Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology) in press
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Research Article

https://doi.org/10.1631/jzus.B2500465

Improved ultrasound diagnosis of lateral lymph node metastasis in papillary thyroid carcinoma through integrated AI-driven multimodal analysis

Jiawei FENG[1*✉], Lu ZHANG[2*], Yuxin YANG[1], Shuiqing LIU[3], Ancheng QIN[4], Yong JIANG[1✉]

[1]*Department of Thyroid Surgery, The Third Affiliated Hospital of Soochow University, Changzhou First People's Hospital, Changzhou, Jiangsu 213000, China*
[2]*Department of Ultrasound, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200020, China*
[3]*Department of Ultrasound, The Third Affiliated Hospital of Soochow University, Changzhou First People's Hospital, Changzhou, Jiangsu 213000, China*
[4]*Department of Thyroid Surgery, Suzhou Municipal Hospital, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou, Jiangsu 215002, China*

Abstract: Objective: Lateral lymph node metastasis (LLNM) in papillary thyroid carcinoma (PTC) significantly impacts PTC prognosis and treatment strategies. This study aimed to develop and validate a multimodal artificial intelligence (AI)-driven integrated model to predict LLNM by combining clinical data, radiomics features, and deep learning-based imaging features. Methods: A cohort of 1,566 patients was divided into training, validation and test sets. To indirectly assess LLNM risk based on tumor aggressiveness characteristics, radiomics features were extracted from ultrasound images of the primary thyroid tumor and selected using Lasso regression to construct a support vector machine (SVM) model. Clinical variables were used for logistic regression. Deep learning features characterizing the primary thyroid nodule were derived from the TresNet model, and these predictions were integrated into a multimodal model and evaluated using ROC curves. Results: The integrated model outperformed individual models, achieving the highest area under the curve in the training (0.984), validation (0.943), and test (0.951) datasets. SHapley Additive exPlanations (SHAP) analysis identified key predictive factors such as TresNet scores, SVM-based radiomics features, tumor size, location, sex, and thyroglobulin antibody levels. After incorporating the model, the diagnostic accuracy enhanced for ultrasound physicians at all levels (junior, medium, and senior), with significant improvements in sensitivity, specificity and confidence. Conclusions: The proposed AI-driven integrated model, which predicts LLNM risk by analyzing primary thyroid tumor characteristics, demonstrated robust predictive performance and clinical utility in LLNM risk stratification for PTC patients. It significantly improves diagnostic accuracy compared to individual models, providing a non-invasive, personalized tool to support clinical decision-making.

Key words: Papillary thyroid carcinoma; Lateral lymph node metastasis; Artificial intelligence; Multimodal model; Deep learning

1 Introduction

Papillary thyroid carcinoma (PTC), the most common endocrine malignancy, generally has a favorable prognosis. However, lateral lymph node metastasis (LLNM), occurring in 12.6%–32.8% of cases, increases

✉Jiawei FENG, 2236305087@qq.com
Yong JIANG, yjiang8888@hotmail.com
* The two authors contributed equally to this work
Jiawei FENG, https://orcid.org/0000-0003-2421-118X
Yong JIANG, https://orcid.org/0000-0003-2502-1353

Received Aug. 4, 2025; Revision accepted Oct. 27, 2025;
Crosschecked xxx. xx, 20xx; Published online xxx. xx, 20xx

recurrence risk and worsens outcomes (Feng et al., 2022). Thus, accurate preoperative LLNM detection is crucial for tailoring surgical strategies and treatment plans.

The 2015 American Thyroid Association (ATA) management guidelines advise against routine prophylactic lateral neck dissection in patients without preoperative evidence of LLNM (Haugen et al., 2016). However, the current diagnostic methods have limitations (Zhao and Li, 2019). Ultrasound, though non-invasive, has low sensitivity for small or subclinical LLNM (Yang et al., 2022). The visualization of lateral lymph nodes is often hindered by surrounding bony structures (clavicle, mandible) and air-filled tissues (trachea, esophagus), particularly limiting the detection of small or deep-seated nodes. Fine-needle aspiration (FNA) cytology is restricted to visible suspicious nodes, leaving many LLNM cases undetected. These problems highlight the need for advanced methods to enhance diagnostic precision (Xing et al., 2020).

Recent advancements in artificial intelligence (AI) offer promising solutions for existing diagnostic challenges (Papadimitroulas et al., 2021). By analyzing ultrasound images of the primary thyroid tumor, AI models can extract features that reflect tumor aggressiveness and indirectly predict LLNM risk, as aggressive tumor characteristics are strongly associated with lymphatic spread. Radiomics extracts subtle imaging features from the primary lesion linked to LLNM beyond human detection (Yu et al., 2021), and deep learning, particularly convolutional neural networks, excels in analyzing complex imaging data of thyroid nodules (Yasaka et al., 2018). Meanwhile, most AI models lack transparency, fail to incorporate clinical variables like age and gender, and are affected by preprocessing distortions, limiting clinical adoption (Baselli et al., 2020). Integrating clinical data with radiomic and deep learning features derived from primary tumor analysis can improve accuracy and interpretability, fostering clinician trust (Brocki and Chung, 2023).

This study integrates clinical data, radiomics and deep learning features extracted from primary thyroid tumor ultrasound images to develop and validate an AI-based multimodal model for preoperative LLNM prediction in PTC patients. The core premise is to predict LLNM based on the primary tumor characteristics rather than directly analyzing lateral lymph nodes, leveraging the established link between tumor aggressiveness and metastatic potential. The proposed model seeks to overcome limitations of traditional diagnostics, improve accuracy, enhance interpretability, and support personalized surgical planning in thyroid cancer management.

2 Materials and methods

2.1 Patients and datasets

The clinical data and ultrasound images of 1,566 patients were retrospectively collected from January 2022 to June 2024 across three hospitals in China: Changzhou First People's Hospital (training and validation sets), Shanghai Ruijin Hospital (test set A), and Suzhou Municipal Hospital (test set B). Patients were divided into training (878), validation (220), test set A (348), and test set B (120) groups (Fig. 1).

Patients aged 18 or older who underwent thyroid surgery with pathological assessment were included. The inclusion criteria were: (1) primary classic PTC confirmed by pathology; (2) high-quality preoperative ultrasound images; (3) complete clinical data; (4) no prior treatment. The exclusion criteria were: (1) non-classic PTC or other thyroid cancers; (2) prior thyroid surgery/ablation; (3) history of other malignancies or familial cancer syndromes; (4) poor ultrasound quality; (5) incomplete data; (6) loss to follow-up; or (7) non-curative surgery with persistent disease within six months.

2.2 Clinical data and ultrasound image collection

Body mass index (BMI) was calculated as weight in kilograms divided by the square of height in meters (kg/m2). Based on WHO guidelines, patients were classified as underweight (BMI<18.5), normal weight ($18.5 \leqslant$ BMI<25), overweight ($25 \leqslant$ BMI<30), or obese (BMI $\geqslant 30$) (Kim et al., 2016). Chronic lymphocytic

thyroiditis (CLT) was defined as diagnosis by elevated thyroid peroxidase antibodies or diffuse heterogeneity on ultrasound. Extrathyroidal extension was defined as >25% tumor abutment to the thyroid capsule on ultrasound (Chung et al., 2020). For multifocal lesions, the largest tumor determined the pathological characteristics. All PTC and lymph node metastases were confirmed by postoperative pathology.

The ultrasound physicians' diagnostic accuracy was evaluated by comparing ultrasound-reported lymph node status with pathological findings. Based on ACR TI-RADS guidelines (Tessler et al., 2017), lymph nodes with suspicious features—such as round shape, absent echogenic hilum, microcalcifications, cystic changes, hyperechogenicity, or peripheral blood flow—were classified as LLNM.

Ultrasound imaging was performed using various pieces of equipment, including Philips, GE Healthcare, Siemens, and Toshiba systems. Two experienced ultrasound physicians with over 10 years of expertise conducted the examinations. For each case, a high-resolution axial grayscale image of the lesion's longest axis was selected and stored in Digital Imaging and Communications in Medicine format for analysis.

2.3 Ultrasound image preprocessing and region of interest segmentation

Raw ultrasound images underwent standardized preprocessing through a uniform pipeline. Images were resampled to isotropic voxel dimensions (1 mm×1 mm×1 mm) and standardized with 25 grayscale intensity discretization bins using 3D-Slicer software (version 4.10.2). The largest tumor cross-section was selected for each patient, with grayscale values normalized to [-1, 1]. Regions of interest (ROI) were cropped and resized to 224 × 224 pixels using nearest-neighbor interpolation. ROIs were manually delineated using 3D-Slicer by two expert ultrasound physicians with extensive thyroid imaging experience. A senior physician (>10 years of experience) performed the primary segmentation for radiomics analysis, blinded to clinical data.

To ensure reliability, intra- and inter-operator variability were assessed. Intra-operator consistency was measured by repeating feature extraction after three days, while inter-operator consistency was evaluated by comparing the results with a second ultrasound physician (>5 years of experience). Intraclass Correlation Coefficient (ICC) values ranged from 0.846 to 0.949 for intra-operator and 0.893 to 0.944 for inter-operator reproducibility.

2.4 Development of the radiomics-based SVM model

Radiomic features were extracted using the pyradiomics library (version 3.0.0), a widely used Python-based tool. A total of 846 features were extracted from the ultrasound images by the delineated ROI. A volcano plot was used to identify radiomic features with significant discriminatory power. Least Absolute Shrinkage and Selection Operator (LASSO) regression was performed to refine feature selection, and a support vector machine (SVM) model was constructed with optimized parameters using 10-fold cross-validation. To address class imbalance, we employed the Synthetic Minority Over-sampling Technique (SMOTE) for the training data during model development. Additionally, class weights were adjusted to be inversely proportional to class frequencies to more heavily penalize the misclassification of minority class samples. All analyses were performed using the "scikit-learn" package (version 1.6.0).

2.5 Development of the deep learning model

A convolutional neural network (CNN) was developed to predict LLNM using ultrasound images. Patient images from Changzhou First People's Hospital were split into training and validation sets (2:1). ROIs were resized to 224 × 224 pixels. Transfer learning with pre-trained ImageNet weights was conducted to improve generalization, and TresNet, the best-performing backbone (Table 3), was used. To mitigate the effects of class imbalance, a weighted cross-entropy loss function was implemented, where the loss weights were inversely proportional to class frequencies (weight for LLNM-positive class = 9.9, weight for LLNM-negative class = 1.0). This approach ensured that the model paid greater attention to correctly classifying the minority LLNM-positive cases during training. To reduce overfitting and enhance the representation of the minority class through data augmentation, the model employed cross-entropy loss, the Adam optimizer (initial learning rate 0.003, decaying

every 100 epochs over 500), a batch size of 32, and image augmentation (random cropping, flipping, rotations [-20°, 20°]).

2.6 Development of the integrated prediction model

The integrated prediction model combines three branches: deep learning branch, radiomics-based SVM branch, and clinical-ultrasound feature branch (Figure 4). The radiomics-based SVM branch captures fine-grained texture and structural patterns, while the deep learning branch, utilizing the TresNet model with frozen parameters, extracts predictive scores from imaging data. Logistic regression integrates outputs from the SVM model, deep learning model and clinical-ultrasound features using 10-fold cross-validation for optimization. The model performance was compared to individual branches and the diagnostic accuracy of ultrasound physicians across independent test sets.

Given the significant class imbalance in our dataset (9.6% LLNM-positive overall), a comprehensive multi-level strategy was implemented across all model components. For the radiomics-based SVM branch, SMOTE was applied to generate synthetic samples of the minority class, combined with class weight adjustment. For the deep learning branch, weighted cross-entropy loss was employed with weights calculated as the inverse of class frequencies, ensuring that the model learned discriminative features for both classes effectively. For the logistic regression integration, class weights were incorporated into the final model fitting. These strategies collectively ensured that our integrated model maintained high sensitivity for detecting LLNM while preserving specificity, which is a critical asset for a clinical screening tool where missed metastases can have significant clinical consequences.

2.7 Evaluation of the clinical utility of the integrated prediction model

In test set B, three groups of ultrasound physicians with varying levels of experience—junior (1–5 years), medium-level (5–10 years), and senior (>10 years)— preoperatively evaluated the cervical lymph node status of PTC patients. All physicians independently reviewed the images without access to any prior diagnostic information. Confidence levels for identifying LLNM were recorded using a 5-point scale: 1 = not confident, 2 = slightly confident, 3 = confident, 4 = very confident, and 5 = almost certain. The results of the integrated model were disclosed to the physicians, and changes in both their diagnostic outcomes and confidence levels were re-evaluated using the same scale.

2.8 Visualization and clinical interpretability of model performance

To enhance the deep learning model's interpretability, t-distributed stochastic neighbor embedding (t-SNE) and gradient-weighted class activation mapping (Grad-CAM) were applied, with Grad-CAM generating saliency maps to highlight key regions influencing predictions. SHapley Additive exPlanation (SHAP) plots were drawn to quantify variable contributions to the integrated model, while Sankey plots visualized diagnostic performance, including true positive (TP), true negative (TN), false positive (FP), and false negative (FN) cases, and confidence level improvements, demonstrating the reliability and clinical utility of the proposed model.

2.9 Statistical analysis

Statistical analysis was performed using SPSS 25.0, R 3.5.3, and Python 3.12.0. Categorical variables were analyzed using the chi-squared or Fisher's exact test, while continuous variables were compared with the t-test or Mann-Whitney U test. Model performance was evaluated with receiver operating characteristic (ROC) curve analysis and the area under the curve (AUC), and DeLong's test. Metrics such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, Kappa value, and loss were reported. A P-value < 0.05 was considered statistically significant.

3 Results

3.1 Baseline characteristics

Figure 1 illustrates the patient recruitment workflow, while Table 1 provides a comprehensive overview of the clinical characteristics of all participants. Collectively, these datasets comprise a total of 1,566 patients diagnosed with PTC, including 151 patients with LLNM and 1,415 patients without LLNM.

The training cohort included 878 patients (213 males, 665 females, mean age (43.3±11.6) years) with 81 LLNM-positive cases and 797 LLNM-negative cases, and the validation cohort included 220 patients (60 males, 160 females, mean age (44.8±11.5) years) with 18 LLNM-positive cases and 202 LLNM-negative cases, with no significant differences between them ($P>0.05$). Test set A consisted of 348 patients (176 males, 172 females, mean age (38.5±11.7) years) with 35 LLNM-positive cases and 313 LLNM-negative cases, and test set B had 120 patients (37 males, 83 females, mean age (42.6±11.0) years) with 17 LLNM-positive cases and 103 LLNM-negative cases.

3.2 Collinearity and logistic regression analysis of risk factors for LLNM

To explore the relationship between LLNM and clinical/ultrasound features, collinearity analysis and logistic regression were conducted (Table 2). Variables like nodular composition and echogenicity, with high variance inflation factor values indicating multicollinearity, were excluded to stabilize the regression model.

Logistic regression identified sex, tumor size, location, and serum thyroglobulin antibody (TG-Ab) levels as independent risk factors for LLNM. Female sex and tumors in the middle, lower, or isthmus regions were associated with lower LLNM risk. Tumor size >2 cm, especially >4 cm, and elevated TG-Ab levels significantly increased the LLNM risk.

3.3 SVM-driven ultrasound radiomics model for LLNM prediction

The open-source Python package “pyradiomics” (version 3.1.0) was employed to extract an extensive array of radiomic features from the ROIs, culminating in a total of 846 features. These features spanned six major categories, including 162 first-order features, 216 gray-level co-occurrence matrix features, 126 gray-level dependence matrix features, 144 gray-level run length matrix features, 144 gray-level size zone matrix features, and 14 shape features.

Through initial analysis, these features were screened for significant differential expression ($P < 0.05$, $|\log2(\text{fold change})| > 1$), identifying 24 features associated with LLNM (Figs. 2a and 2b). The volcano plot (Fig 2A) highlights significantly differential features, while the heatmap (Fig 2B) shows distinct expression patterns, with LLNM cases exhibiting higher expression of specific features compared to non-LLNM cases.

LASSO regression was then applied to reduce redundancy, optimizing feature selection by balancing model simplicity and predictive accuracy (Figs. 2C and 2D). This process refined the 24 features into 6 key predictors, which were subsequently used to construct the SVM model for LLNM prediction. Details of the selected features are provided in Table S1.

3.4 Performance and visualization of deep learning models in LLNM prediction

Table 3 compares the performance of six deep learning models for LLNM prediction. TresNet outperformed others with the highest AUC (0.971 training, 0.877 validation), accuracy (97.0%, 94.2%), and specificity (98.8%, 98.2%), demonstrating strong robustness and generalizability, hence it was identified as the most suitable model (see the detailed TresNet structure in Table S2).

Grad-CAM was further employed to identify areas of interest for TresNet. Figs. 3A and 3B illustrate strong activation in regions associated with LLNM, consistent with diagnostic outcomes, while Figs. 3C and 3D show minimal activation for cases without LLNM. This highlights the ability of TresNet to identify diagnostically relevant regions, supporting its clinical utility.

Additionally, t-SNE visualization (Figs. 3E and 3F) demonstrates clear clustering of metastatic (red) and

non-metastatic (green) lesions in both the training and validation sets, further validating the model's ability to learn distinct features and generalize effectively across datasets.

3.5 Development and integration of the multimodal model in LLNM prediction

The integrated prediction model combines clinical and ultrasound features, radiomics, and deep learning (Fig 4A). Flow 1 uses collinearity detection and logistic regression for clinical and ultrasound features; Flow 2 applies a volcano plot, Lasso regression, and SVM for radiomics; and Flow 3 utilizes TresNet to generate deep learning-based scores. Outputs are integrated via logistic regression for comprehensive LLNM risk stratification.

The SHAP algorithm was used to evaluate feature importance within the model (Fig 4B). TresNet-based deep learning scores had the highest predictive value, followed by SVM-derived radiomics scores. Key clinical variables, including lesion size, location, sex, and TGAb levels, also had significant contributions. The SHAP summary plot visualizes the influence of feature values, with higher values (red) driving positive predictions and lower values (blue) reducing their effect. This integration of clinical, radiomics, and deep learning features enhances the predictive performance and interpretability of the model.

3.6 Visualization and performance comparison among different models

The integrated prediction model showed superior LLNM stratification performance across all datasets (Figs. 5A-C, Table 4). It achieved the highest AUC in the training (0.984), validation (0.943), and test set A (0.951) datasets, with strong sensitivity (88.5%, 76.7%, 85.9%) and specificity (97.6%, 95.5%, 95.6%). The model consistently outperformed TresNet, SVM, and logistic regression.

The Sankey plots (Fig. 5D-F) show that the integrated model achieved higher TP and TN proportions with fewer FP and FN cases compared to TresNet, SVM, and logistic regression, demonstrating its robustness and precision across all datasets.

3.7 Evaluation of the integrated model in clinical practice

The impact of the integrated model on clinical practice was evaluated using Test Set B (Figs. 6A and 6B). Without the model, accuracies for junior, medium and senior physicians were 55.0%, 55.8% and 66.7%. After integration, accuracies increased to 61.4%, 63.2% and 75.1%, with significant improvements in sensitivity and specificity (Table 4). Fig 6C shows the Sankey plot of TN, TP, FP, and FN distributions before and after model integration. The model exhibits improved accuracy across all groups by reducing FP and FN while increasing TP and TN, with the greatest benefit observed for junior and medium-level physicians. Fig 6D illustrates the confidence level improvements after integrating the model. High-confidence levels (conf-4 and conf-5) increased significantly for all physicians: junior (40.27% to 59.77%), medium-level (50.00% to 59.77%), and senior (72.27% to 84.51%), with corresponding reductions in low-confidence levels (conf-1 and conf-2).

4 Discussion

This study presents an AI-driven multimodal model for preoperative LLNM prediction in PTC patients. By combining radiomics, deep learning features, and a logistic regression of clinical and ultrasound data, this model achieves superior predictive performance, leveraging thyroid nodule characteristics to advance precision oncology. Our study addresses the inherent class imbalance challenge (9.6% LLNM prevalence) through a comprehensive multi-tiered approach. We employed SMOTE for synthetic minority class augmentation in the radiomics model, weighted loss functions inversely proportional to class frequencies in the deep learning model, and class-weighted logistic regression for final integration. Utilizing these strategies allowed achieving the observed balanced performance metrics, particularly the high sensitivity (76.7%-88.5%) alongside excellent specificity (95.5%-97.6%), which are essential for a screening tool where missing metastases could lead to

inadequate treatment and disease progression. The low LLNM prevalence in our cohort significantly influences the interpretation of our performance metrics. The high NPV (96.8%-98.8%) is largely attributable to the low disease occurrence, as NPV increases mathematically when disease prevalence decreases. Conversely, the moderate PPV (75.9%-78.4%) reflects the diagnostic challenge inherent in low-prevalence conditions. Clinicians should recognize that in higher-prevalence settings (e.g., tertiary referral centers with high-risk populations), NPV would decrease while PPV would increase. In contrast, the sensitivity and specificity of the model are prevalence-independent metrics, providing more generalizable and reliable performance indicators across different clinical contexts. Despite the class imbalance, our multi-tiered approach ensures that the model provides balanced diagnostic performance. We recommend that clinicians interpret model outputs by considering both the predicted probability and prevalence-independent metrics (sensitivity and specificity), rather than relying solely on predictive values that are influenced by local disease prevalence.

Previous studies have identified age, sex, tumor size, multifocality, serum TG-Ab levels, and ultrasound features such as shape, echogenicity, margin irregularity, and extrathyroidal extension as LLNM risk factors (Feng et al., 2020; Xue et al., 2020; Shao et al., 2023). Meanwhile, our study highlights tumor size, location, sex, and serum TG-Ab levels as key predictors, providing a more focused framework for LLNM risk assessment. Logistic regression enhances the model's simplicity, interpretability and ability to integrate clinical and ultrasound features, addressing the limitations of standalone AI models that may overlook non-image-based information (Christodoulou et al., 2019).

Previous studies have leveraged handcrafted radiomics to extract features like texture, shape, and grayscale patterns for assessing metastasis (Gong et al., 2022; Abbaspour et al., 2024). While effective in CT and MRI imaging, applying this approach in ultrasound faces challenges such as noise, operator dependency, and equipment variability (Gang et al., 2021). This study derived radiomics features from the primary nodule, indirectly reflecting metastatic risk by linking tumor aggressiveness to lymphatic spread. Key radiomics features, including tumor axis length, size uniformity, and gray-level emphasis, were identified as significant predictors of LLNM (Table S1). These features reflect the key morphological and textural characteristics indicative of aggressive tumor behavior, further underscoring their relevance in predicting LLNM. Although radiomics features are interpretable and align with clinical reasoning, their contribution in ultrasound-based models is modest compared to deep learning, likely due to feature overlap and ultrasound limitations.

Deep learning, particularly CNN, provides a data-driven approach to image analysis, extracting high-dimensional features directly from raw images (Pacal et al., 2020). In this study, the CNN characterized thyroid nodules rather than directly analyzing lymph nodes, leveraging the link between primary tumor features and metastatic behavior. This approach overcomes ultrasound variability challenges, such as imaging quality and operator dependency (Stib et al., 2020; Esce et al., 2021). The TresNet model achieved high sensitivity, effectively identifying aggressive nodules. In the validation set, it reached an AUC of 0.877 and a sensitivity of 72.4%, outperforming traditional machine learning models. TresNet's ability to capture complex features like tumor location and texture enhanced LLNM prediction accuracy. Compared to handcrafted radiomics, deep learning performed better by addressing variability and capturing non-linear relationships in high-dimensional data (Wang et al., 2024). CNN identified features like microcalcifications, hypoechoic areas, and tumor-capsule interactions associated with LLNM risk. Grad-CAM visualization enhanced explainability by highlighting key regions, aligning with radiologists' focus on tumor margins and extrathyroidal extension, increasing model trust (Fuentes et al., 2024).

Integrating logistic regression-based clinical and ultrasound features, radiomics, and deep learning synergistically improved diagnostic performance. Logistic regression ensured interpretability, radiomics provided high-specificity morphological features, and deep learning captured high-sensitivity hierarchical patterns (Zhang et al., 2020; Wang et al., 2022). The multimodal model achieved AUCs of 0.943 (validation) and 0.951 (test set A), with sensitivities of 76.7% and 85.9% and specificities of 91.3% and 89.4%. These results substantially exceed conventional imaging performance for lateral CLNM reported in meta-analyses. Zhao et al. (Zhao and Li, 2019) showed that ultrasound alone achieved a pooled sensitivity of 0.70 (95% CI:

0.68-0.72) and specificity of 0.84 (95% CI: 0.82-0.85) with an AUC of 0.88 for lateral CLNM. Albuck et al. (2023) reported that combined CT and ultrasound imaging achieved a sensitivity of 84.5% and specificity of 88.1% with an AUC of 0.919 for lateral CLNM detection. Our integrated model's sensitivity (76.7%-85.9%) is comparable to combined CT+US imaging while achieving superior AUC (0.943-0.951) and specificity (91.3%-89.4%), offering a non-invasive alternative that analyzes primary tumor characteristics rather than requiring dual imaging modalities. This approach significantly enhances LLNM prediction accuracy and addresses challenges like small lymph node size and deep anatomical location, which often reduce ultrasound sensitivity.

The multimodal model developed in this study enhances LLNM detection while reducing unnecessary interventions, such as FNA for benign lymph nodes, especially for less experienced practitioners. Deep learning visualization tools, like class activation maps, improve interpretability by highlighting key diagnostic regions, aiding clinicians in challenging cases. Furthermore, the model boosts the diagnostic confidence of clinicians, supporting more accurate and informed decisions. This framework aligns with precision medicine, offering tailored strategies for LLNM risk assessment and management.

This study has limitations. First, the limited number of centers may restrict generalizability. Large-scale, multicenter studies with diverse populations are needed to confirm robustness. Second, using primary thyroid nodule characteristics as a surrogate for LLNM risk introduces variability, as not all aggressive nodules metastasize. Future research could directly assess lymph nodes using advanced imaging modalities like contrast-enhanced ultrasound or elastography to improve accuracy (Choi et al., 2020; Wang et al., 2021). Third, the relatively low LLNM prevalence affects metric interpretation, particularly predictive values (PPV and NPV), which vary with local disease prevalence, whereas sensitivity and specificity remain stable across different prevalence settings. External validation in populations with varying LLNM prevalence can establish the generalizability of the model across diverse clinical contexts. Finally, while logistic regression improved interpretability, the high-dimensional CNN features remain difficult to explain. Attention-based AI frameworks could enhance clinical understanding and acceptance (Kim et al., 2020; Nasarian et al., 2023).

In conclusion, this study highlights the potential of an AI-based multimodal model for LLNM assessment in PTC patients. By integrating clinical, ultrasound, radiomics, and deep learning features, the proposed model offers a robust, non-invasive tool for personalized risk stratification, improving diagnostic confidence, reducing missed metastases, and minimizing unnecessary surgeries.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Acknowledgments

The authors acknowledge Lei Qin for assistance with English language editing. This research was supported by the Changzhou Science and Technology Bureau under the Angel White Fund Project (CJ20244009) and the Changzhou Talent Program for Young Scientific Researchers (Grant No. Changzhou Science Association [2023] No. 52).

Author contributions

Jia-Wei FENG performed the conceptualization, methodology, software development, formal analysis, and wrote the original draft of the manuscript. Lu ZHANG conducted the investigation, data curation, validation, and contributed to writing the original draft. Yu-Xin YANG performed software development, data curation, and visualization. Shui-Qing LIU provided resources and conducted investigation work. An-Cheng QIN performed validation, formal analysis, and data curation. Yong JIANG contributed to the conceptualization, supervision, project administration, funding acquisition, and writing, reviewing, and editing of the manuscript. All authors read and approved the final manuscript and, therefore, had full access to all the data in the study and take responsibility for the integrity and security of the data.

Compliance with ethics guidelines

Jia-Wei Feng , Lu ZHANG, Yu-Xin YANG, Shui-Qing LIU, An-Cheng QIN, Yong JIANG declare that they have no

conflict of interest. This study was approved by the hospital ethics committees of Changzhou First People's Hospital, Shanghai Ruijin Hospital, and Suzhou Municipal Hospital, and was conducted in accordance with the revised guidelines of the Declaration of Helsinki (2013). The requirement for informed consent was waived for this retrospective analysis at all three institutions due to the observational nature of the study. This article does not contain any studies with human or animal subjects performed by any of the authors.

Table 1. Clinical and ultrasound characteristics of the training, validation, and test sets

Characteristics	Training set (n=878)	Validation set (n=220)	Test set A (n=348)	Test set B (n=120)	*P* value*
Sex					
Male	213 (24.3%)	60 (27.3%)	176 (50.6%)	37 (30.8%)	
Female	665 (75.7%)	160 (72.7%)	172 (49.4%)	83 (69.2%)	0.402
Age, Mean±SD, years	43.3±11.6	44.8±11.5	38.5±11.7	42.6±11.0	0.639
≥55	147 (16.7%)	42 (19.1%)	40 (11.5%)	19 (15.8%)	
<55	731 (83.3%)	178 (80.9%)	308 (88.5%)	101 (84.2%)	0.468
BMI					
Marasmus	41 (4.7%)	10 (4.5%)	30 (8.6%)	8 (6.7%)	
Normal	498 (56.7%)	128 (58.2%)	224 (64.4%)	68 (56.7%)	
Overweight	273 (31.1%)	70 (31.8%)	74 (21.3%)	38 (31.7%)	
Obesity	66 (7.5%)	12 (5.5%)	20 (5.7%)	6 (5.0%)	0.764
BRAF V600E mutation					
Negative	95 (10.8%)	20 (9.1%)	42 (12.1%)	11 (9.2%)	
Positive	783 (89.2%)	200 (90.9%)	306 (87.9%)	109 (90.8%)	0.531
CLT					
Absence	672 (76.5%)	172 (78.2%)	282 (81.0%)	90 (75.0%)	
Presence	206 (23.5%)	48 (21.8%)	66 (19.0%)	30 (25.0%)	0.669
Tumor size, Mean±SD, cm	1.05±0.69	1.09±0.84	1.18±0.77	1.07±0.70	
≤1	582 (66.3%)	149 (67.7%)	212 (60.9%)	77 (64.2%)	0.326
>1 to ≤2	224 (25.5%)	50 (22.7%)	100 (28.7%)	37 (30.8%)	
>2 to ≤4	69 (7.9%)	16 (7.3%)	32 (9.2%)	4 (3.3%)	
>4	3 (7.9%)	5 (2.3%)	4 (1.1%)	2 (1.7%)	0.057
The number of foci					
1	645 (73.5%)	172 (78.2%)	200 (57.5%)	88 (73.3%)	
2	172 (19.6%)	38 (17.3%)	84 (24.1%)	29 (24.2%)	
3 or more	61 (6.9%)	10 (4.5%)	64 (18.4%)	3 (2.5%)	0.142
Location					
Upper	195 (22.2%)	49 (22.3%)	94 (27.0%)	21 (17.5%)	
Middle	461 (52.5%)	111 (50.5%)	122 (35.1%)	70 (58.3%)	
Lower	171 (19.5%)	47 (21.4%)	78 (22.4%)	21 (17.5%)	
Isthmus	51 (5.8%)	13 (5.9%)	54 (15.5%)	8 (6.7%)	0.927
Margin					
Smooth	625 (71.2%)	162 (73.6%)	252 (72.4%)	75 (62.5%)	
Lobulated or irregular	226 (25.7%)	49 (22.3%)	58 (16.7%)	41 (34.2%)	
ETE	27 (3.1%)	9 (4.1%)	38 (10.9%)	4 (3.3%)	0.462
Nodular composition					
Cystic or spongiform	2 (0.2%)	2 (0.9%)	0 (0.0%)	0 (0.0%)	
Mixed cystic and solid	1 (0.1%)	218 (99.1%)	0 (0.0%)	0 (0.0%)	
Solid	875 (99.7%)	1 (0.5%)	348 (100.0%)	120 (100.0%)	0.286
Echogenicity					
Anechoic	1 (0.1%)	1 (0.5%)	0 (0.0%)	0 (0.0%)	
Hyperechoic or isoechoic	9 (1.0%)	219 (99.5%)	10 (2.9%)	1 (0.8%)	
Hypoechoic	858 (99.7%)	0 (0.0%)	334 (96.0%)	118 (98.3%)	

Markedly hypoechoic	10 (1.1%)	0 (0.0%)	4 (1.1%)	1 (0.8%)	0.115
A/T					
< 1	275 (31.3%)	59 (26.8%)	128 (36.8%)	38 (31.7%)	
≥ 1	603 (68.7%)	161 (73.2%)	220 (63.2%)	82 (68.3%)	0.224
Echogenic foci					
None	346 (39.4%)	78 (35.5%)	90 (25.9%)	38 (31.7%)	
Macrocalcifications	45 (5.1%)	8 (3.6%)	12 (3.4%)	4 (3.3%)	
Peripheral calcifications	8 (0.9%)	2 (0.9%)	8 (2.3%)	3 (2.5%)	
Microcalcifications	479 (54.6%)	132 (60.0%)	238 (68.4%)	75 (62.5%)	0.483
LLNM					
Negative	797 (90.8%)	202 (91.8%)	313 (89.9%)	103 (85.8%)	
Positive	81 (9.2%)	18 (8.2%)	35 (10.1%)	17 (14.2%)	0.629
Serum TPO-Ab, Mean±SD, IU/ml	40.12±90.43	24.25±45.68	42.21±93.31	53.28±105.70	0.429
Serum TG, Mean±SD, ng/ml	32.96±78.03	33.04±79.63	28.44±60.43	31.58±65.19	0.153
Serum TG-Ab, Mean±SD, IU/ml	128.78±464.30	78.62±198.18	146.70±541.85	107.05±349.56	0.137

SD standard deviation, *BMI* body mass index, *CLT* chronic lymphocytic thyroiditis, *ETE* extrathyroidal extension, *A/T* Anteroposterior to Transverse ratio, *LLNM* lateral lymph node metastasis, *TG* Thyroglobulin, *TG-Ab* anti-thyroglobulin antibodies, *TPO-Ab* thyroid peroxidase antibody

***** The *P*-value reflects the comparison between the training and validation sets to demonstrate baseline comparability of the development cohorts. Test sets A and B are independent external validation cohorts and were not included in this statistical comparison.

Table 2. Collinearity analysis and logistic regression of clinical and ultrasound characteristics associated with LLNM

Characteristics	VIF	β	Odds Ratio (95% CI)	*P* value
Sex				
Male			Ref	
Female	5.59	-0.609	0.543 (0.313–0.944)	0.030
Age				
≥55			Ref	
<55	2.31	-0.415	0.660 (0.232–1.874)	0.435
BMI				
Marasmus			Ref	
Normal		-0.132	0.876 (0.294–2.615)	0.813
Overweight		-0.634	0.530 (0.164–1.719)	0.291
Obesity	5.511	-1.379	0.252 (0.052–1.230)	0.088
BRAF V600E mutation				
Negative			Ref	
Positive	6.069	-0.334	0.716 (0.325–1.578)	0.408
CLT				
Absence			Ref	
Presence	2.035	0.230	1.259 (0.632–2.506)	0.513
Tumor size, cm				
≤1			Ref	
>1 to ≤2		0.644	1.905 (0.525–6.905)	0.137
>2 to ≤4		2.976	19.608 (6.123–62.789)	<0.001
>4	4.347	5.741	310.600 (37.037–3001.000)	<0.001
The number of foci				
1			Ref	
2		-0.200	0.818 (0.067–10.026)	0.876
3 or more	5.731	0.293	1.340 (0.099–18.077)	0.825
Location				
Upper			Ref	
Middle		-0.755	0.470 (0.264–0.837)	0.010
Lower		-1.048	0.351 (0.159–0.773)	0.009
Isthmus	4.098	-2.239	0.107 (0.024–0.466)	0.003
Margin				
Smooth			Ref	
Lobulated or irregular		-0.093	0.911 (0.503–1.652)	0.759
ETE	1.479	1.098	2.998 (0.974–8.371)	0.056
Nodular composition				
Cystic or spongiform				
Mixed cystic and solid				
Solid	134.787	N/A	N/A	N/A
Echogenicity				
Anechoic				
Hyperechoic or isoechoic				
Hypoechoic				
Markedly hypoechoic	118.690	N/A	N/A	N/A
A/T				
< 1			Ref	
≥ 1	3.782	0.055	1.057 (0.606–1.843)	0.845
Echogenic foci				
None				

Macrocalcifications				
Peripheral calcifications				
Microcalcifications	12.865	N/A	N/A	N/A
Serum TPO-Ab, IU/ml	1.743	0.001	1.001 (0.997–1.003)	0.940
Serum TG, ng/ml	1.360	0.006	1.000 (0.998–1.003)	0.616
Serum TG-Ab, IU/ml	1.267	0.005	1.012 (1.000–1.029)	0.012

VIF Variance inflation factor, *BMI* body mass index, *CLT* chronic lymphocytic thyroiditis, *ETE* extrathyroidal extension, *A/T* Anteroposterior to Transverse ratio, *LLNM* lateral lymph node metastasis, *TG* Thyroglobulin, *TG-Ab* anti-thyroglobulin antibodies, *TPO-Ab* thyroid peroxidase antibody, *CI* confidence interval

Table 3. Performance of the six models in training and validation set

DL models	AUC	AUC 95% CI	ACC	Sensitivity	Specificity	PPV	NPV	Kappa	F1 Score	Loss
Training set										
TresNet	0.971	0.950–0.987	0.970	0.782	0.988	0.871	0.978	0.864	0.876	0.062
ResNet 34	0.881	0.853–0.902	0.890	0.655	0.970	0.851	0.940	0.713	0.750	0.122
ResNet 101	0.871	0.802–0.853	0.922	0.707	0.980	0.811	0.951	0.762	0.752	0.091
ResNet 18	0.621	0.554–0.683	0.881	0.782	0.960	0.659	0.911	0.670	0.720	0.065
Inception 3	0.773	0.625–0.750	0.880	0.548	0.955	0.730	0.922	0.620	0.64	0.035
Validation set										
TresNet	0.877	0.760–0.961	0.942	0.724	0.982	0.733	0.956	0.782	0.817	0.021
ResNet 34	0.782	0.657–0.750	0.862	0.657	0.954	0.551	0.932	0.631	0.677	0.156
ResNet 101	0.750	0.619–0.726	0.851	0.556	0.951	0.655	0.943	0.708	0.658	0.037
ResNet 18	0.730	0.577–0.911	0.930	0.689	0.980	0.498	0.951	0.622	0.741	0.022
Inception 3	0.655	0.522–0.729	0.920	0.540	0.960	0.700	0.930	0.650	0.675	0.023

DL deep learning, *AUC* area under the curve, *CI* confidence interval, *ACC* accuracy, *PPV* positive predictive value, *NPV* negative predictive value

Table 4. Performance comparison of different models and radiologists in prediction of LLNM

Modality	AUC	AUC 95% CI	ACC	Sensitivity	Specificity	PPV	NPV
Training set							
Logistic Regression	0.649	0.590–0.714	0.753	0.423	0.786	0.466	0.931
SVM	0.929	0.843–0.960	0.965	0.680	0.994	0.914	0.969
TresNet	0.971	0.950–0.987	0.970	0.782	0.988	0.871	0.978
Integrated Model	0.984	0.970–0.994	0.967	0.885	0.976	0.784	0.988
Validation set							
Logistic Regression	0.618	0.484–0.752	0.605	0.524	0.613	0.431	0.932
SVM	0.728	0.556–0.873	0.827	0.619	0.847	0.677	0.959
TresNet	0.877	0.760–0.961	0.942	0.724	0.982	0.733	0.956
Integrated Model	0.943	0.874–0.980	0.930	0.767	0.955	0.783	0.968
Test set A							
Logistic Regression	0.638	0.579–0.692	0.638	0.535	0.649	0.431	0.934
SVM	0.773	0.709–0.832	0.660	0.759	0.641	0.694	0.979
TresNet	0.799	0.758–0.835	0.856	0.606	0.881	0.635	0.958
Integrated Model	0.951	0.921–0.974	0.947	0.859	0.956	0.759	0.986
Test set B-before Integrated Model							
Junior	N/A	N/A	0.550	0.546	0.714	0.200	0.923
Medium	N/A	N/A	0.558	0.636	0.760	0.268	0.938
Senior	N/A	N/A	0.667	0.727	0.809	0.320	0.960
Test set B-after Integrated Model							
Junior	N/A	N/A	0.614	0.727	0.782	0.275	0.952
Medium	N/A	N/A	0.632	0.818	0.836	0.353	0.968
Senior	N/A	N/A	0.751	0.909	0.890	0.407	0.986

LLNM lateral lymph node metastasis, *AUC* area under the curve, *CI* confidence interval, *ACC* accuracy, *PPV* positive predictive value, *NPV* negative predictive value

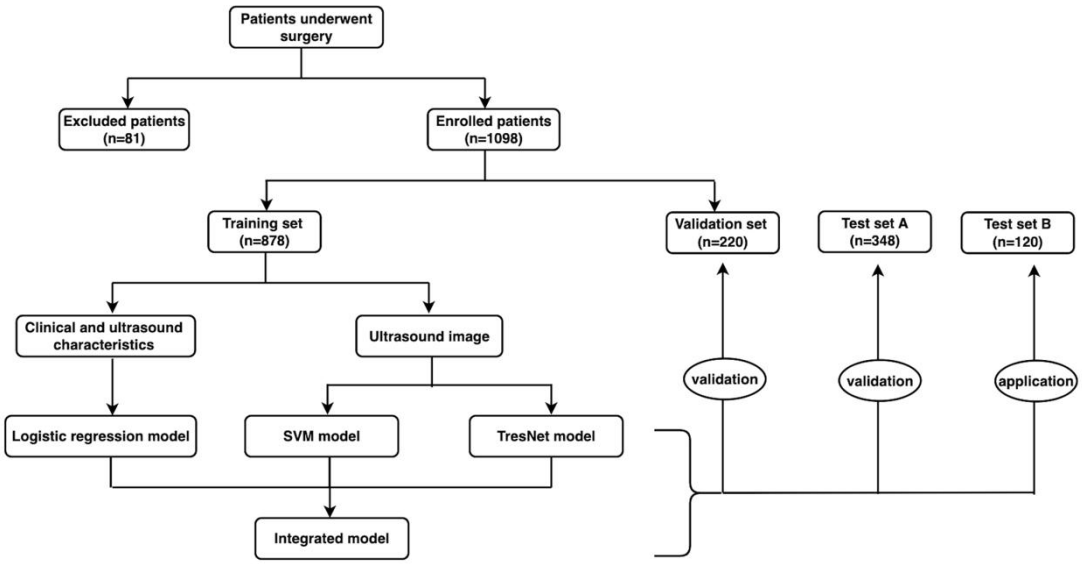

Fig. 1 Flowchart of patient inclusion, dataset splitting, and model development and validation

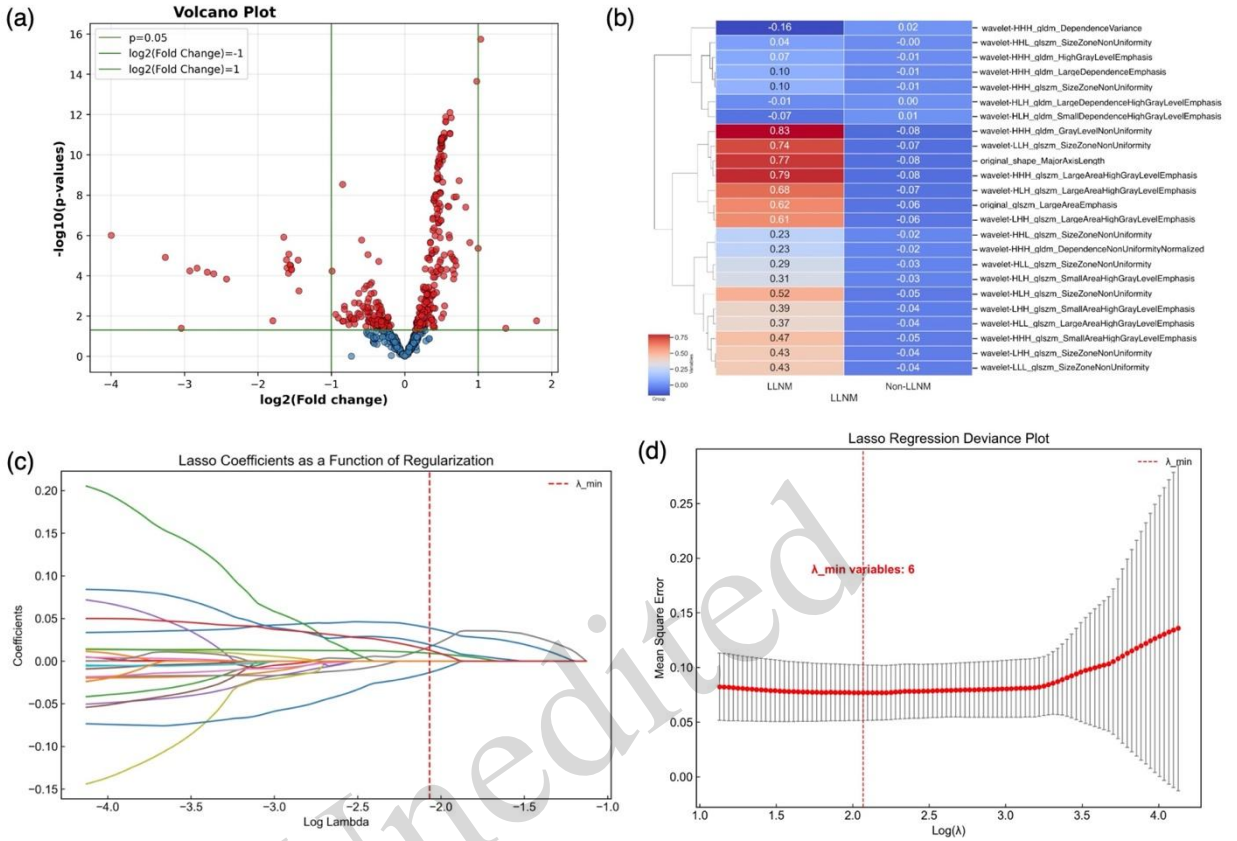

Fig. 2 Ultrasound radiomics feature selection and prediction of LLNM using differential analysis and lasso regression. (a) The volcano plot visualizes the distribution of features based on log2(fold change) and -log10(p-value). The red dots represent significant features the meeting selection criteria. The x-axis denotes log2(fold change) between lateral lymph node metastasis (LLNM) and non-LLNM groups, while the y-axis indicates statistical significance (-log10(p-value)). The green dashed lines mark thresholds for P = 0.05 and |log2(fold change)| = 1. (b) The heatmap displays the expression of 24 selected features. Columns represent samples, rows denote features, and color gradients (blue to red) reflect expression levels. LLNM-specific features show high expression (red) in the LLNM group and low expression (blue) in the non-LLNM group, highlighting distinct patterns between groups. (c) The Lasso coefficient profiles are shown as a function of the regularization parameter (log λ). With increasing λ, coefficients shrink toward zero, removing less relevant features. The red dashed line indicates the optimal λ (λ_min = 0.008562), retaining the most predictive features with minimal mean squared error. (d) The Lasso regression deviance plot illustrates the relationship between log(λ) and mean squared error. The optimal λ (λ_min = 0.008562), marked by the red dashed line, minimizes prediction error, selecting six variables that balance model simplicity and accuracy.

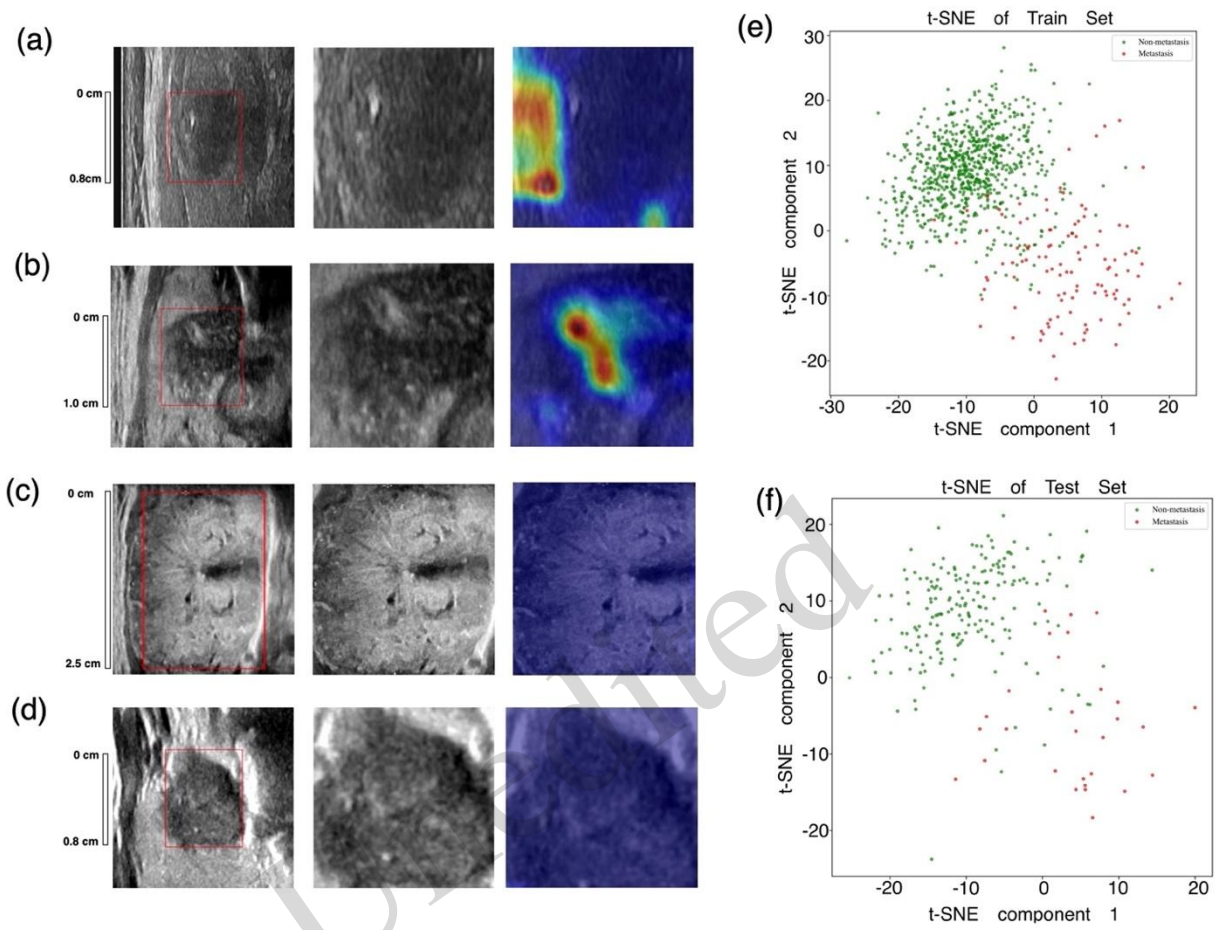

Fig. 3 Gradient-weighted class activation mapping visualization of areas of interest and t-SNE (t-distributed stochastic neighbor embedding) analysis for the TresNet model in lateral lymph node metastasis (LLNM) prediction. (a) A papillary thyroid carcinoma (PTC) patient with LLNM, where the areas of interest highlighted by the model were concentrated in regions associated with metastatic features, indicating strong activation. (b) A PTC patient with LLNM, showing focused activation in metastatic regions, consistent with the detection of LLNM by the model. (c) A PTC case without LLNM, where the areas of interest were diffusely distributed with lower intensity, reflecting the absence of lymph node metastasis. (d) Another PTC case without LLNM, characterized by minimal and diffuse activation patterns, aligning with the model's identification of non-metastatic features. (e) The t-SNE plot of the training set visualizes the distribution of LLNM and non-LLNM cases. Green dots represent non-LLNM cases, while red dots indicate LLNM cases. The clustering pattern demonstrates that the model effectively differentiates between the two groups in the training data. (f) The t-SNE plot of the validation set shows the distribution of LLNM and non-LLNM cases. Similar to the training set, green dots represent non-LLNM cases and red dots indicate LLNM cases. The separation between the two groups validates the ability of the model to distinguish LLNM from non-LLNM in unseen data.

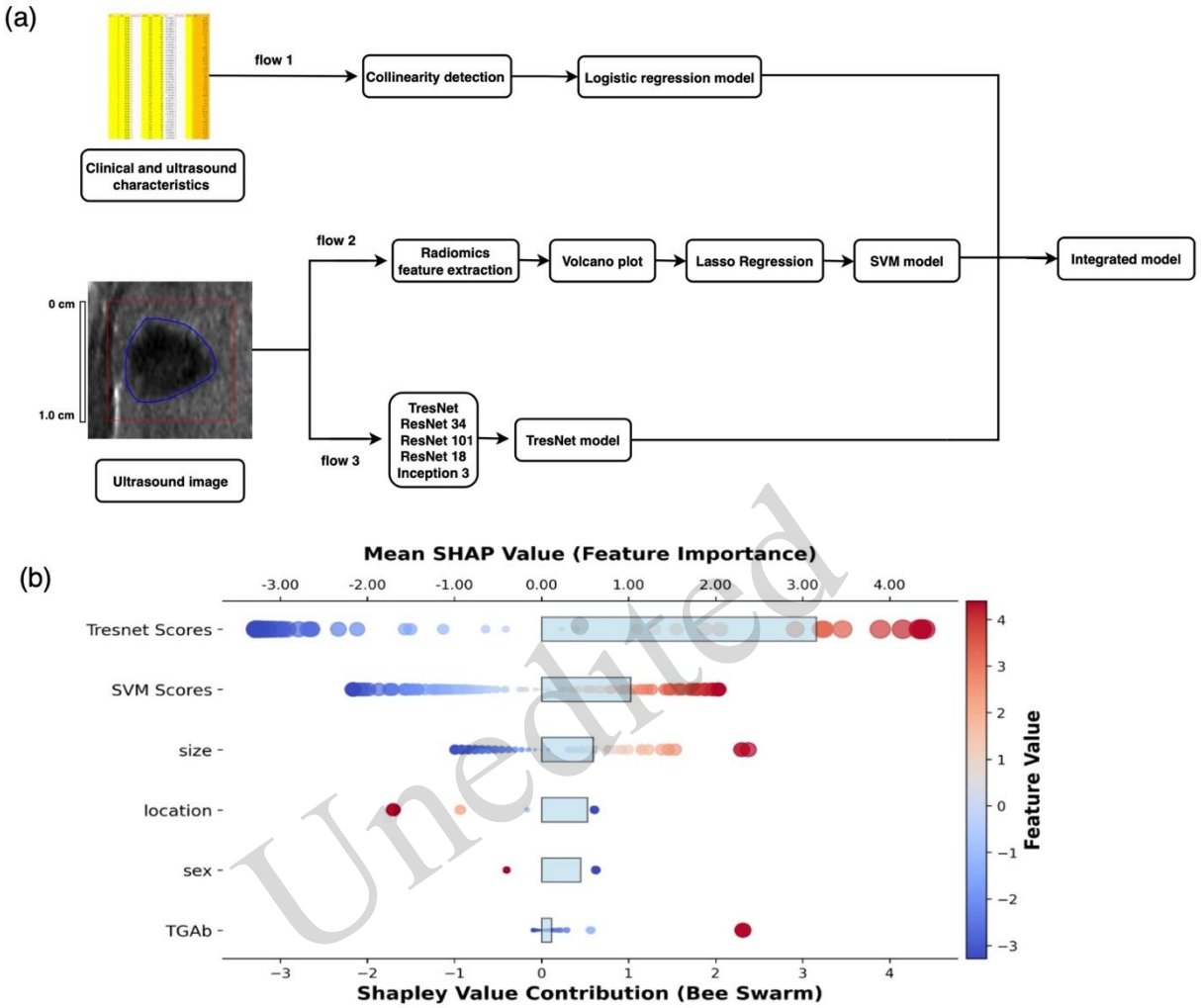

Fig. 4 Flowchart of the integrated model and SHAP (SHapley Additive exPlanations) analysis of feature importance for lateral lymph node metastasis (LLNM) prediction. (a) The flowchart illustrates the construction of the integrated model for LLNM prediction. (b) The SHAP plot combines a bee swarm plot and a bar plot to display feature importance. The bee swarm plot illustrates the distribution of SHAP values for individual features, with dots representing feature contributions across samples. The color gradient (blue: low, red: high) denotes feature values. The bar plot ranks features by their mean absolute SHAP values.

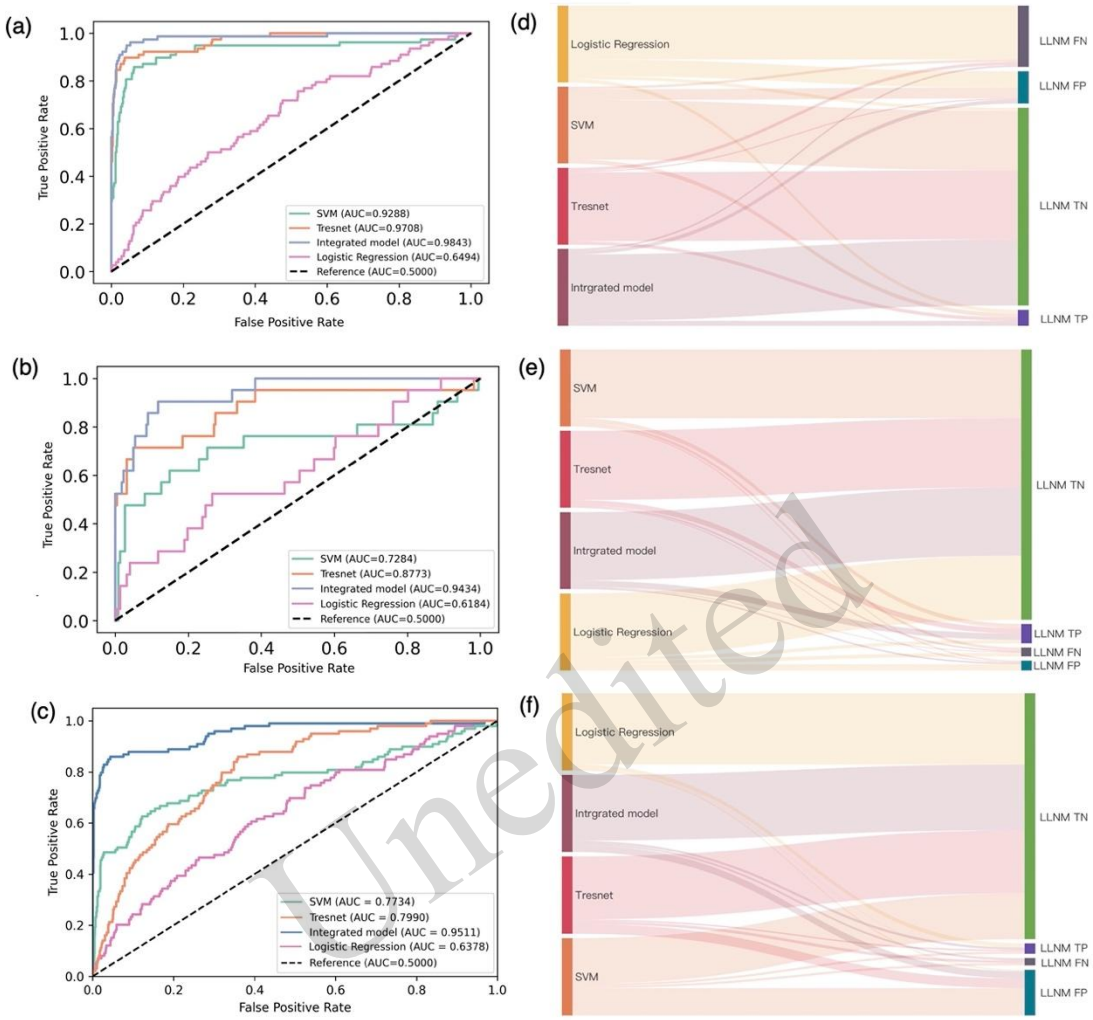

Fig. 5 Performance comparison of integrated and individual models for lateral lymph node metastasis (LLNM) prediction. Receiver operating characteristic curves of four predictive models (integrated model, TresNet, SVM, and logistic regression) for the training dataset (a), validation dataset (b), and test set A (c). Sankey plots showing the distribution of true positive, true negative, false positive, and false negative cases across the four models in the training dataset (d), validation dataset (e), and test set A (f).

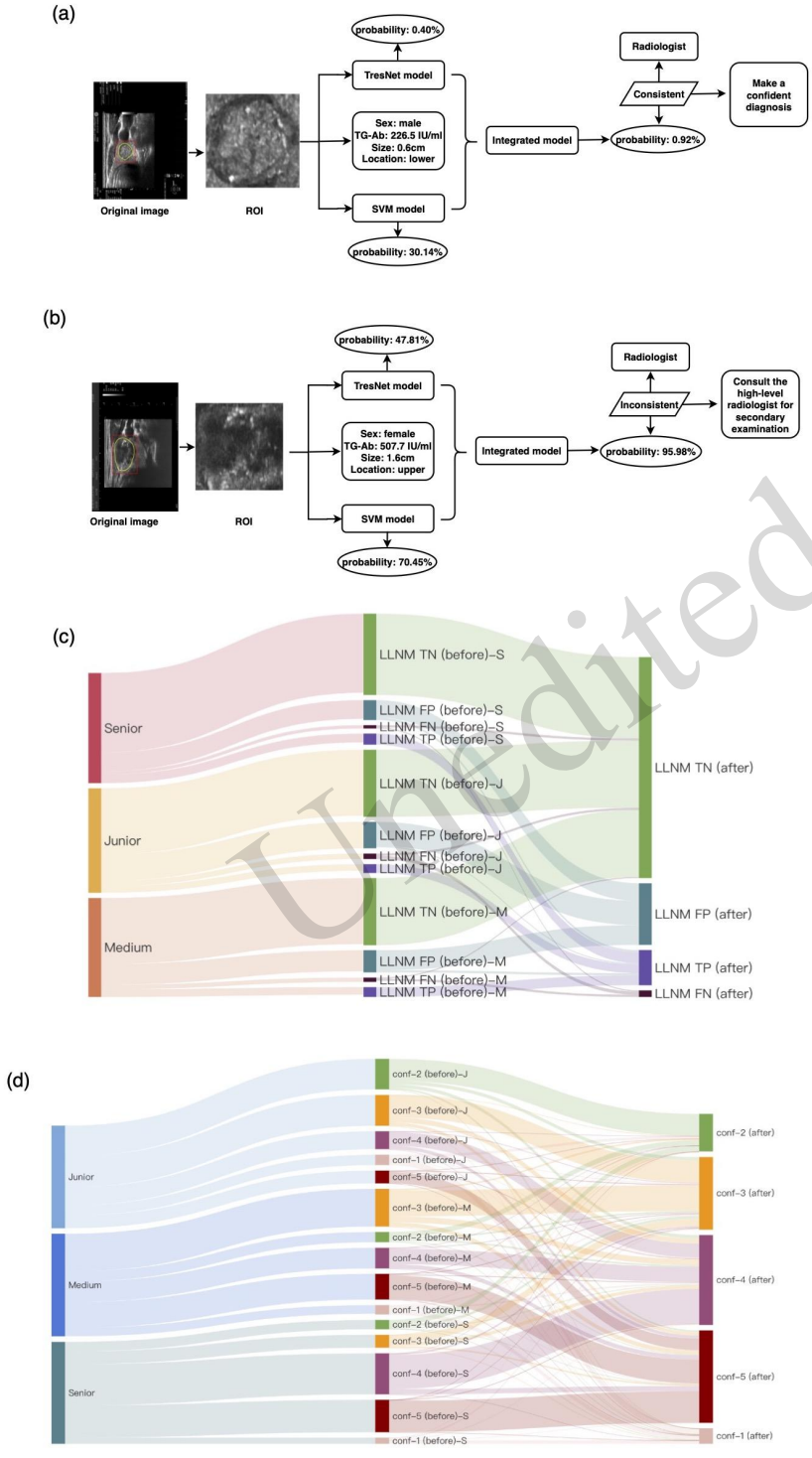

Fig. 6 Performance Evaluation of the Integrated Model in Clinical Practice. (a-b) Representative examples demonstrating how key parameters influence the personalized decision-making process within the integrated model. The examples illustrate scenarios where the integrated model aids in improving diagnostic accuracy and confidence for lateral lymph node metastasis (LLNM) prediction. (c) Sankey plot showing the performance of ultrasound physicians with varying levels of seniority (senior, medium, and junior) in LLNM classification before and after incorporating the integrated model. (d) Sankey plot illustrating the changes in confidence levels (conf-1 to conf-5) for ultrasound physicians before and after incorporating the integrated model.

References

Abbaspour E, Karimzadhagh S, Monsef A, et al., 2024. Application of radiomics for preoperative prediction of lymph node metastasis in colorectal cancer: A systematic review and meta-analysis. Int J Surg, 110(6):3795-3813. https://doi.org/10.1097/js9.0000000000001239

Albuck AL, Issa PP, Hussein M, et al., 2023. A combination of computed tomography scan and ultrasound provides optimal detection of cervical lymph node metastasis in papillary thyroid carcinomas: A systematic review and meta-analysis. Head Neck, 45(9):2173-2184. https://doi.org/10.1002/hed.27451

Baselli G, Codari M, Sardanelli F, 2020. Opening the black box of machine learning in radiology: Can the proximity of annotated cases be a way? Eur Radiol Exp, 4(1):30. https://doi.org/10.1186/s41747-020-00159-0

Brocki L, Chung NC, 2023. Integration of radiomics and tumor biomarkers in interpretable machine learning models. Cancers (Basel), 15(9) https://doi.org/10.3390/cancers15092459

Choi M, Yoon J, Choi M, 2020. Contrast-enhanced ultrasound sonography combined with strain elastography to evaluate mandibular lymph nodes in clinically healthy dogs and those with head and neck tumors. Vet J, 257:105447. https://doi.org/10.1016/j.tvjl.2020.105447

Christodoulou E, Ma J, Collins GS, et al., 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol, 110:12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Chung SR, Baek JH, Choi YJ, et al., 2020. Sonographic assessment of the extent of extrathyroidal extension in thyroid cancer. Korean J Radiol, 21(10):1187-1195. https://doi.org/10.3348/kjr.2019.0983

Esce AR, Redemann JP, Sanchez AC, et al., 2021. Predicting nodal metastases in papillary thyroid carcinoma using artificial intelligence. Am J Surg, 222(5):952-958. https://doi.org/10.1016/j.amjsurg.2021.05.002

Feng JW, Qin AC, Ye J, et al., 2020. Predictive factors for lateral lymph node metastasis and skip metastasis in papillary thyroid carcinoma. Endocr Pathol, 31(1):67-76. https://doi.org/10.1007/s12022-019-09599-w

Feng JW, Ye J, Hong LZ, et al., 2022. Nomograms for the prediction of lateral lymph node metastasis in papillary thyroid carcinoma: Stratification by size. Front Oncol, 12:944414. https://doi.org/10.3389/fonc.2022.944414

Fuentes AM, Milligan K, Wiebe M, et al., 2024. Stratification of tumour cell radiation response and metabolic signatures visualization with raman spectroscopy and explainable convolutional neural network. Analyst, 149(5):1645-1657. https://doi.org/10.1039/d3an01797d

Gang GJ, Deshpande R, Stayman JW, 2021. Standardization of histogram- and gray-level co-occurrence matrices-based radiomics in the presence of blur and noise. Phys Med Biol, 66(7):074004. https://doi.org/10.1088/1361-6560/abeea5

Gong X, Guo Y, Zhu T, et al., 2022. Diagnostic performance of radiomics in predicting axillary lymph node metastasis in breast cancer: A systematic review and meta-analysis. Front Oncol, 12:1046005. https://doi.org/10.3389/fonc.2022.1046005

Haugen BR, Alexander EK, Bible KC, et al., 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. https://doi.org/10.1089/thy.2015.0020

Kim J, Lee S, Hwang E, et al., 2020. Limitations of deep learning attention mechanisms in clinical research: Empirical case study based on the korean diabetic disease setting. J Med Internet Res, 22(12):e18418. https://doi.org/10.2196/18418

Kim SK, Woo JW, Park I, et al., 2016. Influence of body mass index and body surface area on the behavior of papillary thyroid carcinoma. Thyroid, 26(5):657-666. https://doi.org/10.1089/thy.2015.0632

Nasarian E, Alizadehsani R, Acharya U, et al., 2023. Designing interpretable ml system to enhance trust in healthcare: A systematic review to proposed responsible clinician-ai-collaboration framework. Inf Fusion, 108:102412. https://doi.org/10.1016/j.inffus.2024.102412

Pacal I, Karaboga D, Basturk A, et al., 2020. A comprehensive review of deep learning in colon cancer. Comput Biol Med, 126:104003. https://doi.org/10.1016/j.compbiomed.2020.104003

Papadimitroulas P, Brocki L, Christopher Chung N, et al., 2021. Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. Phys Med, 83:108-121. https://doi.org/10.1016/j.ejmp.2021.03.009

Shao L, Wang Z, Dong W, et al., 2023. Risk factors associated with preferential lateral lymph node metastasis in papillary thyroid carcinoma. Cancer Med, 12(22):20670-20676. https://doi.org/10.1002/cam4.6567

Stib MT, Pan I, Merck D, et al., 2020. Thyroid nodule malignancy risk stratification using a convolutional neural network. Ultrasound Q, 36(2):164-172. https://doi.org/10.1097/ruq.0000000000000501

Tessler FN, Middleton WD, Grant EG, 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(5):587-595. https://doi.org/10.1016/j.jacr.2017.01.046

Wang B, Guo Q, Wang JY, et al., 2021. Ultrasound elastography for the evaluation of lymph nodes. Front Oncol, 11:714660. https://doi.org/10.3389/fonc.2021.714660

Wang D, Hu Y, Zhan C, et al., 2022. A nomogram based on radiomics signature and deep-learning signature for preoperative prediction of axillary lymph node metastasis in breast cancer. Front Oncol, 12:940655. https://doi.org/10.3389/fonc.2022.940655

Wang Z, Li X, Zhang H, et al., 2024. Deep learning radiomics based on two-dimensional ultrasound for predicting the efficacy of neoadjuvant chemotherapy in breast cancer. Ultrason Imaging, 46(6):357-366. https://doi.org/10.1177/01617346241276168

Xing Z, Qiu Y, Yang Q, et al., 2020. Thyroid cancer neck lymph nodes metastasis: Meta-analysis of us and ct diagnosis. Eur J Radiol, 129:109103. https://doi.org/10.1016/j.ejrad.2020.109103

Xue S, Han Z, Lu Q, et al., 2020. Clinical and ultrasonic risk factors for lateral lymph node metastasis in papillary thyroid microcarcinoma: A systematic review and meta-analysis. Front Oncol, 10:436. https://doi.org/10.3389/fonc.2020.00436

Yang J, Zhang F, Qiao Y, 2022. Diagnostic accuracy of ultrasound, ct and their combination in detecting cervical lymph node metastasis in patients with papillary thyroid cancer: A systematic review and meta-analysis. BMJ Open, 12(7):e051568. https://doi.org/10.1136/bmjopen-2021-051568

Yasaka K, Akai H, Kunimatsu A, et al., 2018. Deep learning with convolutional neural network in radiology. Jpn J Radiol, 36(4):257-272. https://doi.org/10.1007/s11604-018-0726-3

Yu Y, He Z, Ouyang J, et al., 2021. Magnetic resonance imaging radiomics predicts preoperative axillary lymph node metastasis to support surgical decisions and is associated with tumor microenvironment in invasive breast cancer: A machine learning, multicenter study. EBioMedicine, 69:103460. https://doi.org/10.1016/j.ebiom.2021.103460

Zhang W, Fang M, Dong D, et al., 2020. Development and validation of a ct-based radiomic nomogram for preoperative prediction of early recurrence in advanced gastric cancer. Radiother Oncol, 145:13-20. https://doi.org/10.1016/j.radonc.2019.11.023

Zhao H, Li H, 2019. Meta-analysis of ultrasound for cervical lymph nodes in papillary thyroid cancer: Diagnosis of central and lateral compartment nodal metastases. Eur J Radiol, 112:14-21. https://doi.org/10.1016/j.ejrad.2019.01.006

Supplementary information:
Tables S1 and S2

Table S1. Name of the extracted radiomics feature

Radiomics Features
original_shape_MajorAxisLength
wavelet-HLH_gldm_SmallDependenceHighGrayLevelEmphasis
wavelet-HHH_gldm_GrayLevelNonUniformity
wavelet-HHH_glszm_LargeAreaHighGrayLevelEmphasis
wavelet-HHH_glszm_SmallAreaHighGrayLevelEmphasis
wavelet-LLL_glszm_SizeZoneNonUniformity

Table S2. Structure of the TresNet used in the paper

Name	Output Size	Layer	Parameter Setting	Number of Blocks
Input Layer	224 × 224	Input Image	Grayscale, resized to 224 × 224 pixels	1
Root	112 × 112	Conv + Space-to-Depth	k7, c64, s2, p3	1
	112 × 112	Anti-Aliased Max Pooling	k3, s2, p0, d1	1
Block 1	112 × 112	Residual Block + SE Module	k3, c64, s1, SE included	3
Block 2	56 × 56	Residual Block + SE Module	k3, c128, s2, SE included	4
Block 3	28 × 28	Residual Block + SE Module	k3, c256, s2, SE included	6
Block 4	14 × 14	Residual Block + SE Module	k3, c512, s2, SE included	3
Head Layer	1 × 1	Global Average Pooling + Fully Connected	Output = 2 (binary classification)	1

k: Kernel size; c: Number of output channels; s: Stride of the convolution or pooling operation; p: Padding applied to the input tensor; d: Dilation rate in dilated convolutions; SE: Squeeze-and-Excitation module for channel-wise feature recalibration; Residual Block: A block with skip connections to ease gradient flow and improve training; Global Average Pooling: A pooling operation averaging each feature map; Fully Connected: A

dense layer for classification.