Deep Convolution Neural Network Based Artificial Intelligence Improves Diagnosis of Thyroid Scintigraphy for Thyrotoxicosis: a Dual Center Study

Pei Yang
West China Hospital, Sichuan University https://orcid.org/0000-0002-9323-5137
Yong Pi
Sichuan University
Tao He
Panzhihua Municipal Central Hospital
Ke Zhou
Sichuan University West China Hospital
Xiao Zhong
Sichuan University West China Hospital
Yemei Liu
Sichuan University West China Hospital
Jiangming Sun
Panzhihua Municipal Central Hospital
Jianan Wei
Sichuan University
Yongzhao Xiang
Sichuan University West China Hospital
Lisha Jiang
Sichuan University West China Hospital
Lin Li
Sichuan University West China Hospital
Zhang Yi
Sichuan University
Zhen Zhao
Sichuan University West China Hospital
Huawei Cai (hw.cai@yahoo.com)
https://orcid.org/0000-0003-1341-8417

Original research

Keywords: Artificial intelligence, deep convolution neural network, thyroid scintigraphy, thyrotoxicosis

Posted Date: August 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-56117/v1

Abstract

Background: ^{99m}Tc-pertechnetate thyroid scintigraphy is a valid avenue for distinguishing causes of thyrotoxicosis in the clinic, but the interpretation of thyroid scintigraphic images is subjected with significant variation among different inter-observers. We aim to develop an artificial intelligence (AI) system to improve the diagnosis of thyrotoxicosis.

Materials and methods: We constructed an AI model based on a deep neural network with 2468 thyroid scintigraphic images collected from West China Hospital, and evaluated the diagnostic accuracy for classifying four patterns of thyrotoxicosis: 'diffusely increased,' 'diffusely decreased,' 'focal increased,' and 'heterogeneous uptake.' Then, we compared the diagnostic performance of the AI model and five physicians with 200 testing cohorts from two centers.

Results: We constructed the AI model, which has the best performance in internal database validation based on four kinds of standout pre-trained networks. This AI model achieves satisfactory performance in classifying four patterns of thyrotoxicosis with an overall accuracy of 91.92% for internal and 86.75% for external data validation. In the following contrastive study, the AI model represented improved diagnostic accuracy and consistency than 5 physicians for interpreting data from West China Hospital (88% vs. 66~73%) and Panzhihua Central Hospital (83% vs. 53%~79%), respectively.

Conclusion: Deep convolution neural network based AI model represented considerable performance in classifying four patterns of thyroid scintigraphic images; this may help physicians diagnose causes of thyrotoxicosis and reduced the physicians' error rate.

Background

Artificial intelligence (AI) involves wide aspects in the modern healthcare field. The distinguished advances of AI in big-data retrieval, explicit feature extraction, and satisfactory consistency are strongly beneficial to medical image analysis (1-3). Initially, the AI is mostly utilized in many monotonous, repetitive tasks and heavy workloads, such as evaluating screening mammography and chest X-rays. (4) At present, with the optimization of computer algorithms and the development of deep convolution neural networks (DCNN), AI has applied to more advanced projects about radiodiagnoses, such as automatic nodule detection for lung cancer in CT images and thyroid cancer identification in sonographic images (5, 6). However, there were few efforts have been made in nuclear image interpretation, which also requires a variety of repetitive information as well as proper clinical feature extraction for diagnosis of disease, such as thyroid scintigraphy.(7)

Thyrotoxicosis is a very common endocrine condition with multiple etiologies, and the misdiagnosis might lead to improper medical treatments. Despite blood biochemical detection and ultrasonography are widely used, it is still not enough to obtain a precise diagnosis in several situations. Thyroid scintigraphy with ^{99m}Tc-pertechnetate is a valid avenue to identify the causes of thyrotoxicosis, especially for distinguishing Graves' disease (GD) and toxic multinodular goiter (TMG) when both thyrotropin receptor

antibody was negative or differentiating GD from thyroiditis (8). However, thyroid scintigraphy images are mostly simple planar with limited resolution, which makes the interpretation is a time-consuming, experience-dependent, and subjective work with significant variation among inter-observer measurement (9). Thus, an AI model with the capability of feature extraction, ideal diagnostic accuracy, and consistency for thyroid image interpretation is probably an effective strategy to improve the clinical diagnosis of thyrotoxicosis.

In this study, we collected 2468 thyroid scintigraphic images to establish a deep learning neural network and constructed an automatic AI model for the classification of thyrotoxicosis. Then, we evaluated the diagnostic performance of AI model and compared it with human physicians based on datasets from dual centers.

Methods

Collection, Inclusion, and Exclusion of Patients

This study with retrospective information collection was approved by the Institutional Ethics Committee of West China Hospital in Sichuan University and Panzhihua Central Hospital, respectively. We retrospectively collected cases who were determined as thyrotoxicosis and underwent ^{99m}Tc-pertechnetate thyroid scintigraphy from January 1, 2016 to December 31, 2018 at West China Hospital of Sichuan University (Center 1) and Panzhihua Central Hospital (Center 2). Patients who received anti-thyroid drugs, radioactive iodine therapy or semi/total thyroidectomy were excluded, and images with poor quality or lateral acquisition were also excluded. The thyroid scintigraphy in two hospitals was obtained following the clinical guidelines and manufacturer recommended parameters. Briefly, patients were intravenously injected with 185 MBq of $^{99m}TcO4^-$, and then the images were captured for 100-300×10^3 counts about 5-10min using the gamma cameras, which were both equipped with the low-energy, high-resolution, parallel-hole collimators (GE Discovery NM/CT 670). The energy peak was centered at 140 keV with 15% to 20% windows. All the images were exported as DICOM format for further analysis.

Diagnostic Criteria

Thyroid scintigraphic images were defined as four patterns referring to published criteria (10-13). The ones had homogeneous increased uptake over than the uptake of salivary with enlarged thyroid were defined as ‘Diffusely increased’ (type I); the ones had diminished, and absent uptake was defined as ‘diffusely decreased’ (type II); the ones had focal nodule uptake with suppressed uptake in the surrounding, and contralateral thyroid tissue was defined as ‘local increased’ (type III), and the ones had multiple areas of focal increased and suppressed uptake was defined as ‘heterogeneous uptake’ (type IV). All characteristic performance of these four patters images were shown **in Fig. 1.** For this study, all thyroid scintigraphy images from two centers were independently and blindly classified by three senior nuclear

medicine physicians with more than 10 years of working experience in reading thyroid scintigraphic images. Consensus shall be reached by consulting if there is disagreement.

Construction of AI Model

The images collected from center 1 was defined as the internal dataset for AI construction and internal validation, while the images from center 2 were defined as the external dataset for validation only. The architecture of AI model is illustrated in **Fig. 2**. There are three main steps in the training process: data augmentation, feature extraction, and classification. At the first step, random flipping, rotating, and mix-up are applied to the original image to increase the diversity of the data and improve the robustness of the model in augmentation (14). Then, a feature extraction neural network is employed to extract high-level features from the input image. In this study, we explored four kinds of candidate AI models based on different standout pre-trained networks, including ResNet50 (15), DenseNet169 (16)，InceptionV3 (17), InceptionResNetV2 (18). All these networks have been removed the last fully-connected layer and employed as the feature extraction network. At the final step, a three-layer neural networks are constructed to classify the high-level features into four classes. In the current study, all models were trained using Adam (19) as the optimizer with a weight decay rate of 0.0001 and a learning rate of 0.001 for 300 epochs. The mini-batch size was fixed 12. To reduce the side effect of overfitting, dropout (20) was employed to the last fully connected layer, with a drop probability of 0.8.

Evaluation of Model Performance

The diagnostic accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of four candidate AI models were individually evaluated to select the one with the best performance in the internal dataset validation. Then, the performance of selected AI in the validation of the external dataset was evaluated by areas under the curve (AUC) of receiver operating characteristic (ROC) as well. To further investigate the diagnostic efficiency and accuracy of AI model, we randomly chose two testing cohorts from dual centers containing 100 cases in each and compared the classification performance of our AI and five nuclear medicine physicians from two centers. All physicians were unaware of any patients' clinical details. The overall accuracy of AI model and human physicians were recorded, while the differences of the diagnostic labels and true labels in identifying four thyroid patterns were represented by the confusion matrix of 4 × 4 contingency table.

Statistical analysis

In this study, we calculated the sensitivity, specificity, accuracy, PPV, and NPV of AI

model for classifying four patterns of thyrotoxicosis on thyroid scintigraphy. And we compared overall accuracy between five physicians and AI model. All analyses were performed by using statistical software SPSS 21.0 (SPSS Inc, Chicago, IL, USA).

Results

Patient Characteristics

The distribution of thyroid images used for AI construction was shown in **Table 1**. We collected 2468 cases of thyroid scintigraphic images (2396 female and 72 male; age: 41.24±14.25) as a training cohort and 619 cases (611 female and 8 male; age:41.20 ±14.20) as internal validating cohort from West China Hospital of Sichuan University (center 1). Another 302 cases (214 female and 88 male; age: 44.61±13.68) were obtained from Panzhihua Central Hospital (center 2) as an external validating cohort.

Source	Dataset	Diffusely increased	Diffusely decreased	Focal increased	Heterogeneous uptake	Total
Training	Internal	1040	902	97	429	2468
Validating	Internal	260	226	25	108	619
	External	120	49	13	120	302

Table 1. The distribution of thyroid images used in the current study.

Selection of AI model

The individual performances of AI models from four DCNN methods in internal validation were shown in **Table 2.** The InceptionV3 achieved the highest overall accuracy of 92.73% in classifying four patterns of thyroid images that were selected for further use. As shown in **Fig. 3**, the AUC values of this model was 0.986 for 'diffusely increased,' 0.997 for 'diffusely decreased,' 0.998 for 'focal increased,' and 0.945 for 'heterogeneous uptake' in internal validation, respectively. Accordingly, the AUC performances of the InceptionV3 model also obtained in ideal in external validation were 0.939, 1.000, 0.974, and 0.915, respectively.

Models	Diffusely increased	Diffusely decreased	Focal increased	Heterogeneous uptake	Overall accuracy
InceptionV3	94.51%	99.68%	98.38%	92.89%	92.73%
InceptionResnetV2	94.02%	99.35%	98.06%	92.41%	91.92%
DenseNet169	94.83%	99.03%	95.32%	91.44%	90.30%
ResNet50	93.70%	99.52%	98.55%	92.08%	91.92%

Table 2. The performance of CNN methods in internal cohorts that including InceptionV3, InceptionResnetV2, DenseNet169 and ResNet50

Contrastive Study of AI model and Human

Two interns with three-year experience (physician 1 and 3, from center 1) and three staffs with 5~6 years' experience (physician 2 from center 1, physician 4 and 5 from center 2) participated in the comparative experiment. AI cost 0.54 seconds to accomplish the interpretation, while the human spent an average time of 32.50±13.29 minutes (21-62 minutes) to finish the same work. As shown in **Fig. 4**, the AI model still achieved the highest overall diagnostic accuracy in classifying images in both datasets (88% and 83%, respectively), while the best performances of humans were 73% and 79% in this test. The confusion matrixes indicated individual representative features of AI and human physicians in classification. Notably, in reading the data from center 1 (**Fig. 5**), the disparity in diagnostic accuracy is mainly caused by identifying 'heterogeneous uptake'; AI successfully identified 18 out of 25 cases, while human physicians only identified 5~15 cases. In **Fig. 6**, although AI indicated a certain decline in identifying 'heterogeneous uptake' in the dataset from center 2, it still represented more satisfying overall diagnostic accuracy and consistency than human physicians.

Discussion

According to clinical practice, thyrotoxicosis could be classified into two general situations, thyrotoxicosis with or without hyperthyroidism. Hyperthyroidism involves Grave's disease, or TMG accelerated biosynthesis and secretion of thyroid hormone by the thyroid gland itself, which may require radioactive iodine ablation or anti-thyroid drugs; whereas subacute thyroiditis or autoimmune thyroiditis caused thyrotoxicosis without hyperthyroidism could cure spontaneously (13, 21, 22). Thus, distinguishing the real causes of thyrotoxicosis is essential for therapeutic consultation. Thyroid scintigraphy supplies an effective avenue to represent the status of the thyroid and distinguish the causes of thyrotoxicosis by four-pattern classification. Generally, 'diffusely increased' suggests GD, 'diffusely decreased' pattern suggests thyroiditis, the 'local increased' is suggestive of toxic adenoma, and the 'heterogeneous uptake' is a clue to TMG.(9, 12)

Although there are several guidelines for clinical diagnosis, the superposition of subtle differences between devices, injected drug doses, imaging parameters, and subjective variation of physicians might lead to a different diagnostic conclusions. Thus, we developed an AI model and hoped to help clinical physicians remedy this unsatisfactory situation. In this study, the images were captured for 100×10^3 counts from West China Hospital, which are typical "low-abundant" images, whereas the images from Panzhihua Central Hospital were captured for 300×10^3 counts, which are "high-abundant" examples. Our AI model was constructed from the data in West China Hospital and achieved ideal diagnostic accuracy in identifying the thyroid images from Panzhihua Central Hospital. Beyond the time-saving effect, this AI model is beneficial to classify four common patterns of thyrotoxicosis on thyroid scintigraphy images, especially in identifying the patterns of 'heterogeneous uptake' than human physicians in dual centers,

which is able to reduce the physician's subjective variations in interpreting thyroid images and provide more appropriate management for patients.

Our current AI model was not used to "replace" the physicians, but to "assist" doctors improve the diagnostic accuracy and efficiency, and provide a proper clue for therapeutic selection. However, there are still several limitations in this study. Firstly, a significant decrease of diagnostic accuracy of 'heterogeneous uptake' was found when AI shifted to the new dataset of "high-abundant images," which indicates more optimizations are required in picture feature extraction beyond DCNN. Secondly, although four-pattern classification is valid in general situations, the real status of thyroid function still might be alternative following with different diseases and courses. For example, the patients with Hashitoxicosis could represent all these four patterns during the different courses (23, 24). Thus, the final diagnosis of thyrotoxicosis must be correlated with more information, such as medical history, physical findings, and blood tests, etc. Nevertheless, a new multi-parameter AI model containing both images and hematology index for diagnosis of thyrotoxicosis is under development. Then, further validations by more centers are still required for the optimization of this AI model.

Conclusion

We have successfully constructed an AI model for classifying four common patterns of thyrotoxicosis on thyroid images and achieved considerable diagnostic accuracy in dual centers. With further assessment and validation, this model might be promising in the clinical diagnosis of thyrotoxicosis.

Abbreviations

AI: Artificial intelligence

DCNN: Deep convolution neural networks

GD: Graves' disease

TMG: Toxic multinodular goiter

PPV: Positive predictive value

NPV: Negative predictive value

AUC: Areas under the curve

ROC: Receiver operating characteristic

Declarations

Availability of data and material

The datasets generated and analyzed during the current study are not publicly available but available from the corresponding author upon reasonable request.

Acknowledgments

Not applicable

Funding

This project was financially supported by the National Major Science and Technology Projects of China (2018AAA0100201), the Sichuan Provincial Science and Technology Project of the Health Planning (19PJ79), the Sichuan Science and Technology Program of China (2020JDRC0042) and “1.3.5 project for disciplines of excellence in West China Hospital (ZYGD18016).

Author information

Affiliations

Department of Nuclear Medicine, West China Hospital, Chengdu, 610041, P.R.China.

Pei Yang, Ke Zhou, Xiao Zhong, Yemei Liu, Yongzhao Xiang, Lisha Jiang, Lin Li, Zhen Zhao and Huawei Cai.

Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, P. R. China.

Yong Pi, Jianan Wei, and Zhang Yi

Department of Nuclear Medicine, Panzhihua Central Hospital, Panzhihua, 617067, P.R.China.

Tao He and Jiangming Sun

Contributions

All the authors participated in the study. Cai H, Zhao Z, Pi Y and Yi Z designed this study and completed the drafting of manuscript. Yang P, He T, Xiang Y, and Jiang L carried out the clinical data collection. Pi Y, Wei J, and Yi Z participated in the construction of AI model. Yang P, Zhong X, Liu Y, Sun J, Zhao Z, Li L, and Cai H completed the validation of AI performance.

Corresponding author

Correspondence to Huawei Cai and Zhen Zhao

Ethics declarations

Ethics approval and consent to participate

The study protocol was approved by the by the Institutional Ethics Committee of West China Hospital in Sichuan University and Panzhihua Central Hospital ethics committee.As this study was of retrospective nature, a consent form was waived by the local ethics committee.

Consent for publication

Not applicable

Competing interests

The authors declare there they have no conflict of interest.

References

1. Dong M, Huang X, Xu B. Unsupervised speech recognition through spike-timing-dependent plasticity in a convolutional spiking neural network. PLoS One. 2018; 13(11):e0204596.
2. Frank DA, Chrysochou P, Mitkidis P, Ariely D. Human decision-making biases in the moral dilemmas of autonomous vehicles. Sci Rep. 2019; 9(1):13080.
3. Moravcik M, Schmid M, Burch N, et al. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. Science. 2017; 356(6337):508-13.
4. Choy G, Khalilzadeh O, Michalski M, et al. Current Applications and Future Impact of Machine Learning in Radiology. Radiology. 2018; 288(2):318-28.
5. Zhang G, Jiang S, Yang Z, et al. Automatic nodule detection for lung cancer in CT images: A review. Computers in biology and medicine. 2018; 103:287-300.
6. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. The Lancet Oncology. 2019; 20(2):193-201.
7. Ma L, Ma C, Liu Y, Wang X. Thyroid Diagnosis from SPECT Images Using Convolutional Neural Network with Optimization. Comput Intell Neurosci. 2019; 2019:6212759.
8. Giovanella L, Avram AM, Iakovou I, et al. EANM practice guideline/SNMMI procedure standard for RAIU and thyroid scintigraphy. Eur J Nucl Med Mol Imaging. 2019; 46(12):2514-25.
9. Patel KA, Warren R, Brooke A, et al. Interpretation of thyroid scintigraphy is inconsistent among endocrinologists. J Endocrinol Invest. 2017; 40(10):1155-7.
10. Intenzo CM, dePapp AE, Jabbour S, Miller JL, Kim SM, Capuzzi DM. Scintigraphic manifestations of thyrotoxicosis. Radiographics. 2003; 23(4):857-69.
11. Smith JR, Oates E. Radionuclide imaging of the thyroid gland: patterns, pearls, and pitfalls. Clin Nucl Med. 2004; 29(3):181-93.
12. Sharma A, Stan MN. Thyrotoxicosis: Diagnosis and Management. Mayo Clinic proceedings. 2019; 94(6):1048-64.

13. Ross DS, Burch HB, Cooper DS, et al. 2016 American Thyroid Association Guidelines for Diagnosis and Management of Hyperthyroidism and Other Causes of Thyrotoxicosis. Thyroid : official journal of the American Thyroid Association. 2016; 26(10):1343-421.
14. Zhang H, Cisse M, Dauphin YN, Lopezpaz DJaL. mixup: Beyond Empirical Risk Minimization. 2017.
15. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. computer vision and pattern recognition2016; 770-8.
16. Huang G, Liu Z, Der Maaten LV, Weinberger KQ. Densely Connected Convolutional Networks. computer vision and pattern recognition2017; 2261-9.
17. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. computer vision and pattern recognition2016; 2818-26.
18. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. national conference on artificial intelligence2016; 4278-84.
19. Kingma DP, Ba JJaL. Adam: A Method for Stochastic Optimization. 2014.
20. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov RJJoMLR. Dropout: a simple way to prevent neural networks from overfitting. 2014; 15(1):1929-58.
21. Cooper DS. Hyperthyroidism. Lancet (London, England). 2003; 362(9382):459-68.
22. Teelucksingh S, Motilal MS, Bailey H, et al. MANAGEMENT OF THYROTOXICOSIS AMONG GENERAL PRACTITIONERS IN TRINIDAD COMPARED WITH 2016 AMERICAN THYROID ASSOCIATION GUIDELINES FOR HYPERTHYROIDISM. Endocrine practice : official journal of the American College of Endocrinology and the American Association of Clinical Endocrinologists. 2019; 25(7):657-62.
23. Charles M. Intenzo M. Radiographics-Scintigraphic features of autoimmune thyroiditis. Radiographics. 2001; 21:957-64.
24. Meier DA, Kaplan MM. Radioiodine uptake and thyroid scintiscanning. Endocrinol Metab Clin North Am. 2001; 30(2):291-313.

Figures

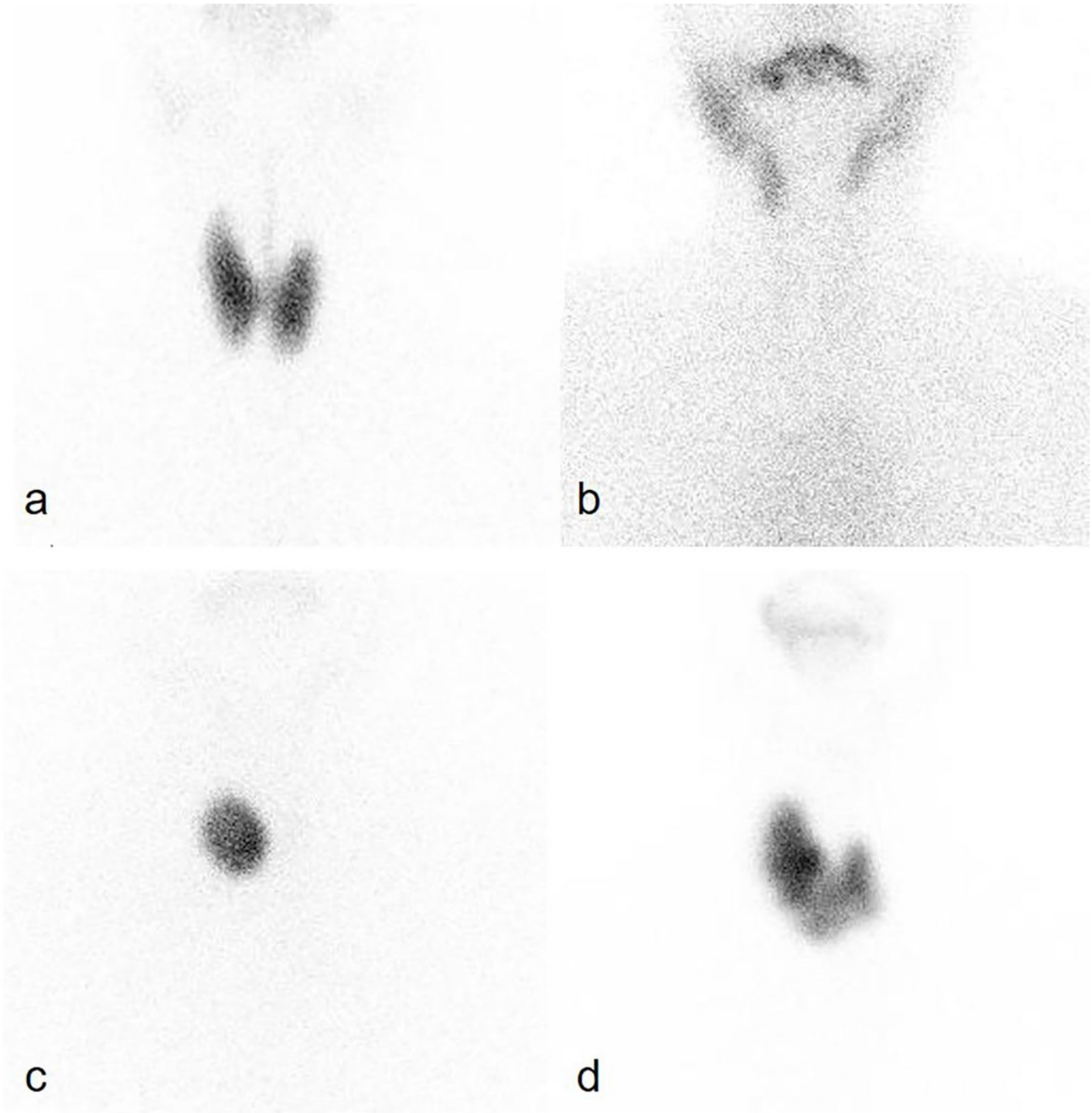

Figure 1

The characteristic performance of ‘diffusely increased’ (a), ‘diffusely decreased’ (b), ‘local increased’ (c) and ‘heterogeneous uptake’ (d)

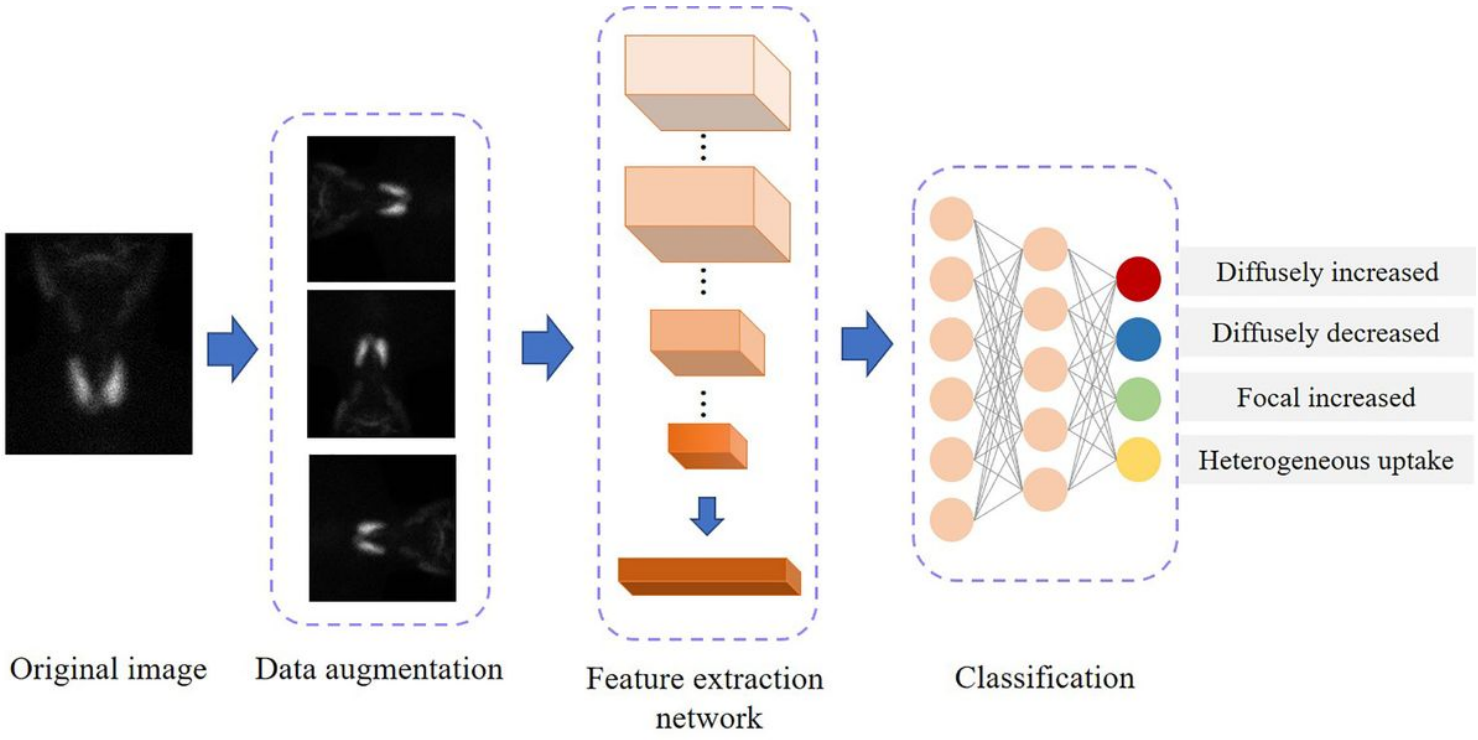

Figure 2

The architecture process of AI model

a

Metrics	Dataset	Diffusely increased	Diffusely decreased	Focal increased	Heterogeneous uptake
		InceptionV3			
Accuracy	Internal	94.51%	99.68%	98.38%	92.89%
	External	89.74%	99.34%	98.68%	87.75%
Sensitivity	Internal	90.77%	99.56%	100.00%	81.48%
	External	90.00%	95.92%	76.92%	83.33%
Specificity	Internal	97.21%	99.75%	98.32%	95.30%
	External	89.56%	100.00%	99.65%	90.66%
PPV	Internal	95.93%	99.56%	71.43%	78.57%
	External	85.04%	100.00%	90.91%	85.47%
NPV	Internal	93.57%	99.75%	100.00%	96.06%
	External	93.14%	99.22%	98.97%	89.19%

b

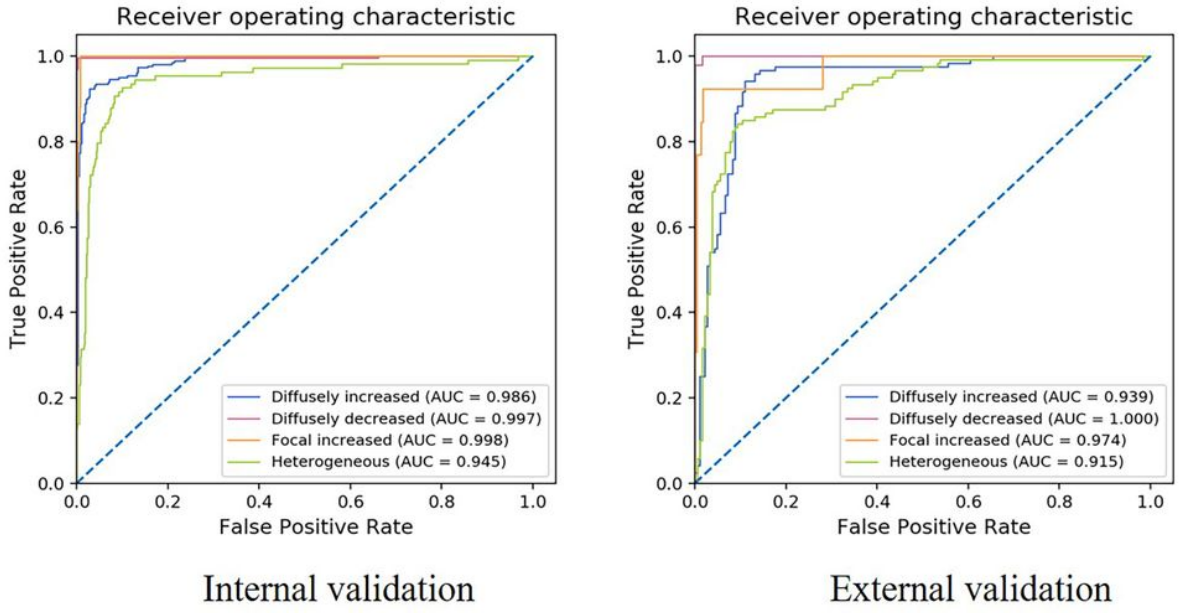

Figure 3

The performance (a) and AUC calculation (b) of InceptionV3 AI model in internal and external validation in classifying fours patterns of thyrotoxicosis

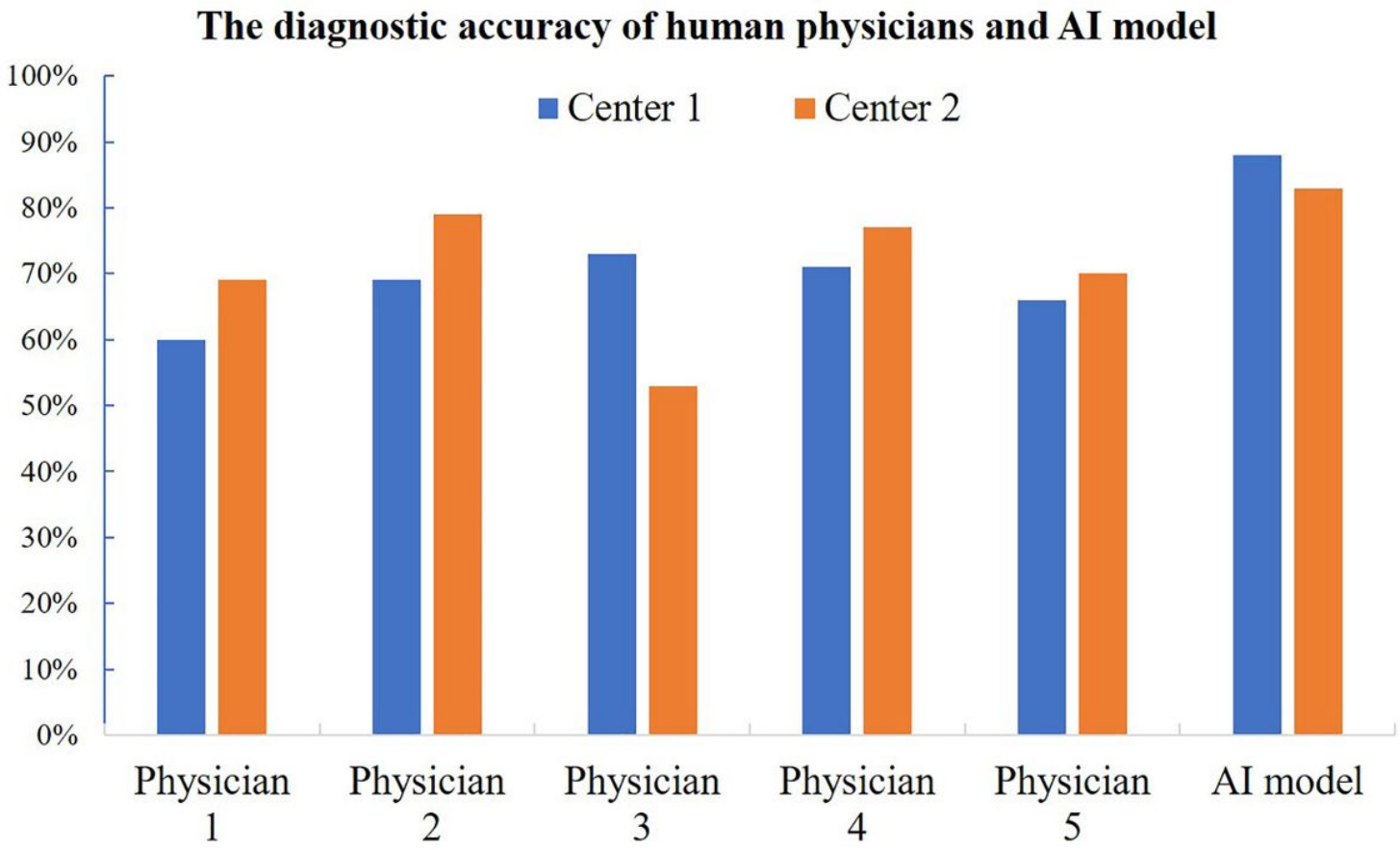

Figure 4

The diagnostic performance of human physicians and AI model

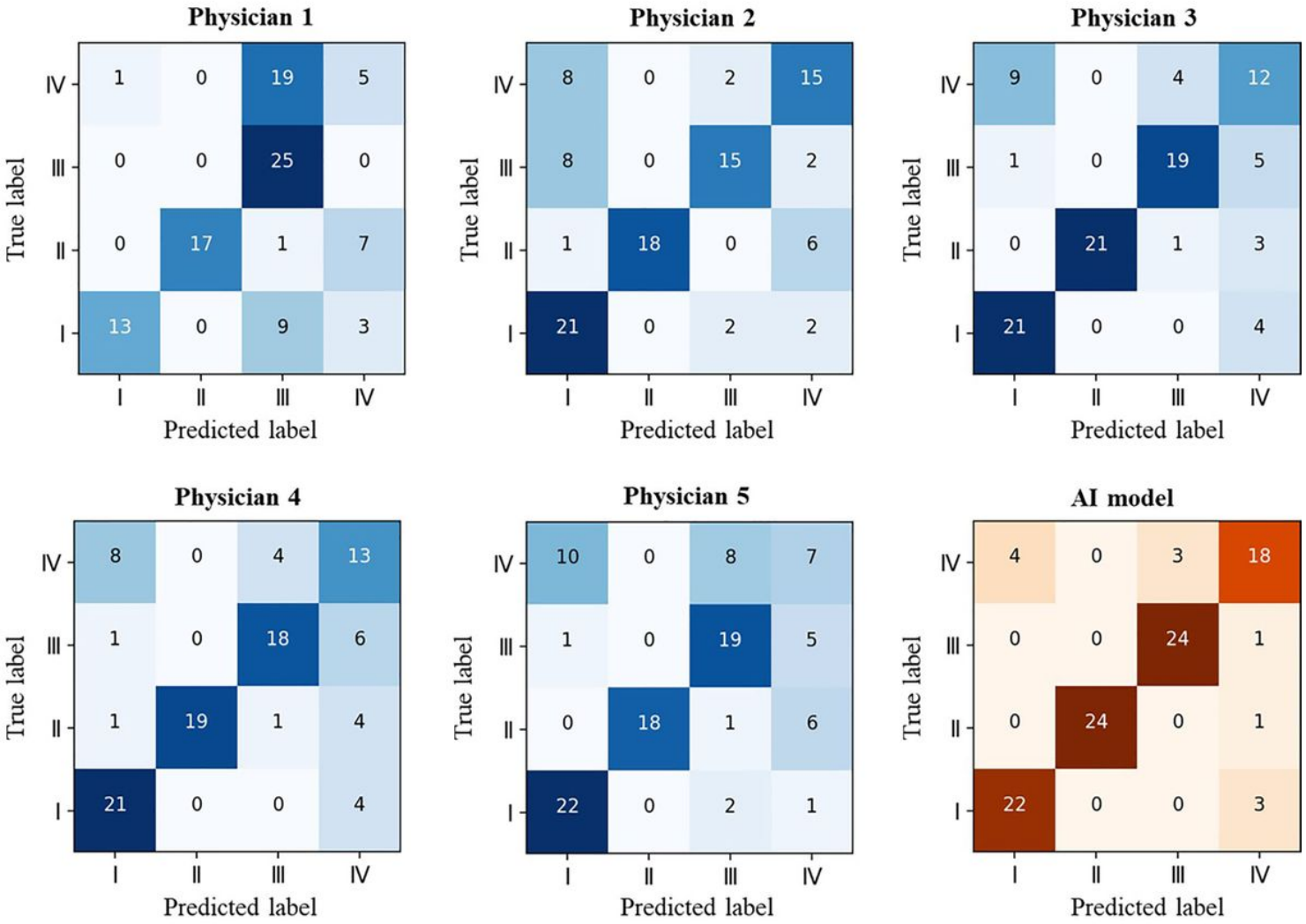

Figure 5

The results of confusion matrix between five physicians and AI model in Center 1. Type Ⅰ: diffusely increased; Type Ⅱ: diffusely decreased; Type Ⅲ: local increased; Type Ⅳ: heterogeneous uptake

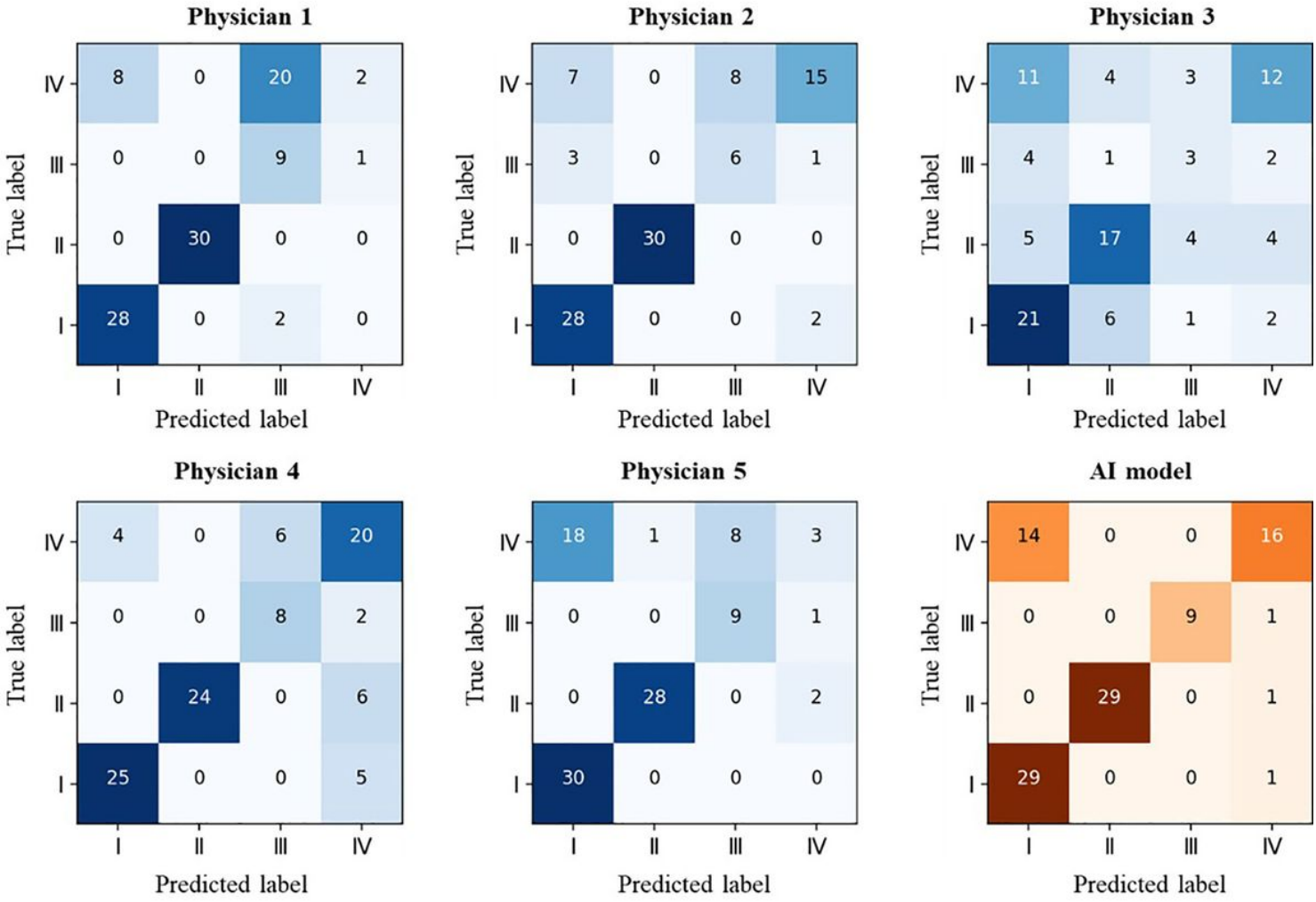

Figure 6

The confusion matrix between five physicians and AI model in Center 2. Type Ⅰ: diffusely increased; Type Ⅱ: diffusely decreased; Type Ⅲ: local increased; Type Ⅳ: heterogeneous uptake