11) (2004) 1455–1464.

[11] J.R. Wienke, W.K. Chong, J.R. Fielding, K.H. Zou, C.A. Mittelstaedt, Sonographic features of benign thyroid nodules: interobserver reliability and overlap with malignancy, J. Ultrasound Med. 22 (10) (2003) 1027–1031.

[12] M.C. Frates, C.B. Benson, J.W. Charboneau, E.S. Cibas, O.H. Clark, B.G. Coleman, J.J. Cronan, P.M. Doubilet, D.B. Evans, J.R. Goellner, I.D. Hay, B.S. Hertzberg, C.M. Intenzo, R.B. Jeffrey, J.E. Langer, P.R. Larsen, S.J. Mandel, W.D. Middleton, C.C. Reading, S.I. Sherman, F.N. Tessler, Society of Radiologists in Ultrasound, Management of thyroid nodules detected at US: Society of Radiologists in Ultrasound consensus conference statement, Radiology 237 (3) (2005) 794–800.

[13] H.J. Aerts, E.R. Velazquez, R.T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun. 5 (2014) 4006.

[14] R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: Images Are More than Pictures, They Are Data, Radiology 278 (2) (2016) 563–577.

[15] Y.Q. Huang, C.H. Liang, L. He, J. Tian, C.S. Liang, X. Chen, Z.L. Ma, Z.Y. Liu, Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer, J Clin Oncol 34 (18) (2016) 2157–2164.

[16] J. Shu, Y. Tang, J. Cui, R. Yang, X. Meng, Z. Cai, J. Zhang, W. Xu, D. Wen, H. Yin, Clear cell renal cell carcinoma: CT-based radiomics features for the prediction of Fuhrman grade, Eur. J. Radiol. 109 (2018) 8–12.

[17] X.D. Min, M. Li, D. Dong, Z.Y. Feng, P.P. Zhang, Z. Ke, H.J. You, F.F. Han, H. Ma, J. Tian, L. Wang, Multi-parametric MRI-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: Cross-validation of a machine learning method, Eur. J. Radiol. 115 (2019) 16–21.

[18] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, M. Eramian, Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network, J. Digit Imaging 30 (4) (2017) 477–486.

[19] J.W. Jiang, J. Shi, Q. Zhang, M. Chen, Computer aided diagnosis of Lymphoma based on dual-mode ultrasound radiomics, Ultrasound in Medicine and Biology 45 (S18) (2019).

[20] Q. Zhang, Y. Xiao, J.F. Suo, J. Shi, J.H. Yu, Y. Guo, Y.Y. Wang, H.R. Zheng, Sonoelastomics for Breast Tumor Classification: A Radiomics Approach with Clustering-Based Feature Selection on Sonoelastography, Ultrasound in Medicine and Biology 43 (5) (2017) 1058–1069.

[21] B. Zhang, J. Tian, S. Pei, Y. Chen, X. He, Y. Dong, L. Zhang, X. Mo, W. Huang, S. Cong, S. Zhang, Machine Learning-Assisted System for Thyroid Nodule Diagnosis, Thyroid 29 (6) (2019) 858–867.

[22] K. Wang, X. Lu, H. Zhou, Y. Gao, J. Zheng, M. Tong, C. Wu, C. Liu, L. Huang, T. Jiang, F. Meng, Y. Lu, H. Ai, X.Y. Xie, L.P. Yin, P. Liang, J. Tian, R.Q. Zheng, Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study, Gut 68 (4) (2018) 729–741.

[23] V.C. Nitesh, W.B. Kevin, O.H. Lawrence, SMOte: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[24] H.C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298.

[25] Z. Bolei, K. Aditya, L. Agata, O. Aude, T. Antonio, Learning Deep Features for Discriminative Localization, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, June, 2016, pp. 27–30.

[26] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing areas under two or more correlated reciever operating characteristics curves: a nonparamentric approach, Biometrics 44 (3) (1988) 837–845.