

Deep learning for intelligent diagnosis in thyroid scintigraphy

Journal of International Medical Research

49(1) 1–11

© The Author(s) 2021

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0300060520982842

journals.sagepub.com/home/imr



Tingting Qiao^{1,*}, Simin Liu^{1,*}, Zhijun Cui^{2,*},
Xiaqing Yu¹, Haidong Cai¹, Huijuan Zhang³,
Ming Sun¹, Zhongwei Lv^{1,*}  and Dan Li^{1,*}

Abstract

Objective: To construct deep learning (DL) models to improve the accuracy and efficiency of thyroid disease diagnosis by thyroid scintigraphy.

Methods: We constructed DL models with AlexNet, VGGNet, and ResNet. The models were trained separately with transfer learning. We measured each model's performance with six indicators: recall, precision, negative predictive value (NPV), specificity, accuracy, and F1-score. We also compared the diagnostic performances of first- and third-year nuclear medicine (NM) residents with assistance from the best-performing DL-based model. The Kappa coefficient and average classification time of each model were compared with those of two NM residents.

Results: The recall, precision, NPV, specificity, accuracy, and F1-score of the three models ranged from 73.33% to 97.00%. The Kappa coefficient of all three models was >0.710. All models performed better than the first-year NM resident but not as well as the third-year NM resident in terms of diagnostic ability. However, the ResNet model provided “diagnostic assistance” to the NM residents. The models provided results at speeds 400 to 600 times faster than the NM residents.

Conclusion: DL-based models perform well in diagnostic assessment by thyroid scintigraphy. These models may serve as tools for NM residents in the diagnosis of Graves' disease and subacute thyroiditis.

Keywords

Intelligent diagnosis, deep learning, thyroid scintigraphy, nuclear medicine residents, thyroid disease, Graves' disease, subacute thyroiditis, diagnostic performance

Date received: 19 August 2020; accepted: 30 November 2020

¹Department of Nuclear Medicine, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China

²Department of Medicine Imaging, the Chongming Branch of Shanghai Tenth People's Hospital, Tongji University, Shanghai, China

³School of Software Engineering, Tongji University, Shanghai, China

*These authors contributed equally to this work.

Corresponding author:

Dan Li, Department of Nuclear Medicine, Shanghai Tenth People's Hospital, Tongji University, 301 Yanchangzhong Road, Shanghai 200072, China.
Emails: plumredlinda@163.com



Introduction

The incidence of thyroid functional diseases has increased in recent years, and such diseases have become the second most common endocrine disorders. In China, 40 million people have been diagnosed with hypothyroidism and more than 10 million people have been diagnosed with hyperthyroidism.¹ Unlike ultrasonography and computed tomography (CT), which are commonly used to identify thyroid nodules, thyroid scintigraphy with (99m)Tc-pertechnetate is used to determine the functional status of the thyroid gland or thyroid nodules. As a powerful tool that provides information regarding the location, size, shape, and functional status of the thyroid gland, thyroid scintigraphy is a necessary part of the clinical workup of thyroid functional diseases. Numerous studies have shown that thyroid scintigraphy with (99m)Tc-pertechnetate is the most effective way to distinguish Graves' disease from thyroiditis, the two most common causes of thyrotoxicosis.²⁻⁵ However, the interpretation of thyroid scintigraphy results largely relies on the experience of nuclear medicine (NM) residents, making diagnostic reporting somewhat subjective.

In recent years, deep learning (DL) has been introduced into computer-aided diagnosis (CAD) systems to improve the accuracy of medical imaging diagnosis, save time, and explore new directions and opportunities in radiology.^{6,7} Applying CAD systems may not only reduce radiologists' workload but also lessen subjective and ambiguous reporting. Many CAD studies on thyroid imaging have been performed,⁸⁻¹² including the application of CAD to ultrasound images for the discrimination of benign and malignant thyroid nodules^{8,9} and CT-based CAD^{10,11} for the detection of thyroid abnormalities.

Few reports have described the use of CAD in thyroid scintigraphy, especially

for the diagnosis of thyroid disease. Ma et al.¹² used deep convolutional neural network (DCNN) models to assess thyroids. Because of the small number of thyroid nodule images and significant variations across those images, the authors did not perform DL assessment of the thyroid nodules.¹²

This study was performed to construct several DL-based CAD systems for the detection of Graves' disease and subacute thyroiditis. Three DCNNs (AlexNet, VGGNet, and ResNet) were used to extract imaging features and classify the thyroid scintigraphy images according to the diagnostic results. We also compared the diagnostic performances of a first-year NM resident (C.Z.J.) and third-year NM resident (L.S.M.) with assistance from the best-performing DL model to assess the model's efficacy as a "diagnostic assistant." Notably, the residents were not permitted access to the true diagnostic results.

Materials and methods

Datasets

All thyroid scintigraphy images were obtained using a Discovery NM/CT 670 scanner (GE Healthcare, Chicago, IL, USA). The imaging conditions and methods were as follows. First, 370 MBq of (99m)Tc-pertechnetate was injected through the cubital vein before thyroid scintigraphy. After 15 to 30 minutes, a low-energy, high-resolution parallel-hole collimator was used for thyroid plane imaging. Upon completion of the planar imaging, a low-energy, high-resolution pinhole collimator was used for local magnification of the thyroid. The matrix was 128×128 , and the acquisition count was 5×10^5 .

We conducted a retrospective analysis of 1430 patients who underwent thyroid scintigraphy in the Department of Nuclear Medicine at Shanghai Tenth People's

Hospital from May 2016 to June 2019. The dataset included 175 patients with no thyroid disease, 834 patients with Graves' disease, and 421 patients with subacute thyroiditis during hyperthyroid status. All samples used met the following two criteria:

1. All diagnoses were confirmed through clinical history and auxiliary examinations including thyroid function tests (thyroid-stimulating hormone, free triiodothyronine, and free thyroxine), radioiodine uptake tests, and ultrasonography. Many cases were confirmed following treatment at follow-up.
2. The imaging results of all thyroid scintigraphy scans corresponded to the final clinical diagnosis.

System architecture

DL-based CAD performance is superior to traditional CAD system performance.⁸⁻¹² DL-based CAD takes thyroid

scintigraphy data as an input and automatically outputs the diagnostic results, as shown in Figure 1(a). Building a DCNN model involves two aspects: data augmentation and model training. We used progressive auxiliary classifier generative adversarial networks to augment the samples within each class, so that each class contained 1000 cases.

Initialization using transfer features can improve performance, and it is therefore a universally useful technique for improving the performance of deep neural networks.¹³ We used the pre-trained ImageNet to perform transfer learning by fine-tuning the last fully connected layer and freezing the previous network layers, as shown in Figure 1(b).

We split the dataset at a ratio of 7:3, with 70% of the data used for the training set and 30% of the data used for the validation set. We iteratively trained 1000 epochs with the training set for each network. We used

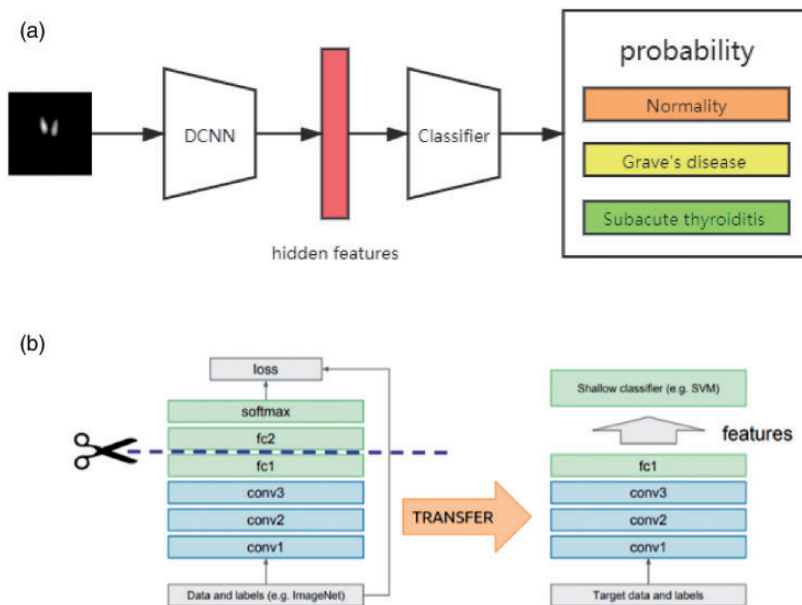


Figure 1. (a) Architecture of the CAD system. (b) Schematic diagram of transfer learning CAD, computer-aided diagnosis; Fc: fully connected layer; conv: convolutional layers.

the trained network to classify the validation set, and the output results were used to calculate the evaluation metric.

Data augmentation

To reduce the negative effect of class imbalance on the classifier, we generated synthetic samples with auxiliary classifier generative adversarial networks,¹⁴ which generated the specified type of samples. We introduced the concept of progressive growth¹⁴ to improve the quality of the synthetic images. The progressive growth process of the network is shown in Figure 2(a).

The network starts by generating low-resolution images and gradually increases the resolution of the generated images. By learning the pixel features of the images at different resolutions, the synthetic images contain more details. The synthetic images at different resolutions are shown in Figure 2(b), which presents the progressive transition of generated images from low resolution to high resolution.

Model training

AlexNet. AlexNet is composed of five convolutional layers followed by three fully

connected layers,¹⁶ as shown in Figure 3 (a). For the nonlinear portion, AlexNet uses a rectified linear unit instead of a tanh or sigmoid function, which was the earlier standard for traditional neural networks. The rectified linear unit trains much faster than the sigmoid because the derivative of the sigmoid becomes very small in the saturating region, and therefore the updates to the weights almost vanish. Additionally, AlexNet reduces the overfitting by using a dropout layer after every fully connected layer. The dropout layer has a probability associated with it, and it is applied at every neuron on the response map separately, as shown in Figure 3(b).

VGGNet. VGGNet¹⁷ offers an improvement over AlexNet by replacing large kernel-sized filters with multiple 3×3 -kernel-sized filters, one after another (Figure 3 (c)). Within a given receptive field, multiple stacked smaller kernels are more effective than larger ones because multiple nonlinear layers increase the depth of the network. This enables the network to learn more complex features at a lower cost. Blocks with the same filter size are applied multiple

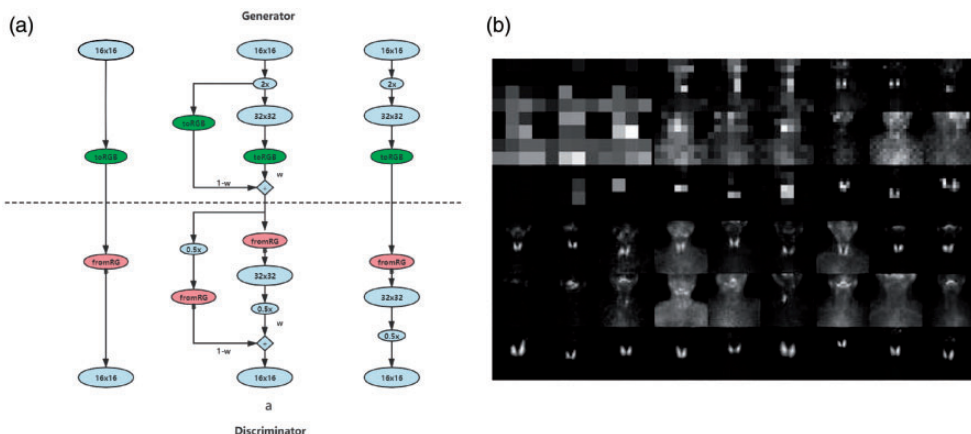


Figure 2. (a) Progressive auxiliary classifier generative adversarial network growth process. (b) Synthetic thyroid scintigraphy at different resolutions.

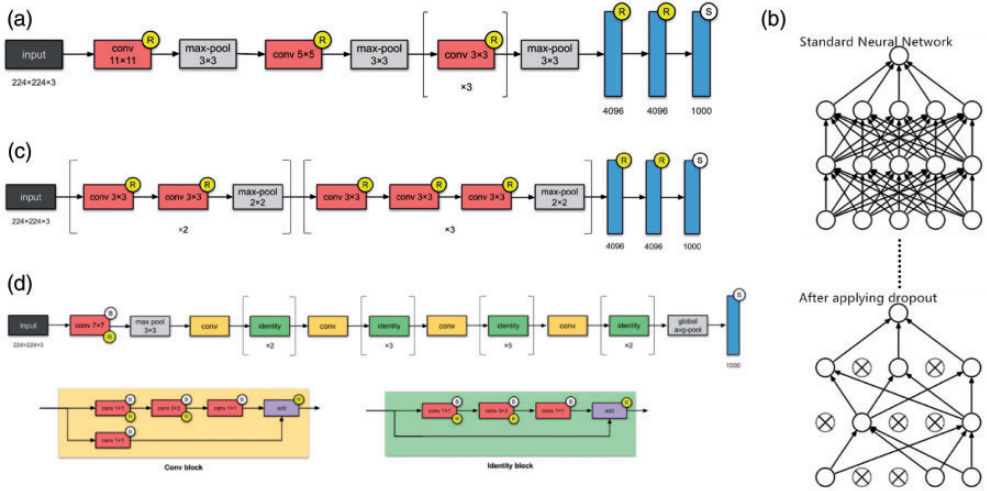


Figure 3. (a) Structure of AlexNet. (b) Network before and after dropout. (c) Structure of VGGNet. (d) Structure of ResNet.

times to extract more complex and representative features.

ResNet. Deep network structures can increase accuracy,¹⁸ although this results in a vanishing gradient and degradation problems. Residual networks^{19,20} allow training of such deep networks by constructing the network through modules called residual modules, as shown in Figure 3(d). The architecture is similar to VGGNet, consisting mostly of 3×3 filters. Using the VGGNet setup, a shortcut connection as described above is inserted to form a residual network.

Classifier. We chose the Softmax classifier to solve the multiclass classification problem. Softmax is given by:

$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

The Softmax classifier is the binary logistic regression classifier's generalization to multiple classes. It gives a slightly more

intuitive output (normalized class probabilities) and has a probabilistic interpretation. Its final conclusion reflects the max probability category.

Experimental environment. Computing was performed on the 8 NVIDIA GPUs model V100 PCIe (32 GB) hardware. The Compute Unified Device Architecture (CUDA) version 10.1 and the CUDA Deep Neural Network library (cuDNN) version 10.1 were used. The DCNN experimental code was developed based on the DL framework PyTorch 1.4.0 and Python 3.7.

Performance evaluation. The confusion matrix is a situation analysis that summarizes the predicted results of a classification model. The confusion matrix includes true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs). To further standardize the evaluation of the classification performance of the models, the following six indicators were taken from the confusion matrix: precision, recall, negative predictive value (NPV),

specificity, accuracy, and F1-score. They were defined as follows:

$$\begin{aligned} \text{Precision} &= \text{TPs}/(\text{TPs} + \text{FPs}), \\ \text{Recall} &= \text{TPs}/(\text{TPs} + \text{FNs}), \\ \text{NPV} &= \text{TNs}/(\text{TNs} + \text{FNs}), \\ \text{Specificity} &= \text{TNs}/(\text{FPs} + \text{TNs}), \\ \text{Accuracy} &= (\text{TPs} + \text{TNs})/(\text{Ps} + \text{Ns}), \\ \text{F1-score} &= 2 \times \text{TPs}/ \\ & (2 \times \text{TPs} + \text{FPs} + \text{FNs}). \end{aligned}$$

Additionally, the diagnostic performances of the three models were evaluated using receiver operating characteristic curve analysis, with the area under the curve tabulated for each model.

The Kappa coefficient is an effective evaluation method for multi-classification models in machine learning. It is used in consistency testing and can also be used to measure classification accuracy. P-values of <0.05 were considered statistically significant.

We also tested the accuracy of the ResNet-based CAD system when used in conjunction with an NM resident. NPV, specificity, accuracy, and the Kappa coefficient were used to describe the diagnostic performance.

Average diagnostic time. The total time required for each of the three models as well as the NM residents to diagnose the 300 cases was measured and then divided by 300 to determine the average time required for diagnosis of a single case.

Ethics

The protocol of Research on Automatic Diagnosis Report Generation Technology for Diversified Pathological Information Based on Machine Learning (SHSY-IEC-KY-3.0/18-147/01) was reviewed and approved by the ethics committee of Shanghai Tenth People's Hospital. All patients provided verbal informed consent to participate.

Results

Diagnostic performance of the three DCNN models

We summarized and compared the classification performance evaluation metrics of the three DCNN models, including the confusion matrices, recall, precision, NPV, specificity, accuracy, and F1-score, as shown in Figure 4 and Table 1. The consistency between the classification results and the clinical diagnostic results was tested with the Kappa coefficient. The three DCNN

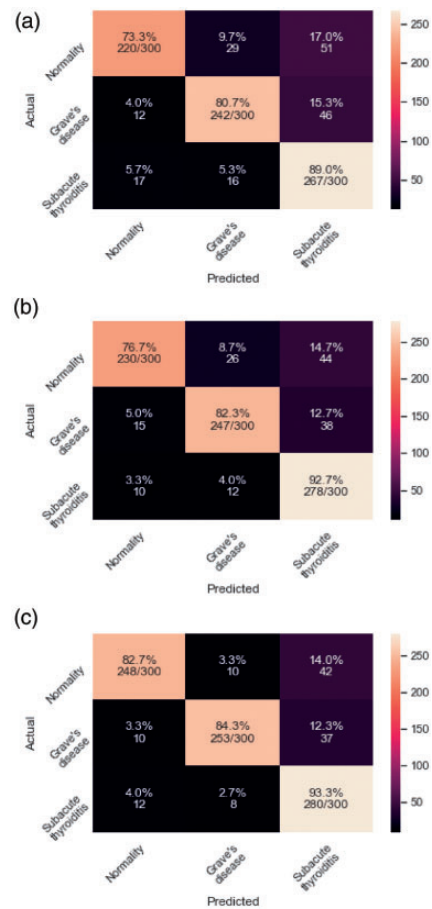


Figure 4. Confusion matrices for the three models. (a) AlexNet confusion matrix. (b) VGGNet confusion matrix. (c) ResNet confusion matrix.

Table 1. Comparison of six classification performance indexes (recall, precision, NPV, specificity, accuracy, and FI-score) of different models.

Model	Class	Recall	Precision	NPV	Accuracy	Specificity	FI-score	Kappa coefficient
AlexNet	Normality	73.33%	88.35%	87.71%	87.89%	95.17%	80.15%	0.715
	Graves' disease	80.67%	84.32%	90.54%	88.56%	92.50%	82.45%	
	Subacute thyroiditis	89.00%	73.35%	93.84%	85.56%	83.83%	80.42%	
VGGNet	Normality	76.67%	90.20%	89.15%	89.44%	95.83%	82.88%	0.758
	Graves' disease	82.33%	86.67%	91.38%	89.89%	93.67%	84.44%	
	Subacute thyroiditis	92.67%	77.22%	95.93%	88.44%	86.33%	84.24%	
ResNet	Normality	82.67%	91.85%	91.75%	91.78%	96.33%	87.02%	0.802
	Graves' disease	84.33%	93.36%	92.53%	92.78%	97.00%	88.62%	
	Subacute thyroiditis	93.33%	77.99%	96.30%	89.00%	86.83%	84.98%	

NPV, negative predictive value.

models achieved areas under the receiver operating characteristic curves ranging from 0.85 to 0.90 for the diagnosis of Graves' disease, subacute thyroiditis, and absence of thyroid disease (Figure 5).

Diagnostic performance of NM residents

We used the NPV, specificity, accuracy, and Kappa coefficient to characterize the diagnostic performance of the first- and third-year NM residents, both with and without the assistance of the ResNet-based CAD system, as shown in Tables 2 and 3.

Diagnostic time

The average diagnostic time for all three models (AlexNet, VGGNet, and ResNet) to classify one case in the test set was <1 s. The average diagnostic time for the first- and third-year NM residents to diagnose one case was 180 s and 120 s, respectively.

Discussion

The incidence of thyroid functional diseases has significantly increased in recent years. Graves' disease and subacute thyroiditis are often insidious because the early symptoms can be nonspecific. This can lead to delays in the diagnosis. Thyroid scintigraphy is

sensitive for the detection of thyroid functional disorders. However, limited medical resources have led to low diagnostic efficiency. Furthermore, variations in NM physicians' recognition of and the lack of defined standardized features on the images also contribute to variability in thyroid disease diagnosis. Machine learning will be an effective solution to the shortage of medical resources,²¹ especially with regard to automatic diagnosis of thyroid scintigraphy studies.²²⁻²⁵ The principal purpose of this paper is to describe the design of classifier models using different DL algorithms for the diagnosis of Graves' disease and subacute thyroiditis by thyroid scintigraphy.

DL has greatly improved the accuracy of machine learning methods in image recognition, demonstrating many achievements²⁶⁻²⁸ in the field of artificial intelligence and triggering an upsurge in research and development. However, few studies have focused on DL technology for the intelligent recognition of thyroid scintigraphy images.¹² The thyroid has a simple morphology with low variability and requires only simple image feature recognition, which makes it amenable to artificial intelligence diagnosis. Thus, we created three DCNN models and determined which model offers the best diagnostic accuracy.

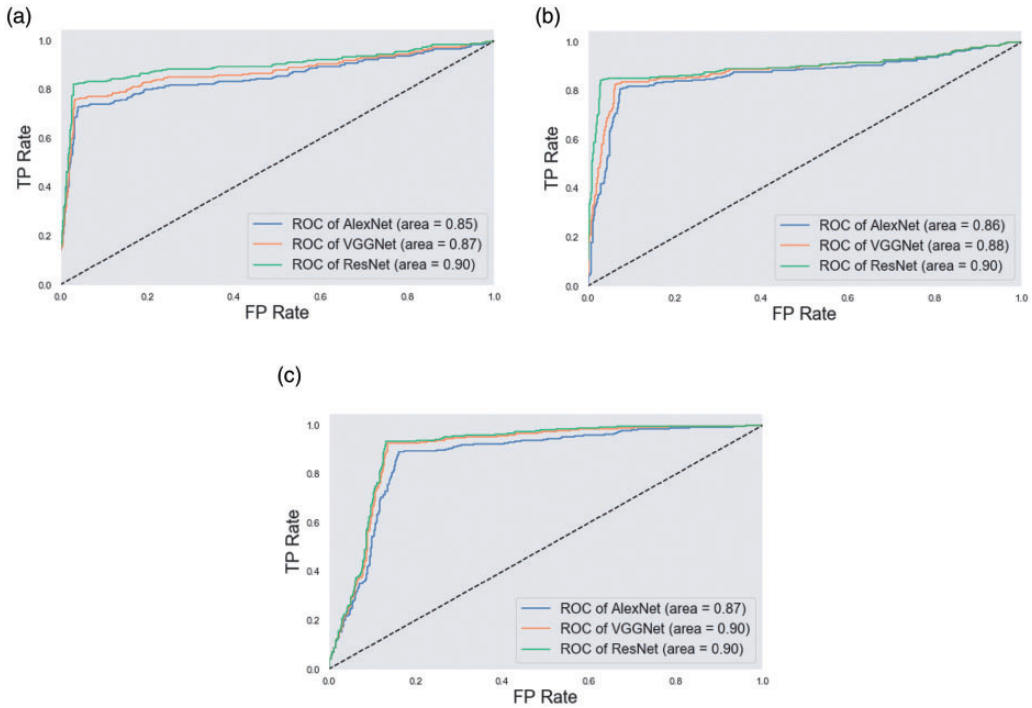


Figure 5. Receiver operating characteristic (ROC) curve comparison across the three models for the diagnosis of Graves' disease, subacute thyroiditis, and absence of thyroid disease. (a) ROC curves for patients without thyroid disease. (b) ROC curves for patients with Graves' disease. (c) ROC curves for patients with subacute thyroiditis.

Table 2. Diagnostic performances of two NM residents without CAD system assistance.

Reader	Class	NPV	Accuracy	Specificity	Kappa coefficient
First-year resident	Normality	86.55%	84.67%	91.17%	0.675
	Graves' disease	89.58%	87.33%	91.67%	
	Subacute thyroiditis	91.70%	84.67%	84.67%	
Third-year resident	Normality	92.48%	92.33%	96.33%	0.820
	Graves' disease	91.38%	89.89%	93.67%	
	Subacute thyroiditis	97.06%	90.33%	88.17%	

NM, nuclear medicine; CAD, computer-aided diagnosis; NPV, negative predictive value.

Our dataset is class-imbalanced and limited because thyroid scintigraphy is not part of the routine physical examination. We augmented the amount of thyroid scintigraphy data to ensure that the DCNN models had sufficient data to train. In addition, we used the pre-trained ImageNet to perform transfer learning and therefore achieve

better-performing models. Nevertheless, for multiclassification tasks in class-imbalanced datasets, accuracy alone is not sufficient to characterize the model's abilities. Therefore, we measured more thorough evaluation metrics, including recall, precision, NPV, specificity, accuracy, and F1-score. The accuracy of all three models

Table 3. Diagnostic performances of two NM residents with CAD system assistance.

Reader	Class	NPV	Accuracy	Specificity	Kappa coefficient
First-year resident	Normality	93.53%	91.67%	94.00%	0.810
	Graves' disease	93.01%	92.11%	95.33%	
	Subacute thyroiditis	94.50%	90.89%	91.67%	
Third-year resident	Normality	94.58%	93.67%	96.00%	0.853
	Graves' disease	93.48%	94.11%	98.00%	
	Subacute thyroiditis	97.51%	92.67%	91.33%	

NM, nuclear medicine; CAD, computer-aided diagnosis; NPV, negative predictive value.

assessed in this study reached at least 85.56% for the detection of Graves' disease, subacute thyroiditis, and absence of thyroid disease. The comprehensive evaluation metric (F1-score) for all three models was >80.15%, and the DL-based models accurately classified the validation set with areas under the curve ranging from 0.85 to 0.90. This indicated very good overall performance of the models. The Kappa coefficient for all three models was >0.715, indicating a high degree of consistency between the model prediction results and the clinical diagnostic results. These results suggest that the three models are effective in discriminating Graves' disease, subacute thyroiditis, and absence of thyroid disease from one another.

ResNet consistently showed better performance than AlexNet and VGGNet. ResNet was able to solve the degradation and vanishing gradient problems through the application of residual learning architecture and a batch normalization technique. Because the algorithm of VGGNet is better than that of AlexNet,¹⁷ it was not surprising that the classification performance results of VGGNet were better than those of AlexNet. With this in mind, ever-improving algorithms should lead to continuously improving classification performances.

The DCNN models significantly outperformed the first-year NM resident. Inexperienced NM physicians may overestimate their knowledge and experience,

making decisions on the basis of a limited number of imaging features. As such, they may fail to assess all imaging features systematically. Alternatively, DCNN models systematically and thoroughly assess all features within an image and readily perform adjustments in the context of noisy data. Therefore, DCNN models may effectively facilitate decision-making for inexperienced NM physicians. However, even the optimal ResNet model did not outperform the third-year resident. In a further assessment, we compared the residents' abilities with assistance from the ResNet model. With assistance from the model, the diagnostic performance of the first-year resident greatly improved, but the beneficial impact on the third-year resident was relatively small. This suggests that the benefit provided by DCNN models to experienced doctors is limited. The DCNN models provided results 400 to 600 times faster than the residents, indicating that the models may be helpful in streamlining clinical work, providing a reference for physicians' reports, and reducing work pressure on doctors.

This study has some limitations. First, we retrospectively gathered relatively typical images of patients with Graves' disease, subacute thyroiditis, and absence of thyroid disease to train the models rather than using thyroid images collected at random. Second, because of insufficient samples and class imbalances, some indistinctive image features that were regarded as suspicious

were neglected and deleted from the model constructions. This affected the FPs and FNs of the models. Third, images of more types of thyroid disease, especially thyroid nodules, need to be gathered, and greater participation by experienced professional physicians would further optimize the models. Translating technical success to significant clinical impact is the ultimate challenge.

In conclusion, we constructed three DCNN models to diagnose Graves' disease and subacute thyroiditis. Our results demonstrate that such models can serve as a "diagnostic assistant" to improve the diagnostic performance of NM residents. After the NM physician loads the image, the model can identify the image type and automatically generate a preliminary diagnostic result, helping to improve diagnostic efficiency, accuracy, and consistency. However, establishing the clinical utility of these models requires further clinical research.

Acknowledgements

The authors wish to thank the anonymous referees and editors of this special issue for their constructive comments. The authors also thank Prof. Russell Oliver Kosik for his support.

Declaration of conflicting interest

The authors declare that there is no conflict of interest.

Funding

This study was supported by the Major Project of Tongji University (Grant No. 15012150066); Program of Shanghai Academic/Technology Research Leader (Grant No. 18XD1403000); Major Issues of Production, Study and Research of Shanghai Science and Technology Commission (Grant No. 14DZ1940606); The National Natural Science Foundation of China (Grant No. 81671716); Shanghai 2018 "Science and Technology Innovation Action Plan" Science and Technology Support Project in Biomedicine (Grant No. 18441903500); and

Shanghai Leading Talent program sponsored by Shanghai Human Resources and Social Security Bureau (Grant No. 03.05.19005).

ORCID iD

Zhongwei Lv  <https://orcid.org/0000-0003-3370-5560>

References

1. Shan Z, Chen L, Lian X, et al. Iodine status and prevalence of thyroid disorders after introduction of mandatory universal salt iodization for 16 years in China: a cross-sectional study in 10 cities. *Thyroid* 2016; 26: 1125–1130.
2. Scappaticcio L, Trimboli P, Keller F, et al. Diagnostic testing for Graves' or non-Graves' hyperthyroidism: a comparison of two thyrotropin receptor antibody immunoassays with thyroid scintigraphy and ultrasonography. *Clin Endocrinol (Oxf)* 2020; 92: 169–178.
3. Uchida T, Suzuki R, Kasai T, et al. Cutoff value of thyroid uptake of (99m)Tc-pertechnetate to discriminate between Graves' disease and painless thyroiditis: a single center retrospective study. *Endocr J* 2016; 63: 143–149.
4. Zuhur SS, Ozel A, Kuzu I, et al. The diagnostic utility of color Doppler ultrasonography, Tc-99m pertechnetate uptake, and TSH-receptor antibody for differential diagnosis of Graves' disease and silent thyroiditis: a comparative study. *Endocr Pract* 2014; 20: 310–319.
5. Ramos CD, Zantut Wittmann DE, Etchebere ECSDC, et al. Thyroid uptake and scintigraphy using 99mTc pertechnetate: standardization in normal individuals. *Sao Paulo Med J* 2002; 120: 45–48.
6. Sorin V, Barash Y, Konen E, et al. Creating artificial images for radiology applications using generative adversarial networks (GANs) - a systematic review. *Acad Radiol* 2020; 27: 1175–1185.
7. Richardson ML, Garwood ER, Lee Y, et al. Noninterpretive uses of artificial intelligence in radiology. *Acad Radiol* 2020; S1076-6332(20)30039-8.

8. Buda M, Wildman-Tobriner B, Hoang JK, et al. management of thyroid nodules seen on US images: deep learning may match performance of radiologists. *Radiology* 2019; 292: 695–701.
9. Ippolito AM, De Laurentiis M, La Rosa GL, et al. Neural network analysis for evaluating cancer risk in thyroid nodules with an indeterminate diagnosis at aspiration cytology: identification of a low-risk subgroup. *Thyroid* 2004; 14: 1065–1071.
10. Lee JH, Ha EJ, Kim DY, et al. Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT: external validation and clinical utility for resident training. *Eur Radiol* 2020; 30: 3066–3072.
11. Chen A, Niermann KJ, Deeley MA, et al. Evaluation of multiple-atlas-based strategies for segmentation of the thyroid gland in head and neck CT images for IMRT. *Phys Med Biol* 2012; 57: 93–111.
12. Ma L, Ma C, Liu Y, et al. Thyroid diagnosis from SPECT images using convolutional neural network with optimization. *Comput Intell Neurosci* 2019; 2019: 6212759.
13. Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks? *Adv Neural Inf Process Syst* 2014; 27: 3320–3328.
14. Odena A, Olah C, and Shlens J. Conditional image synthesis with auxiliary classifier GANs. Proceedings of the 34th International Conference on Machine Learning. *Proc Mach Learn Res* 2017; 70: 2642–2651.
15. Karras T, Aila T, Laine S, et al. Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.
16. Krizhevsky A, Sutskever I and Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017; 60: 84–90.
17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
18. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2015; 2015: 1–9.
19. He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks. *Comput Vis ECCV Springer, Cham*, 2016: 630–645.
20. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2016: 2016; 770–778.
21. Do HM, Spear LG, Nikpanah M, et al. Augmented radiologist workflow improves report value and saves time: a potential model for implementation of artificial intelligence. *Acad Radiol* 2020; 27: 96–105.
22. Chan HP, Hadjiiski LM and Samala RK. Computer-aided diagnosis in the era of deep learning. *Med Phys* 2020; 47: e218–e227.
23. Thompson RF, Valdes G, Fuller CD, et al. Artificial intelligence in radiation oncology: a specialty-wide disruptive transformation? *Radiother Oncol* 2018; 129: 421–426.
24. Fazal MI, Patel ME and Tye J. The past, present and future role of artificial intelligence in imaging. *Eur J Radiol* 2018; 105: 246–250.
25. Xing L, Krupinski EA and Cai J. Artificial intelligence will soon change the landscape of medical physics research and practice. *Med Phys* 2018; 45: 1791–1793.
26. Hu Z, Li Y, Zou S, et al. Obtaining PET/CT images from non-attenuation corrected PET images in a single PET system using Wasserstein generative adversarial networks. *Phys Med Biol* 2020; 65: 215010.
27. Peng J, Kang S, Ning Z, et al. Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from CT imaging. *Eur Radiol* 2020; 30: 413–424.
28. Gong K, Berg E, Cherry SR, et al. Machine learning in PET: from photon detection to quantitative image reconstruction. *Proc IEEE* 2020; 108: 51–68.