# Deep learning radiomics for non-invasive diagnosis of benign and malignant thyroid nodules using ultrasound images

Zhou, Hui, Wang, Kun, Tian, Jie

Hui Zhou, Kun Wang, Jie Tian, "Deep learning radiomics for non-invasive diagnosis of benign and malignant thyroid nodules using ultrasound images," Proc. SPIE 11319, Medical Imaging 2020: Ultrasonic Imaging and Tomography, 1131908 (16 March 2020); doi: 10.1117/12.2549433

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

# Deep learning radiomics for non-invasive diagnosis of benign and malignant thyroid nodules using ultrasound images

Hui Zhou[a, b], Kun Wang[a, b], Jie Tian*[a, b, c]
[a]CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China; [b]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China; [c] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, China

## ABSTRACT

**Background**: The differential diagnosis of benign and malignant thyroid nodules from ultrasound (US) images remained challengeable in clinical practice. We aimed to develop and validate a highly automatic and objective diagnostic model named deep learning Radiomics of thyroid (DLRT) for the differential diagnosis of benign and malignant thyroid nodules from US images. **Methods**: We retrospectively enrolled US images and corresponding fine-needle aspiration biopsies from 1645 thyroid nodules. A basic convolutional neural network (CNN) model, a transfer learning 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). **Results**: AUCs of DLRT were 0.96 (95% confidence interval [CI]: 0.94-0.98) and 0.95 (95% confidence interval [CI]: 0.93-0.97) in the training and validation cohort, respectively, for the differential diagnosis of benign and malignant thyroid nodules, which were significantly better than other deep learning models ($P < 0.05$) and human observers ($P < 0.05$). **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.

**Keywords:** Thyroid Nodules, Thyroid Ultrasound, Deep Learning, Ultrasound Radiomics, Diagnosis

## 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.

In clinic, Fine-needle aspiration (FNA) biopsy has been treated as the golden standard for the diagnosis of benign and malignant thyroid nodules[1, 5]. However, it is invasive and limited by specimen collection and operator experience[6]. Currently, ultrasound (US) is the first clinical choice of thyroid nodules screening, because of its non-radioactivity, easy-to-operate, and rapid diagnostic work-up[4]. Therefore, US features can be utilized to differentiate malignancies from benign thyroid nodules[7, 8].

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[9]. Radiomics has been widely used for analyzing CT and MR images with impressive effectiveness[10-13], but its applications in US are still rarely reported[14-17]. 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.

Medical Imaging 2020: Ultrasonic Imaging and Tomography, edited by Brett C. Byram, Nicole V. Ruiter,
Proc. of SPIE Vol. 11319, 1131908 · © 2020 SPIE · CCC code: 1605-7422/20/$21 · doi: 10.1117/12.2549433

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.

## 2. METHODOLOGY

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.

### 2.1 Basic CNN model

The CNN architecture contains input, convolution, activation, pooling, forward computation, and back propagation (Fig. 1). Its details are explained as following.

***Input.*** The CNN model starts with the input layer, i.e.

$$C_0 = X \tag{1}$$

Where X stands for the input image data, $C_0$ stands for the output of the input layer.

**Convolution.** The result of convolution between matrix $C_0$ and matrix $W_1$ (the weights, size of $3\times3$) is to let matrix $W_1$ slides on the matrix $C_0$, in other words, matrix $W_1$ and all of the $3\times3$ continuous submatrix of $C_0$ will perform the operation of "corresponding sum of product of elements".

**Activation.** After the operation of convolution, the result will be activated by an activation function, here we adopted the "ReLU" function $f(x)=\max(0,x)$, when the input is negative, the output of the activation function will be zero, and when the input is positive, the result will be equal to the input. Then we could get the output of the first hidden layer

$$C_1 = \partial_1(W_1 * C_0 + b_1) \tag{2}$$

Where $\partial_1$ is the activation function ReLU, $b_1$ denotes the bias, * denotes the convolution operation, $C_1$ is the output of the first convolution layer, called feature map.

**Pooling.** The pooling operation of $C_1$ is as follows

$$P_1 = \text{maxpooling}(C_1) \tag{3}$$

Where $P_1$ is the pooled map. The operation "maxpooling" means only the maximum value of the matrix $C_1$ within disjoint $2\times2$ small matrixes will be kept for next step. All the convolution, activation and pooling operations will be repeated in all four hidden layers

**Forward computation.** The input of the first convolutional layer is the raw data matrix $X$ with a size of $250\times250$, after an operation of convolution which contains a number of 16 filters with the size of $3\times3$, the output of this

convolutional layer will be a number of 16 feature maps with the size of $248\times248$, and then the result will be activated with the activation function of "ReLU".

After the first operation of convolution, there will be a pooling operation. The size of the pooling window is $2\times2$, and we will adopt the max-pooling strategy here. Then the $248\times248$ feature map will be transferred to a $124\times124$ pooled map.

A total of four times of convolution, activation and pooling operations will be executed to complete the computation in turn. When it comes to the last fully connected layer and the output layer, the result will be the possibility.

**Back propagation.** Assuming the loss function of the whole network is $J$, and $J(w,b)=\frac{1}{q}\sum_{j=1}^{q}\frac{1}{2}\left[y^j-p_{(w,b)}(x^j)\right]^2$, where $j$ represents the order of neuron, $q$ means the number of neuron. The most important parameters in the network are the weights $w$ between two neurons and the bias $b$ between two layers. And $x$ is the input of a neuron, $p(x)$ means the actual output of the neuron, while $y$ is the expected output of the neuron.

Therefore, $J$ represents the sum of squared error between the actual output and the expected output. In the end, our task is to make $J$ as small as possible, and to achieve this goal, we need to acquire suitable parameters $w$ and $b$ through learning process from data, here we will use gradient descent strategy,2 then $w^l:=w^l-\alpha\frac{\partial J}{\partial w^l}$ and $b^l:=b^l-\alpha\frac{\partial J}{\partial b^l}$ to fine-tuning $w$ and $b$ ,where $l$ means the order of the layer, $\alpha$ means the learning rate.

The parameters $w$ and $b$ will continue to improve at the end of each iteration of the whole training process. When the loss function tends to decrease and be stable, the CNN model is considered as having completed the training process, which means the CNN model is ready to predict new data.

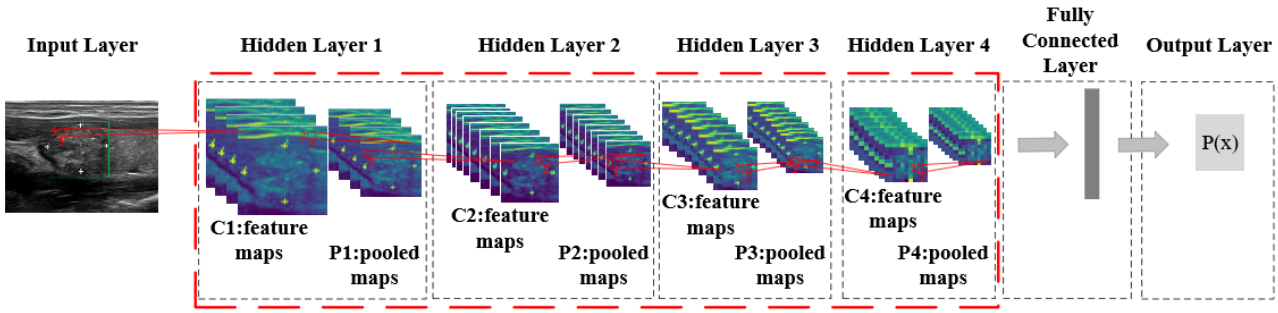


Figure 1. Illustration of basic convolutional neural network (CNN) model flowchart.

## 2.2 Transfer learning model

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on another task. Here we adopted the pre-trained model which was transferred from one of our previous studies followed by a strategy called fine-tuning. This approach consists of selecting source model, reusing model and fine-tuning model (Fig. 2).

***Select Source Model.*** A pre-trained source model which was developed for another ultrasound image classification task[18] was chosen for the DLRT base model.

***Reuse Model.*** The pre-trained model could then be used as the starting point for a model on the DLRT task. This part involved using the first three hidden layers of the model.

***Tune Model.*** Lastly, the model need to be adapted or refined on the input-output pair data available for the task of DLRT. This part involved freezing the first three hidden layers and fine-tuning the rest part of the model.

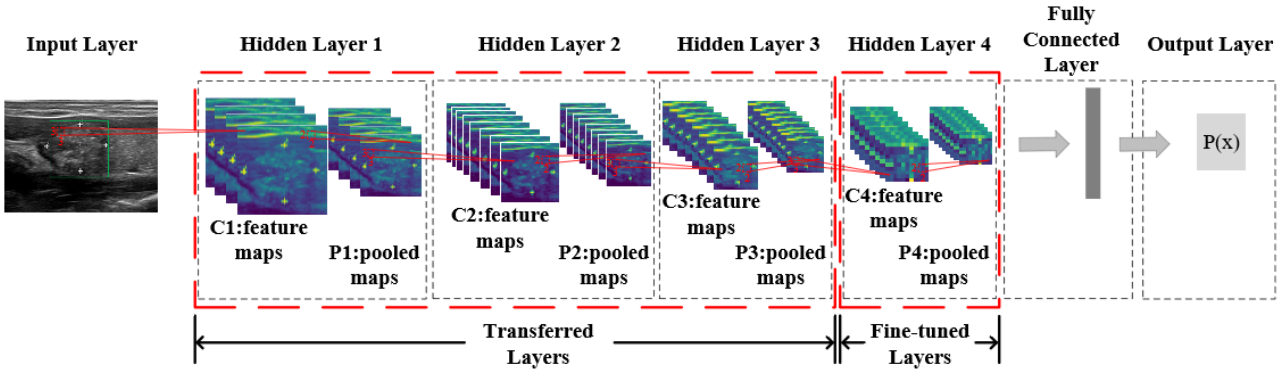


Figure 2. Illustration of transfer learning model flowchart.

## 2.3 DLRT model

For applying DLRT, we designed a simple manual initiation by defining multiple region-of-interests (ROIs). 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 center area. Then, the corresponding three cropped images were used as input layers to trigger the DLRT model (Fig. 3). DLRT adopted the CNN architecture and transfer learning strategy. It consisted of four hidden layers. The first three layers were transferred from one of our previous studies without any modification[18], 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.

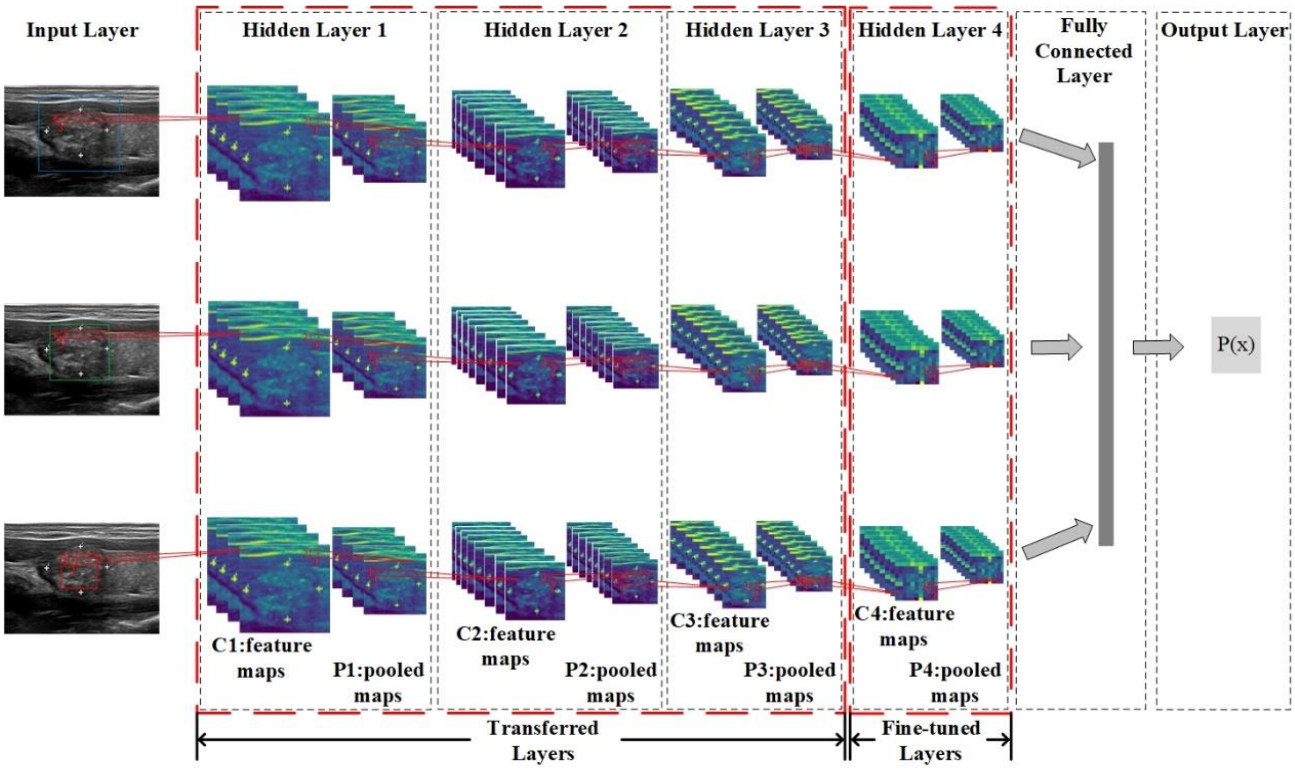


Figure 3. Illustration of deep learning radiomics of thyroid (DLRT) model flowchart.

### 2.4 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 transfer learning 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 transfer learning model employed transferred parameters for the first three layers from another study[18] 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. 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 as the input for the basic CNN and transfer learning models.

### 2.5 Comparison between radiomics and human observers

Thyroid nodule images from 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 transfer learning model.

## 3. DATA

From January 2017 to March 2018, 2179 consecutive thyroid patients who underwent US examination and US-guided FNA biopsy were recruited at our hospital. 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 < 5 mm, (2) unqualified histology with ambiguous diagnostic findings (too few cells or atypical pathology), and (3) follicular neoplasm or suspicious for a

follicular neoplasm. Demographic information, imaging examination, and clinical baseline characteristics were collected from the hospital PACS workstation.

A total of 2212 nodules from 2179 potentially eligible patients were retrospectively enrolled in this study. Among them, 423 patients were excluded due to too few cells from US-guided FNA for pathology, atypical pathology, or follicular neoplasm. Another 127 patients were excluded, because their nodule diameter was less than 5 mm. Finally, 1629 patients with 1645 thyroid nodules were enrolled, and we only employed one US image from each nodule for analysis.

After randomization of enrolled 1629 patients (1645 nodules), 1097 nodules (428 malignant, 39.0%) were assigned to the training cohort, and the other 548 nodules (214 malignant, 39.1%) composed the validation cohort.

## 4. RESULTS

### 4.1 Diagnostic accuracy of DLRT, the basic CNN, and the transfer learning model

Both in training and validation 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 and B). Differences of AUCs were all statistically significant ($P < 0.05$). AUCs of DLRT reached 0.96 and 0.95 in training and validation cohorts, respectively, which were 0.09 and 0.10 higher than these of the transfer learning model who offered the second highest AUCs. The basic CNN model offered the worst AUCs in both training and validation cohorts, which were 0.82 and 0.81, respectively.

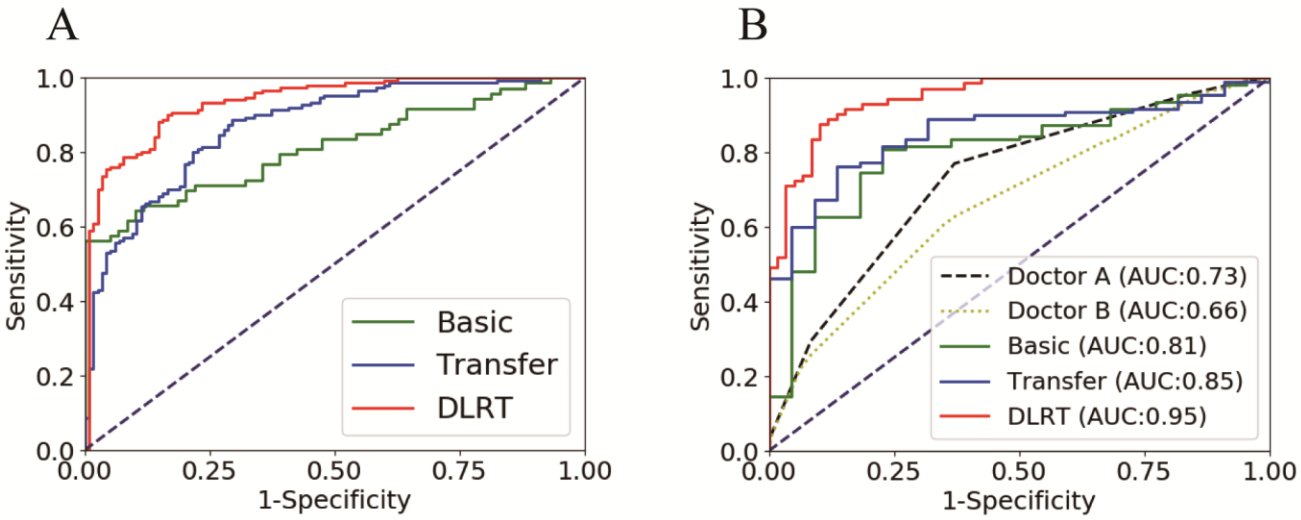


Figure 4. Comparison of receiver operating characteristic (ROC) curves, area under the curve (AUC) between radiomics models (DLRT, the basic CNN, and the transfer learning model), and human observers (a senior and a junior US Radiologist) for the differential diagnosis of benign and malignant thyroid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

### 4.2 Comparison between Radiomics and human observers

A senior and a junior US Radiologist who were blind to pathology data performed differential diagnosis using US images from the validation cohort. Without surprise, the senior observer (Doctor A) outperformed the junior one (Doctor B) with a significant AUC improvement of 0.07 ($P < 0.05$). However, both human observers provided lower AUC than all three Radiomics models (Fig. 4B). The gap of AUC between the best (DLRT) and worst (Doctor B) researched 0.29 ($P < 0.001$).

## 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. It holds a good potential for improving the overall diagnostic efficacy in routine thyroid US examinations.

## 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 Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer," Thyroid. Papers 26(1), 1-133 (2015).

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

[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. Papers 104(11) 5665-5672 (2019).

[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. Papers 19(11) 1167-1214 (2009).

[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. Papers 22(10) 996-1001 (2012).

[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. Papers 247(3) 762-770 (2008).

[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. Papers 262(3) 1002-1013 (2012).

[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. Papers 25(5) 538-550 (2015).

[9] 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. Papers 5 4006 (2014).

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

[11] 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. Papers 34(18) 2157-2164 (2016).

[12] 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. Papers 109 8-12 (2018).

[13] 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. Papers 115 16-21 (2019).

[14] 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) 477-486 (2017).
[15] 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. Papers 45(S18) (2019).
[16] 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. Papers 43(5) 1058-1069 (2017).
[17] 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. Papers 29(6) 858-867 (2019).
[18] 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 multicenter study," Gut. Papers 68(4) 729-741 (2018).