



Deep learning based classification of ultrasound images for thyroid nodules: a large scale of pilot study

Qing Guan^{1,2#}, Yunjun Wang^{1,2#}, Jiajun Du³, Yu Qin³, Hongtao Lu³, Jun Xiang^{1,2}, Fen Wang^{2,4}

¹Department of Head and Neck Surgery, Fudan University Shanghai Cancer Center, Shanghai 200032, China; ²Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, China; ³Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200240, China; ⁴Department of Ultrasonography, Fudan University Shanghai Cancer Center, Shanghai 20032, China

Contributions: (I) Conception and design: J Xiang, F Wang; (II) Administrative support: H Lu; (III) Provision of study materials or patients: Q Guan; (IV) Collection and assembly of data: Y Wang; (V) Data analysis and interpretation: J Du, Y Qin; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Jun Xiang, MD, PhD. Department of Head and Neck Surgery, Fudan University Shanghai Cancer Center, 270 Dong'an Road, Shanghai 200032, China. Email: junxiang82@163.com.

Background: To explore the ability of the deep learning network Inception-v3 to differentiate between papillary thyroid carcinomas (PTCs) and benign nodules in ultrasound images.

Methods: A total of 2,836 thyroid ultrasound images from 2,235 patients were divided into a training dataset and a test dataset. Inception-v3 was trained and tested to crop the margin of the images of nodules and provide a differential diagnosis. The sizes and sonographic features of nodules were further analysed to identify the factors that may influence diagnostic efficiency. Statistical analyses included χ^2 and Fisher's exact tests and univariate and multivariate analyses.

Results: There were 1,275 PTCs and 1,162 benign nodules in the training group and 209 PTCs and 190 benign nodules in the test group. A margin size of 50 pixels and an input size of 384×384 showed the best outcome after training, and these parameters were selected for the test group. In the test group, the sensitivity and specificity for Inception-v3 were 93.3% (195/209) and 87.4% (166/190), respectively. Inception-v3 displayed the highest accuracy for 0.5–1.0 cm nodules. The accuracy differed according to the margin description ($P=0.024$). Taller nodules were more accurately diagnosed than were wider nodules ($P=0.015$). Microcalcification [odds ratio (OR) =0.254, 95% confidence interval (CI): 0.076–0.847, $P=0.026$] and taller shape (OR =0.243, 95% CI: 0.073–0.810, $P=0.021$) were negatively associated with misdiagnosis rate.

Conclusions: Inception-v3 can achieve an excellent diagnostic efficiency. Nodules that are 0.5–1.0 cm in size and have microcalcification and a taller shape can be more accurately diagnosed by Inception-v3.

Keywords: Ultrasound; papillary thyroid cancer; deep learning; Inception-v3

Submitted Nov 25, 2018. Accepted for publication Mar 04, 2019.

doi: 10.21037/atm.2019.04.34

View this article at: <http://dx.doi.org/10.21037/atm.2019.04.34>

Introduction

Epidemiologic studies show that thyroid cancer is the fifth most common cancer in women, and its incidence has increased rapidly in recent years (1). Papillary thyroid carcinoma (PTC) is the predominant pathologic subtype of thyroid cancer. According to the Surveillance,

Epidemiology, and End Results (SEER) cancer registry, the incidence of PTC has increased 3.7-fold, from 3.4 to 12.5 per 100,000 individuals, from 1975 to 2009; meanwhile, the proportion of thyroid cancer nodules that are less than a centimetre in size has increased from 23% [1983] to 36% [2009] (2). Ultrasound (US) is widely accepted as the

primary diagnostic tool for screening thyroid nodules and preoperatively evaluating PTCs (3). Sonographic features such as microcalcification, solid composition, taller shape and irregular margin are considered typical for PTC. The biggest limitation of US is operator dependence, as the accuracy of diagnosis varies among radiologists with different levels of experiences. Usually, an inexperienced radiologist needs a superior to double-check the images to reduce the misdiagnosis rate, especially for cases with a Thyroid Imaging Reporting and Data System (TI-RADS) score of 4. This strategy is practiced at our institution, but not every healthcare facility has the most experienced radiologist for thyroid malignancies, and thus, an unbiased and consistent method to provide valuable second opinion and assist inexperienced radiologists is required.

Computer-aided diagnosis has been rapidly developing. Most recently, machine learning (ML) has been introduced in US imaging-mediated diagnosis. ML is defined as a set of methods that automatically detect patterns in data and then utilize the patterns to predict future data or enable decision-making under uncertain conditions (4). Deep learning is a part of ML and is a special type of artificial neural network that resembles the multilayered human cognition system. Deep learning algorithms such as convolutional neural network (CNN) and Keras neural network are currently utilized in multiple aspects of healthcare, especially in imaging-based diagnosis and prognostic analysis in cancer (5-7).

In this study, we employed the deep learning algorithm Inception-v3 to distinguish PTCs from benign thyroid nodules using images captured by US. Inception-v3 employs inception modules composed of several small convolutional layers. The inception modules increase the layer depth with relatively few parameters; thus, Inception-v3 displays better performance during image classification tasks than do other deep learning algorithms.

Methods

Patients and US images

This study was approved by the Ethical Committee of Fudan University Cancer Center. Oral and written informed consent was obtained from all patients after the nature of the procedures had been fully explained. The study patients were selected from January 2014 to December 2016. In total, 1,795 nodules in 1,359 consecutive patients (mean age, 41.7 ± 15.6 years; age

range, 17–72 years) were histopathologically confirmed to be PTC, 341 of which were excluded due to the lack of qualified US images. Simultaneously, 1,363 benign thyroid nodules in 1,090 patients (mean age, 42.6 ± 13.4 years; age range, 15–72 years) were selected. Benign thyroid nodules were diagnosed based on the following criteria: (I) surgical specimen confirmed by pathology; (II) fine-needle aspiration (FNA) specimen confirmed by cytology; or (III) US findings of very low suspicion (spongiform, pure cystic or partially cystic without any suspicious pattern in the solid portion) according to the American Thyroid Association (ATA) guidelines (3).

Every image was rated by three experienced radiologists using the TI-RADS (8). TI-RADS scores of 2, 3, 4a, 4b and 4c were denoted as not suspicious, probably benign, one suspicious feature, two suspicious features, and three or more suspicious features, respectively. Sonographic features and measurements, including composition, echogenicity, calcification, margin, and shape, were also determined by the three radiologists based on the TI-RADS proposed by Kwak *et al.* (8). Composition was categorized as solid, cystic or spongiform. Spongiform composition indicated that tiny cystic spaces occupied most of the nodule. Echogenicity was categorized as hyperechoic, isoechoic, hypoechoic or markedly hypoechoic by comparing nodules to normal thyroid tissues. A nodule was considered to have marked hypoechogenicity if the echogenicity was lower than that of the surrounding strap muscle. Microcalcification was defined as punctate or “dot-like” foci without posterior acoustic artefacts <1 mm in diameter. Otherwise, calcification was categorized into macrocalcification or triangular reverberation artefacts with a decreasing width of deeper echoes referred to as comet-tail artefacts. The margin was categorized as being smooth, irregular, lobulated, ill-defined, or halo or showing extrathyroidal extension (ETE). The shape was categorized as wider or taller. The accuracy was then compared across the categories for each feature. PTC displays some typical features on US images such as microcalcification, solid composition, taller shape and irregular margin. Univariate and multivariate analyses were utilized to determine the relationship between each feature and the misdiagnosis rate.

Data distribution

There were 2,836 images in total, among which 1,484 were of PTCs, and 1,352 were of benign thyroid nodules, including multinodular disease and adenomas. The majority

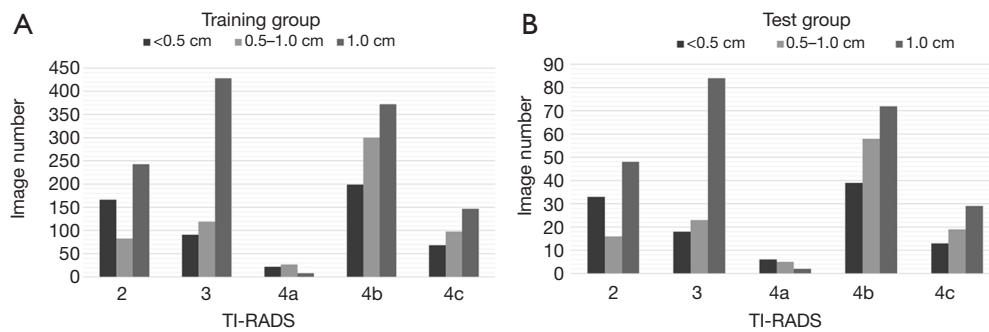


Figure 1 Distribution of size and TI-RADS score in two groups. (A) Training group; (B) test group. TI-RADS, Thyroid Imaging Reporting and Data System.

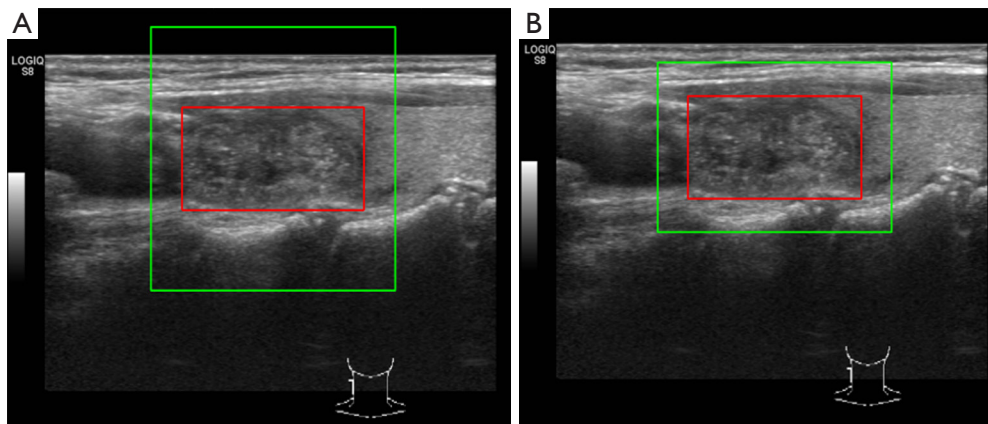


Figure 2 Different shape for cropping with 50 pixel margin. (A) Square; (B) rectangle.

of images including 1,275 PTCs and 1,162 benign nodules were assigned to the training group, and the remaining images including 209 PTCs and 190 benign nodules were assigned to the test group. The images were randomly assigned, provided the distribution of size and TI-RADS scores was similar between the two groups. The images were further divided into three groups based on the size: <0.5, 0.5–1.0, and >1.0 cm. The distribution of TI-RADS scores was uniform among groups based on size (Figure 1).

Image preparation and cropping

The annotated rectangular box only covered the nodule, while visual diagnosis depended on information from both the nodule and surrounding tissue. Therefore, we cropped the original images with margins, which denoted the distances between the left and right boundaries of the cropped image. The lesion boundary is shown in Figure 2. As

clearly shown in both Figure 2A,B, the boundary denoted the distance between the left edge of the larger green rectangle and the left edge of the smaller red rectangle. The sizes of different lesions varied because of differences between their real size and observed size. Therefore, we chose random margins from 0 to 100 pixels for each image in every training iteration. We then cropped the image as a square by setting the height equal to the width and resized the image into 384×384 to match the input size of the network. In this study, we employed bilinear interpolation algorithms to resize the image. We chose a square box rather than a rectangular box as the cropping shape because the nodule shape is important for diagnosis. The input shape of the network is square, indicating that the cropped image should also be a square. The two cropping shapes are clearly shown in Figure 2, and the corresponding resized images are shown in Figure 3. The original shape of the nodule was an ellipse. Figure 3 shows that the resized result of a square (Figure 3A) cropping

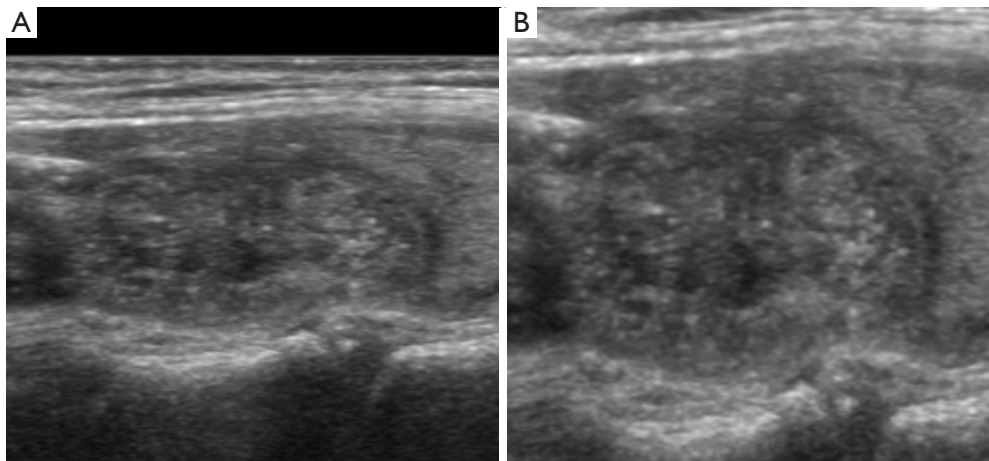


Figure 3 Resized results of different cropping shape. (A) Square; (B) rectangle.

window maintained length-width ratio of the original image of the nodule but that the rectangular (*Figure 3B*) cropping window changed the shape of the nodule in the image to a circle. Therefore, we chose square as the cropping shape because it maintained the shape features of the nodule across the scaling process. The input size of the network was fixed, while the size of nodules differed. Thus, the real input sizes of the nodules depended on their margin. A margin of 50-pixels displayed the highest performance based on nonparametric receiver operating characteristic (ROC) analysis. Therefore, we set the default margin size at 50 pixels.

Network construction

We chose Inception-v3, which was pre-trained on ImageNet database, and fine-tuned the network for our analysed thyroid nodules. The size of the input layer was adjusted to 384×384 based on the validation results. The corresponding size of fully connected layers was modified according to the input layer. Finally, as there were only 2 categories of nodules, benign and malignant, we reduced the output dimension to 2. The network structure of Inception-v3 is shown in *Figure 4*. Inception-v3 is composed of 3 kinds of Inception modules, namely, A, B and C. *Figure 4A,B,C* show the corresponding structures of Inception modules A, B and C. The Inception modules are all composed of several small convolutional layers and pooling layers.

Statistical analysis

ROC analysis was performed to calculate the best cutoff value

for the best margin size. The areas under ROC curves were used to measure the relative predicted accuracy based on the margin size. Categorical data were summarized as frequencies and percentages. χ^2 and Fisher's exact tests were used for categorical variables. Moreover, univariate and multivariate analyses were performed to determine the predictive value of sonographic features using logistic regression represented by odds ratios (ORs) and 95% confidence intervals (CIs). A P value less than 0.05 was considered significant. Statistical analyses were performed using SPSS 19.0 for Windows (SPSS Inc., Chicago, IL, USA).

Margin size and input size

We modified the cropping margin size and the network input size to obtain the best performance, which was evaluated by accuracy, sensitivity, specificity and A_z value. The optimal margin was selected between 0 and 100 pixels, and the network input size was adjusted with a fixed cropping margin. Then, these two hyperparameters were applied to the training process. We present performances based on margin sizes of 0, 25, 50, 75 and 100 pixels in *Table 1*, and a margin size of 50 pixels showed the greatest advantage across all four evaluation terms. The corresponding ROC curves are shown in *Figure 5*. According to the performance, we finally fixed the margin size to 50 pixels. The network input size was adjusted from 224×224 to 480×480, with a fixed margin of 50 pixels. Preliminary data proved that the input size of 384×384 leads to the best result. Thus, we set the input size to 384×384.

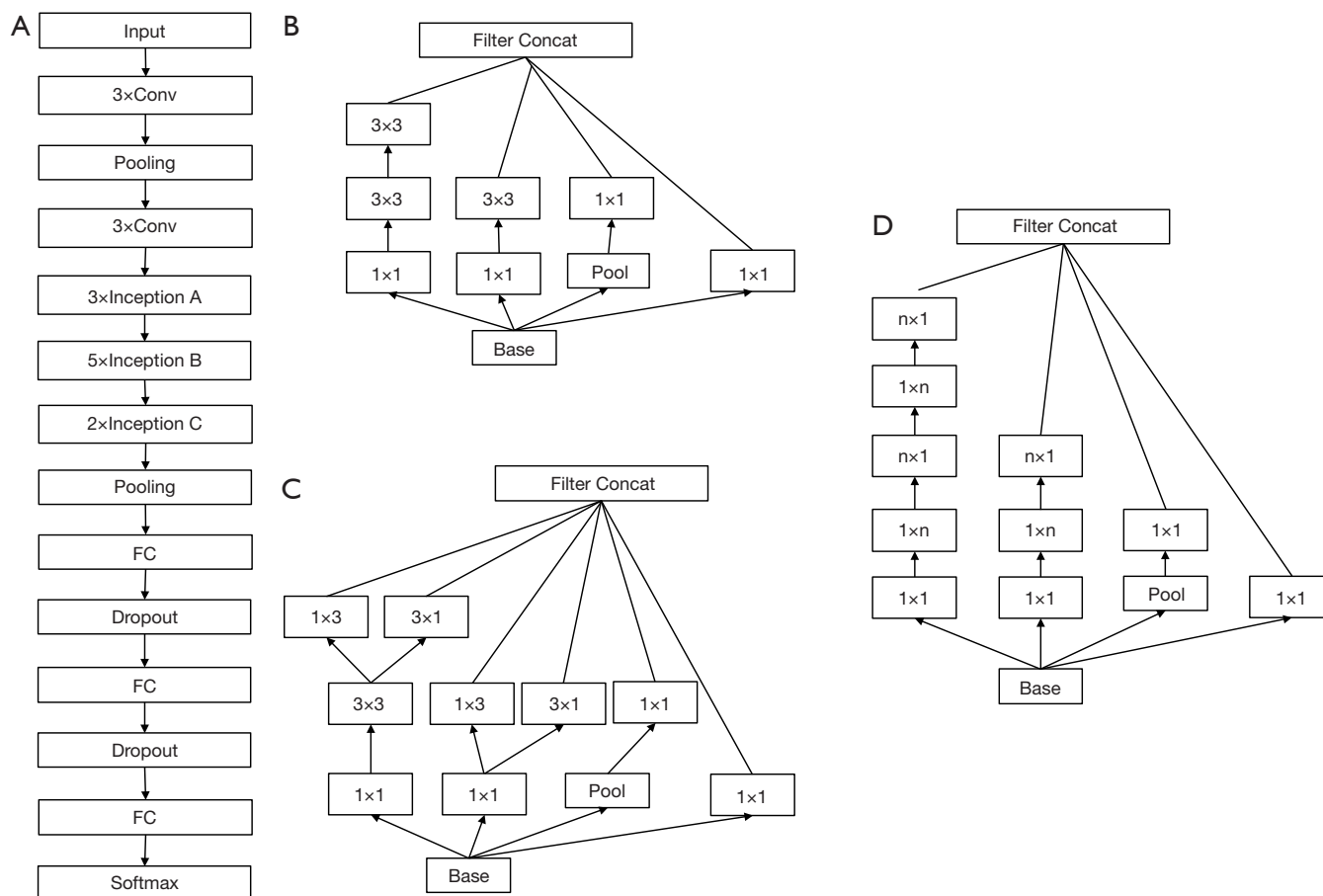


Figure 4 (A) The structure of Inception-v3. Inception-v3 is composed of 3 kinds of Inception modules, namely, A, B and C; (B) inception module A; (C) inception module B; and (D) inception module C. The Inception modules are all composed of several small convolutional layers and pooling layers.

Table 1 Validation of performance based on different margin sizes

Margin size (pixel)	Accuracy (%)	Sensitivity (%)	Specificity (%)	A _z value
0	84.2	88.0	80.0	0.921
25	87.7	90.9	84.2	0.947
50	90.5	93.3	87.4	0.956
75	89.0	90.9	86.8	0.954
100	88.5	90.9	85.8	0.944

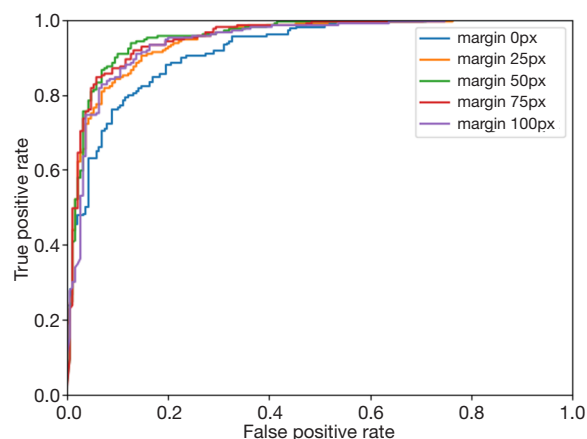


Figure 5 ROC curves for margin sizes of 0, 25, 50, 75, and 100 pixels. ROC, receiver operating characteristic.

Table 2 Sonographic features associated with the diagnostic efficiency of Inception-v3

Sonographic feature	PTC				Benign nodules			
	No.	True (%)	False (%)	P	No.	True (%)	False (%)	P
Total	209	195	14		190	166	24	
Composition				0.788				0.200
Solid	208	194 (93.3)	14 (6.7)		128	108 (84.4)	20 (15.6)	
Cystic	0	0	0		1	1 (100.0)	0	
Spongiform	1	1	0		61	57 (93.4)	4 (6.6)	
Echogenicity				0.099				0.814
Hyperechoic	0	0	0		3	3 (100.0)	0	
Isoechoic	15	12 (80.0)	3 (20.0)		97	86 (88.7)	11 (11.3)	
Hypoechoic	193	182 (94.3)	11 (5.7)		81	69 (85.2)	12 (14.8)	
Markedly hypoechoic	1	1 (100.0)	0		9	8 (88.9)	1 (11.1)	
Calcification				0.260				0.659
Micro	92	89 (96.7)	3 (3.3)		1	1 (100.0)	0	
Macro	11	10 (90.9)	1 (9.1)		2	2 (100.0)	0	
Comet-tail	6	6 (100.0)	0		3	2 (66.7)	1 (33.3)	
None	100	90 (90.0)	10 (10.0)		184	166 (87.4)	24 (12.6)	
Margin				0.024				0.383
Smooth	17	13 (76.5)	4 (23.5)		124	111 (89.5)	13 (10.5)	
Irregular	24	24 (100.0)	0		1	1 (100.0)	0	
Lobulated	6	6 (100.0)	0		9	7 (77.8)	2 (22.2)	
Ill-defined	109	104 (95.4)	5 (4.6)		54	46 (85.2)	8 (14.8)	
Halo	4	3 (75.0)	1 (25.0)		2	1 (50.0)	1 (50.0)	
ETE	49	45 (91.8)	4 (8.2)		0	0	0	
Shape				0.015				0.110
Wider	114	102 (89.5)	12 (10.5)		188	165 (87.8)	23 (12.2)	
Taller	95	93 (97.9)	2 (2.1)		2	1 (50.0)	1 (50.0)	

PTC, papillary thyroid carcinoma.

Nodule size

The hyper-parameters described above were modified in the subgroup data based on nodule size and were then applied to train the network. The optimal parameters of the network were used to classify the thyroid nodules into three size groups, namely, <0.5, 0.5–1 and >1 cm. The 0.5–1 cm size group showed the best results among the three groups, with a specificity of 93.9%, a sensitivity of 94.4%, and an A_z value of 0.971. The <0.5 and >1 cm size groups demonstrated

sensitivities, specificities and A_z values of 100%, 81.4% and 0.962 and 88.8%, 87.7% and 0.943, respectively.

Sonographic features

We next investigated which sonographic features affected the diagnostic ability of Inception-v3. The results are shown in *Table 2*. There was no significant difference among the images of benign nodules; however, for the images of PTCs, the accuracy differed according to the margin description.

Table 3 Typical sonographic features of PTCs associated with the misdiagnosis rate of Inception-v3

Variable	Univariate			Multivariate		
	P value	OR	95% CI for OR	P value	OR	95% CI for OR
Composition (solid)	0.374	1.627	0.556–4.760			
Hypoechoic	0.254	0.666	0.331–1.338			
Calcification						
Micro	0.026	0.254	0.076–0.847	0.050	0.296	0.088–0.998
Macro	0.720	0.761	0.170–3.403	0.788	0.813	0.180–3.679
Margin						
ETE	0.457	0.648	0.207–2.030			
Ill-defined	0.361	0.710	0.341–1.479			
Irregular	0.998	0	0			
Lobular	0.886	1.122	0.233–5.408			
Shape (taller)	0.021	0.243	0.073–0.810	0.040	0.282	0.084–0.944

PTC, papillary thyroid carcinoma; OR, odds ratio; CI, confidence interval.

Nodules with an irregular, lobulated, ill-defined, or ETE margin were more accurately diagnosed than were nodules with a smooth or halo margin ($P=0.024$). Additionally, taller nodules were more easily diagnosed by Inception-v3 than were wider nodules ($P=0.015$). Furthermore, as shown in *Table 3*, the typical features of PTCs including solid composition, hypoechogenicity, the presence of microcalcification, nonsmooth margin and taller shape were analysed to identify their association with the misdiagnosis rate of Inception-v3. Microcalcification (OR =0.254, 95% CI: 0.076–0.847, $P=0.026$) and taller shape (OR =0.243, 95% CI: 0.073–0.810, $P=0.021$) were negatively associated with misdiagnosis rate in both univariate and multivariate analyses.

Discussion

A number of studies have attempted to develop deep learning networks for the differentiation of thyroid nodules using US images; however, obtaining a large number of images from one facility is not easy. Fudan University Cancer Center is a tertiary referral hospital treating thousands of thyroid cancer patients each year and has thus enabled us to obtain a very large dataset including 2,836 images from 2,235 patients. Most importantly, all the PTC nodules were surgically removed and confirmed by pathology. Based on this dataset, this study has provided solid evidence to show that Inception-v3 has a similar accuracy to that of

experienced radiologists in differentiating PTCs from benign nodules, confirming the potential of Inception-v3 to provide a second opinion, especially when radiologists are inexperienced.

To optimize this system, we further investigated the parameters that could affect the accuracy of this system. Training experiments demonstrated that a margin size of 50 pixels and input size of 384×384 led to the best diagnostic efficiency; therefore, these parameters were selected for the test group. The test group had 399 images, and the sensitivity and specificity for Inception-v3 were 93.3% (195/209) and 87.4% (166/190), respectively. For comparison, the sensitivity and specificity for radiologists were 84.7% (177/209) and 97.9% (186/190), respectively. Although the overall sensitivity and specificity were close, Inception-v3 was more accurate in diagnosing PTCs but less accurate in diagnosing benign nodules than were experienced radiologists, indicating that the features of PTC were more easily delineated by Inception-v3.

Since the learning process of Inception-v3 has not been elucidated, we analysed the features of the images to determine the learning process. Nodule size was our first concern. The images were divided into three groups based on nodule size, and the 0.5–1 cm size group showed the best specificity (93.9%) and A_z value (0.9712), while the <0.5 cm size group showed the best sensitivity (100%). These results indicated that Inception-v3 was advantageous for diagnosing PTC nodules <1 cm, which are also known

as papillary thyroid microcarcinomas (PTMCs). Since the incidence of PTMC is rapidly increasing, and the 2015 ATA management guidelines have endorsed active surveillance for low-risk PTMC using US (3), the application of Inception-v3 could be promising in this area. Previous studies using deep learning for differential diagnosis have placed more emphasis on the diagnostic efficiency, feature extraction process and optimization of the models (9,10). One study compared the diagnostic performance and agreement of US characteristics between an experienced radiologist and a deep learning model (11); however, very few studies have investigated the correlation between the image features and diagnostic efficiency of deep learning networks. In this study, our analysis of sonographic features showed that nodules with an irregular, lobulated, ill-defined, or ETE margin were more accurately diagnosed than were nodules with a smooth or halo margin, which indicated that nodules with smooth margins may appear as benign nodules to Inception-v3; in other words, the diagnosis of PTC by Inception-v3 considerably relies on the margin. Furthermore, microcalcification and taller shape were proven to be negatively associated with the misdiagnosis rate of Inception-v3 in both univariate and multivariate analyses. Microcalcification is a well-known risk factor for PTC, with an OR of 11.6 (12). Mussa *et al.* suggested that in the mere presence of microcalcifications, an FNA biopsy is warranted (13). A taller-than-wide shape is also considered a feature suggestive of malignancy (14-16). Both microcalcification and taller shape are typical sonographic features of PTC, and the fact that these features influenced the diagnostic accuracy of Inception-v3 suggested that Inception-v3 had successfully recognized PTC images by these special properties. One of the limitations of this study is that we did not include the Doppler images which reflect the vascularity of the nodule, future study design should take this into consideration.

Conclusions

In summary, we propose that the deep learning network Inception-v3 can be applied to facilitate the differentiation of PTCs in the clinic. After being trained on a large dataset, the deep learning framework Inception-v3 could achieve an excellent diagnostic efficiency. We have demonstrated that thyroid nodules of a 0.5–1.0 cm size that have microcalcification and a taller shape could be more accurately diagnosed by Inception-v3.

Acknowledgements

Funding: This study was funded by Shanghai municipal planning commission of science and research fund for young scholar (award number 20154Y0050).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: The study was approved by the Ethical Committee of Fudan University Cancer Center. Oral and written informed consent was obtained from all patients after the nature of the procedures had been fully explained.

References

1. Pellegriti G, Frasca F, Regalbuto C, et al. Worldwide increasing incidence of thyroid cancer: update on epidemiology and risk factors. *J Cancer Epidemiol* 2013;2013:965212.
2. La Vecchia C, Malvezzi M, Bosetti C, et al. Thyroid cancer mortality and incidence: a global overview. *Int J Cancer* 2015;136:2187-95.
3. Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid* 2016;26:1-133.
4. Murphy KP. *Machine learning: a probabilistic perspective*, 1st ed. Cambridge. The MIT Press, 2012.
5. Sato M, Horie K, Hara A, et al. Application of deep learning to the classification of images from colposcopy. *Oncol Lett* 2018;15:3518-23.
6. Men K, Chen X, Zhang Y, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Front Oncol* 2017;7:315.
7. Matsuo K, Purushotham S, Moeini A, et al. A pilot study in using deep learning to predict limited life expectancy in women with recurrent cervical cancer. *Am J Obstet Gynecol* 2017;217:703-5.
8. Kwak JY, Han KH, Yoon JH, et al. Thyroid imaging reporting and data system for us features of nodules: a step in establishing better stratification of cancer risk. *Radiology* 2011;260:892-9.

9. 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;30:477-86.
10. Yu Q, Jiang T, Zhou A, et al. Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images. *Eur Arch Otorhinolaryngol* 2017;274:2891-7.
11. Choi YJ, Baek JH, Park HS, et al. A Computer-Aided diagnosis System Using Artificial Intelligence for the Diagnosis and Characterization of Thyroid Nodules on Ultrasound: Initial Clinical Assessment. *Thyroid* 2017;27:546-52.
12. Smith-Bindman R, Lebda P, Feldstein VA, et al. Risk of Thyroid Cancer Based on Thyroid Ultrasound Imaging Characteristics: Results of a Population-Based Study. *JAMA Intern Med* 2013;173:1788-96.
13. Mussa A, De Andrea M, Motta M, et al. Predictors of Malignancy in Children with Thyroid Nodules. *J Pediatr* 2015;167:886-92.
14. Moon WJ, Jung SL, Lee JH, et al. Benign and malignant thyroid nodules: US differentiation--multicenter retrospective study. *Radiology* 2008;247:762-70.
15. Moon HJ, Kwak JY, Kim EK, et al. A Taller-Than-Wide Shape in Thyroid Nodules in Transverse and Longitudinal Ultrasonographic Planes and the Prediction of Malignancy. *Thyroid* 2011;21:1249-53.
16. Chen SP, Hu YP, Chen B, et al. Taller-Than-Wide Sign for Predicting Thyroid Microcarcinoma: Comparison and Combination of Two Ultrasonographic Planes. *Ultrasound Med Biol* 2014;40:2004-11.

Cite this article as: Guan Q, Wang Y, Du J, Qin Y, Lu H, Xiang J, Wang F. Deep learning based classification of ultrasound images for thyroid nodules: a large scale of pilot study. *Ann Transl Med* 2019;7(7):137. doi: 10.21037/atm.2019.04.34