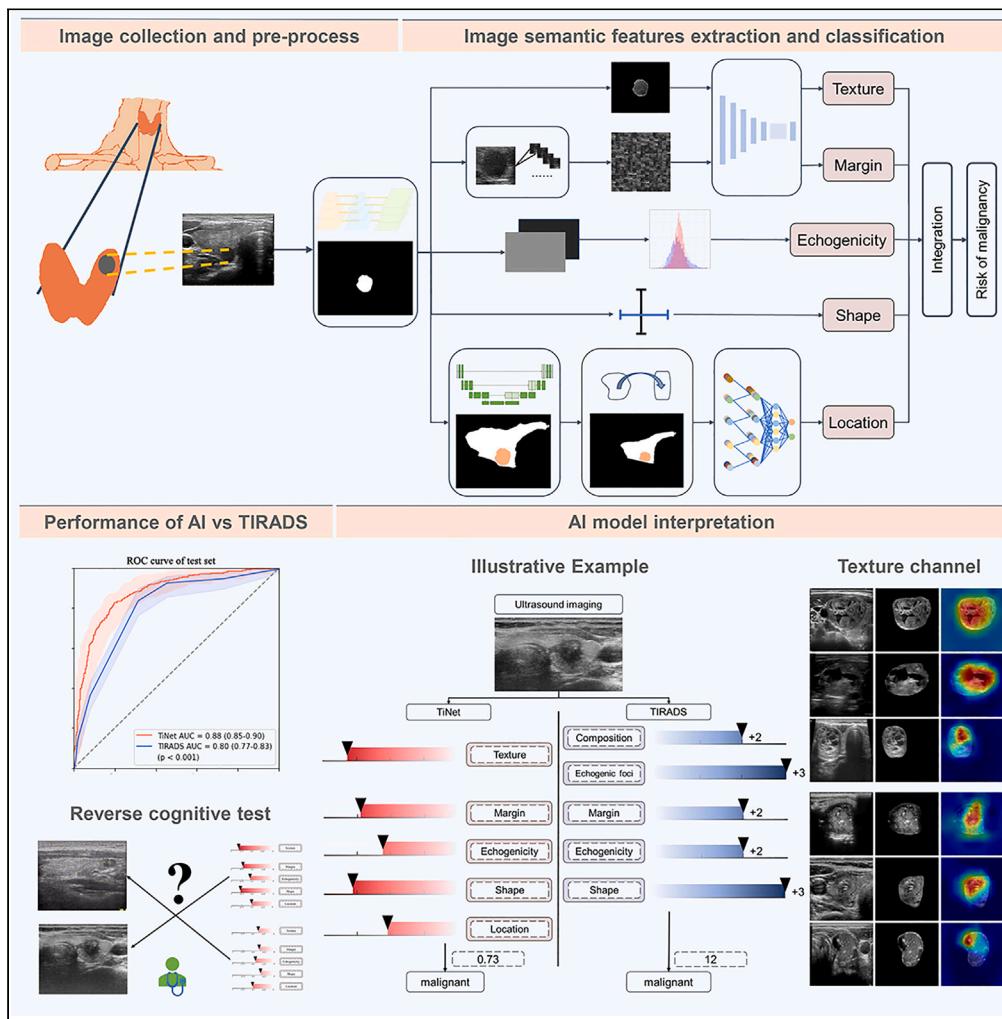


Article

Human understandable thyroid ultrasound imaging AI report system — A bridge between AI and clinicians



Siqiong Yao,
Pengcheng Shen,
Tongwei Dai, Fang
Dai, Yun Wang,
Weituo Zhang, Hui
Lu

wangyun31415926@163.com
(Y.W.)
zhangweituo@sjtu.edu.cn
(W.Z.)
huilu@sjtu.edu.cn (H.L.)

Highlights
TiNet is a human
understandable AI version
thyroid cancer ultrasound
prediction system

TiNet can provide
quantitative explanations
for prediction factors of
thyroid cancer

TiNet can serve as a
bridge between AI and
clinicians to assist medical
decision-making

Article

Human understandable thyroid ultrasound imaging AI report system — A bridge between AI and clinicians

Siqiong Yao,^{1,6} Pengcheng Shen,^{1,6} Tongwei Dai,¹ Fang Dai,¹ Yun Wang,^{3,*} Weituo Zhang,^{4,*} and Hui Lu^{1,2,5,7,*}

SUMMARY

Artificial intelligence (AI) enables accurate diagnosis of thyroid cancer; however, the lack of explanation limits its application. In this study, we collected 10,021 ultrasound images from 8,079 patients across four independent institutions to develop and validate a human understandable AI report system named TiNet for thyroid cancer prediction. TiNet can extract thyroid nodule features such as texture, margin, echogenicity, shape, and location using a deep learning method conforming to the clinical diagnosis standard. Moreover, it offers excellent prediction performance (AUC 0.88) and provides quantitative explanations for the predictions. We conducted a reverse cognitive test in which clinicians matched the correct ultrasound images according to TiNet and clinical reports. The results indicated that TiNet reports (87.1% accuracy) were significantly easier to understand than clinical reports (81.6% accuracy; $p < 0.001$). TiNet can serve as a bridge between AI-based diagnosis and clinicians, enhancing human–AI cooperative medical decision-making.

INTRODUCTION

In 2020, over 586,000 cases were recorded to be caused by thyroid cancer worldwide, making it the most pervasive endocrine cancer.¹ Ultrasonography is currently the most commonly used method for detecting thyroid cancer, and with improvements in resolution, detection rates of thyroid cancer keep increasing.¹ According to previous studies, artificial intelligence (AI) on ultrasound images using deep learning models could achieve higher diagnostic performance for thyroid cancer than human experts.^{2–8}

However, AI models have been extensively criticized for their black-box nature; that is, they can provide a prediction, but fail to interpret themselves as a human understandable reasoning process.^{9–11} The following several aspects are the major concerns of the uninterpretable AI from clinicians. First, clinicians are interested in understanding how AI achieves its higher diagnostic accuracy compared to human experts. A specific question is whether AI models find some novel features unrecognized by human experts before, or does AI only optimize known features with more precise digitization (e.g., converting categorical features into numerical features).^{12,13} Second, in real-world healthcare settings, clinicians must combine AI predictions with their own ultrasound findings and other types of clinical information of patients to achieve medical-related decision-making. However, this task would be difficult to perform without interpreting AI results, especially when AI predictions contradict clinicians' own assessment.¹⁴ Third, current AI models are mostly trained and validated in cleaned and selective data. Hence, unexpected errors likely occur in complicated real-world healthcare settings.^{15,16} For example, if a clinician mistakenly inputs an X-ray image into an ultrasound AI model, the AI will still output a prediction anyway. Without a human understandable interpretation of AI, clinicians would find it difficult to be aware of the occurrence of such error, which could potentially harm the patients.¹⁷ Therefore, the aim of this study is to design, develop, and validate an AI report system for ultrasonographical thyroid cancer diagnosis that, in addition to malignancy prediction, provides human understandable features that AI used to derive its diagnosis.

The development of interpretable or explainable AI (XAI) has garnered significant attention from scholars across various disciplines in recent years.^{18–21} However, no consensus exists on the type of explanation that is truly desired and how these explanations are evaluated, especially in medicine. As most methodologies are invented by computer scientists, these methodologies are more suitable for engineering purposes.

¹Joint International Research Laboratory of Metabolic & Developmental Sciences, Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

²JJTU-Yale Joint Center of Biostatistics and Data Science, National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai, China

³Department of Hepatobiliary pancreatic center, Xuzhou City Central Hospital, The Affiliated Hospital of the Southeast University Medical School (Xuzhou), The Tumor Research Institute of the Southeast University (Xuzhou), Xuzhou, Jiangsu, China

⁴Hong Qiao International Institute of Medicine, Shanghai Tong Ren Hospital and Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China

⁵Shanghai Engineering Research Center for Big Data in Pediatric Precision Medicine, Center for Biomedical Informatics, Shanghai Children's Hospital, Shanghai, China

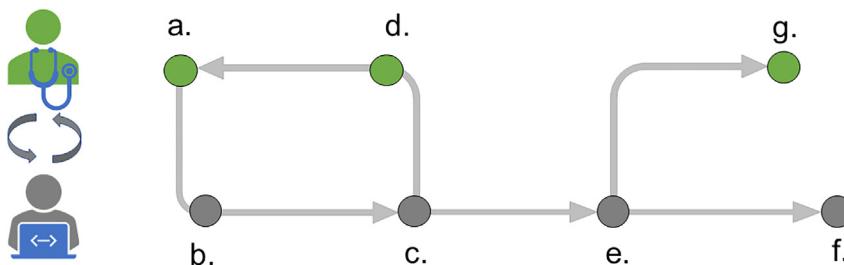
⁶These authors contributed equally

⁷Lead contact

*Correspondence:
wangyun31415926@163.com
(Y.W.),
zhangweituo@sjtu.edu.cn
(W.Z.),
huilu@sjtu.edu.cn (H.L.)

<https://doi.org/10.1016/j.isci.2023.106530>





- a. Identify the desired semantic features for AI interpretation
- b. Design the algorithm to retrieve the desired features from the images
- c. Visualize the retrieved features
- d. Confirm that the retrieved features agree with the desired semantics
- e. Construct the AI model (TiNet) to report the prediction and interpretable features
- f. Evaluate the predictive performance of AI
- g. Evaluate the understandability of AI

Figure 1. Loop of the interpretable AI model design with the collaboration of medical experts and computer scientists

The green and gray dots in the figure represent the step conducted by clinicians and computer scientists, respectively.

Some studies have attempted to combine deep learning with the Thyroid Imaging Reporting and Data System (TIRADS)²² features. For example, Deng et al.²³ designed a network to obtain a qualitative TIRADS result to predict malignancy. Zhang et al.²⁴ used human expert-extracted TIRADS features as an additional predictor of an AI model. Manh et al.²⁵ trained models to predict TIRADS features, in addition to malignancy prediction. However, these studies have only explained AI predictions that were attributable to characteristics already known by clinicians; that is, the novel features discovered by the AI model remained in the black box, which were likely to be the models wherein AI outperformed human experts. An ideal interpretable AI model should attempt to present the extracted features beyond the boundaries of existing medical knowledge in a human understandable manner, which was the purpose of our study.

In this study, we aimed to develop and validate an interpretable AI model (TiNet) for predicting thyroid cancer. TiNet could extract the affecting factors for thyroid cancer diagnosis and quantitatively analyzed the contribution of them to the overall prediction. These factors were corresponding basically to the TIRADS which was used in clinical diagnosis, such as the texture, margin, echogenicity, shape and the location of thyroid nodules. In addition, we designed a paradigm called the reverse cognitive test to evaluate whether AI results could be understood by clinicians.

RESULTS

Loop of the interpretable AI model design

We collaborated with both medicine experts and computer scientists in the loop of model design (Figure 1). The close collaboration ensured the AI interpretation met the clinical demands (see STAR Methods section).

Data collection

We collected thyroid ultrasound image sets for 10 years (from January 2012 to January 2022) to do a retrospective study from four top-class hospitals in China, where each image was marked by a doctor with more than 5 years of ultrasound experience. We included 10,386 patients who met the following inclusion criteria: (1) Patients with thyroid cancer confirmed by pathology, (2) patients were benign confirmed by pathology, or patients who have not undergone surgery and were diagnosed as benign, but with a diagnosis report of a radiologists with more than 5 years of experience, and one year follow-up record, (3) there are no other major diseases that affect the survival rate. Finally, we confirmed 10,021 images of 8,079 patients as research subjects (Table 1) based on the following exclusion criteria: (1) No preoperative ultrasound report or partial report loss, and poor quality of ultrasound images, (2) pathological diagnosis results are controversial, (3) history of thyroidectomy or coexistence of other head and neck carcinoma. We trained and validated the model using 9,019 images of 7,460 patients from one hospital, and external tested the model on

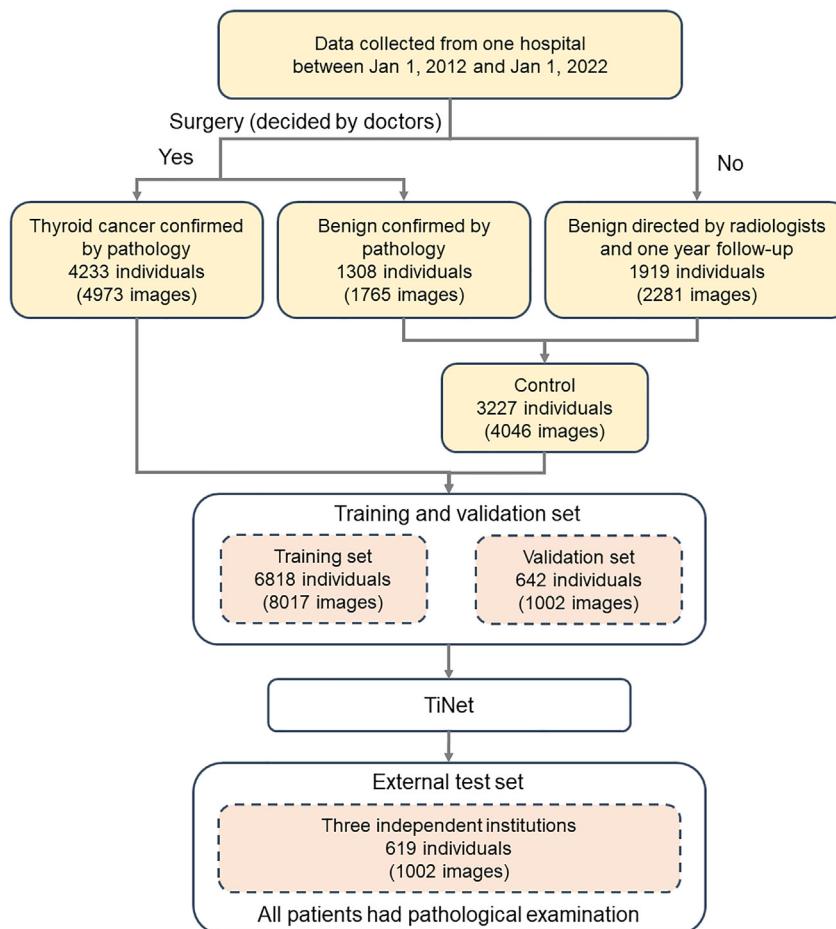
Table 1. Summary of the study population and images

	Total	Training set	Validation set	External test set
Patients	8079	6818	642	619
Benign	3470	2976	251	243
Malignant	4609	3842	391	376
Pictures	10021	8017	1002	1002
Benign	4496	3596	450	450
Malignant	5525	4421	552	552

1002 images of 619 patients from the other three independent institutions (Figure 2). Our study design and subject selection ensured a robust and representative dataset for model development and validation.

Development of the TiNet model

TiNet was designed with five channels to extract features for texture, margin, echogenicity, shape, and location of thyroid nodules and nodule segmentation was a prerequisite for analysis (Figure 3). However, unlike natural images, thyroid ultrasound images had more blurred and complex backgrounds, making it difficult for nodule recognition tasks. To address this, we employed a foreground optimization segmentation network which took the feature pyramid network (FPN)²⁶ as the backbone to enhance the contextual relevance of the foreground by calculating the global similarity matrix. In texture and margin channels, we

**Figure 2. Flow diagram summarizing the inclusion of patients and images**

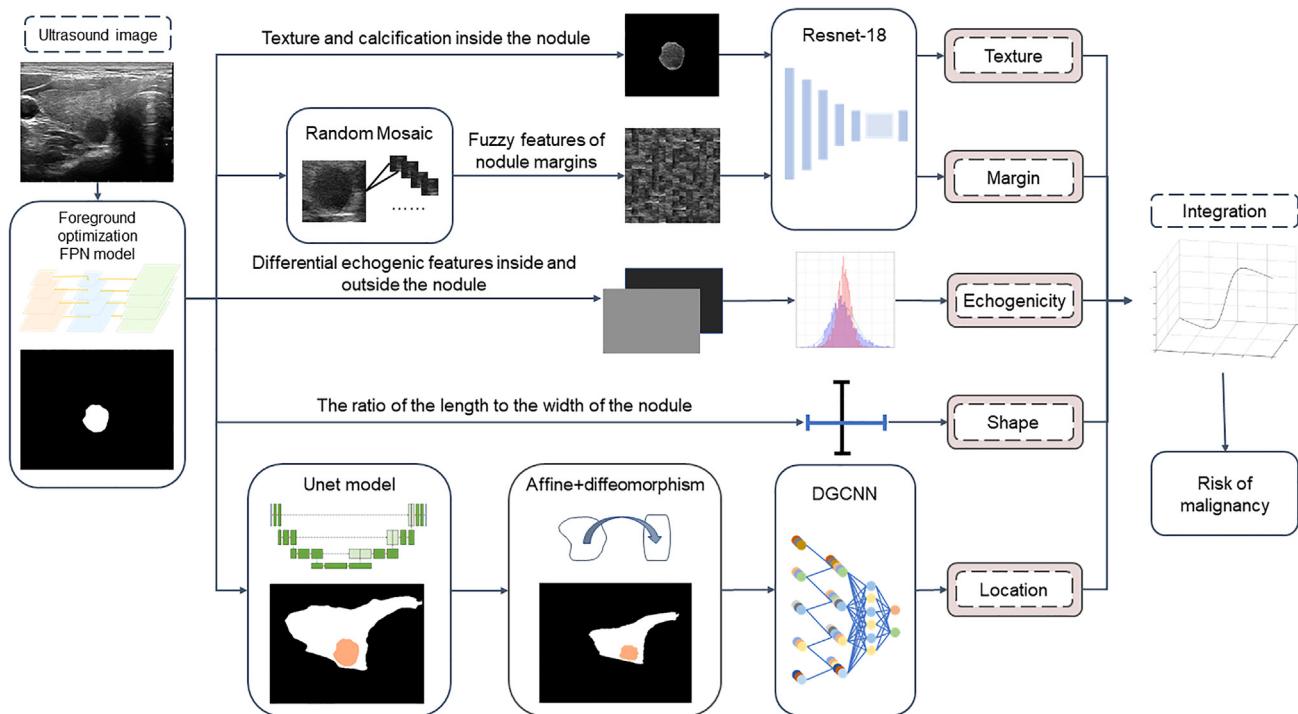


Figure 3. Development of the TiNet model

Five channels were used to extract features including texture, margin, echogenicity, shape, and location of thyroid nodules. Each channel utilized distinct algorithms to extract the corresponding feature and provided a feature-specific score for a given instance. The final AI prediction for malignancy risk was derived from these five scores.

used Resnet-18²⁷ as the framework for feature extraction. Echogenicity and shape channels were calculated based on echo difference and aspect ratio respectively. In the location channel, we segmented the thyroid using the U-Net²⁸ and creatively used a differential homeomorphism-based registration²⁹ method to perform homogeneous positioning of nodules in ultrasound images of various patients, and then we designed a deep graph convolutional neural network (DGCNN)^{30,31} to location information of nodules for thyroid cancer prediction. Five features were integrated to achieve the final prediction and analyzed the contribution of each factor to the overall prediction based on analysis of variance (ANOVA). More details of the model were provided in the [STAR Methods](#) section.

Prediction performance of the models

We evaluated and compared the predictive performance of TIRADS, TiNet, and its five individual channels ([Table 2](#)). TiNet demonstrated significantly higher performance (AUC 0.88) compared with TIRADS (0.80) in the external test set ($p < 0.001$, by De-Dong's test; [Figure 4](#)). We also observed that TiNet obtained a higher performance on each evaluation index than TIRADS ([Tables 2](#) and [3](#)). Particularly, the specificity of TIRADS was relatively low. Furthermore, we have confirmed the contribution of five features to the prediction were the textured feature (65.23%), edge margin feature (12.29%), echogenicity feature (8.43%), shape feature (7.67%), and growth position information (6.38%). [Figure 5](#) presented the performance of the thyroid and nodule segmentation modules.

Model interpretation of TiNet

In [Figure 6](#), we provide an illustrative example of TiNet comparable to the TIRADS report. In the analysis of five features of TiNet output, the texture feature corresponded to composition and echogenic foci in TIRADS, whereas the margin, echogenicity, and shape features had a similar semantic with their counterpart in TIRADS. The location information of nodules was an AI-unique feature that did not have a counterpart in TIRADS. Therefore, clinical experts could further verify the impact of nodule location information on prediction in future research. The results showed that each feature risk degree for TiNet prediction was basically consistent with the score of TIRADS evaluation criterion.³²

Table 2. Prediction performance of the models and channels

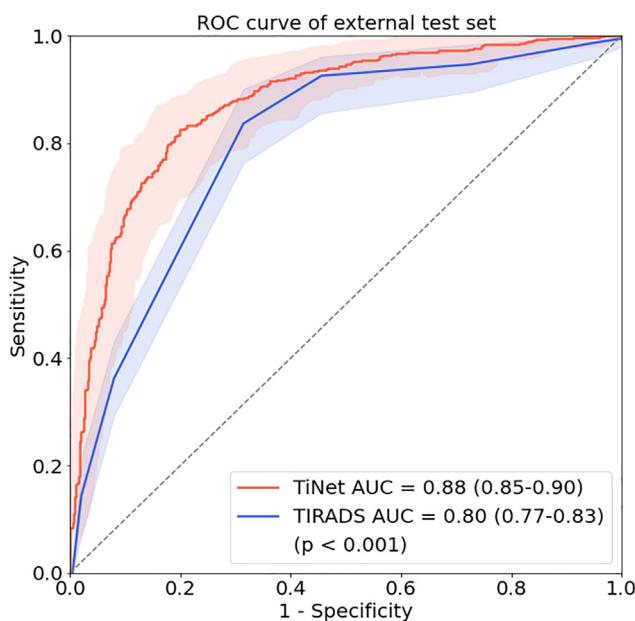
Model & Channel	AUC (95% CI)			
	Training set (N = 8017)	Validation set (N = 1002)	External test set (N = 1002)	Contribution Degree
TIRADS	0.83 (0.82–0.84)	0.82 (0.79–0.85)	0.80 (0.77–0.83)	–
TiNet	0.89 (0.88–0.89)	0.87 (0.84–0.89)	0.88 (0.85–0.90)	–
Texture	0.85 (0.84–0.86)	0.84 (0.81–0.86)	0.83 (0.80–0.86)	65.2%
Margin	0.81 (0.80–0.82)	0.73 (0.70–0.76)	0.71 (0.68–0.74)	12.3%
Echogenicity	0.63 (0.62–0.64)	0.62 (0.58–0.65)	0.66 (0.62–0.69)	8.4%
Shape	0.63 (0.62–0.65)	0.64 (0.60–0.68)	0.66 (0.62–0.69)	7.7%
Location	0.65 (0.64–0.67)	0.63 (0.60–0.66)	0.65 (0.62–0.68)	6.4%
TiNet (patient level)	0.91 (0.90–0.91)	0.91 (0.88–0.93)	0.90 (0.87–0.92)	–

In addition, we utilized an attention heatmap to further interpret the focus areas of network prediction. [Figure 7](#) presented the heatmap of malignant nodules in the texture channel, showing that the colors in the cystic and calcified areas were darker. This indicated that the texture channel predicted thyroid cancer through these two features, which was consistent with the diagnosis basis of TIRADS.

We used t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the unsupervised clustering of ultrasound images ([Figure 8](#)).³³ The results showed different types of nodules can be clustered based on image features. Particularly, benign nodules clearly cluster, most of which are cystic nodules, and distinctly aggregate malignant nodules containing prominently calcified, high-grade nodules also cluster. The cluster contains micro-calcified nodules with more malignant nodules, and the nodules in the middle position are hypoechoic nodules, whose malignancy is difficult to distinguish. This cluster contains more hyperechoic nodules with more benign nodules.

Reverse cognitive test

To quantitatively evaluate whether the TiNet report was understood by clinicians, we designed and conducted a reverse cognitive test ([Figure 9](#) and [Table 4](#)). We recruited 10 senior clinicians who had more

**Figure 4. ROCs of TiNet and TIRADS in the external test set**

Specifically, TiNet exhibits better predictive performance in terms of AUC than TIRADS. $p < 0.001$, which was calculated by De-Long's test.

Table 3. Prediction performance of TiNet and TIRADS

	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
TIRADS			
Training set	0.76 (0.75–0.77)	0.82 (0.80–0.83)	0.65 (0.62–0.68)
Validation set	0.76 (0.74–0.78)	0.80 (0.78–0.82)	0.64 (0.61–0.68)
External test set	0.74 (0.72–0.76)	0.78 (0.76–0.79)	0.62 (0.59–0.65)
TiNet			
Training set	0.81 (0.80–0.82)	0.83 (0.82–0.85)	0.77 (0.75–0.78)
Validation set	0.79 (0.76–0.81)	0.81 (0.78–0.84)	0.76 (0.72–0.80)
External test set	0.81 (0.78–0.83)	0.83 (0.80–0.87)	0.77 (0.73–0.81)

than 5 years of experience on thyroid ultrasound examination and each clinician was asked to perform three groups of tasks. Group 1 was asked to match 100 pairs of randomly selected ultrasound images with the corresponding TiNet AI reports. Groups 2 and 3 were the control groups, wherein the TIRADS reports or single AI risk scores were provided, rather than AI reports. Note that images for different tasks and clinicians were different. Experiment proved that the average matching accuracy of the TiNet group was 87.1%, higher than that of the TIRADS group (81.6%) and significantly higher than that of the AI risk score group (68.2%, $p<0.001$). This result of the reserve cognitive test demonstrated that TiNet, as an AI version of the TIRADS report, could be well-understood by clinicians and ready to be combined with clinical expertise to improve future clinical decisions.

DISCUSSION

In this study, we developed and validated TiNet as an AI version of the TIRADS report system for thyroid cancer diagnosis. Unlike previous AI models, TiNet provides five human understandable features with conventional AI risk scores. The five features include the texture, margin, echogenicity, shape, and location information of the thyroid nodules. In our training, validation, and external test datasets, the TiNet model achieved a significantly higher prediction performance than TIRADS which was similar to current performance in published literature.^{18–21,34} We also showed, in the reverse cognitive test, that TiNet was well understood by human experts and significantly more interpretable than single AI in terms of the risk score.

AI interpretability is widely concerned issue and full of controversies. Accordingly, we have two important innovations. First, we suggested that the interpretability of AI in medicine should be initiated from the clinical demand, rather than the engineering perspective. To ensure that we could provide a clinically desired AI interpretation, we conducted a pipeline of the AI model design that engaged clinicians and computer scientists. We believe that this concept and pipeline could be applied in many other areas of medicine.

Second, quantitatively evaluating the interpretability of the AI model was very difficult. To address this problem, we considered the “AI interpretability” from a cognitive science perspective and designed a paradigm, a reverse cognitive test. A conventional cognitive test was often used to measure the ability

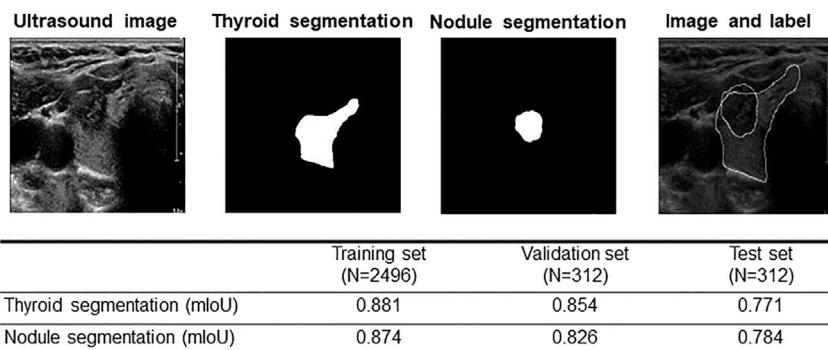
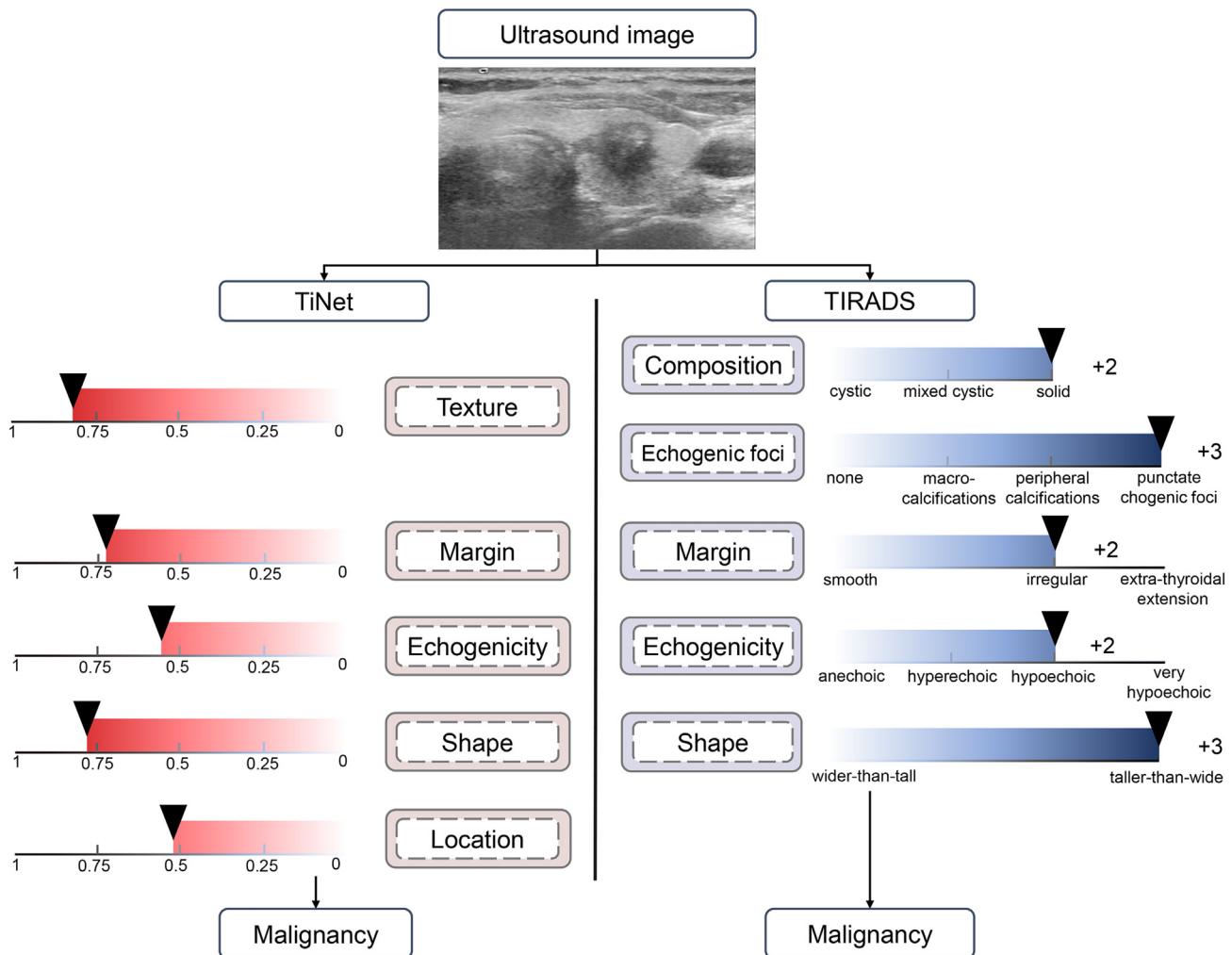


Figure 5. Performance of thyroid and nodule segmentation

**Figure 6. An illustrative example of TiNet compared with TIRADS**

The texture feature in TiNet corresponded to the composition and echogenicity foci factors in TIRADS, and the margin, echogenicity, and shape features could all correspond to TIRADS, and location information of nodules was a unique factor of TiNet. The results have showed that the prediction results of TiNet and TIRADS were mainly the same.

or function of the receiver. Meanwhile, the reverse cognitive test was conducted in a similar manner, but the aim was to measure the ability of the information giver (that was why it was called "reverse"). Therefore, it would be a generalizable approach to evaluate the interpretability of the AI model.

Our model could also partially answer the questions of clinicians: why AI could outperform human experts in diagnostic tasks. We could observe from the model structure that the higher prediction accuracy of the TiNet model than that of TIRADS was because of two aspects: The features with the counterparts in TIRADS were more precisely digitized, and the AI-unique features "location" that were conventionally not considered in the TIRADS system. We also provided AI interpretability in a conventional meaning. In this study, the effects of five factors on cancer (ANOVA) and cluster analysis (t-SNE) were obtained in the entire population. According to the actual requirements of clinical applications, we provided individual benign or malignant prediction for each patient and performed further quantitative analysis of the risk factors. Therefore, TiNet could better serve the clinic disease diagnose as an individualization tool.

Before this study, several studies have combined ultrasound AI for thyroid cancer prediction with TIRADS. Chen et al.³⁵ developed an AI model based on ACR TIRADS, which evaluated TIRADS features to predict malignancy. Therefore, the prediction power of the model is limited by the traditional TIRADS features. Yu

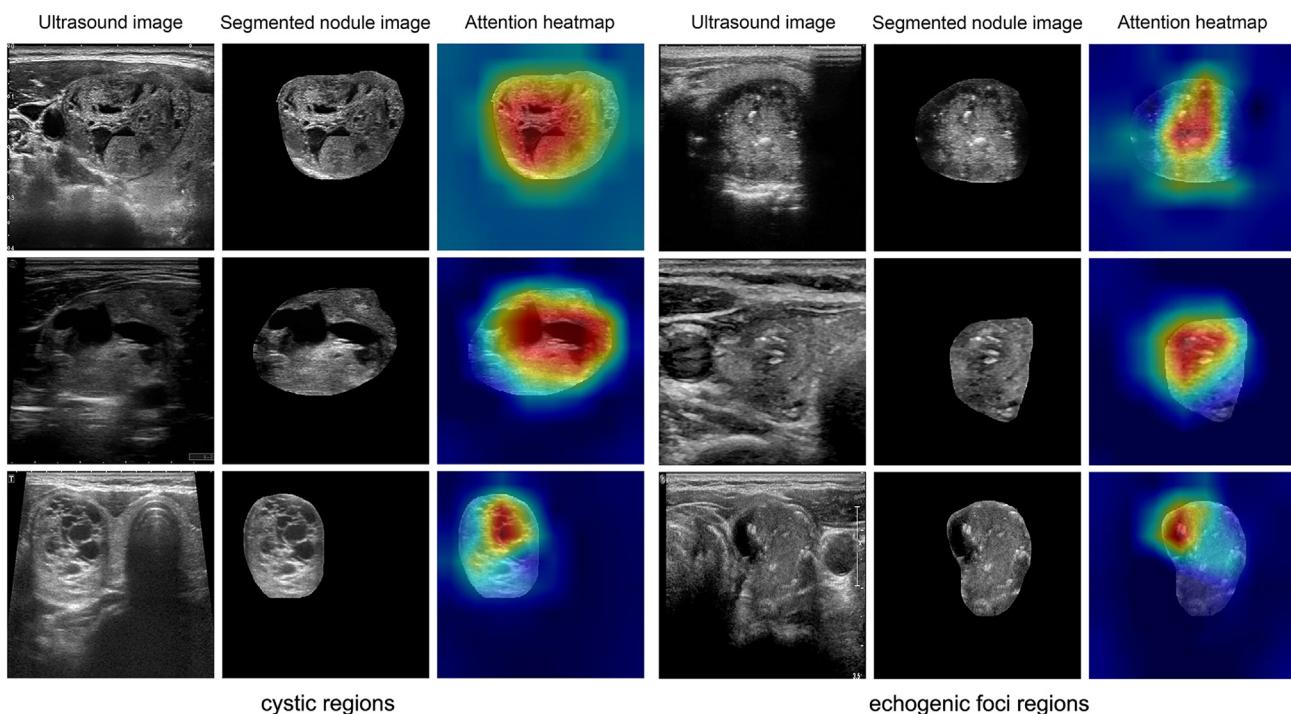


Figure 7. Visualized interpretation of the TiNet texture channel

The columns present the ultrasound images, corresponding segmented nodule regions, and AI attention heatmaps of the TiNet texture channel. The hot spot area in the heatmap shows the AI's focused region. All nodules are malignant in figure. In typical examples displayed in cystic regions panel, the AI focused on the cystic regions of the nodule, whereas the AI focused on the echogenic foci regions in the case of cystic nodules in echogenic foci regions panel. The visualized interpretation shows that the texture channel of TiNet extracts features that correspond to the composition (cystic/solid) and echogenic foci in TIRADS.

et al.³⁶ developed an AI model leveraging features discovered by deep learning and tried to interpret the AI model with traditional TIRADS features. However, their model interpretation was incomplete because the extra features discovered by deep learning could not be explained by TIRADS. Unlike previous studies, we decomposed the AI prediction into AI generated TIRADS-like features instead of traditional TIRADS. As such, the more features discovered by deep learning were also explained in this system and we proved this AI-version of TIRADS report was still understandable by clinicians. This AI-version TIRADS report was an innovative approach to bridge the clinician and AI model.

Limitations of the study

Our research still had some limitations. First, this work was retrospective and the data collection was limited. We need to verify our conclusions in prospective clinical data and large-scale validation population. Second, the data that we collected were mainly single papillary thyroid nodules, so that the diagnostic efficacy of multiple nodules or other histological subtype of thyroid cancer may be insufficient. We consider expanding more diversified data to upgrade our version in the future. Third, in this study, TiNet only extracted five features related to clinical diagnosis of thyroid cancer. We have reason to believe that more valuable information about thyroid cancer prediction can be tried to find based on the architecture-design of TiNet.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- [KEY RESOURCES TABLE](#)
- [RESOURCE AVAILABILITY](#)
 - Lead contact
 - Materials availability
 - Data and code availability
- [METHOD DETAILS](#)

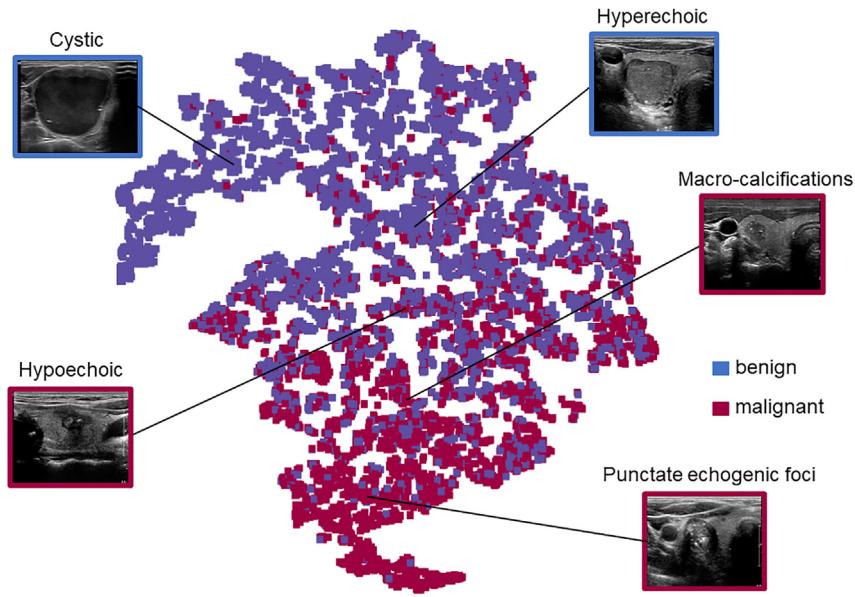


Figure 8. Clustering of the thyroid nodule by t-SNE

Benign and malignant nodules are marked with blue and red, respectively, and these nodules were largely separated by this clustering. Typical types of nodules were also clustered; for example, the cystic nodules were clustered in a branch the most distant from malignancy, and the nodules with punctate echogenic foci are aggregated in a malignant region.

- Loop of the AI model design
- Deep learning framework
- Visual analysis of the prediction model
- Reverse cognitive test
- Statistical analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

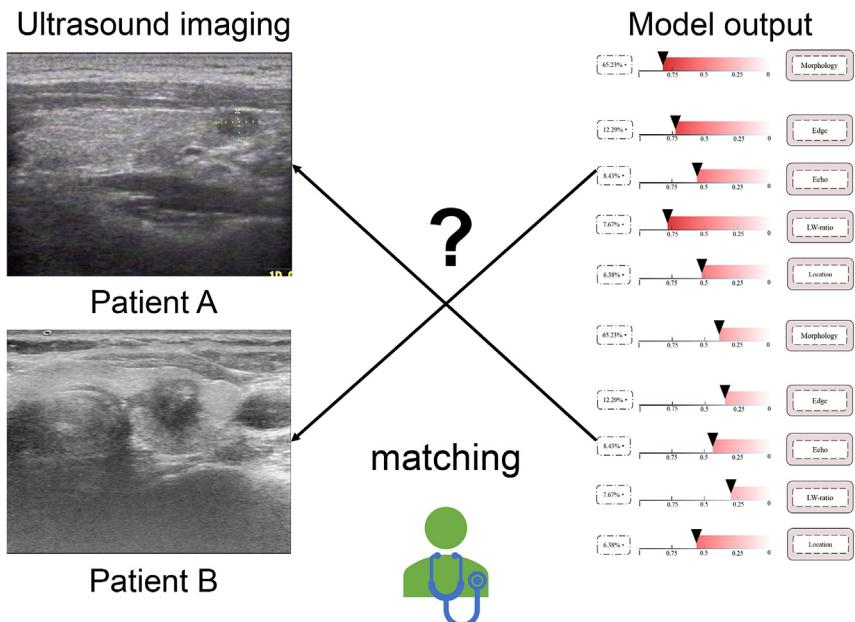


Figure 9. Process of the reverse cognitive test

Ten clinicians were asked to repeatedly match two randomly picked ultrasound images with their corresponding TiNet AI reports, TIRADS, or AI risk scores.

Table 4. Results of the reverse cognitive test

Clinicians	Accuracy of clinicians' judgment (%)		
	TiNet report	TIRADS report	Traditional AI risk score
1	89	84	70
2	79	80	65
3	84	77	67
4	93	85	64
5	89	87	71
6	81	82	67
7	92	81	73
8	82	79	68
9	94	85	74
10	88	76	63
Mean	87.1	81.6	68.2

TiNet report was more easily to understand by clinicians, leading to the highest improvement in clinical auxiliary diagnosis among the three reports.

Note: Experimental method: Clinicians matched the corresponding unknown ultrasound images based on information from three different reports.

Purpose of the test: To evaluate the clinicians' understanding of the reported information, and the guiding effect of the reported information on clinical judgment.

ACKNOWLEDGMENTS

This work was partially supported by the Neil Shen's SJTU Medical Research Fund, SJTU Trans-med Awards Research STAR20210106, NSFC81903417, Science and Technology Development Fund of Xuzhou Medical University (XYFY2020045), and the Innovative Research Team of High-Level Local Universities in Shanghai (SHSMU-ZDCX20212200).

AUTHOR CONTRIBUTIONS

S.Y., W.Z., and H.L. developed the concept for the manuscript. S.Y. and P.S. contributed to drafting of the manuscript. S.Y., P.S., and T.D. designed the model and analysis the data. W.Z., H.L., and F.D. contributed to critical revision of the manuscript. Y.W. contributed to provide medical data and clinical advice. S.Y. and P.S. contributed equally to this work.

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: September 19, 2022

Revised: January 8, 2023

Accepted: March 24, 2023

Published: March 31, 2023

REFERENCES

- Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **71**, 209–249. <https://doi.org/10.3322/caac.21660>.
- Buda, M., Wildman-Tobriner, B., Castor, K., Hoang, J.K., and Mazurowski, M.A. (2020). Deep learning-based segmentation of nodules in thyroid ultrasound: improving performance by utilizing markers present in the images. *Ultrasound Med. Biol.* **46**, 415–421.
- Kumar, V., Webb, J., Gregory, A., Meixner, D.D., Knudsen, J.M., Callstrom, M., Fatemi, M., and Alizad, A. (2020). Automated segmentation of thyroid nodule, gland, and cystic components from ultrasound images using deep learning. *IEEE Access* **8**, 63482–63496. <https://doi.org/10.1109/ACCESS.2020.2982390>.

4. Ma, J., Wu, F., Jiang, T., Zhao, Q., and Kong, D. (2017). Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. *Int. J. Comput. Assist. Radiol. Surg.* 12, 1895–1910.
5. Song, J., Chai, Y.J., Masuoka, H., Park, S.-W., Kim, S.-j., Choi, J.Y., Kong, H.-J., Lee, K.E., Lee, J., Kwak, N., et al. (2019). Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. *Medicine* 98, e15133.
6. Yao, S., Yan, J., Wu, M., Yang, X., Zhang, W., Lu, H., and Qian, B. (2020). Texture synthesis based thyroid nodule detection from medical ultrasound images: interpreting and suppressing the adversarial effect of in-place manual annotation. *Front. Bioeng. Biotechnol.* 8, 599. <https://doi.org/10.3389/fbioe.2020.00599>.
7. Zheng, B., Liu, J., Gu, J., Du, J., Wang, L., Gu, S., Cheng, J., Yang, J., and Lu, H. (2016). Classification of benign and malignant thyroid nodules using a combined clinical information and gene expression signatures. *PLoS One* 11, e0164570.
8. Zheng, B., Liu, J., Gu, J., Lu, Y., Zhang, W., Li, M., and Lu, H. (2015). A three-gene panel that distinguishes benign from malignant thyroid nodules. *Int. J. Cancer* 136, 1646–1654. <https://doi.org/10.1002/ijc.29172>.
9. Zhang, Y., Tiño, P., Leonardis, A., and Tang, K. (2021). A survey on neural network interpretability. *IEEE Trans. Emerg. Top. Comput. Intell.* 5, 726–742.
10. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29.
11. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215.
12. Fan, F.-L., Xiong, J., Li, M., and Wang, G. (2021). On interpretability of artificial neural networks: a survey. *IEEE Trans. Radiat. Plasma Med. Sci.* 5, 741–760.
13. Lipton, Z.C. (2018). The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 31–57.
14. Higgins, D., and Madai, V.I. (2020). From bit to bedside: a practical framework for artificial intelligence product development in healthcare. *Advanced intelligent systems* 2, 2000052.
15. Nguyen A., Yosinski J., Clune J. Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images. in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. p. 427–436.
16. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I.J., Fergus R. Intriguing Properties of Neural Networks. arXiv. 2014; (Preprint at) <http://arxiv.org/abs/1312.6199>.
17. Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V.I.; Precise4Q Consortium (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* 20, 310–319.
18. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115.
19. Liao, W., Zou, B., Zhao, R., Chen, Y., He, Z., and Zhou, M. (2020). Clinical interpretable deep learning model for glaucoma diagnosis. *IEEE J. Biomed. Health Inform.* 24, 1405–1412.
20. Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.-M., von Tengg-Kobligk, H., Summers, R.M., and Wiest, R. (2020). On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol. Artif. Intell.* 2, e190043.
21. Sha, Y., and Wang, M.D. (2017). Interpretable Predictions of Clinical Outcomes with an Attention-Based Recurrent Neural Network. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 233–240.
22. Tessler, F.N., Middleton, W.D., and Grant, E.G. (2018). Thyroid imaging reporting and data system (TI-RADS): a user's guide. *Radiology* 287, 29–36. <https://doi.org/10.1148/radiol.2017171240>.
23. Deng, P., Han, X., Wei, X., and Chang, L. (2022). Automatic classification of thyroid nodules in ultrasound images using a multi-task attention network guided by clinical knowledge. *Comput. Biol. Med.* 150, 106172.
24. Zhang, S., Du, H., Jin, Z., Zhu, Y., Zhang, Y., Xie, F., Zhang, M., Tian, X., Zhang, J., and Luo, Y. (2020). A novel interpretable computer-aided diagnosis system of thyroid nodules on ultrasound based on clinical experience. *IEEE Access* 8, 53223–53231.
25. Manh, V.T., Zhou, J., Jia, X., Lin, Z., Xu, W., Mei, Z., Dong, Y., Yang, X., Huang, R., and Ni, D. (2022). Multi-attribute attention network for interpretable diagnosis of thyroid nodules in ultrasound images. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* 69, 2611–2620.
26. Zheng Z., Zhong Y., Wang J., Ma A. Foreground-Aware Relation Network for Geospatial Object Segmentation In High Spatial Resolution Remote Sensing Imagery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. p. 4096–4105.
27. Guo M., Du Y. Classification of Thyroid Ultrasound Standard Plane Images Using ResNet-18 Networks. *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*. 2019. p. 324–328.
28. Ronneberger O., Fischer P., Brox T. U-net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-assisted Intervention*. 2015. p. 234–241.
29. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., and Gee, J.C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 54, 2033–2044.
30. Welling M., Kipf T.N. (2016). Semi-supervised Classification with Graph Convolutional Networks. *J. International Conference on Learning Representations*.
31. Zhang M., Cui Z., Neumann M., Chen Y. (2018). An End-to-End Deep Learning Architecture for Graph Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*.
32. Tessler, F.N., Middleton, W.D., Grant, E.G., Hoang, J.K., Berland, L.L., Teeffey, S.A., Cronan, J.J., Beland, M.D., Desser, T.S., Frates, M.C., et al. (2017). ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. *J. Am. Coll. Radiol.* 14, 587–595.
33. Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9: 2579–2605.
34. Li, X., Zhang, S., Zhang, Q., Wei, X., Pan, Y., Zhao, J., Xin, X., Qin, C., Wang, X., Li, J., et al. (2019). Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol.* 20, 193–201. [https://doi.org/10.1016/S1470-2045\(18\)30762-9](https://doi.org/10.1016/S1470-2045(18)30762-9).
35. Chen, Y., Gao, Z., He, Y., Mai, W., Li, J., Zhou, M., Li, S., Yi, W., Wu, S., Bai, T., et al. (2022). An artificial intelligence model based on ACR TI-RADS characteristics for US diagnosis of thyroid nodules. *Radiology* 303, 613–619. <https://doi.org/10.1148/radiol.211455>.
36. Yu, H., Li, J., Sun, J., Zheng, J., Wang, S., Wang, G., Ding, Y., Zhao, J., and Zhang, J. (2022). Intelligent diagnosis algorithm for thyroid nodules based on deep learning and statistical features. *Biomed. Signal Process Control* 78, 103924. <https://doi.org/10.1016/j.bspc.2022.103924>.
37. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., and Aerts, H.J.W.L. (2018). Artificial intelligence in radiology. *Nat. Rev. Cancer* 18, 500–510.
38. Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *2017/07 (IEEE Computer Society)*, pp. 1800–1807.
39. Lin G., Milan A., Shen C., Reid I. Refinenet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017. p. 1925–1934.
40. Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J.,

- Roth, H., Myronenko, A., Xu, D., and Xu, Z. (2020). Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Trans. Med. Imaging* 39, 2531–2540.
41. Jang W.-D., Lee C., Kim C.-S. Primary Object Segmentation in Videos via Alternate Convex Optimization of Foreground and Background Distributions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 696–704.
42. Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 2117–2125.
43. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 770–778.
44. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252.
45. Yu, J., Deng, Y., Liu, T., Zhou, J., Jia, X., Xiao, T., Zhou, S., Li, J., Guo, Y., Wang, Y., et al. (2020). Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics. *Nat. Commun.* 11, 4807–4810.
46. Beg, M.F., Miller, M.I., Trouvé, A., and Younes, L. (2005). Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int. J. Comput. Vis.* 61, 139–157.
47. Gargya, R., and Leng, T. (2017). Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* 124, 962–969.
48. Liu, T., Guo, Q., Lian, C., Ren, X., Liang, S., Yu, J., Niu, L., Sun, W., and Shen, D. (2019). Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. *Med. Image Anal.* 58, 101555.
49. Xu, P., Ding, J., Zhang, H., and Huang, H. (2019). Discernible image mosaic with edge-aware adaptive tiles. *Comput. Vis. Media* 5, 45–58.
50. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans. Med. Imaging* 35, 1299–1312.
51. Alexander, E.K., Marqusee, E., Orcutt, J., Benson, C.B., Frates, M.C., Doubilet, P.M., Cibas, E.S., and Atri, A. (2004). Thyroid nodule shape and prediction of malignancy. *Thyroid* 14, 953–958.
52. Xie, C., Cox, P., Taylor, N., and LaPorte, S. (2016). Ultrasonography of thyroid nodules: a pictorial review. *Insights Imaging* 7, 77–86.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
PyCharm	JetBrains	https://www.jetbrains.com/pycharm
Python version 3.8	Python Software Foundation	https://www.python.org
TiNet	This paper	https://github.com/Snowinbio/TiNet-iScience2023
Other		
Thyroid nodule images dataset as training and validation set	The Tumor Research Institute of the Southeast University (Xu zhou)	N/A
Thyroid nodule images dataset as external test set	Shanghai Jiao Tong University School of Medicine	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Hui Lu (huilu@sjtu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The image dataset used in this study are subject to privacy restrictions, but may be anonymized and made available upon reasonable request to the [lead contact](#) (Prof. Hui Lu via huilu@sjtu.edu.cn).
- The code for model development is available in the github repository (<https://github.com/Snowinbio/TiNet-iScience2023>).
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

METHOD DETAILS

Loop of the AI model design

We designed a replicable workflow to engage clinicians into the loop of model design and to ensure that the model design is consistent with the desired semantic concept (Figure 1). This model design pipeline is driven by the fuzzy trace theory from psychology and the TIRADS system from medical radiology. Medical experts identified the desired semantic features for AI interpretation, whereas computer scientists designed the algorithm to retrieve and visualize the corresponding semantic features from the images for the clinicians. If the clinicians confirmed that the retrieved features agree with the desired semantics, the computer scientists would construct the AI model (TiNet) to report the prediction and interpretable features. After integrating the model, the computer scientists would evaluate the predictive performance, and the clinicians would evaluate the understandability of AI.

Deep learning framework

To clarify the contributions of different types of ultrasonographical features in the prediction, we elaborately designed a deep learning framework (Figure 3). In this experiment, first five features in ultrasound images were separated by channels. The channels were based on the segmentation of the nodules using the foreground optimization FPN model,²⁶ and the framework of the texture and margin channels were ResNet-18.²⁷ Meanwhile, the echogenicity channel was calculated based on the echo difference, and

the shape channel used an aspect ratio. The location channel was constructed using the thyroid segmentation model (U-Net), registration with affine, and DGCNN.^{30,31} Subsequently, the prediction probability of each channel was quantified into the value interval as the range of each treatment factor in the prediction. Finally, a comprehensive prediction model and the contribution degree of factors were obtained by integrating each factor channel. The integration method was logistic regression. In practical applications, a single image would obtain the prediction results of five channels, their respective interval values, the overall prediction results, and the contribution of various factors in the overall results. The following were the methods used.

First, to analyze the influence of the characteristics of the nodules, we must effectively segment the nodules, which could be used as the input for the classification. However, unlike natural images or other medical images, thyroid ultrasound images had the following characteristics: large-scale change, low contrast between organs, fuzzy background, relatively small detection targets, and a small number.³⁷ More serious misjudgment can occur due to the large intra class variance of the background, resulting in great challenges to the segmentation task.³⁸⁻⁴⁰ This study used the foreground optimization method to accurately segment thyroid nodules.^{26,41} The foreground optimization FPN in this method could address the multi-scale input problem.⁴² The foreground-background relationship module could improve the discrimination of the nodule itself by correlating the surrounding semantic information related to the nodule morphological pixels. The aim is to enhance the difference between the nodule and surrounding background. To focus on the classification error samples misjudged as nodules in the background and to adjust the descending direction of the gradient for improving the recognition rate, the foreground optimization module was embedded in the model. Further, many models generate jagged edges when segmenting the thyroid. Considering the segmentation performance and to reduce jagged edges, we used the U-Net network as the thyroid segmentation module after many model comparison experiments.

Among them, the foreground optimization network was expressed as follows. We start the network by providing definitions:

$$f_i = \begin{cases} M(C_i) + N(f_{i+1}), & i = 2, 3, 4, 5 \\ M(C_i), & i = 6 \end{cases} \quad (\text{Equation 1})$$

where f_i is the i -th layer in pyramidal feature map, $C_i (i = 2, 3, 4, 5)$ denotes the set of feature maps from FPN with same number of channels d , M denotes the lateral connection by a learnable convolutional layer, and N denotes a nearest neighbor upsampling. We attached a branch on C_5 to generate scene feature C_6 via global context aggregation.

To transform the feature map, we have scale-aware projection function φ_{θ_i} in [Equation 2](#).

$$\varphi_{\theta_i} : R^{d \times H \times W} \rightarrow R^{d_u \times H \times W} \quad (\text{Equation 2})$$

where θ_i is the learnable parameters of $\varphi_{\theta_i}(\cdot)$. Then, we obtained the transformed feature map f_i^u :

$$f_i^u = \varphi_{\theta_i}(f_i) \quad (\text{Equation 3})$$

To obtain the relation of maps, we applied $I(\cdot)$ on C_6 to compute the scene embedding vector v .

$$v = I(C_6) \quad (\text{Equation 4})$$

Here, I is a projection function implemented by a learnable convolutional layer, and vector v is shared for each pyramid. Thus, we can obtain the relation map r_i by similar estimation implemented by pointwise inner product $\text{sim}(\cdot)$.

$$r_i = \text{sim}(v, f_i^u) \quad (\text{Equation 5})$$

For the encoder, the input pyramidal feature map f_i can be recoded to obtain the re-encoding feature map f_i^ω . The encoder has learnable parameter ω_i :

$$f_i^\omega = \text{re_encoding}(f_i, \omega_i) \quad (\text{Equation 6})$$

Finally, we obtained the relation enhanced foreground feature map g_i in [Equation 7](#).

$$g_i = \frac{e^{r_i} f_i^\omega}{1 + e^{r_i}} \quad (\text{Equation 7})$$

Based on the above principles, we implemented a foreground-optimized segmentation model to extract the nodule information.

We collected 3120 pieces of data with nodule and thyroid segmentation labels and obtained the nodule and thyroid segmentation modules through training and verification. Subsequently, we used the segmentation modules to segment the remaining data. We annotated all grand truth segmentation labels by doctors with more than 10 years of thyroid ultrasound examination experience. Figure 5 shows the segmentation accuracy results of the segmentation model. For the nodule segmentation module, the baseline of optimization-based foreground network was the ResNet-50 model,⁴³ which was pre-trained on ImageNet.⁴⁴ In this module, sizes of the convolution kernel comprised 4×4 , 8×8 , 16×16 , and 32×32 . The learning rate was 0.007, and the maximum training iterations setting was 60000. While training the U-Net network, for the thyroid segmentation module, we set batch-size to 8, used the Adam optimizer, set the learning rate to 0.001, and set the maximum epoch to 300.

- 1) Analysis of internal textured characteristics of thyroid nodules: Finally, we modeled data obtained from the segmentation and used Resnet-18⁴⁵ to predict risk of malignancy based on texture, and obtained the texture channel.
- 2) Analyzing the growth location information of thyroid nodules: To analyze the influence of the growth location of thyroid nodules for malignant prediction, we needed to quantify only the overall relationship between both nodule and thyroid margin information. We first used the U-Net method²⁸ and foreground optimization method²⁶ to segment thyroids and nodules, respectively, and obtained these boundary information (white lines without background texture). The location channel then predicted malignancy risk based on this boundary information.

While using ultrasound to detect nodules, extrusion and multi-scale and multi-resolution problems would become serious. We needed to statistically compare the growth location relationship of nodules relative to the thyroid in the same coordinate system. Hence, we chose a unified thyroid template to register the segmented thyroid and nodule ultrasound images into the template. According to the deformation field obtained by thyroid registration, we also mapped each nodule into the thyroid template to automatically obtain the location information in the current registered image. Specifically, we used the software package ANTs based on the symmetric diffeomorphism method to do the registration.²⁹ ANTs assumes that the diffeomorphism, φ , is defined on the image domain, Ω , and preserves an affine transform at the boundary such that $\varphi(\partial\Omega) = A(\text{Id})$ where $A(\text{Id})$ is an affine mapping applied to the identity transformation. The map φ , over time, parameterizes a family of diffeomorphisms, $\varphi(x, t) : \Omega \times t \rightarrow \Omega$, which can be generated by integrating a (potentially) time-dependent, smooth velocity field, $v : \Omega \times t \rightarrow \mathbb{R}^d$, through the ordinary differential equation (o.d.e.) as follows:

$$\frac{d\varphi(x, t)}{dt} = v(\varphi(x, t), t), \varphi(x, 0) = x \quad (\text{Equation 8})$$

The deformation field yielded by φ is $u(x) = \varphi(x, 1) - x$.

If we decompose φ into two components φ_1 and φ_2 , and define, $int \in [0, 0.5]$, $v(x, t) = v_1(x, t)$, $v(x, t) = v_2(x, 1-t)$ when $t \in [0.5, 1]$, the generation of the standard large deformation diffeomorphic metric matching (LDDMM)⁴⁶ equation will be as follows:

$$\begin{aligned} \{v_1^*, v_2^*\} = \operatorname{argmin}_{v_{1,2}} \left\{ & \int_0^{0.5} \|Lv_1(x, t)\|^2 + \int_0^{0.5} \|Lv_2(x, t)\|^2 dt \right. \\ & \left. + \lambda \int_{\Omega} \Pi_{\sim}(\mathcal{J} \circ \varphi_1(x, 0.5), \mathcal{J} \circ \varphi_2(x, 0.5)) d\Omega \right\} \end{aligned} \quad (\text{Equation 9})$$

The method of registration is 'syn', and the method of interpolation is linear interpolation. The optimization metric in the registration process is mutual information, while the final evaluation index is mean square error (MSE).

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (\text{Equation 10})$$

$$MSE = \frac{1}{MN} \sum_{y=1}^M \sum_{x=1}^N [J(x, y) - J'(X, Y)]^2 \quad (\text{Equation 11})$$

where J is the thyroid image before registration, J' is the thyroid image after registration, and m and N represent the width and height of the image, respectively.

After the segmentation and registration operations, the thyroid could be mapped to a fixed position. Moreover, the actual physical position information of the nodule and thyroid could be transformed into the relative position information in the current registration picture. The location channel was required to predict risk of malignancy according to positional relationship. It was a graph convolution network structure, which could build location relationship in a higher dimension and more effectively explore the implied correspondence in the information. Therefore, this paper proposed that the positional relationship between the thyroid and the nodule should be transformed into a graph structure, and the distance between the nodule and the thyroid could be characterized by edge weight in the graph. Second, we used the depth graph convolutional network^{30,31} to extract the structural features of the map (i.e., the relationship between the overall position of the nodule and the thyroid was comprehensively considered). Finally, the channel could get the location feature-specific score.

- 3) Nodule margin analysis: The feature of nodule margins is vital for diagnosing thyroid cancer. However, either TIRADS or experienced doctors cannot provide precise criteria on how to filter margin feature from an entire ultrasonography image. Computer vision methods are often designed for common images instead of radiography images. Moreover, losing important medical information is possible using these methods to extract features from ultrasonography data.^{47,48} Hence, we wanted to preserve details from original images at most using a new method called "random mosaic"⁴⁹ to retain margin information. RM extracted massive random patches from the margin area of a thyroid nodule and jointed them into an entire image. We determined the margin area by segmentation task, which has presented its good performance. The method converted margin feature to an image that can be processed by many deep learning methods. Moreover, it guaranteed the consistent size of extracted feature in the case of inconsistent size of different nodules. In our experiment, we set every patch's size as 16×16 pixels, and every mosaic image had 14×14 patches. We then used ResNet-18 to predict malignancy risk, whose inputs were mosaic images. Moreover, it was trained with the parameters of ImageNet pre-trained model.^{27,50}
- 4) Nodule shape and echo prediction: According to segmented nodule regions, we calculated the ratio of edge length to width to obtain value of prediction interval.⁵¹ This channel was the shape channel. We calculated the nodule echo difference, which was the difference between the echo average of the thyroid area outside the nodule and echo average inside the nodule. We then fit the distribution of the echo difference through the Gaussian function to obtain the prediction interval.⁵² This channel was the echogenicity channel.
- 5) Comprehensive prediction model integrating five models: We obtained the value range prediction interval of each group of patients based on five channels. For the texture, location, and margin channels, the prediction value obtained by deep learning method would be helpful for getting the value range of the group. We obtained the group value range of the aspect ratio for shape feature. We obtained the group range of echo difference for the echogenicity feature. Furthermore, we combined outputs of texture, margin, echogenicity, shape, and location channel to predict whether a nodule was benign or malignant from the perspective of these features. Using the logistic method, we took the predicted probability values of these five features and their interaction terms as predictors. The model form was as follows:

$$\text{logit}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \quad (\text{Equation 12})$$

Here, x_1 , x_2 , x_3 , x_4 , and x_5 were the predicted malignant probability of texture, echogenicity, margin, shape, and location channels, respectively. Moreover, y was the prediction result of TiNet model. In this prediction model, the training, validation and external test sets in five channels were also used. The training set was used to train the logistic regression model, and the effect was evaluated on the validation and external test sets.

We used the logistic method. Its form was the same as above, but the fitting model used whole data. Currently, we focus on the significance of coefficients β_1 , β_2 , β_3 , β_4 , and β_5 in the model.

Visual analysis of the prediction model

- 1) Interpretations for individual patient's pathology: We provided predicted outcome interpretations for each individual patient's pathology. When the ultrasound image was used in the TiNet system, the visualization results of five influencing factors would be the output including the risk degree of each of the five factors. (The system quantified the value between 0 and 1. The closer to 0, the more normal it would be. The closer to 1, the degree of malignancy would be larger.) and the contribution of each factor to the overall result. Doctors could use these quantitative results to adjust diagnosis results to be more objective and accurate.
- 2) Attention heatmaps provided clinical interpretation for texture features: The TiNet system could provide a certain degree of clinical interpretation for texture features in images through output heatmaps. The heatmap could be generated by calculating network parameters using the gradcam++ method, which was used to provide a cue of the area having the greatest relationship with prediction. For example, by analyzing the internal texture of nodules, whether the model paid more attention to the part with calcification could be verified.
- 3) Visualizing results using t-SNE: We used t-distributed stochastic neighborhood embedding (t-SNE)³³ to visualize the results of the five models and was dependent on the scale and location of input data as a dimension reduction method. T-SNE was a non-linear embedding technique and could create a lower dimensional map of data while preserving the local structure to the extent possible. To use t-SNE, we had to set the perplexity value as 30, which determined the variances of Gaussian kernels. In a high-dimensional space, we used variance to calculate the similarity.

Reverse cognitive test

We had a reverse cognitive test to verify the understandability of the TiNet model. We provided clinicians with two ultrasound images (including a benign and a malignant) and one of the prediction results of TIRADS or TiNet. Next, the clinicians should match images with benign or malignant according to the prediction results. After 100 tests, we calculated the average accuracy of clinicians.

Statistical analysis

- 1) Confidence interval estimation: Bootstrapping was used only to estimate 95% confidence intervals (CI) for the performance metrics of our classification results (i.e., AUC and accuracy). We applied n-out-of-n bootstrap with replacement at image level from our datasets. For each bootstrap sample, we calculated and reserved performance metrics for that bootstrap sample and bootstrap sampling was repeated 1000 times. We then estimated the 95% CI by using the 2.5 and 97.5 percentiles of the empirical distribution of the corresponding metrics.
- 2) Accuracy, sensitivity, specificity, and AUC calculations: Accuracy, representing the fraction of correct classifications, was calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (\text{Equation 13})$$

where TP, TN, FP, and FN meant true positive, true negative, false positive, and false negative, respectively.

Sensitivity (also called the true-positive rate and recall) is the proportion of positives that are correctly identified as follows:

$$\frac{\text{Sensitivity}}{\text{TPR}} = \frac{TP}{TP + FN}. \quad (\text{Equation 14})$$

Specificity (also called the true negative rate), which measures the proportion of correctly identified negatives, was calculated as follows:

$$\frac{\text{Specificity}}{\text{TNR}} = \frac{TN}{TN + FP}. \quad (\text{Equation 15})$$

AUC, which means the area under the ROC curve, measuring the entire two-dimensional area underneath the entire ROC curve (integral calculus) from (0,0) to (1,1), was calculated as follows:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \quad (\text{Equation 16})$$

- 3) IoU calculations: IoU is the intersection over union, and mIoU is the mean of IoU. Given two bounding boxes b_1 and b_2 , their IoU could be computed as

$$IoU(b_1, b_2) = \frac{A(b_1 \cap b_2)}{A(b_1 \cup b_2)} = \frac{A(b_1 \cap b_2)}{A(b_1) + A(b_2) - A(b_1 \cap b_2)} \quad (\text{Equation 17})$$

where $A(\cdot)$ was the area of the shape. Therefore, the calculation of IoU could be formulated as a problem that involves computing the area of each spherical rectangle and the intersection of two spherical rectangles.

QUANTIFICATION AND STATISTICAL ANALYSIS

In our study, statistical data were analyzed using python (version 3.8; <https://www.python.org>). ANOVA was used to compare the contribution of different factors to the results, and $p < 0.01$ was considered statistically significant. For normally distributed data, and the comparison between prediction results of TiNet and TIRADS was calculated by De-Long's test, and $p < 0.001$ was considered statistically significant. Methods for constructing CI are found in the [statistical analysis](#) section. All comparisons used a two-sided $\alpha = 0.05$ for significance testing, except where otherwise indicated.