

Responses Letter

Journal: World Journal of Radiology

Article Title: Non-contrast CT radiomics model to predict benign and malignant thyroid nodules with lobe segmentation: A dual-center study

Submission ID: 106682

Date: 2025-04-18

Dear Yun-XiaoJian Wu, Assistant Editor and Reviewers,

We sincerely appreciate the time and effort you have dedicated to reviewing our manuscript. Your thoughtful comments and constructive suggestions have significantly contributed to the improvement of our work. In response to your feedback, we have carefully revised the manuscript. Below is a detailed point-by-point response to the reviewers' comments, highlighting the changes made to enhance the manuscript.

Sincerely,

Correspondence:

Jian HE MD, PhD

Address: No. 321 Zhongshan Road, Nanjing, Jiangsu Province, China. 210008

Tel: +86 189-5193-2498

Fax: +86 025-83106666

E-Mail: hjxueren@126.com

Reviewer 1, ID: 03731871

This study aims to develop a radiomics machine learning model based on NCCT, utilizing thyroid lobe segmentation. It evaluates the model's clinical utility in diagnosing thyroid nodules, aiming to provide a low-risk and efficient diagnostic tool for distinguishing between benign and malignant thyroid nodules. The author proposes using only standard CT images to differentiate between benign and malignant thyroid nodules, which can lower radiation exposure compared to enhanced CT scans. This is the main advantage of the manuscript. However, this manuscript has the following shortcomings:

Comment 1: In the abstract section, the description of methods and results is not concise and clear enough, lacking logical coherence;

Reply 1: Thank you very much for your valuable comments! We have carefully revised the abstract section to make it more concise, clear and logically coherent. The following are the main changes:

“Methods: This multicenter retrospective study included 272 patients with thyroid nodules (376 thyroid lobes) from Center A (May 2021–April 2024), using histopathology as the reference standard. The dataset was divided into training (264 lobes) and validation (112 lobes) cohorts, supplemented by prospective temporal (97 lobes, May–August 2024, Center A) and external multicenter (81 lobes, Center B) test cohorts. Thyroid lobes were segmented along the isthmus midline, with reliability assessed via intraclass correlation coefficient ($ICC \geq 0.80$). Radiomics signatures were extracted using Pearson correlation and LASSO regression with 10-fold cross - validation. Seven machine learning algorithms were evaluated, with performance assessed via AUC, Brier score, decision curve analysis, and DeLong’s test against

radiologists' assessments. Model interpretability was analyzed using SHAP.

Results: The XGB model showed superior diagnostic stability across datasets, achieving AUC values of 0.899 (95% CI: 0.845–0.932) in training, 0.803 (95% CI: 0.715–0.890) in validation, 0.855 (95% CI: 0.775–0.935) in temporal testing, and 0.802 (95% CI: 0.664–0.939) in external testing, all significantly higher than radiologists' (AUC: 0.596/0.529/0.558/0.538, $P < 0.001$). SHAP analysis highlighted key predictive factors: radscore, age, tumor size group, calcify, and cystic. The model also demonstrated strong calibration (Brier score: 0.125–0.144) and superior clinical net benefit at decision thresholds $>20\%$."

Simplify the description of the methodology: We simplified the description of the methodology section, highlighting the key steps and logical flow of the study, and removing redundant details so that readers can quickly understand the design and methodology of the study.

Optimizing the presentation of results : In the results section, we have reorganized the data to present the model's performance metrics in a clearer way, and highlighted the model's strengths and key findings to make the results more logical and readable.

We believe that these changes can better meet your requirements and make the abstract section more in line with the norms and standards of academic papers. Thank you again for your valuable comments and support!

Comment 2: In the introduction section, the author's rationale for selecting NCCT images of the thyroid gland for research is insufficient.

Reply 2: Thank you very much for your valuable comments! We have carefully revised the introduction section to add the reasons and rationale for choosing non-enhanced CT (NCCT) as the study subject. The following are the main revisions:

"NCCT offers key advantages for thyroid nodule imaging. It eliminates iodine contrast agent, with lower radiation than CECT, and it provides enough thyroid nodule morphology details, especially calcification features,

aiding benign - malignant distinction.”

We believe these changes better meet your requirements and make the introduction section more complete and compelling, while in the discussion section we elaborate on the value of adopting the NCCT. Thank you again for your valuable comments and support!

Comment 3: The exclusion criteria include the size of nodules, but the inclusion criteria do not include the size of nodules.

Reply 3: Thank you very much for your valuable comments! Your question is very pertinent, and regarding the description of nodule size in the inclusion and exclusion criteria, we provide the following explanation:

In this study, we took nodule size into consideration. Nodules smaller than 2 mm were explicitly excluded from the exclusion criteria for the following reason: thyroid nodules smaller than 2 mm are difficult to accurately identify and measure in non-enhanced CT images, which may have a greater impact on the accuracy of the study results.

In order to better reflect the inclusion and exclusion criteria of the study, we will make further clarifications and additions to the relevant content in the revised manuscript to ensure the transparency and logical consistency of the study methodology.

“Eligible patients met the following inclusion criteria: (1) Histopathological confirmation of thyroid nodules, with malignant cases classified as papillary carcinoma and benign cases as adenoma/nodular goiter; (2) Preoperative neck CT performed within 15 days prior to surgery; (3) Thyroid nodules with a maximum diameter of ≥ 2 mm.”

Thank you again for your valuable comments and support!

Comment 4: The author did not describe how to extract the specimen tissue of thyroid nodules in the main text;

Reply 4: Thank you for your valuable comments. We have added to the

pathological assessment section a detailed description of how the thyroid nodule specimens were obtained. Our study used tissue specimens obtained after thyroidectomy (total or partial), which were systematically sampled and processed in representative sections to ensure accuracy and reliability of pathological assessment.

“Histopathological evaluation was conducted on formalin-fixed, paraffin-embedded surgical specimens obtained from thyroidectomy procedures (total or hemithyroidectomy). The specimens were systematically sampled, with representative sections taken from each nodule, including areas of suspicious calcification or cytological atypia. These sections were independently reviewed by two board-certified thyroid pathologists with 15 and 10 years of experience, respectively. Malignancies were classified as PTC, while benign lesions were categorized as adenomas or nodular goiters.”

Thank you again for your comment!

Comment 5: The author describes the diagnosis of malignant thyroid nodules based on the following four conditions. Do malignant thyroid nodules identified on CT scans meet all four conditions at the same time, or only some of them? Additionally, it's important to note that some thyroid nodules with regular morphology can still be malignant. As a result, this diagnostic criterion may yield inaccurate results.

Reply 5: Thank you for your valuable comments. We have further elaborated on the description of the diagnostic criteria:

In this study, the diagnosis of malignant thyroid nodules was based on the following features: (1) irregular nodule morphology; (2) extraperitoneal invasion or invasion into the surrounding tissues; (3) nodule density lower than that of the surrounding thyroid parenchyma; (4) the presence of crushed-stone-like microcalcifications within the lesion; and (5) metastasis to the cervical lymph nodes. It should be noted that malignant thyroid nodules may show only some, but not all, of the above features on CT scan. In

addition, some thyroid nodules with regular morphology may still be malignant. Therefore, these diagnostic criteria have limitations and may not identify all malignant thyroid nodules with complete accuracy. The data for the construction of our model were obtained from cases that were confirmed by pathological findings to ensure the accuracy of the diagnosis. In the diagnosis of human imaging physicians, the judgement is mainly based on the above five features, but due to the limitations of non-enhanced CT in displaying certain subtle features, there are some difficulties in the diagnostic results of physicians, resulting in a low diagnostic accuracy (AUC values of 0.596/0.529/0.558/0.538). This suggests that the diagnosis of benign and malignant thyroid nodules based on non-enhanced CT is still challenging, and more effective auxiliary tools are needed to improve the diagnostic accuracy. In conclusion, we have interpreted the description of the diagnostic criteria to more accurately reflect the reality of the study, and thank you again for your comment.

Comment 6: In the table 4, the AUC , Accuracy, Sensitivity, Specificity, PPV, NPV, Precision Recall, and F1 of RF and KNN were 1 in the training cohort, the authors needs to explain it.

Reply 6: Thank you for your valuable comments. In machine learning models, some algorithms (e.g. RF and KNN) are indeed prone to overfitting on the training set, which may result in metrics such as AUC being close to or equal to 1.000. This phenomenon is usually due to the over-complexity of the model or the relatively small amount of training data, which results in the model remembering the characteristics of the training data, including the noise, and failing to learn the generalized laws behind the data. To avoid overfitting, we used a validation set, a time-test set, and an external test set for a comprehensive assessment of model performance. Through this multi-stage validation approach, we were able to more accurately assess the generalization ability of the models and ultimately filter out the XGB models

that performed most consistently across datasets. In the Discussion section, we provide additional descriptions and emphasize the importance of avoiding overfitting through multi-dataset validation.

“In the meantime, we note that overfitting can occur in machine learning models like RF and KNN, especially when the model is too complex or the training data is limited. This may cause nearly perfect training metrics (e.g., AUC = 1). To avoid overfitting, we used multiple datasets for validation, including validation, temporal test, and external test cohorts. This approach helps assess the models' generalizability and ensures the chosen model (XGB) performs well across different data settings, which is crucial for clinical application.”

Thank you again for your comment!

Comment 7: In the discussion section, the author failed to adequately address the results, leaving key aspects of the findings underexplored. The clarity of the discussion was lacking, making it difficult for readers to grasp the main points effectively. Additionally, the focus of the discussion was not well-defined, which diminished its overall impact. Furthermore, there was a notable absence of comparative analysis with relevant previous studies, which could have provided valuable context and contributed to a more robust understanding of the results.

Reply 7: Thank you for your valuable comments. We have revised the discussion section to make additions and improvements in the following areas:

In-depth presentation of findings: We explain in detail the performance of the XGB model in each dataset, including its superior diagnostic stability and advantages over other models. A more in-depth analysis of the model's AUC values in different datasets is also provided to help readers better understand the clinical significance of the study results.

Clarifying the focus of the discussion: We have reorganised the structure of

the discussion section to focus on the assessment of the performance of the model, its comparison with other diagnostic methods, and the potential of the model for clinical application. Through this restructuring, we aim to make the discussion section more focused and organised, so that the reader can more clearly grasp the core content of the article.

Addition of comparisons with previous studies: We added comparative analyses with previous relevant studies, especially with those that used other imaging methods or machine learning models for thyroid nodule diagnosis. Through this comparison, we not only provide a broader context for the current study, but also highlight the innovations and contributions of this study.

“Our XGB model demonstrated robust diagnostic performance across all datasets, achieving AUC values of 0.899 (training), 0.803 (validation), 0.855 (temporal test), and 0.802 (external test), which were significantly higher than radiologist assessments (AUCs: 0.529–0.596, $P < 0.001$). This aligns with prior studies on CT-derived RadScore, such as Kong et al. (AUC 0.84 with arterial-phase CT^[30]) and Lin et al. (AUC 0.92 with multiphase-enhanced CT^[31]). However, unlike these contrast-dependent approaches, our study introduces an NCCT-based RadScore. NCCT, with its high spatial resolution, effectively captures 3D morphologic and textural details without hemodynamic data^[32]. Compared to CECT, NCCT offers advantages like reduced radiation exposure and no iodinated contrast-related risks, such as allergic reactions and nephrotoxicity^[33-34]. The hemilobar segmentation used in our study also enhances model robustness by enabling volumetric assessment of the entire thyroid lobe, reducing biases from multinodular localization and segmentation variability^[35].”

Thank you again for your valuable comments and support!

Summary of Revisions

Section	Changes
Abstract	Improved logic and conciseness of presentation in the methodology and results sections of the summary
Introduction	Additional reasons for choosing NCCT
Methods	Additional inclusion of exclusionary conditions and correspondence
Methods	Supplementary pathology results for detailed information
Methods	Additional description of radiologists' criteria for diagnosing malignant thyroid nodules
Results	A note on overfitting of machine learning algorithms in training sets
Discussion	Adjustments to the discussion section to optimise presentation and focus on findings

Reviewer 2, ID: 08548531

Reviewer Comments: The manuscript titled “Non-contrast CT radiomics model to predict benign and malignant thyroid nodules with lobe segmentation: A dual-center study” presents a well-executed study that develops and validates a machine learning model based on non-contrast CT (NCCT) radiomics to preoperatively classify thyroid nodules. The use of lobe segmentation, dual-center validation, and the XGB model’s performance across multiple cohorts demonstrate a robust approach with potential clinical relevance. The SHAP analysis adds interpretability, enhancing its applicability. The paper is clearly structured, with detailed methodology and comprehensive results. However, minor revisions are required to address

specific methodological clarifications, presentation inconsistencies, and figure optimization before it can be accepted for publication in the World Journal of Radiology. I recommend acceptance following these minor revisions.

Comments and Suggestions for Revision:

Comment 1: Class Imbalance in the Training Cohort The training cohort exhibits an imbalance between benign ($n = 80$) and malignant ($n = 184$) samples. While the Synthetic Minority Oversampling Technique (SMOTE) was employed to address this, the manuscript does not adequately discuss its potential impact on model performance, such as possible bias in predicting benign cases. I suggest adding a brief discussion in the “Discussion” section (e.g., 1-2 sentences) to acknowledge this limitation and its implications, such as: “Although SMOTE was applied to mitigate class imbalance, its effect on the model’s ability to accurately predict benign nodules may be limited by the original sample size, warranting further validation with a larger benign cohort.”

Reply 1: We sincerely appreciate your valuable and insightful comment. Your guidance has been instrumental in helping us improve the quality of our manuscript.

In the revised version of our paper, we have incorporated the following discussion in the “Limitations” section (page 17, paragraph 2):

“Although SMOTE was applied to mitigate class imbalance in the training cohort, its effect on the model’s ability to accurately predict benign nodules may be limited by the original sample size, warranting further validation with a larger benign cohort.”

Once again, we are extremely grateful for your time and effort in reviewing our work and providing such constructive feedback.

Comment 2: Manual Segmentation Feasibility The reliability of manual segmentation is supported by an intraclass correlation coefficient ($ICC \geq 0.80$), yet the manuscript does not address its practical feasibility or time cost in a

clinical setting. I recommend briefly mentioning the potential of automated segmentation (e.g., using U-Net) as a future direction to improve scalability. This could be added to the “Limitations” or “Future Directions” section, for example: “While manual segmentation ensured consistency, its time-intensive nature may limit clinical adoption; automated approaches could enhance efficiency in future iterations.”

Reply 2: We agree that manual segmentation may pose challenges for clinical scalability. In the revised limitation section (page 17, paragraph 2), we now state:

"While manual segmentation ensured consistency in this study, its time-intensive nature may limit clinical adoption. Future work will prioritize automated segmentation frameworks (e.g., U-Net) to improve efficiency and reproducibility, thereby facilitating seamless integration into radiological workflows."

Thanks again for your comment!

Comment 3: Statistical Notation Consistency Throughout the manuscript, the statistical symbol P should be consistently presented as capitalized and italicized (i.e., P) in accordance with standard statistical notation. Currently, the usage varies (e.g., lowercase “ p ” in some instances and uppercase “ P ” in others). Please revise the text, tables, and figure legends to ensure uniformity (e.g., $P < 0.001$).

Reply 3: We apologize for this oversight. All instances of statistical notation (e.g., p-values) in the text, tables, and figure legends have been revised to italicized uppercase “ P ” (e.g., $P < 0.001$). This includes Tables 1–5 and Figures 4–7.

Comment 4: Table 3 Formatting In Table 3 (“Weights of LASSO selected features and training set Z-score parameters”), the numerical values under the “Average” and “Variance” columns display inconsistent decimal places

(e.g., 68429444 vs. 388518307.355696). I suggest standardizing the number of decimal places (e.g., to 3 or 4 digits) for clarity and uniformity, unless scientific precision dictates otherwise for specific features.

Reply 4: Thank you very much for your suggestion. For the sake of clarity and consistency, we have standardized all values in Table 3 to three decimal places, and the revised table ensures consistent formatting. Thank you again!

Comment 5: Table 4 and Redundant Figures Table 4 (“Multi-cohort predictive performance of various models”) provides a comprehensive summary of model performance across cohorts but is not cited in the main text. Additionally, its content overlaps significantly with Figure 5 (“Performance evaluation of seven machine learning models”) and Figure 6 (“Radar plot comparing the diagnostic performance of the XGB model and radiologists”). The figures, while visually appealing, lack clarity due to low resolution and small font sizes, making the data less accessible than the table. I recommend retaining Table 4, citing it explicitly in the “Results” section (e.g., “Model performance across cohorts is summarized in Table 4”), and removing Figures 5 and 6 to avoid redundancy and improve readability.

Reply 5: Thank you very much for your suggestion, it was an oversight in citing Table 4 and incorrectly wrote Table 4 as Table 3 in the first paragraph on page 14, which has now been corrected and clarified that Model performance across cohorts is summarized in Table 4.

Fig. 5 and Fig. 6 do overlap with Table 4 in the content, after further consideration, we deleted Fig. 6, but kindly ask if Fig. 5 can be retained, ROC, calibration curves and DCA are the important display graphs of clinical prediction models^[1]. Due to the need to insert the image into the word in the submission system resulting in a lower image resolution, therefore, we attach the high-resolution vector image of Fig. 5 in the attachment of the revised manuscript again, and it must be able to satisfy the reading requirements after the typesetting. It can meet the reading requirements. Thank you again for

your valuable comments!

Reference:

[1] Binuya, M.A.E., Engelhardt, E.G., Schats, W. *et al.* Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. *BMC Med Res Methodol* **22**, 316 (2022). <https://doi.org/10.1186/s12874-022-01801-8>

Comment 6: Figure Presentation Several figures (e.g., Figure 2: “Workflow of model development,” Figure 3: “LASSO feature selection and comparison,” and Figure 7: “SHAP analysis and clinical application”) contain text that is too small to be legible, hindering effective communication of results. I suggest revising these figures by increasing font sizes and optimizing layout to ensure all labels, annotations, and legends are clearly readable. High-resolution versions should be provided in the revised submission.

Reply 6: Thank you very much for your suggestion! Fig. 2, Fig. 3 and Fig. 6 (original Fig. 7) do have more textual information in the diagrams, due to the requirements of the submission system and the limitations of word uploading, the images in the initial version of the manuscript are compressed resulting in difficulties in recognizing them, and we are deeply sorry for that. In the revised version, we have uploaded a high-definition vector image and enlarged the font size appropriately, in the hope that the information in the figure can be more clearly and effectively recognized in the adjustment of the editing system. Thank you again for your suggestions!

Summary of Revisions

Section	Changes
Limitation	Description of the limitations of adding SMOTE
Limitation	Adding a note on the limitations of manual segmentation

n	
full text	Changing the format of the statistical symbol P
Table 3	Standardise the number of decimal places
Results	Clarify Table 4 references
Figs	Delete Fig 6
Figs	Adjust fonts in Fig and upload high resolution vector artwork

Reviewer 3, ID: 08380617

Overall Evaluation This study presents a well-designed, dual-center investigation developing an NCCT-based radiomics-clinical fusion model for preoperative differentiation of thyroid nodules. The integration of lobe segmentation, multicenter validation, and SHAP-driven interpretability strengthens the clinical relevance. The manuscript is logically structured, methodologically rigorous, and addresses a significant gap in thyroid nodule diagnostics. However, several issues require clarification and improvement to

enhance scientific validity and readability.

Major Comments:

Comment 1: Methodological Concerns - Exclusion Criteria: Excluding nodules <2 mm and those in the isthmus/pyramidal lobe may limit clinical applicability. Justification for these exclusions (e.g., technical limitations or clinical irrelevance) should be provided. In addition, some of the reasons for the selection of exclusion criteria may need to be further elaborated and added to the discussion, such as what is the reason for the exclusion of < 18-year-olds? Does it lead to selection bias?

Reply 1: Thank you very much for your suggestion! Among the exclusion criteria, we excluded thyroid nodules smaller than 2 mm and those located in the isthmus cone lobes. Our images were performed on the basis of CT plain images of the thyroid gland, which lacked the contrast of iodine contrast agent, and although a resampling of 0.5*0.5*1mm was used to try to maintain the size of the thyroid images, nodules smaller than 2mm did present some difficulty in recognizing them on the plain images^[1-2],so thyroid nodules smaller than 2mm were excluded from the study.

In image segmentation, we used the segmentation of the right and left hemilobar thyroid lobes, which reduces the difference in segmentation as well as avoids localisation for multiple nodules, but there are segmentation difficulties for nodules located exactly in the median isthmus, which fortunately accounted for a very small percentage of patients in our study data (n=4, 0.905%).

The exclusion of minors younger than 18 years of age was mainly considered for two reasons, firstly, patients younger than 18 years of age are minors, whose informed consent and decision-making ability in terms of medical decision-making and research participation is limited, and need to be represented by their legal guardians, which may involve more ethical and legal issues^[3]. Second, the need for stricter protection of minors' rights and interests during the research process, and the adoption of special measures

and considerations in all aspects of study design, implementation and data management to comply with ethical and relevant legal requirements, also pose certain limitations and challenges^[4]. In summary, we excluded these small amounts of patient data from the study to increase the credibility and standardization of the study. Thank you again for your suggestions!

Reference:

[1] Deep learning model for diagnosis of thyroid nodules with size less than 1 cm: A multicenter, retrospective study. *European Journal of Radiology Open*. 2024;33:100390. doi:10.1016/j.ejro.2024.100390.

[2] A Multi-View Deep Learning Model for Thyroid Nodules Detection and Characterization in Ultrasound Imaging. *Diagnostics*. 2024;14(13):2066. Published 2024 Jun 25. doi:10.3390/diagnostics14132066

[3] Involvement and Autonomy of Minors in Medical Settings: The Perspectives of Minors and Parents. *Children (Basel)*. 2023 Oct 17;10(12):1844. DOI:10.3390/children10121844.

[4] NYU Grossman Medical Ethics Teens and Children in Clinical Research. *Children (Basel)*. 2022 Oct 20;9(10):853. DOI:10.3390/children9100853.

Comment 2: Class Imbalance: Despite SMOTE, the training cohort had a 70:30 malignant-to-benign ratio. The impact of imbalance on model generalizability warrants discussion.

Reply 2: Thank you very much for your comments! As our study included pre-surgical CT images, the number of benign nodules was significantly lower than that of malignant nodules, and the data included in the study was not large, so the SMOTE oversampling algorithm was used in the model training, in the hope of minimising the malignant prediction probability due to the difference between benign and malignant. However, since SMOTE is generated data, it is inevitable that there will be bias. In the revised version of our paper, we have incorporated the following discussion in the “Limitations” section (page 17, paragraph 2):

“Although SMOTE was applied to mitigate class imbalance in the training cohort, its effect on the model’s ability to accurately predict benign nodules may be limited by the original sample size, warranting further validation with a larger benign cohort.”

In future studies, the expectation is to include more data and balance the benign-malignant ratio to improve the generalisability of the model. Thank you again for your suggestions!

Comment 3: Consistency between institutions: Although the authors mention in the final limitations section that there may be differences due to differences in institutional protocols, it would be useful to provide further information on what measures have been taken to reduce bias as a result.

Reply 3: Thank you very much for your comments. In the training set, validation set and time test set, our images were all from one device (Philips IQon Spectral CT), while the external test set data were from another CT (GE Revolution 256 CT) in another centre, and the two centres, although with different devices and scanning protocols, since they were in the same area, our patient preparation and training were essentially the same (calm breathing, head tilt, etc.), while in with the processed images we used a consistent resampling scheme (0.5*0.5*1mm) to reduce the differences between images, as well as a Z-score to reduce the differences between features. Inevitably, however, the performance of the external test set is still degraded compared to the other datasets. We made the following modifications in the limitations section:

“Despite the implementation of identical pre-processing procedures, such as resampling, it is unavoidable that the external test set will exhibit bias due to the utilisation of disparate CT equipment.”

Thanks again for your comment!

Comment 4: Statistical and Technical Validation - AUC Interpretation: The

XGB model's AUC in the validation cohort (0.803) is moderate. The authors should discuss whether this performance meets clinical utility thresholds and compare it to existing ultrasound-based models (e.g., TI-RADS).

Reply 4: Thank you very much for your valuable comments! In the present study, the AUC of the XGB model in the validation test cohort was 0.802-0.855. Although this performance is moderate, it still has some clinical utility from the point of view of exploring new methods, and can be a powerful complement to assist physicians in decision-making and to existing models and diagnostic protocols.

In a study of ACR TI-RADS 4 - 5 thyroid nodules diagnosed by multimodal ultrasound imaging histology techniques, the AUC of its multimodal ultrasound TI-RADS combined diagnostic model in the validation set was 0.890, which was only slightly higher than that of the XGB model in this study [1].

In conclusion, although the XGB model in this study is of medium level, its performance is close to that of the ultrasound-based TI-RADS model, and it has certain clinical utility in the diagnosis of thyroid nodules, which can be used as a reference for assisting clinical diagnosis. Thank you again for your comments!

Reference:

[1] Li HJ, Sui GQ, Teng DK, Lin YQ, Wang H. Incorporation of CEUS and SWE parameters into a multivariate logistic regression model for the differential diagnosis of benign and malignant TI-RADS 4 thyroid nodules. *Endocrine*. 2024;83(3):691-699. doi:10.1007/s12020-023-03524-2

Comment 5: Clinical Relevance - Comparison with Radiologists: While the model outperformed radiologists (AUC 0.596–0.538), the radiologists' diagnostic criteria (e.g., TI-RADS categories) and experience levels (junior vs. senior) need elaboration. A direct comparison with established guidelines (e.g., ATA or NCCN) would strengthen clinical relevance.

Reply 5: Your comments are greatly appreciated. The two radiologists who participated in the study had 15 (junior) and 25 (senior) years of diagnostic radiological experience, and in the second paragraph on page eight, we have added the following:

“Radiologist A (junior, 15 years' experience) and Radiologist B (senior, 25 years' experience)”

In the present study, we diagnosed benign and malignant thyroid nodules based on an imaging histology multimodal model of non-enhanced CT images with the following diagnostic criteria: (1) irregular nodule morphology; (2) extraperitoneal invasion or invasion into surrounding tissues; (3) nodule density lower than that of the surrounding thyroid parenchyma; (4) presence of crushed stone-like microcalcifications within the lesion; and (5) presence of metastatic cervical lymph nodes.

The ATA guidelines emphasise the use of ultrasound features in the assessment of the risk of malignancy in thyroid nodules, including features such as microcalcifications, irregular nodule morphology, irregular borders, and extraperitoneal invasion, which echo the features we observe in CT images. For example, the ATA guidelines state that features such as nodule morphology, margins, and calcification are important factors in assessing the risk of malignancy in thyroid nodules, which corresponds to features such as irregular nodule morphology and microcalcifications that we observe in CT images [2].

The NCCN guidelines also focus on ultrasound features of thyroid nodules such as microcalcifications, irregular nodule morphology, irregular borders, and extraperitoneal invasion, which are similar to those we observed in the CT images. The NCCN guidelines also emphasise the importance of these features in the assessment of the risk of malignancy of thyroid nodules, which is somewhat similar to the diagnostic criteria that we used in our study [3].

In conclusion, we believe that the diagnostic criteria based on non-enhanced CT images proposed in our study are consistent with the existing ultrasound

TI-RADS and ATA and NCCN guidelines in terms of diagnostic goals and assessment of features. Although our study used a different imaging modality, its diagnostic criteria fit with the core principles of the existing guidelines, which provides new perspectives and complements the multimodal diagnosis of thyroid nodules. Thank you again for your comments!

Reference:

[1] Tessler F N, Middleton W D, Grant E G. Thyroid imaging reporting and data system (TI-RADS): a user's guide. *Radiology*. 2018;287:29-36.

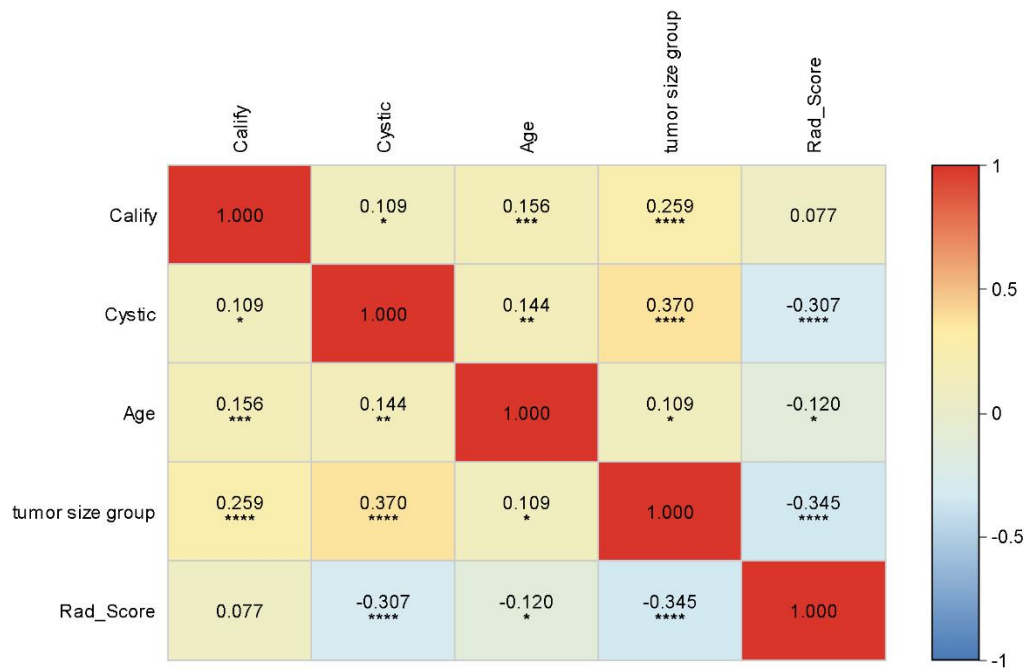
[2] Haugen, B. R., Alexander, E. K., Bible, K. C., Doherty, G. M., Mandel, S. J., Nikiforov, Y. E., ... & Wartofsky, L. (2016). 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid*, 26(1), 1-133.

[3] Haddad, R. I., Bischoff, L., Ball, D., Bernet, V., Blomain, E., Busaidy, N. L., ... & Darlow, S. (2022). Thyroid carcinoma, version 2.2022, NCCN clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 20(8), 925-951.

Comment 6: SHAP Interpretability: The clinical translation of SHAP-derived feature importance (e.g., "radiomics score" as top predictor) requires clearer linkage to histopathological correlates (e.g., microcalcifications, cellular density).

Reply 6: Thank you very much for your professional question. Due to the imperfection of our data, there is no way to correlate histopathological correlates such as cell density for the time being. Here, we add a correlation matrix (including the number of training and validation sets) that was incorporated into the final model, and we found that Radscore was significantly negatively correlated with cystic and tumor size with age, which is also consistent with our findings: less cystic and smaller tumor size

combined with younger age in PTC. In future studies, we will definitely strive to include more pathological histological information to provide more valuable information for clinical translation and model interpretation.



Comment 7: Limitations - The retrospective design and single histopathological subtype (PTC) limit generalizability to other malignancies (e.g., follicular carcinoma). Minor Comments 1. Language and Clarity - Ensure consistent terminology (e.g., "calcify" vs. "calcification," "cystic" vs. "cystic degeneration"). 2. Tables and Figures - Figure 7: Include arrows/annotations in case illustrations to highlight key radiomic features.

Reply 7: Thank you very much for your suggestion, although PTC accounts for 80-90% of all malignant thyroid cancers, a single pathology type does limit the amount of model application, and we will gradually expand the inclusion of other malignant thyroid cancers in future studies. In the meantime, we corrected the terminology throughout the text and unified the expression as calcify and cystic. In the case presentation of Fig. 7, the radiological features of

the two cases: calcify and cystic were negative, and the thyroid tissue was smaller in the image, so we added the text in the figure annotation. Thank you again for your suggestions!

Summary of Revisions

Section	Changes
Methods	Explanation for exclusion of <2mm, <18 years and isthmus nodules
Limitation	Explain the limitations of SMOTE
Methods	Explain that the external is preprocessed using different equipment
Results	Explaining differences between model performance and TI-RADS
Methods	Explain the identity of the radiologist and the consistency between manual diagnosis and ATA,NCCN
Results	Additional description of the correlation between radscore and features
other	Harmonized presentation, supplemented by graphical notes

Revision reviewer 1

Comment: I appreciate the authors' thorough revisions in response to the previous comments. The manuscript has been significantly improved in both clarity and scientific rigor. All concerns raised during the initial review have been adequately addressed, with additional experiments/ data and textual refinements strengthening the overall quality of the work. The study now presents a well-structured narrative, supported by robust methodology and clear results. The conclusions are justified by the data, and the discussion appropriately contextualizes the findings within the field. I recommend acceptance of the manuscript in its current form.

Reply: Thanks for your comments.

Revision reviewer 2

Comment: The authors have addressed my previous concerns satisfactorily. The manuscript is now close to being suitable for publication, pending minor revisions. I recommend acceptance after the following issues are addressed: Many of the figures still contain fonts that are too small, particularly the axis labels on both the x- and y-axes. This significantly compromises the readability of the visual data and may hinder comprehension for the reader. The authors should ensure that all textual elements in the figures are clearly legible when printed at journal size. Revisions should be made accordingly.

Reply: We understand the importance of having editable figures. Although the original figure document was not editable, we have carefully recreated Figures using PowerPoint to ensure it matches the original image and is fully editable. This allows for any necessary adjustments to the graphics, arrows, or text.