Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study

Xiangchun Li†, Sheng Zhang*†, Qiang Zhang*, Xi Wei*, Yi Pan, Jing Zhao, Xiaojie Xin, Chunxin Qin, Xiaoqing Wang, Jianxin Li, Fan Yang, Yanhui Zhao, Meng Yang, Qinghua Wang, Zhiming Zheng, Xiangqian Zheng, Xiangming Yang, Christopher T Whitlow, Metin Nafi Gurcan, Lun Zhang, Xudong Wang, Boris C Pasche, Ming Gao, Wei Zhang†, Kexin Chen†*

Summary

Background The incidence of thyroid cancer is rising steadily because of overdiagnosis and overtreatment conferred by widespread use of sensitive imaging techniques for screening. This overall incidence growth is especially driven by increased diagnosis of indolent and well-di erentiated papillary subtype and early-stage thyroid cancer, whereas the incidence of advanced-stage thyroid cancer has increased marginally. Thyroid ultrasound is frequently used to diagnose thyroid cancer. The aim of this study was to use deep convolutional neural network (DCNN) models to improve the diagnostic accuracy of thyroid cancer by analysing sonographic imaging data from clinical ultrasounds.

Methods We did a retrospective, multicohort, diagnostic study using ultrasound images sets from three hospitals in China. We developed and trained the DCNN model on the training set, 131 731 ultrasound images from 17 627 patients with thyroid cancer and 180 668 images from 25 325 controls from the thyroid imaging database at Tianjin Cancer Hospital. Clinical diagnosis of the training set was made by 16 radiologists from Tianjin Cancer Hospital. Images from anatomical sites that were judged as not having cancer were excluded from the training set and only individuals with suspected thyroid cancer underwent pathological examination to confirm diagnosis. The model's diagnostic performance was validated in an internal validation set from Tianjin Cancer Hospital (8606 images from 1118 patients) and two external datasets in China (the Integrated Traditional Chinese and Western Medicine Hospital, Jilin, 741 images from 154 patients; and the Weihai Municipal Hospital, Shandong, 11 039 images from 1420 patients). All individuals with suspected thyroid cancer after clinical examination in the validation sets had pathological examination. We also compared the specificity and sensitivity of the DCNN model with the performance of six skilled thyroid ultrasound radiologists on the three validation sets.

Findings Between Jan 1, 2012, and March 28, 2018, ultrasound images for the four study cohorts were obtained. The model achieved high performance in identifying thyroid cancer patients in the validation sets tested, with area under the curve values of 0·947 (95% CI 0·935–0·959) for the Tianjin internal validation set, 0·912 (95% CI 0·865–0·958) for the Jilin external validation set, and 0·908 (95% CI 0·891–0·925) for the Weihai external validation set. The DCNN model also showed improved performance in identifying thyroid cancer patients versus skilled radiologists. For the Tianjin internal validation set, sensitivity was 93·4% (95% CI 89·6–96·1) versus 96·9% (93·9–98·6; p=0·003) and specificity was 86·1% (81·1–90·2) versus 59·4% (53·0–65·6; p<0·0001). For the Jilin external validation set, sensitivity was 84·3% (95% CI 73·6–91·9) versus 92·9% (84·1–97·6; p=0·048) and specificity was 86·9% (95% CI 77·8–93·3) versus 57·1% (45·9–67·9; p<0·0001). For the Weihai external validation set, sensitivity was 84·7% (95% CI 77·0–90·7) versus 89·0% (81·9–94·0; p=0·25) and specificity was 87·8% (95% CI 81·6–92·5) versus 68·6% (60·7–75·8; p<0·0001).

Interpretation The DCNN model showed similar sensitivity and improved specificity in identifying patients with thyroid cancer compared with a group of skilled radiologists. The improved technical performance of the DCNN model warrants further investigation as part of randomised clinical trials.

Funding The Program for Changjiang Scholars and Innovative Research Team in University in China, and National Natural Science Foundation of China.

Introduction

The incidence of thyroid cancer has been increasing worldwide over the past two decades, including in the USA, where a decrease in the incidence of many other cancer types has been reported.[1]

Institute ( ), Department of Diagnostic and Therapeutic Ultrasonography ( ), Department of Maxillofacial and Otorhinolaryngology Oncology ( ), Department of Pathology ( ), Department of Epidemiology and Biostatistics ( ), and Department of Thyroid and Neck Cancer ( ), National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy of Tianjin, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China; Department of Thyroid and Breast Surgery ( ) and Department of Ultrasonography ( ), Weihai Municipal Hospital, Shandong, China; Department of Ultrasonography, Affiliated Hospital of Chifeng University, Inner Mongolia, China ( ); Department of Ultrasonography, Integrated Traditional Chinese and Western Medicine Hospital, Jilin, China ( ); Department of Ultrasonography, Dezhou Municiple Hospital, Shandong, China ( ); and Departments of Radiology and

thyroid nodule, according to the Thyroid Imaging, Reporting and Data System (TI-RADS) guidelines. The American College of Radiology (ACR) TI-RADS,[3] European TI-RADS,[4] and American Thyroid Association guidelines[5] propose multiple criteria to interpret sonographic images. Among these criteria, solid aspect, hypoechogenicity, taller-than-wide shape, irregular margin, extrathyroidal extension, calcification, and punctate echogenic foci are clinically relevant features associated with suspicion of malignant disease.[3–8] Patients with suspected thyroid cancer undergo fine-needle aspiration biopsy or surgical resection, which is assessed by pathological examination (the gold standard for diagnosis). Therefore, diagnosis of thyroid cancer is a time-consuming and often subjective process requiring substantial experience and expertise of radiologists.

There are four main subtypes of thyroid cancer: papillary, follicular, medullary, and anaplastic.[7] The 5-year

Published online December 21, 2018 http://dx.doi.org/10.1016/S1470-2045(18)30762-9

detection. In this study, we aimed to ascertain the capability of deep learning models for automated diagnosis of thyroid cancer using real-world sonographic data from clinical thyroid ultrasound examinations. We compared results with pathological examination reports (the diagnostic gold standard). This study encompassed model development with a cohort of more than 300 000 images, and validation of the model in three validation datasets.

Methods

Study design and participants

We did a retrospective, multicohort, diagnostic study using ultrasound images sets from three hospitals in China. We obtained ultrasound images for the training set (312 399 images from 42 952 patients) from the thyroid imaging database at Tianjin Cancer Hospital, Tianjin, China. We obtained images for validation sets from thyroid imaging databases at Tianjin Cancer Hospital (internal validation set, 8606 images from 1118 patients), the Integrated Traditional Chinese and Western Medicine Hospital, Jilin, China (Jilin external validation set, 741 images from 154 patients), and Weihai Municipal Hospital, Shandong, China (Weihai external validation set, 11 039 images from 1420 patients).

We included adult patients aged 18 years or older. Clinical diagnosis of the training set was made by 16 radiologists from Tianjin Cancer Hospital, according to TI-RADS guidelines.[3–5] All patients with thyroid cancer and 5651 negative control individuals in the training set, and all individuals in the three validation sets, underwent pathological examination. Pathological examination reports were provided by the pathology department at Tianjin Cancer Hospital. All ultrasound images and pathological examination reports were deidentified before they were transferred to investigators.

This study was approved by the institutional review board (IRB) of Tianjin Cancer Hospital and undertaken according to the Declaration of Helsinki. Informed consent from patients with thyroid cancer and controls was exempted by the IRB because of the retrospective nature of this study.

Procedures

All thyroid ultrasound images extracted from the thyroid imaging database at all three hospital sites were in jpeg format. Ultrasound equipment manufactured by Philips, Toshiba, and GE Healthcare (various models) was used to generate ultrasound images.

Image quality control was performed for the training set; we removed images from thyroid cancer patients if the anatomical sites did not have cancer as per the pathological review report, according to the location sign on the image. For example, if the image available was from the left lobes of the thyroid but pathology data were for the isthmus of the thyroid, the image was considered not suitable for training. For the validation sets, all images were included. Sonographic images with lymph nodes were also included in both training and validation sets.

	Training set* (n=42 952)	Tianjin internal validation set (n=1118)	Jilin external validation set (n=154)	Weihai external validation set (n=1420)
[illegible]	17 627 (41%)	563 (50%)	70 (45%)	542 (38%)
[illegible]	131 731	4491	347	4818
[illegible]	5651 (13%)	555 (50%)	84 (55%)	878 (62%)
[illegible]	51 255	4115	394	6221
[illegible]	19 674 (46%)	0	0	0
[illegible]	129 413	0	0	0
[illegible]	10 832 (25%)	261 (23%)	34 (22%)	282 (20%)
[illegible]	78 768	1785	154	1992
[illegible]	32 032 (75%)	866 (77%)	120 (78%)	1138 (80%)
[illegible]	233 268	6830	587	9047
[illegible]	44 (36 54)	47 (24 41)	51 (45 59)	50 (41 59)
[illegible]	2009 (5%)	112 (10%)	1 (<1%)	24 (2%)
[illegible]	8823 (21%)	146 (13%)	33 (21%)	258 (18%)
[illegible]	5830 (14%)	381 (34%)	5 (3%)	76 (5%)
[illegible]	26 202 (61%)	479 (43%)	115 (75%)	1062 (75%)

[illegible]

Table 1: Baseline characteristics

A DCNN classification model, in which image input features (eg, image pixels) are mapped to the corresponding output label (eg, benignity or malignancy), was used to train the deep learning algorithm. The DCNN algorithm can learn hierarchical representations from the input imaging data. Such a trained model can make predictions on input data. We used the ResNet model[18] with 50 layers (ResNet-50) and the Darknet model[19] with 19 layers (Darknet-19) for image classification. Layers are functional units of neural network and can have different functions in that they learn and store abstract features of the input image. The ResNet-50 and Darknet-19 models were first trained iteratively for classification of patients with thyroid cancer (using 131731 images) and controls (using 180 668 images). We next combined these two deep learning models by weighting their performance (measured by area under the curve [AUC]) and assessed the ensemble DCNN model with the internal and external validation sets.

Darknet-19 was proposed as the backbone for the object detection algorithm[19] because it is more computationally efficient than ResNet-50 (in that Darknet-19 has fewer arithmetic operations compared with ResNet-50) and achieved performance metrics with ImageNet data[19] that were comparable with those obtained with ResNet-50 (appendix p 4). The weights of ResNet-50 and Darknet-19 were initialised from the same network that had been trained to classify 1000 objects in the ImageNet dataset,[20] except the last layer. The weights of last layer were

[illegible] Online

on-the-fly data augmentation[12,21] for each image during training to avoid overfitting. On-the-fly augmentation generates more training images through image processing such as random cropping, rotation, horizontal or vertical flipping, scaling, translations, and adjustment of the saturation and exposure, which mimic the data diversity observed in the real world, avoiding model overfitting. Image augmentation was not done for the validation sets. Additionally, a weight decay rate of 0·0005 was also set to additionally combat for overfitting. Weight decay can prevent the weights of neural network from growing too large.

To quantify the contribution of the pixels that most influence the DCNN model's prediction, we generated a class activation map[22] by using global average pooling in

randomly initialised and the output unit was changed to two for matching the number of classes in our study (ie, thyroid cancer control). We trained the network with stochastic gradient descent running on an NVIDIA graphic processing unit (GPU) with a GTX 1080Ti graphics card (NVIDIA, Beijing, China). We also applied

Published online December 21, 2018 http://dx.doi.org/10.1016/S1470-2045(18)30762-9

with thyroid nodules displaying malignant characteristics at clinical examination (17 627 [76%] of 23 278 individuals). The remaining 19 674 individuals were used as negative controls. All individuals in the three validation sets had pathological examination results. Pathological assessment was done by board-certified pathologists at individual sites according to WHO Classification of Tumors of Endocrine Organs. All pathological assessments were based on haematoxylin and eosin-stained whole-slide images.

Statistical analysis For classification of purposes, we used

used to assess the performance of the ensemble DCNN model versus the group of six skilled thyroid ultrasound radiologists (table 3). Radiologist 1 read 4483 images (n=654 individuals), radiologist 2 read 5967 images (n=774), radiologist 3 read 3734 images (n=500), radiologist 4 read 2982 images (n=482), radiologist 5 read 741 images (n=154), and radiologist 6 read 2233 images (n=274). The entire image set for every selected patient was shown to and read by the radiologists. Radiologists' manual interpretation results were aggregated and the classification accuracy, sensitivity, and specificity were calculated and compared with that of deep learning models.

Among the radiologists, for the Tianjin internal validation set, accuracy ranged from 78·0% (95% CI 74·1–81·6; 390 of 500 individuals) to 79·6% (75·8–83·0; 398 of 500 individuals), sensitivity ranged from 94·1% (90·5–96·7; 241 of 256 individuals) to 98·4% (96·0–99·6; 252 of 256 individuals), and specificity from 57·0% (50·5–63·3; 139 of 244 individuals) to 62·3% (55·9–68·4; 152 of 244 individuals). For the Jilin external validation set, accuracy ranged from 70·8% (95% CI 62·9–77·8; 109 of 154 individuals) to 74·7% (67·0–81·3; 115 of 154 individuals), sensitivity from 85·7% (75·3–92·9; 60 of 70 individuals) to 97·1% (90·1–99·7; 68 of 70 individuals), and specificity from 51·2% (40·0–62·3; 43 of 84 individuals) to 63·1% (51·9–73·4; 53 of 84 individuals). For the Weihai external validation set, accuracy ranged from 72·6% (66·9–77·8; 199 of 274 individuals) to 81·8% (76·7–86·1; 223 of 274 individuals), sensitivity from 85·6% (77·9–91·4; 101 of 118 individuals) to 94·1% (88·2–97·6; 111 of 118 individuals), and specificity from 62·2% (54·1–69·8; 97 of 156 individuals) to 78·8% (71·6–85·0; 123 of 156 individuals). The inter-radiologist agreement rate was 86·4% (95% CI 83·1–89·3; 432 of 500 individuals; Fleiss' Kappa 0·79) in the Tianjin internal validation set, 76·6% (69·1–83·1; 118 of 154 individuals; Fleiss' Kappa 0·65) in the Jilin external validation set, and 69·7% (63·9–75·1; 191 of 274 individuals; Fleiss' Kappa 0·59) in the Weihai external validation set.

Compared with the skilled radiologists, the ensemble DCNN model achieved high performance in identifying thyroid cancer patients. For the Tianjin internal validation set, accuracy was 89·8% (95% CI 86·8–92·3; 994 of 1118 individuals) with the DCNN model versus 78·8% (75·0–82·3; 394 of 500 individuals; p<0·0001) with the radiologists, sensitivity was 93·4% (95% CI 89·6–96·1; 519 of 563 individuals) versus 96·9% (93·9–98·6; 248 of 256 individuals; p=0·003), and specificity was 86·1% (95% CI 81·1–90·2; 475 of 555 individuals) versus 59·4% (53·0–65·6; 145 of 244 individuals; p<0·0001). For the Jilin external validation set, accuracy was 85·7% (95% CI 79·2–90·8; 132 of 154 individuals) versus 72·7% (65·0–79·6%; 112 of 154 individuals; p<0·0001), sensitivity was 84·3% (95% CI 73·6–91·9%; 59 of 70 individuals) versus 92·9% (84·1–97·6; 65 of 70 individuals; p=0·048), and specificity was 86·9% (95% CI 77·8–93·3; 73 of 84 individuals) versus 57·1% (45·9–67·9%; 48 of 84 individuals; p<0·0001). For the Weihai external

Published online December 21, 2018 http://dx.doi.org/10.1016/S1470-2045(18)30762-9

Darknet-19 model, and the ensemble DCNN model are

malignancy and, thus, were more difficult to differentiate. The improvement in accuracy and specificity reported with the DCNN model might lead to a reduction in unnecessary fine-needle aspiration biopsy procedures. However, clinical diagnostic validity needs to be assessed in future randomised clinical trials against current standard procedures.

The trained DCNN model could correctly pinpoint malignant thyroid nodules in a weakly supervised manner through class activation map analysis. DCNN models and machine learning approaches based on conventional feature extraction have previously been investigated for discrimination of malignancy of thyroid nodules from ultrasound images. For example, Ma and colleagues[25] used DCNN and analysed 8148 manually annotated thyroid nodules and obtained an accuracy of 83·0% (95% CI 82·3–83·7) in thyroid nodule diagnosis; however, data from this study are not available so we could not assess them with our artificial intelligence system. Xia and colleagues[26] achieved an accuracy of 87·7% in differentiating malignant and benign nodules by applying extreme machine learning to radiologist-collected features that were obtained from 203 ultrasound images of 187 patients with thyroid cancer. Pereira and colleagues[27] reported an accuracy of 83% achieved by a DCNN model in distinguishing between malignant and benign thyroid nodules from 946 images of 165 patients, which was substantially higher than machine learning algorithms based on conventional feature extraction. However, these studies were limited by small sample sizes and no external validation sets. We do not know if the improvement in accuracy we reported in our study relates to the machine learning method used or to the much larger training dataset.

Our study has some limitations. We did not include training data from other hospitals, and we did not do sensitivity analyses with respect to tumour size and subtypes of malignant disease. 5651 (13%) of 42 952 individuals in the training set were true negatives, with the assumption that patients who did not undergo surgery would be mainly negative diagnoses. The performance of our artificial intelligence system is expected to increase by

 Published online December 21, 2018 http://dx.doi.org/10.1016/S1470-2045(18)30762-9

Support Grant from the National Cancer Institute to the Comprehensive Cancer Center of Wake Forest Baptist Medical Center (P30 CA012197).

References

1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. *CA C/ J C* 2017; 67: 7–30.

2 Chen W, Zheng R, Baade PD, Zhang S, Zeng H. Cancer statistics in China, 2015. *CA C/ J C* 2016; 66: 115–32.

3 Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *J A C /* 2017; 14: 587–95.

4 Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. *E . J* 2017; 6: 225–37.

5 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. . 2016; 26: 1–133.

6 Li Q, Lin X, Shao Y, Xiang F, Samir AE. Imaging and screening of thyroid cancer. / C N . A 2017; 55: 1261–71.

7 Tamhane S, Gharib H. Thyroid nodule update on diagnosis and management. *C D / E* 2016; 2: 17.

8 Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS. The diagnosis and management of thyroid nodules: a review. *JAMA* 2018; 319: 914–24.

9 American Cancer Society. Thyroid cancer survival rates, by type and stage. April 15, 2016. https://www.cancer.org/cancer/thyroid-cancer/detection-diagnosis-staging/survival-rates.html (accessed Nov 29, 2018).

10 Jegerlehner S, Bulliard J-L, Aujesky D, et al. Overdiagnosis and overtreatment of thyroid cancer: a population-based temporal trend study. *PL . O* 2017; 12: e0179387.

11 Park S, Oh C-M, Cho H, et al. Association between screening and the thyroid cancer "epidemic" in South Korea: evidence from a nationwide study. *BMJ* 2016; 355: i5745.

12 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *N/* 2017; 542: 115–18.

13 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; 316: 2402–10.

14 Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017; 318: 2211–23.

15 Kermany DS, Goldbaum M, Cai W, Lewis MA. Identifying medical diagnoses and treatable diseases by image-based deep learning. *C* 2018; 172: 1122–31.

16 Chang Y, Paul AK, Kim N, et al. Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: a comparison with radiologist-based assessments. *M P.* 2016; 43: 554–67.

17 LeCun Y, Bengio Y, Hinton G. Deep learning. *N/* 2015; 521: 436–44.

18 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *P IEEE C C P/* 2016; published online Dec 12. DOI:10.1109/CVPR.2016.90.

19 Redmon J, Farhadi A. YOLO9000: better, faster, stronger. *P IEEE C C P/* 2017; published online Nov 9. DOI:10.1109/CVPR.2017.690.

ussakovsky O, TfZhaJ587.[(u H6698 Tm(I R)2NEMC ETBT/T1_3 1 Tf7 0 0 7 276.t)2767-21143 Tdt laren S,PR.32J7 (