

106682_Auto_Edited - 副本.docx

WORD COUNT

3534

TIME SUBMITTED

09-MAY-2025 03:58PM

PAPER ID

116136763

Name of Journal: *World Journal of Radiology*

Manuscript NO: 106682

Manuscript Type: ORIGINAL ARTICLE

Retrospective Study

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

Wang H *et al.* NCCT radiomics for thyroid nodules

Abstract

BACKGROUND

Accurate preoperative differentiation of benign and malignant thyroid nodules is critical for optimal patient management. However, conventional imaging modalities present inherent diagnostic limitations.

AIM

To develop a non-contrast computed tomography-based machine learning model integrating radiomics and clinical features for preoperative thyroid nodule classification.

METHODS

This multicenter retrospective study enrolled 272 patients with thyroid nodules (376 thyroid lobes) from center A (May 2021-April 2024), using histopathological findings as the reference standard. The dataset was stratified into a training cohort (264 lobes) and an internal validation cohort (112 lobes). Additional prospective temporal (97 lobes, May-August 2024, center A) and external multicenter (81 lobes, center B) test cohorts were incorporated to enhance generalizability. Thyroid lobes were segmented along the isthmus midline, with segmentation reliability confirmed by an intraclass correlation

coefficient (≥ 0.80). Radiomics feature extraction was performed using Pearson correlation analysis followed by least absolute shrinkage and selection operator regression with 10-fold cross-validation. Seven machine learning algorithms were systematically evaluated, with model performance quantified through the area under the receiver operating characteristic curve (AUC), Brier score, decision curve analysis, and DeLong test for comparison with radiologists interpretations. Model interpretability was elucidated using SHapley Additive exPlanations (SHAP).

RESULTS

The extreme gradient boosting model demonstrated robust diagnostic performance across all datasets, achieving AUCs of 0.899 [95% confidence interval (CI): 0.845-0.932] in the training cohort, 0.803 (95%CI: 0.715-0.890) in internal validation, 0.855 (95%CI: 0.775-0.935) in temporal testing, and 0.802 (95%CI: 0.664-0.939) in external testing. These results were significantly superior to radiologists assessments (AUCs: 0.596, 0.529, 0.558, and 0.538, respectively; $P < 0.001$ by DeLong test). SHAP analysis identified radiomic score, age, tumor size stratification, calcification status, and cystic components as key predictive features. The model exhibited excellent calibration (Brier scores: 0.125-0.144) and provided significant clinical net benefit at decision thresholds exceeding 20%, as evidenced by decision curve analysis.

CONCLUSION

The non-contrast computed tomography-based radiomics-clinical fusion model enables robust preoperative thyroid nodule classification, with SHAP-driven interpretability enhancing its clinical applicability for personalized decision-making.

Key Words: Papillary thyroid carcinoma; Thyroid nodules; Radiomics; Machine learning; Non-contrast computed tomography

Wang H, Wang X, Du YS, Wang Y, Bai ZJ, Wu D, Tang WL, Zeng HL, Tao J, He J. Non-contrast computed tomography radiomics model to predict benign and malignant thyroid nodules with lobe segmentation: A dual-center study. *World J Radiol* 2025; In press

Core Tip: This study introduces a novel non-contrast computed tomography-based machine learning model integrating radiomics and clinical features with lobe segmentation for preoperative differentiation of benign and malignant thyroid nodules. Leveraging dual-center data and thyroid lobe segmentation, the extreme gradient boosting model demonstrated superior diagnostic accuracy and stability across diverse cohorts, outperforming traditional radiologist assessments. Key predictors, including radiomic score, age, and tumor size group, calcify and cystic, were showed through SHAP analysis, enhancing model interpretability. The approach offers a robust, non-invasive tool for personalized preoperative decision-making, with the potential to improve clinical management of thyroid nodules.

INTRODUCTION

Thyroid nodules, the most prevalent endocrine tumors, have been ranked as the third most common malignancy in thyroid cancer according to the 2022 China Cancer Burden Report[1]. Among these, papillary thyroid carcinoma (PTC) represents the predominant malignant subtype, accounting for 85%-90% of all thyroid malignancies[2]. Although PTC generally exhibits a favorable prognosis, regional lymph node metastasis occurs in 30%-80% of patients, complicating diagnosis and increasing the risk of adverse outcomes[3,4]. Clinical management strategies vary significantly depending on malignancy status: Benign nodules are typically monitored conservatively, whereas malignant cases necessitate radical surgery, often supplemented by radioactive iodine therapy and targeted treatments[5]. This highlights the critical need for early and accurate differentiation between benign and malignant nodules to optimize therapeutic strategies and improve patient prognosis[6].

The National Comprehensive Cancer Network guidelines emphasize the essential role of imaging in thyroid nodule evaluation[7]. Ultrasound (US) is the primary diagnostic modality due to its noninvasive nature, absence of ionizing radiation, and ease of operation[8]. However, its diagnostic accuracy remains suboptimal (60%-70%) due to spatial resolution limitations and operator-dependent variability, particularly in detecting microcalcifications and occult lesions[9]. While US-guided fine-needle aspiration is the diagnostic gold standard, its invasive nature poses risks of bleeding and infection. Additionally, the American Thyroid Association discourages fine-needle aspiration for nodules < 1 cm unless local infiltration or lymph node metastasis is suspected[10]. Contrast-enhanced computed tomography (CECT) offers superior spatial resolution and 3D reconstruction for preoperative assessment of tumor invasiveness, but its clinical utility is constrained by radiation exposure, iodine contrast risks, and limited quantitative analytical capacity[11]. These limitations underscore the need for a more accessible, low-risk diagnostic approach with improved efficacy.

Recent advancements in artificial intelligence have revolutionized medical imaging analysis, with machine learning algorithms demonstrating significant potential in complex image classification by leveraging large-scale data training, thereby enhancing diagnostic accuracy and consistency[12]. The field of radiomics, first introduced by Lambin *et al*[13] in 2012, enables noninvasive tumor characterization through high-throughput extraction of imaging features, transforming visual data into quantifiable metrics[13-15]. Current radiomics research on thyroid nodules has predominantly focused on US and CECT, while non-contrast computed tomography (NCCT) remains underutilized[16,17]. NCCT presents several advantages for thyroid nodule assessment. Unlike CECT, it eliminates the need for iodine-based contrast agents while delivering lower radiation exposure. Additionally, NCCT provides sufficient morphological detail of thyroid nodules, particularly in characterizing calcification patterns - a critical feature for differentiating benign from malignant lesions. Moreover, many studies rely on single-nodule segmentation, which may introduce bias due to manual lesion localization and margin delineation in multinodular cases[18]. This study aims to

develop an NCCT-based radiomics machine learning model utilizing thyroid lobe segmentation and to evaluate its clinical utility in thyroid nodule diagnosis, providing a low-risk, efficient diagnostic tool for early detection.

MATERIALS AND METHODS

Patient's selection

This retrospective study was conducted in accordance with the Declaration of Helsinki and received formal approval from the Institutional Review Board of Nanjing Medical University's Fourth Affiliated Hospital, approval No. 20240628-K077). The ethics committee granted an exemption from informed consent requirements, as per institutional policies for retrospective data analysis. The study population comprised patients who underwent thyroid resection between May 2021 and April 2024. 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 computed tomography (CT) performed within 15 days prior to surgery; and (3) Thyroid nodules with a maximum diameter of ≥ 2 mm. Exclusion criteria included: (1) Diffuse thyroid pathology (Hashimoto's or granulomatous thyroiditis; $n = 8$); (2) Nodules < 2 mm ($n = 21$); (3) Non-diagnostic image quality ($n = 19$); (4) Nodules located in the isthmus or pyramidal lobe ($n = 4$); (5) History of malignancy or prior radiotherapy ($n = 12$); and (6) Age < 18 years ($n = 1$). After excluding 65 ineligible patients, 376 thyroid lobes from 272 patients (104 bilateral cases) were included and stratified into training and validation cohorts (7:3 ratio) using the thyroid isthmus midline as the anatomical boundary. Additionally, the study incorporated a temporal test cohort (May-August 2024; 97 lobes from 75 patients) and an external test cohort (Zhongda Hospital; January-June 2024; 81 lobes from 75 patients) for independent evaluation. The study adhered to the CLEAR checklist[19], with patient selection workflows illustrated in Figures 1 and 2.

Imaging acquisition

Patients enrolled in the training, validation, and temporal test cohorts received both NCCT and CECT examinations through the IQon Spectral CT (Philips Medical Systems, Netherlands). The imaging protocol encompassed anatomical regions extending from the skull base through the superior mediastinal compartment. Subjects were positioned in a standardized supine orientation with cervical extension and shoulder depression to reduction of cervical sclerotic bundle artifacts while maintaining normal steadily breathing throughout image acquisition. Acquisition parameters included a 120 kVp tube voltage, DoseRight auto-adjusting tube current (mean 145 mAs, Index 23), 64 × 0.625 mm collimation, 250 mm field of view, pitch 0.969, matrix 512 × 512, iDose4 iterative reconstruction, and 1 mm isotropic resolution (350/60 window settings)[28]. CECT involved dual-phase imaging at 25 seconds (arterial) and 60 seconds (venous) post-iodhexol injection (320 mgI/mL, 3 mL/s), with DICOM data archived in the PACS system. The external test cohort underwent imaging on a Revolution CT scanner (GE Healthcare, United states) using distinct acquisition parameters: 120 kVp tube voltage, automated tube current modulation, 256 mm × 0.625 mm collimation, 250 mm field of view, pitch 0.992, matrix 512 × 512, ASIR-V (50%), and 0.625 mm slice thickness (350/50 window settings). Notably, contrast-phase data were excluded from all analyses to maintain NCCT specificity.

Pathological results, imaging features, and clinical variable collection

Histopathological assessment was performed on formalin-fixed, paraffin-embedded surgical specimens obtained *via* total or hemithyroidectomy. Systematic sampling was conducted, with representative sections harvested from each nodule - particularly targeting regions exhibiting suspicious calcifications or cytological atypia. All specimens underwent blinded evaluation by two fellowship-trained thyroid pathologists (with 15 years and 10 years of subspecialty experience, respectively). Malignant lesions were histologically confirmed as PTC, while benign pathologies were classified as follicular adenomas or nodular goiters. radiologist A (junior, 15 years' experience) and radiologist B (senior, 25 years' experience) independently evaluated CT

imaging features through PACS workstations, including tumor size group [small (≤ 5 mm), intermediate (5-10 mm), large (> 10 mm)], multiple nodules, calcify, and cystic. Interobserver discrepancies (18 cases, 3.25%) primarily involved sub-centimeter nodule measurements (6 cases) and microcalcification detection (12 cases) and were adjudicated by a third radiologist (25 years of experience). Calcify and cystic exhibited perfect inter-rater agreement due to standardized diagnostic criteria. Demographic and biochemical parameters - including age, sex, body mass index, free triiodothyronine, free thyroxine, thyroid-stimulating hormone, thyroglobulin antibody, and thyroid peroxidase antibody - were extracted from the hospital information system.

For human-machine comparative analysis, both radiologists performed blinded NCCT evaluations on the training and validation cohorts following a one-month washout period and standardized diagnostic training. The assessment protocol restricted access to clinical data, allowing only laterality information while concealing patient identifiers, cohort allocation, and prior interpretations. Malignancy classification was based on the following criteria: (1) Irregular nodule morphology; (2) Extracapsular invasion or infiltration into adjacent tissues; (3) Lower nodule density relative to surrounding thyroid parenchyma; (4) Gravelly microcalcifications within the lesion; and (5) Presence of metastatic cervical lymph nodes[20].

Lobe thyroid tissue segmentation

Segmentation was conducted using 3D Slicer 5.7.0 (<https://www.slicer.org>) with isotropic voxel resampling ($0.5 \text{ mm}^3 \times 0.5 \text{ mm}^3 \times 1 \text{ mm}^3$) to enhance spatial resolution for small thyroid structures. A radiographer with 14 years of CT imaging expertise performed segmentation of the thyroid lobe region of interest on axial CT slices, guided by pathological diagnosis results. The segmentation adhered to anatomical boundaries defined by the midline of the thyroid isthmus. Radiologist A subsequently conducted a meticulous layer-by-layer review to verify delineation accuracy.

Radiomics feature extraction and consistency validation

The radiomics extraction process was conducted utilizing the PyRadiomics toolkit (<https://github.com/Radiomics/pyradiomics>), which resulted in the generation of 1130 features. These features encompassed 14 shape descriptors, 18 first-order histogram-based indices, and 75 textural characteristics derived through grey-level matrix transformations (including GLCM, GLDM, GLRLM, GLSZM, and NGTDM analyses). In addition, 279 Log-filtered features and 744 features obtained through wavelet decomposition were identified. To assess segmentation reproducibility, 40 randomly selected thyroid lobes underwent duplicate manual segmentation by radiologist A, followed by an independent segmentation by radiographer A one week later. The intraclass correlation coefficient (ICC) quantified feature consistency across intra- and inter-observer assessments, with features demonstrating $ICC \geq 0.8$ retained to ensure robustness against segmentation variability.

Feature engineering and establishment of radiomics score

All radiomic features were standardized using Z-score normalization ($Z = [X - \mu]/\sigma$, where X represents the feature value, μ the mean, and σ the standard deviation) to mitigate scale disparities. To counteract class imbalance in the training cohort, where benign samples were underrepresented, synthetic minority oversampling was applied at a 1:1.3 ratio to equalize benign and malignant cases. Feature selection involved Pearson correlation analysis, retaining a single representative feature per cluster exhibiting $r > 0.7$ to minimize multicollinearity. Least absolute shrinkage and selection operator (LASSO) regression with 10-fold cross-validation determined the optimal regularization parameter (λ minimum) and identified non-zero coefficient features, culminating in the construction of the radiomic score (RadScore) through linear combination: $RadScore = \sum (\text{feature_weight} \times \text{feature_value}) + \text{intercept}$.

Machine learning model construction and evaluation

Seven machine learning models were constructed based on differential analysis of clinical features in the training cohort, integrating statistically significant variables

alongside RadScore. These models included decision trees, random forests (RF), logistic regression, support vector machines (SVM), extreme gradient boosting (XGB), K-nearest neighbors, and light gradient boosting machines. Model performance was evaluated by comparing the area under the receiver operating characteristic curve (AUC), with the optimal algorithm determined by the highest AUC values and further visualized through confusion matrices. Calibration curves assessed predictive accuracy, while decision curve analysis quantified clinical net benefit across probability thresholds. Feature importance was interpreted using SHapley Additive exPlanations (SHAP) values to determine variable contributions to model predictions.

Statistical analysis

All analyses were conducted in R software (v4.2.1), utilizing the glmnet package for LASSO regression and the rms package for nomogram construction and calibration curve generation. The normality of data distribution was assessed *via* the Kolmogorov-Smirnov test, with non-normally distributed data expressed as median (interquartile range) and compared using the Mann-Whitney *U* test. Categorical variables were presented as frequencies (percentages) and analyzed *via* the χ^2 test. The performance metrics of the model included AUC, accuracy, sensitivity, specificity, positive and negative predictive values (positive predictive value/negative predictive value), precision, recall, F1 score, and Brier score. AUC comparisons across models were conducted using DeLong's test, with statistical significance defined as $P < 0.05$.

RESULTS

Baseline characteristics

The retrospective cohort comprised 272 patients [60 males (22.1%), 212 females (77.9%); mean age, 48.2 ± 13.7 years] with 376 thyroid lobes, including 104 bilateral cases. Histopathological evaluation identified 270 PTC and 106 benign lobes. Patients were randomly allocated into a training cohort (271 lobes) and a validation cohort (105 lobes), with no significant baseline differences between groups (all $P > 0.05$). The temporal test

cohort included 75 patients (97 lobes, 22 bilateral), while the external test cohort consisted of 75 patients (81 lobes, 6 bilateral) from an independent institution (Table 1). Univariate analysis identified four malignancy-associated variables: Patient age ($P = 0.016$), nodule size category ($P < 0.001$), calcify patterns ($P = 0.003$), and cystic ($P = 0.012$). Comparative statistics between benign and malignant characteristics are detailed in Table 2.

Feature screening and radscore calculation

Radiomic feature extraction using PyRadiomics yielded 1130 features per thyroid lobe, subjected to a multi-stage refinement process. Inter- and intra-observer reliability analysis ($ICC \geq 0.8$) retained 612 stable features, followed by Pearson correlation filtering ($r \geq 0.7$), which removed 51 redundant parameters. LASSO regression with ten-fold cross-validation (optimal $\lambda = 0.01236625$) identified 27 non-zero discriminative predictors. These features were linearly combined using LASSO-derived coefficients to generate the RadScore, which demonstrated significant diagnostic stratification between benign and malignant lesions, as detailed in Table 3 and illustrated through feature selection dynamics in Figures 3 and 4.

Model construction and evaluation

Five clinically significant predictors - RadScore, age, tumor size group, calcify pattern, and cystic - were integrated into seven machine learning algorithms. Comprehensive evaluation across multiple validation metrics identified the XGB model as the optimal classifier, demonstrating superior discriminative performance. The XGB model achieved an AUC of 0.889 [95% confidence interval (CI): 0.845-0.932] in the training cohort, 0.803 (95%CI: 0.715-0.890) in the validation cohort, 0.855 (95%CI: 0.775-0.935) in the temporal test, and 0.802 (95% CI: 0.664-0.939) in the external test. Robust accuracy (0.696-0.845) and F1 scores (0.790-0.852) were maintained across all datasets, model performance across cohorts is summarized in Table 4. The model exhibited high calibration fidelity (Brier scores: 0.121-0.144) and substantial clinical utility, as indicated

by decision curve analysis with net benefit thresholds exceeding 20% probability (Figure 5). These results collectively validate the XGB model's generalizability and reliability in thyroid nodule characterization.

Model performance and interpretability

In the validation cohort, the XGB model outperformed both human radiologists ($P < 0.001$) and other machine learning models, as confirmed by the DeLong test (decision trees: $P < 0.05$; RF: $P < 0.01$), as detailed in Table 5. SHAP analysis ranked RadScore as the most influential predictor, followed by age, tumor size group, calcify, and cystic. This interpretability framework was further corroborated through representative case illustrations (Figure 6), highlighting the distinct predictive contributions of each feature in one benign and one malignant case.

DISCUSSION

This retrospective study developed a machine learning model integrating NCCT radiomics and clinical features for thyroid nodule malignancy prediction. The XGB algorithm exhibited superior stability and generalizability across all datasets. Leveraging NCCT-derived radiomic signatures, the model enables noninvasive differentiation of thyroid nodules, presenting a novel framework for precision diagnostics. SHAP-based interpretability further enhances its clinical applicability by providing transparent decision-support insights.

PTC, the most prevalent thyroid malignancy, is characterized by *BRAFV600E* mutations and RET/PTC rearrangements, both of which drive tumor progression *via* sustained MAPK/ERK pathway activation[21]. In contrast, benign nodules typically maintain follicular architecture, exhibit homogeneous colloid distribution, and lack infiltrative margins[22]. Conventional imaging modalities, including US, which detects microcalcifications with 91% specificity, and CT, which assesses capsular integrity with 85% sensitivity, remain limited in their capacity to quantify tumor microenvironment heterogeneity or molecular features[23]. Radiomics, a high-throughput analytical

approach, addresses these limitations by facilitating machine learning-driven differentiation of malignant phenotypes[24]. Feature analysis revealed multidimensional distinctions: Shape irregularity correlated with invasive growth patterns[25], first-order statistics reflected grayscale dispersion associated with cellular density heterogeneity[26], and texture features captured microstructural disorganization characteristic of malignancy[27]. Additionally, Log and wavelet transformations enhanced sensitivity to microcalcifications and angiogenic patterns[28]. Beyond diagnostic applications, these imaging biomarkers provide mechanistic insights into thyroid carcinogenesis and may inform targeted therapeutic strategies[29].

Our XGB model demonstrated robust and consistent diagnostic performance across all evaluation cohorts, 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 superior to radiologists assessments (AUC range: 0.529-0.596; $P < 0.001$ by DeLong test). These results corroborate previous CT-based radiomics studies including Kong *et al*[30] (AUC = 0.84 using arterial-phase CT) and Lin *et al*[31] (AUC = 0.92 with multiphase-enhanced CT), while introducing the novel application of NCCT-derived RadScore that leverages high spatial resolution to capture comprehensive 3D morphologic and textural features without requiring hemodynamic data[32]. Compared to CECT, our NCCT-based approach provides distinct clinical advantages including significantly reduced radiation exposure and elimination of iodinated contrast-associated risks such as allergic reactions and nephrotoxicity[33,34]. The implementation of hemilobar segmentation further enhances model robustness by facilitating complete volumetric assessment of thyroid lobes while minimizing potential biases from multinodular localization and segmentation variability[35]. We observed that certain machine learning algorithms (*e.g.*, RF and K-nearest neighbors) were prone to overfitting, as evidenced by near-perfect training performance (AUC = 1.000) but poor generalization, which we mitigated through a rigorous multi-tiered validation strategy incorporating independent validation, prospective temporal testing, and multicenter external

validation cohorts to ensure the selected XGB model maintains strong performance across diverse clinical settings - a critical prerequisite for successful clinical translation.

Several key clinical predictors emerged from the analysis. Malignant nodules were associated with younger patients (median age: 49.0 *vs* 54.5 years, $P < 0.001$), consistent with established thyroid cancer epidemiology[36]. The predominance of malignancy in nodules ≤ 10 mm (63% *vs* 31% > 10 mm in the benign group) may reflect the high prevalence of *BRAFV600E* mutations in microcarcinomas[37]. Calcify were significantly more frequent in malignant lesions (32% *vs* 18%, $P = 0.028$), aligning with the presence of psammoma bodies in PTC[38], while cystic was more common in benign nodules (29% *vs* 7%, $P < 0.001$), a pattern indicative of follicular adenoma degeneration[39]. These findings corroborate established sonographic malignancy markers, emphasizing the value of multimodal characterization.

Several important limitations should be acknowledged in this study. Despite employing standardized pre-processing procedures including image resampling, inherent biases persist in the external test set due to variations in CT scanner manufacturers and acquisition protocols across institutions. While synthetic minority oversampling was implemented to address class imbalance in the training cohort, its effectiveness in improving benign nodule prediction may be constrained by the original limited sample size of benign cases, necessitating future validation with expanded benign cohorts. Although we comprehensively evaluated multiple machine learning algorithms, the maximum achieved AUC below 0.9 indicates potential for improvement through larger, more diverse datasets. The manual segmentation approach, while ensuring consistency through rigorous inter-reader ICC verification, remains subject to inherent observer variability and represents a time-intensive process that may hinder clinical scalability. Future research directions will prioritize the development of automated segmentation pipelines leveraging U-Net architectures, coupled with the integration of advanced deep learning models (ResNet, DenseNet, Vision Transformer) within multicenter validation studies to enhance both predictive accuracy and clinical applicability while maintaining rigorous methodological standards.

CONCLUSION

In conclusion, this study developed an NCCT-based XGB model leveraging thyroid lobe segmentation for noninvasive differentiation of benign and malignant thyroid nodules. When coupled with SHAP interpretability, this model holds promise for clinical decision support in preoperative evaluation.