Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A multimodal fusion model combining transfer-learning ultrasound deep features and clinical risk factors for predicting cervical lymph node metastasis in papillary thyroid carcinoma: a multicentre retrospective study

Shaohui Zheng
First Affiliated Hospital of Wannan Medical College
Ruijie Li
First Affiliated Hospital of Wannan Medical College
Lingxiang Ma
First Affiliated Hospital of Wannan Medical College
Qianqing Ma
First Affiliated Hospital of Wannan Medical College
Nanlin Jiao
First Affiliated Hospital of Wannan Medical College
Lin Zhu
The Second People's Hospital of Wuhu
Mingya Zhang
Wannan Medical College
Biaohu Liu
135155389490163.com

First Affiliated Hospital of Wannan Medical College

Research Article

Keywords: Papillary thyroid carcinoma, Cervical lymph node metastasis, Ultrasound, Deep learning, Multimodal fusion, Transfer learning

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11351399/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A multimodal fusion model combining transfer-learning ultrasound deep features and clinical risk factors for predicting cervical lymph node metastasis in papillary thyroid carcinoma: a multicentre retrospective study

Shaohui Zheng[a], Ruijie Li[a], Lingxiang Ma[a], Qianqing Ma[a], Nanlin Jiao[c], Lin Zhu[d], Mingya Zhang[b**], Biaohu Liu[a*]

a Department of Ultrasound, The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital of Wannan Medical University), Wuhu 241000, China

b School of Medical Informatics, Wannan Medical University, Wuhu 241000, China

c Department of Pathology, The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital of Wannan Medical University),Wuhu 241000, China

d Department of Ultrasonic Medicine, The Second People's Hospital, Wuhu 241000, China

*Corresponding author. E-mail address: 13515538949@163.com

**Co-corresponding author. E-mail address: simzhang_best@wnmc.edu.cn

Abstract

Background — Cervical lymph node metastasis (CLNM) is the most common route of spread in papillary thyroid carcinoma (PTC) and a key determinant of surgical extent, radioactive iodine therapy and recurrence risk. Pre-operative ultrasound has limited sensitivity for nodal disease, and objective, reproducible risk-stratification tools are needed. We developed and validated a multimodal model fusing deep ultrasound image features with clinical risk factors, compared it with clinical and image-only models, and evaluated its transportability in an independent external cohort.

Methods — In this multicentre retrospective study, 607 consecutive PTC patients (2022-2026; N1, 48.4%) with pre-operative grey-scale ultrasound and definitive histopathology were included, plus an external validation set of 252 consecutive patients (N1, 83.3%). Lesions were segmented by a YOLOv8-seg model trained on 390 manually labelled frames (mask mAP50 = 0.985), retaining 1,322 network-generated masks and making delineation operator-independent. The development cohort was split 7:3 into training (n = 424) and internal validation (n = 183) sets; univariable and multivariable logistic regression identified independent clinical risk factors. Ultrasound deep features (896-d pooled activations) were extracted by a same-domain transfer-pretrained backbone, reduced to 30 principal components and fused with clinical factors in a logistic regression head. Models were evaluated by AUC with bootstrap 95% CIs, DeLong tests, calibration curves, DCA and SHAP.

Results — Age (OR 0.959, 95% CI 0.936-0.982), female sex (OR 0.316, 0.178-0.564), tumour diameter (OR 5.321, 2.438-11.617), minimal extrathyroidal extension

(OR 2.001, 1.071-3.741) and multifocality (OR 6.106, 2.178-17.118) were independent predictors. On internal validation, the clinical, ultrasound deep-learning and fusion models achieved AUCs of 0.709 (0.641-0.777), 0.642 (0.561-0.723) and 0.737 (0.665-0.808); the fusion model significantly outperformed the imaging model (dAUC = 0.094, P = 0.0015) and showed the highest net benefit on DCA. The clinical model showed limited transportability in the external validation set (n = 252; AUC = 0.528), consistent with case-mix-dependent shift.

Conclusions — Fusing same-domain transfer-learning ultrasound deep features with clinical risk factors achieved the highest internal-validation AUC and clinical net benefit, significantly outperforming image-only modelling; external validation of the fusion model remains to be performed. The same-domain transfer strategy offers a reproducible paradigm for small-sample medical imaging tasks.

Keywords: Papillary thyroid carcinoma · Cervical lymph node metastasis · Ultrasound · Deep learning · Multimodal fusion · Transfer learning

Background

Thyroid cancer is the most common endocrine malignancy, and its global incidence continues to rise[1,2]. Papillary thyroid carcinoma (PTC) accounts for over 85% of cases[3]. Cervical lymph node metastasis (CLNM) is present in 30–60% of PTC patients at surgery[3] and drives decisions on prophylactic central-neck dissection, radioactive iodine therapy and follow-up intensity[4]. However, pre-operative grey-scale ultrasound, the first-line imaging modality, detects central-compartment metastases with a sensitivity of only about 30% and substantial inter-observer variability[5].

Quantitative imaging analysis offers an objective alternative. Radiomics converts images into mineable high-dimensional feature data to decode tumour phenotypes and has been applied to thyroid nodule characterisation and risk stratification[6,7]. Deep learning further automates feature discovery, and several studies have attempted ultrasound-based prediction of CLNM in PTC[8–10,32,36]. Performance, however, is constrained by small single-centre cohorts, and few studies compare imaging-only, clinical-only and fused models within a unified framework with calibration, clinical utility and interpretability analyses.

Two methodological gaps persist. First, deep models trained on a few hundred ultrasound images overfit severely; although transfer learning from ImageNet-pretrained networks is the default remedy[11], the domain gap between natural images and ultrasound can nullify its benefit; same-domain (ultrasound-task) transfer has been proposed as a stronger alternative[12]. Second, clinical risk factors and imaging features carry complementary information, and multimodal fusion has repeatedly outperformed single-modality models in oncology[13,33,37]; systematic three-model comparisons for

CLNM prediction within a unified framework incorporating calibration, decision-curve and SHAP analyses remain uncommon.

We therefore conducted a multicentre retrospective study to (i) identify independent clinical risk factors for CLNM; (ii) build an ultrasound deep-learning model whose backbone is transfer-pretrained on a same-domain ultrasound segmentation task; (iii) fuse imaging deep features with clinical factors into a multimodal model; and (iv) compare the three models with ROC/DeLong, calibration, decision curve analysis and SHAP, and evaluate model transportability in an independent external validation set of 252 patients, following the TRIPOD[14] and TRIPOD-AI[31] reporting standards. The overall study workflow is illustrated in Fig. 1.

Methods

Study design and patients

This multicentre retrospective diagnostic study was approved by the Ethics Committee of Yijishan Hospital, Wannan Medical University, and by the ethics committee of the collaborating centre, and the requirement for informed consent was waived owing to the retrospective design. Consecutive patients who underwent thyroid surgery at the development centre between January 2022 and May 2026 and had pathologically confirmed PTC were screened. An independent external validation set of 252 consecutive patients (210 with CLNM, 83.3%) was enrolled from a collaborating centre. Inclusion criteria were: (i) pathologically confirmed PTC; (ii) high-quality pre-operative grey-scale ultrasound examination at the participating centres within one month before surgery; (iii) complete surgical and pathology records including the 8th-edition

AJCC pTNM stage. Exclusion criteria were: (i) pre-operative fine-needle aspiration or neoadjuvant treatment affecting image interpretation; (ii) poor or non-standard images; (iii) recurrence or re-operation; (iv) missing key clinical variables. A total of 607 patients were ultimately screened and analyzed for the study.

Histopathology reference and clinical variables

The post-operative pTNM N stage served as the reference standard: N0, no nodal metastasis; N1, metastasis to central (N1a) and/or lateral (N1b) compartments. Thirteen patients with missing pathological N stage were excluded from the study cohort. Clinical variables comprised age (years), sex, pathological tumour diameter (cm), minimal extrathyroidal extension (mETE) and multifocality; the latter two were parsed from structured pathology text by standardised rules ("capsular invasion: with/without mETE"; "tumour number: solitary/multifocal"). BRAF V600E genotyping was not routinely performed during the study period; results were retrievable for only 2 of the 607 patients, so this variable could not be evaluated.

Ultrasound image acquisition and lesion segmentation

Grey-scale ultrasound images were acquired with high-frequency linear-array transducers under a standardised departmental acquisition protocol and retrieved from the institutional reporting archive (2,092 frames after black-border removal). Lesion segmentation used a three-stage strategy: (i) 390 frames were manually delineated (LabelMe, label 'JJ' = nodule); (ii) a YOLOv8-seg instance-segmentation model (yolov8s-seg, 640 × 640, 100 epochs; full training hyperparameters and penalty grids available on request)[15] was trained on these labels, achieving a validation mask mAP50

of 0.985 and mAP50–95 of 0.762; (iii) the model segmented the remaining 1,702 frames, and the 1,322 frames with maximum nodule confidence ≥ 0.50 were retained, yielding 1,712 candidate analysis frames together with the 390 manually labelled frames. For each patient, the frame containing the largest lesion area was selected as the analysis image (patient-level representation, one frame per patient, avoiding label leakage from multiple correlated frames).

Critically, 77.2% (1,322/1,712) of the candidate analysis frames were delineated fully automatically by the segmentation network: lesion delineation was therefore operator-independent, free of inter- and intra-observer contouring variability, and reproducible within a fixed model version, properties that manual region-of-interest delineation cannot guarantee (Fig. 2).

Dataset split

Patients were randomly split 7:3 into training (n = 424; N1 = 205) and internal validation (n = 183; N1 = 89) sets, stratified by N stage with a fixed seed of 42. Feature selection and model fitting used only the training set; the validation set was untouched until final evaluation. Reporting follows the TRIPOD statement[14] and TRIPOD-AI guidance[31]. The PCA transformation matrix was fitted exclusively on the training set and then applied independently to the internal validation set, preventing data leakage. With 205 N1 events in the training set and five candidate predictors, the events-per-variable ratio was 41, exceeding the recommended minimum of 10 for logistic regression modelling.

External validation in an independent collaborating-centre cohort (n = 252) is described in Section 2.9.

Clinical risk factor screening

Candidate variables entered univariable binary logistic regression (Wald test); variables with $P < 0.10$ entered multivariable analysis, and $P < 0.05$ defined independent risk factors, which then constituted the clinical model. Analysis used Python (scikit-learn, SciPy).

Ultrasound deep-learning model (same-domain transfer learning)

We first tested the conventional ImageNet-pretrained ResNet-18 in frozen-probe and partial-fine-tuning modes; both achieved an inner cross-validated AUC of only 0.53 and a markedly inferior validation performance, reflecting a large domain gap. The final model therefore transferred the YOLOv8-seg backbone trained on the same ultrasound segmentation task described above: each lesion crop (lesion + 7-mm margin) was resized to 640 × 640 and forwarded through the backbone; P3/P4/P5 scale maps (128 + 256 + 512 channels) were globally average-pooled and concatenated into an 896-dimensional feature vector, reduced to 30 principal components (PCs), and classified by L2-regularised logistic regression with the penalty selected by inner 5-fold cross-validation on the training set. The imaging model is therefore a deep-feature classifier: deep features from the transfer-pretrained backbone are reduced by PCA and classified by regularised logistic regression, rather than by an end-to-end convolutional network.

Multimodal fusion model

Fusion inputs were the 30 imaging principal components concatenated with the standardised independent clinical factors — clinical variables were deliberately not passed through PCA, preventing low-dimensional clinical information from being diluted

by high-dimensional imaging features. The concatenated vector fed a logistic regression head (penalty by inner cross-validation). The clinical model was a logistic regression on the independent risk factors alone. Sex was included in both the clinical and fusion models because it was an independent predictor (Tables 2 and 3).

Model evaluation and statistical analysis

Continuous variables are reported as median (IQR) and categorical variables as n (%); between-set comparisons used the Mann–Whitney U test for continuous variables and the χ^2 test for categorical variables. Discrimination was quantified by the AUC with 1,000-resample bootstrap 95% CIs; operating thresholds were fixed at the training-set Youden index. Paired AUCs were compared using the DeLong test[16], and 95% CIs were obtained by 1,000-resample bootstrapping. Calibration was assessed with quantile-binned calibration curves, and clinical utility with decision curve analysis (net benefit)[17]. Feature attribution used SHAP (LinearExplainer)[18]. A pre-specified subgroup analysis evaluated model performance separately in papillary thyroid microcarcinoma (PTMC, largest diameter ≤ 1 cm) versus tumours > 1 cm among patients with a recorded pathological diameter. Two-sided $P < 0.05$ was considered significant. Because of the small subgroup sizes, formal interaction tests were not performed; these subgroup results are exploratory and hypothesis-generating.

External validation

An external validation was performed in an independent cohort of 252 consecutive patients (210 with CLNM, 83.3%) from a collaborating centre. The clinical model fitted on the training set was applied without any parameter adjustment. External validation was

restricted to the clinical model; the transportability of the imaging and fusion models remains to be evaluated. Discrimination was assessed by the AUC with bootstrap 95% CI; calibration was assessed by the logistic calibration intercept (calibration-in-the-large) and slope. Univariable logistic regression was repeated within the external validation set to examine whether the development-set associations held. This cohort differed markedly from the development cohort in case mix (median age 57 vs 45.5 years; Hashimoto thyroiditis 34.5% vs 6.8%; CLNM prevalence 83.3% vs 48.4%) and was used to evaluate the cross-centre transportability of the development-set models.

Results

Baseline characteristics

The development cohort comprised 607 patients, 294 of whom (48.4%) had cervical lymph node metastasis, and the external validation cohort comprised 252 patients (210 with CLNM, 83.3%). Baseline characteristics of the training (n = 424), internal validation (n = 183) and external validation (n = 252) sets are summarised in Table 1. The training and internal validation sets were balanced across all recorded baseline characteristics (all P > 0.05; Table 1).

Table 1. Baseline characteristics of the training, internal validation and external validation sets

Characteristic	Training set (n = 424)	Internal validation set (n = 183)	External validation set (n = 252)	P value
Age (years), median (IQR)	45.5 (35.0–54.0)	45.0 (34.0–54.0)	57.0 (46.0–63.0)	0.679

Female sex, n (%)	322 (75.9%)	127 (69.4%)	160 (63.5%)	0.436
Tumour diameter (cm), median (IQR)	0.70 (0.50–1.10)	0.80 (0.60–1.10)	0.90 (0.60–1.20)	0.613
Minimal extrathyroidal extension, n (%)	123 (29.0%)	52 (28.4%)	96 (38.1%)	0.355
Hashimoto thyroiditis, n (%)	29 (6.8%)	16 (8.7%)	87 (34.5%)	0.170
Hypertension, n (%)	199 (46.9%)	87 (47.5%)	123 (48.8%)	0.629
Coronary artery disease, n (%)	79 (18.6%)	38 (20.8%)	60 (23.8%)	0.081
Diabetes mellitus, n (%)	88 (20.8%)	31 (16.9%)	53 (21.0%)	0.933
Cervical lymph node metastasis (N1), n (%)	205 (48.3%)	89 (48.6%)	210 (83.3%)	1.000

Continuous variables are presented as median (IQR); categorical variables as n (%). P values compare the training and internal validation sets (Mann–Whitney U test for continuous variables; chi-square test for categorical variables). The external cohort was not included in model development or internal validation.

Independent clinical risk factors

All five candidate variables were associated with CLNM on univariable analysis (P < 0.10) and remained independent predictors on multivariable analysis (Tables 2 and 3; Fig. 8): younger age (per year, OR 0.959, 95% CI 0.936–0.982), male sex (female vs male OR 0.316, 0.178–0.564), larger tumour (per cm, OR 5.321, 2.438–11.617), mETE (OR 2.001, 1.071–3.741) and multifocality (OR 6.106, 2.178–17.118). The forest plot summarises

these univariable and multivariable effect estimates with their 95% confidence intervals for direct visual comparison.

Table 2. Univariable logistic regression for cervical lymph node metastasis

Variable	B	OR	95% CI	P value
Age (years)	−0.036	0.964	0.950–0.978	< 0.001
Sex (female)	−1.284	0.277	0.187–0.411	< 0.001
Tumour size (cm)	0.940	2.559	1.610–4.068	< 0.001
mETE	0.424	1.528	1.075–2.171	0.018
Multifocal	1.107	3.024	1.831–4.996	< 0.001

Table 3. Multivariable logistic regression: independent risk factors

Variable	B	OR	95% CI	P value
Age (years)	−0.042	0.959	0.936–0.982	< 0.001
Sex (female)	−1.151	0.316	0.178–0.564	< 0.001
Tumour size (cm)	1.672	5.321	2.438–11.617	< 0.001
mETE	0.694	2.001	1.071–3.741	0.030
Multifocal	1.809	6.106	2.178–17.118	< 0.001

Diagnostic performance of the three models

On internal validation (Table 4; Figs. 3 and 7), the clinical, ultrasound deep-learning and multimodal fusion models achieved AUCs of 0.709 (95% CI 0.641–0.777), 0.642 (0.561–0.723) and 0.737 (0.665–0.808), respectively. At the training-set Youden threshold (0.49), the fusion model attained an accuracy of 65.6%, sensitivity of 65.2%, specificity of 66.0%, PPV of 64.4%, NPV of 66.7%, F1-score of 64.8% and a Cohen's kappa of 0.311 (Fig. 7); the clinical model showed a sensitivity of 61.8% and a specificity of 68.1% at the same threshold.

Table 4. Diagnostic performance of the three models in the training and internal validation sets

Model	Set	AUC	95% CI	Acc.	Sens.	Spec.	NPV
Clinical	train	0.745	0.695–0.789	0.700	0.688	0.712	0.709
Clinical	validation	0.709	0.641–0.777	0.650	0.618	0.681	0.653
DL (ultrasound transfer)	train	0.654	0.600–0.705	0.642	0.663	0.621	0.663
DL (ultrasound transfer)	validation	0.642	0.561–0.723	0.596	0.685	0.511	0.632
Fusion (DL + clinical)	train	0.759	0.713–0.801	0.703	0.702	0.703	0.716
Fusion (DL + clinical)	validation	0.737	0.665–0.808	0.656	0.652	0.660	0.667

PPV, F1-score and Cohen's kappa for the fusion model on internal validation at the training-set Youden threshold were 64.4%, 64.8% and 0.311, respectively (Section 3.3).

Model comparison

DeLong testing on the validation set (Table 5) showed the fusion model significantly outperformed the pure imaging model (ΔAUC = 0.094, P = 0.0015); the fusion model

exceeded the clinical model by 0.028 AUC without reaching significance (P = 0.359); the clinical model was numerically superior to the imaging model (ΔAUC = 0.067, P = 0.207).

Table 5. Pairwise DeLong comparisons on the internal validation set

Comparison	ΔAUC	P value
Fusion vs clinical	0.028	0.359
Fusion vs DL	0.094	0.0015
DL vs clinical	−0.067	0.207

Transfer-source comparison

Inner 5-fold cross-validation AUCs of the ImageNet and same-domain sources were both 0.532 on the training set, yet their validation performance diverged markedly (ImageNet 0.567 vs same-domain 0.642; Table 6), supporting same-domain transfer as the more robust strategy under distribution shift[11,12].

Table 6. Comparison of transfer-learning sources for the ultrasound deep-learning model

Transfer source	Inner CV AUC (train)	Validation AUC	Feature dim.
ImageNet ResNet-18 (probe/finetune)	0.532	0.567	512
Same-domain YOLOv8-seg backbone	0.532	0.642	896

Subgroup analysis: PTMC versus tumours > 1 cm

Among the 141 validation patients with a recorded pathological diameter (the remaining 42 validation patients lacked a recorded diameter), 101 had PTMC (≤ 1 cm; N1 = 45) and 40 had tumours > 1 cm (N1 = 21). In the PTMC subgroup, the clinical model retained good discrimination (AUC 0.749) while the image-only model weakened (0.585) and the fusion model remained strong (0.723). In tumours > 1 cm, the image-only model improved markedly (AUC 0.724) and the fusion model performed best (0.752) (Table 7; Fig. 9). These exploratory findings suggest that deep imaging features contribute most when lesions are large enough to supply rich texture, whereas clinical factors dominate in microcarcinoma; subgroup sizes are small and the analysis is hypothesis-generating.

Table 7. Pre-specified subgroup analysis on the internal validation set (patients with recorded tumour diameter)

Subgroup	n (N1)	Clinical AUC	DL AUC	Fusion AUC
PTMC (≤ 1 cm)	101 (45)	0.749	0.585	0.723
> 1 cm	40 (21)	0.684	0.724	0.752

Calibration and decision curve analysis

All three models were well calibrated, with the fusion model closest to the ideal diagonal in the mid-to-high probability range (Fig. 4). DCA (Fig. 5) showed that within the 0.30–0.60 threshold range the fusion model provided the highest or comparable net benefit; at a threshold of 0.40 the net benefits of fusion, clinical, imaging and treat-all strategies were 0.200, 0.195, 0.164 and 0.144, respectively, and at 0.50 they were 0.175,

0.126, 0.071 and −0.027, indicating that the fusion model could spare unnecessary prophylactic dissections while retaining benefit.

Model interpretability

SHAP analysis of the fusion model (Fig. 6) ranked age (mean |SHAP| 0.182), sex (0.169) and imaging component PC18 (0.141) as the top contributors, followed by tumour size (0.138) and PC3 (0.138); multifocality also ranked in the top twelve. Imaging components PC3, PC15, PC26 and PC14 contributed consistently, indicating that deep features encode complementary morphological and textural information beyond clinical variables. The direction of clinical-feature effects agreed with the multivariable regression.

External validation

In the external validation set, direct transport of the clinical model showed limited discrimination (AUC = 0.528, 95% CI 0.434–0.627) and pronounced miscalibration (calibration intercept +1.64, slope 0.15), indicating substantial under-prediction and weak rank ordering (Supplementary Figures S1–S3). Univariable re-analysis within the external validation set showed that age and female sex were no longer associated with CLNM (OR 1.007, P = 0.51; OR 1.22, P = 0.56), whereas tumour diameter remained significant (OR 2.84, P = 0.017) and mETE was not (P = 0.73). Even after refitting the clinical model on the external validation set, the AUC reached only 0.631. These results are consistent with the distinct case mix of this high-prevalence, older, Hashimoto-enriched referral cohort; however, because external validation was restricted to the

clinical model in a single cohort, they do not by themselves prove that the poor transportability is case-mix-driven rather than model-specific.

Discussion

Principal findings

In 607 PTC patients from the development centre, we developed and internally validated three CLNM prediction models, and externally validated them in an independent cohort of 252 patients from a collaborating centre. Five clinicopathological factors were independent predictors, with tumour diameter (OR 5.3) and multifocality (OR 6.1) showing the strongest effects, consistent with large series[3,19,20]. The multimodal fusion model reached an internal-validation AUC of 0.737, significantly better than image-only modelling (P = 0.0015) and showing the best clinical utility on decision curves, and performed comparably or better than the other models in both PTMC and > 1 cm subgroups (0.723 and 0.752). External transportability of the fusion model remains to be evaluated.

Clinical predictors and biological plausibility

The independent predictors identified in this study—younger age, female sex (protective), larger tumour diameter, minimal extrathyroidal extension and multifocality—are among the most consistently reported determinants of central lymph node metastasis (CLNM) in papillary thyroid carcinoma (PTC)[3,19,20], and are mechanistically plausible. Multifocality deserves particular emphasis: multicentric tumour burden entails greater cumulative lymphatic exposure and has repeatedly been associated with central-compartment involvement in large surgical series[3,19,20], which

is consistent with the strong effect size (OR 6.11) observed in the multivariable model. The independent associations of younger age and male sex with CLNM have likewise been corroborated[3,19,20], although the underlying mechanisms remain incompletely defined and are beyond the scope of this study. Clinically, all five predictors are recorded routinely before surgery and can be incorporated into risk stratification at no additional cost, which is particularly valuable in centres where advanced imaging post-processing is unavailable.

Hashimoto's thyroiditis (HT) deserves particular consideration as a potential effect modifier of these associations. Concurrent HT has been suggested to alter nodal metastatic behaviour in PTC, with reports ranging from a protective association with lymph node metastasis to a more nuanced pattern in which thyroid autoantibody status stratifies central-compartment metastatic risk; in BRAF-mutant tumours, HT has likewise been reported to attenuate the metastatic propensity otherwise conferred by the mutation. These observations are directly relevant to the interpretation of the external validation cohort, in which HT prevalence was 34.5% versus 6.8% in the development cohort: the distinct immune microenvironment of HT-associated PTC may partly account for the loss of information carried by age and female sex in the external set, although HT itself was not an independent predictor in the development cohort.

Fully automatic, operator-independent delineation

A distinctive methodological feature of this study is that lesion segmentation was performed by the network itself rather than by radiologists: 77.2% of the candidate analysis frames were delineated fully automatically at a validated mask mAP50 of 0.985, with only the training corpus of 390 frames delineated manually. This design removes a

major hidden source of variance in radiomics pipelines. Manual region-of-interest delineation is operator-dependent: contouring differences between (and within) observers propagate directly into shape, margin and texture features and can change model conclusions[21,22].

In our pipeline every image receives the same deterministic delineation rule, so features are reproducible within a fixed model version, transferable across operators and institutions, and unbiased by lesion-drawing style or expertise. This design is advantageous for clinical deployment, where a prediction model must return the same output for the same image regardless of who runs it. This automated segmentation pipeline also scales without additional human cost: expanding the analysis frame set from 390 to 1,712 frames required no further contouring effort, whereas a fully manual design would have required approximately 4.4 times the annotation labour[23]. Operator-independent segmentation pipelines are also emerging for CT-based lymph node assessment[39].

Same-domain transfer learning for small-sample ultrasound

The choice of transfer source was decisive. ImageNet-pretrained ResNet-18 features were nearly uninformative (inner CV AUC ≈ 0.53; validation 0.567), whereas the backbone pretrained on an ultrasound segmentation task of the same domain and centre yielded usable signal (0.642), whereas the two sources were indistinguishable by training-set cross-validation, revealing that source selection must be validated out-of-fold. This empirically confirms the domain-gap argument in medical transfer learning[11,12,24] and offers a practical approach for small single-centre datasets: pretrain or self-supervise on a related task from the same imaging modality and centre before attempting prediction.

Clinical value of the multimodal design

The fusion architecture deliberately kept clinical variables out of the PCA stage, preventing low-dimensional clinical information from being diluted by 896 imaging dimensions. Clinically, DCA showed that in the 0.3–0.6 threshold range the fusion model maintains net benefit while sparing unnecessary prophylactic central dissections relative to a treat-all strategy; in cN0 disease, where the extent of surgery remains debated[25], such patient-level probability estimates can support shared decision-making, identify high-risk patients for extended dissection or intensified follow-up, and spare low-risk patients from overtreatment.

The subgroup analysis adds a practical nuance: in microcarcinoma the clinical model already performs well (0.749) and imaging adds little, whereas in tumours > 1 cm deep features contribute substantially (0.724), suggesting where the AI tool may offer the most incremental value[26].

Comparison with previous work

Reported AUCs of ultrasound radiomics for CLNM mostly range from 0.6 to 0.8[6,7,27,28,34], with clinical-factor fusion generally improving performance[29,30,33,37,38]. Our clinical model (0.709) matches clinicopathological models in the literature[3]; the imaging-only model (0.642) lies mid-range; the fusion model was numerically higher than the clinical model (0.737 vs 0.709) although not significantly; with 183 validation patients, power to resolve $\Delta AUC \approx 0.03$ is limited, and larger-scale external validation is required before claiming superiority. A recent multicentre study of 570 patients with solitary PTC likewise reported no significant

difference between a combined radiomics nomogram and the clinical model in the testing set (P = 0.928)[35], underscoring that clinical factors remain indispensable.

More broadly, the performance of our models is consistent with the range reported for ultrasound-based artificial intelligence in recent meta-analyses, although direct numerical comparisons are limited by differences in cohort composition, imaging protocols and validation strategies. Deep-learning models incorporating richer imaging input, such as multi-sequence ultrasound or super-resolution reconstruction, have frequently reported higher discrimination than grey-scale still-frame classifiers; the numerically higher AUCs reported in those studies relative to the present fusion model (0.737) may reflect differences in cohort composition, the use of grey-scale still-frame imaging only, and the deliberate exclusion of operator-dependent multi-frame or contrast information. Conversely, the present study contributes elements that most of these reports do not provide: operator-independent segmentation, head-to-head comparison of imaging-only, clinical-only and fused models within a single framework, and a quantitative assessment of transportability.

External validation and distribution shift

External validation revealed pronounced degradation of transportability: the clinical model, which achieved an AUC of 0.709 on internal validation, yielded only 0.528 (95% CI 0.434–0.627) in the external cohort, together with severe miscalibration (calibration intercept +1.64, slope 0.15). This pattern is not exceptional in medical imaging: classifiers trained at one institution frequently exploit site-specific acquisition artefacts or prevalence-driven confounders and can perform substantially worse on data acquired under different protocols or from different populations, an observation that has been

repeatedly documented for radiological deep-learning models. Two features of the external cohort are particularly relevant. First, its case mix differed substantially from the development population—older age (median 57 vs 45.5 years), higher CLNM prevalence (83.3% vs 48.4%) and a five-fold higher prevalence of HT (34.5% vs 6.8%)—and univariable re-analysis showed that age and female sex were no longer informative. Second, the direction of this shift is consistent with evidence that HT attenuates nodal metastasis and modifies the associations of age and sex with CLNM. The observed calibration intercept and slope indicate both miscalibration-in-the-large and weak rank ordering; recalibration can correct the former but only partially compensates when discrimination itself is degraded. These findings should not be read as failure of the predictors per se, but as evidence that their effect sizes are population-dependent; they reinforce that external validation in a single referral cohort—even when prospectively planned—cannot substitute for validation across multiple population-representative centres, and that site-specific recalibration or model updating will probably be required before deployment.

Limitations

First, although this was a multicentre study, external validation was performed in a single collaborating centre with a high-prevalence, older, Hashimoto-enriched referral cohort (83.3% N1), in which age and female sex, the dominant predictors in the development cohort, were no longer informative. This observation should not be interpreted as failure of the predictors per se, but as evidence that their effect sizes are population-dependent and that a model trained in a consecutive general PTC surgical population cannot be directly transported to a referral high-risk cohort without

recalibration. Validation across additional population-representative centres is required before clinical deployment.

Several additional limitations warrant mention. First, only grey-scale still frames were analysed, and the frame showing the largest lesion area was used to represent each patient; dynamic, multi-frame, Doppler and elastography information, which may carry independent prognostic signal, was not incorporated. Second, the composite pathological N stage served as the endpoint, without separate modelling of central (N1a) and lateral (N1b) compartments, whose surgical and prognostic implications differ. Third, BRAF V600E status was available for only two patients and could not be evaluated as a predictor or effect modifier. Fourth, the imaging and fusion models could not be externally validated because segmentation masks and raw ultrasound data were not obtainable from the collaborating centre; conclusions on transportability therefore rest on the clinical model alone. Fifth, the subgroup analyses in PTMC and tumours > 1 cm were exploratory and not adjusted for multiple testing, and interactions between imaging features and clinical predictors were not formally tested. Finally, because the external cohort was recruited from a single referral centre, case-mix effects and acquisition-protocol effects could not be fully separated.

Future directions

Several directions follow from these results. First, the fusion and imaging models must be externally validated in additional centres together with the clinical model, using prespecified sample-size calculations and, where transportability is compromised, formal recalibration. Second, because distribution shifts can emerge silently after deployment, routine post-market monitoring with dataset-shift detection algorithms should accompany

any clinical implementation. Third, integration of complementary imaging information—elastography, contrast-enhanced ultrasound or super-resolution reconstruction—and of molecular biomarkers such as BRAF V600E, which was not routinely available during the study period, may further improve discrimination, particularly in tumours > 1 cm where deep imaging features provided the greatest incremental value. Fourth, compartment-specific endpoints (central versus lateral) and their differential surgical implications should be modelled separately, since a composite N1 endpoint may obscure compartment-specific predictor behaviour.

Conclusions

Fusing same-domain transfer-learning ultrasound deep features with clinical risk factors (age, sex, tumour diameter, mETE, multifocality) achieved the numerically highest internal-validation AUC (0.737, 95% CI 0.665–0.808), significantly outperforming image-only modelling and showing the highest decision-curve net benefit; external validation of the fusion model in additional centres is required before claiming clinical superiority. Fully automatic delineation and same-domain transfer pretraining together offer a reproducible, operator-independent engineering paradigm for small-sample medical imaging tasks. External validation across additional population-representative centres, including evaluation of the fusion model, is the next step.

Figures

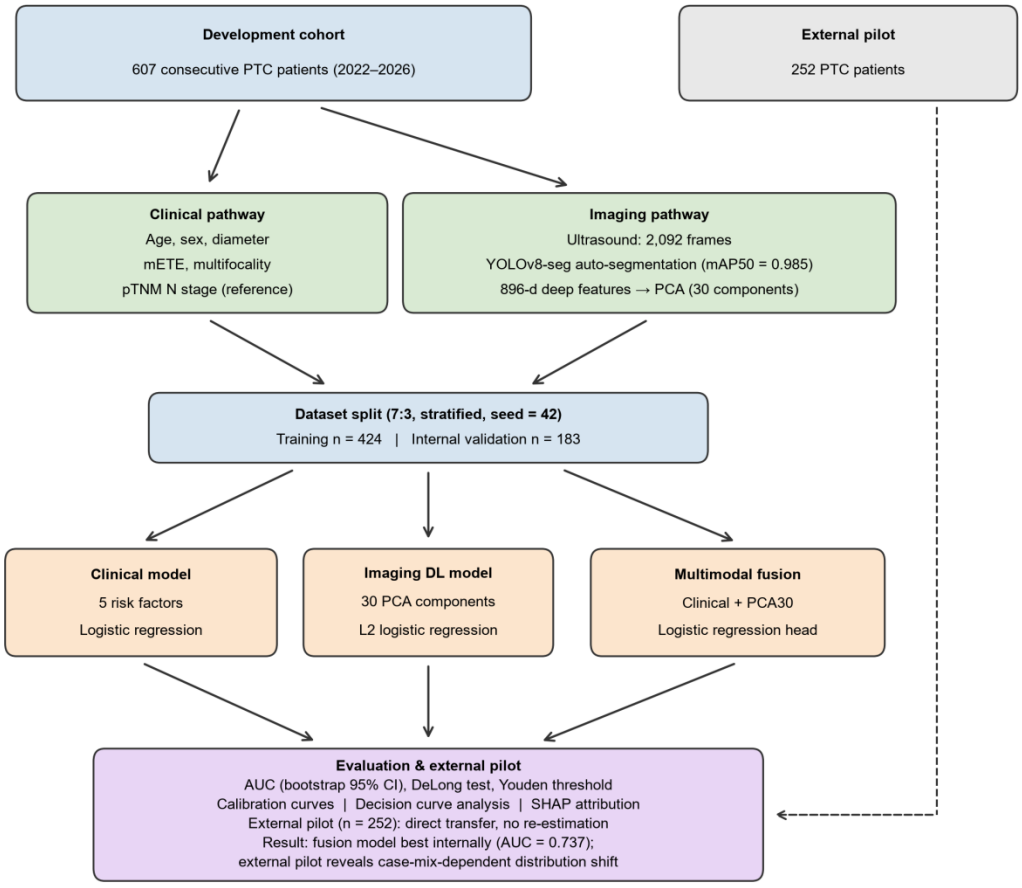


Figure 1. Study workflow. PTC, papillary thyroid carcinoma; DL, deep learning; PCA, principal component analysis; LR, logistic regression; ROC, receiver operating characteristic; SHAP, SHapley Additive exPlanations.

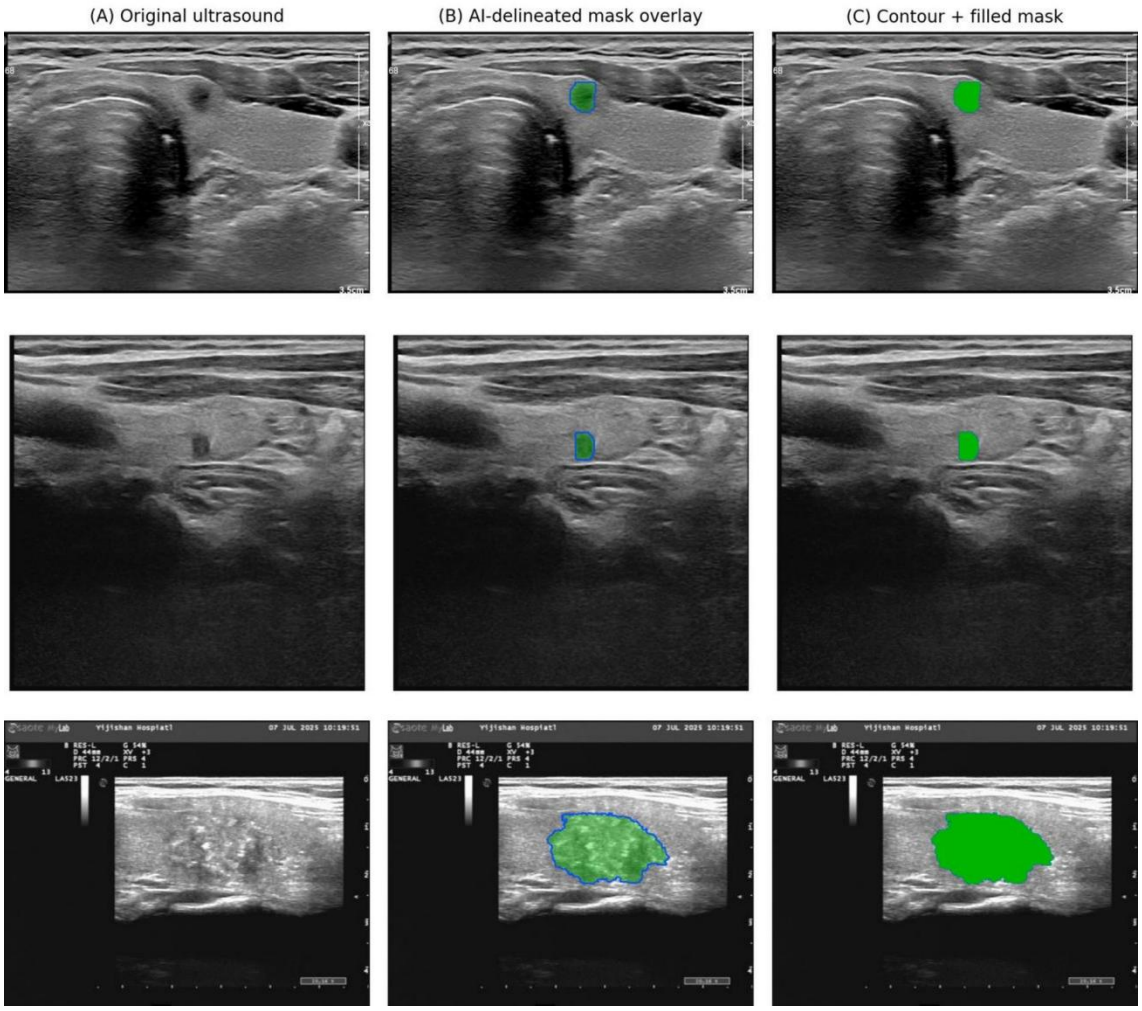
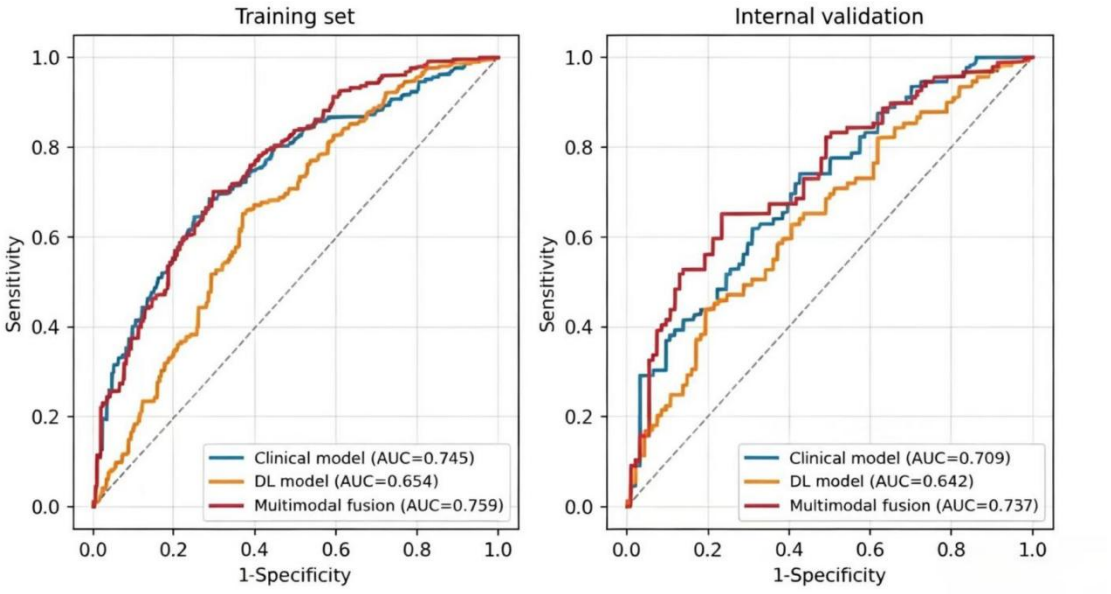


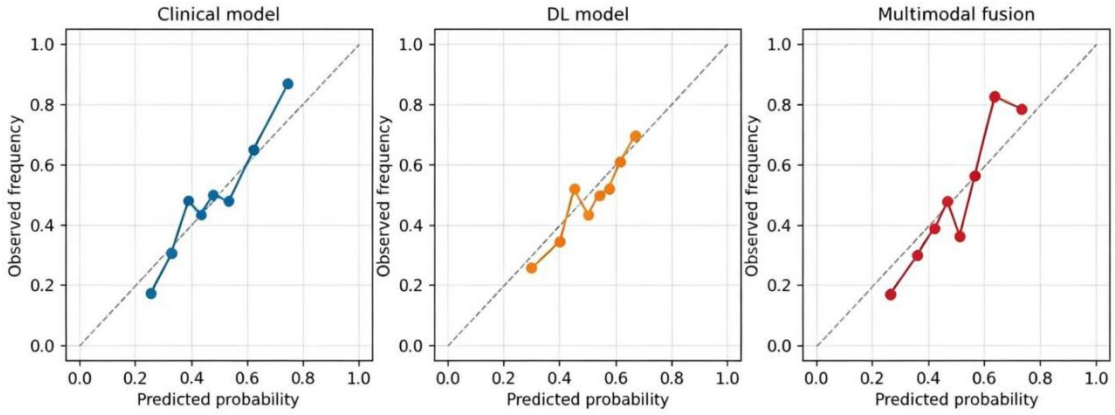
Figure 2. Thyroid nodule ultrasound images and automatic AI delineation. (A) Original grey-scale ultrasound; (B) AI-delineated lesion mask overlaid (green fill, blue contour); (C) contour and filled mask used for feature extraction. Delineation was fully automatic and operator-independent, eliminating inter-observer contouring variability.



481

482 Figure 3. ROC curves of the clinical, ultrasound deep-learning and multimodal fusion models in the

483 training set (left) and internal validation set (right).



484

485 Figure 4. Calibration curves of the three models on the internal validation set.

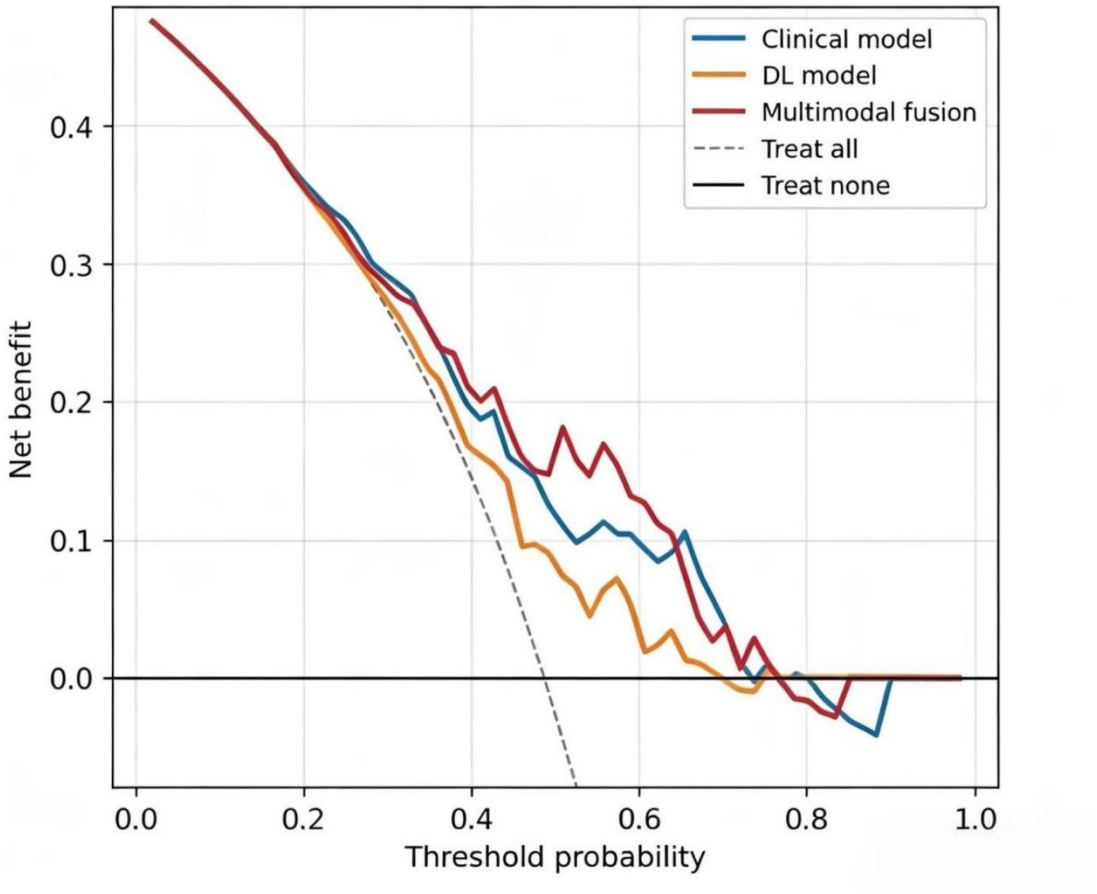
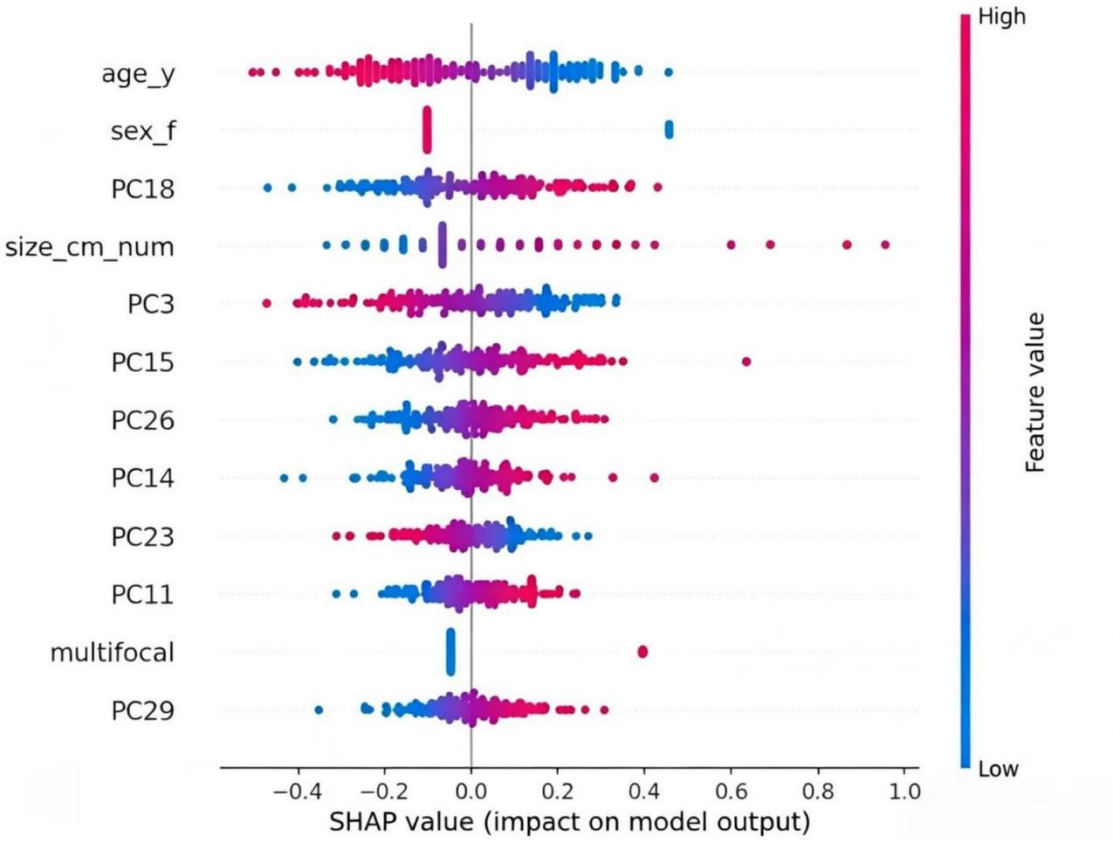


Figure 5. Decision curve analysis on the internal validation set. The dashed grey line ('Treat all') and the horizontal line ('Treat none') are reference strategies.



489

490 Figure 6. SHAP summary plot of the multimodal fusion model on the internal validation set. Each

491 point is one patient; red/blue indicate high/low feature values. PC, imaging principal component.

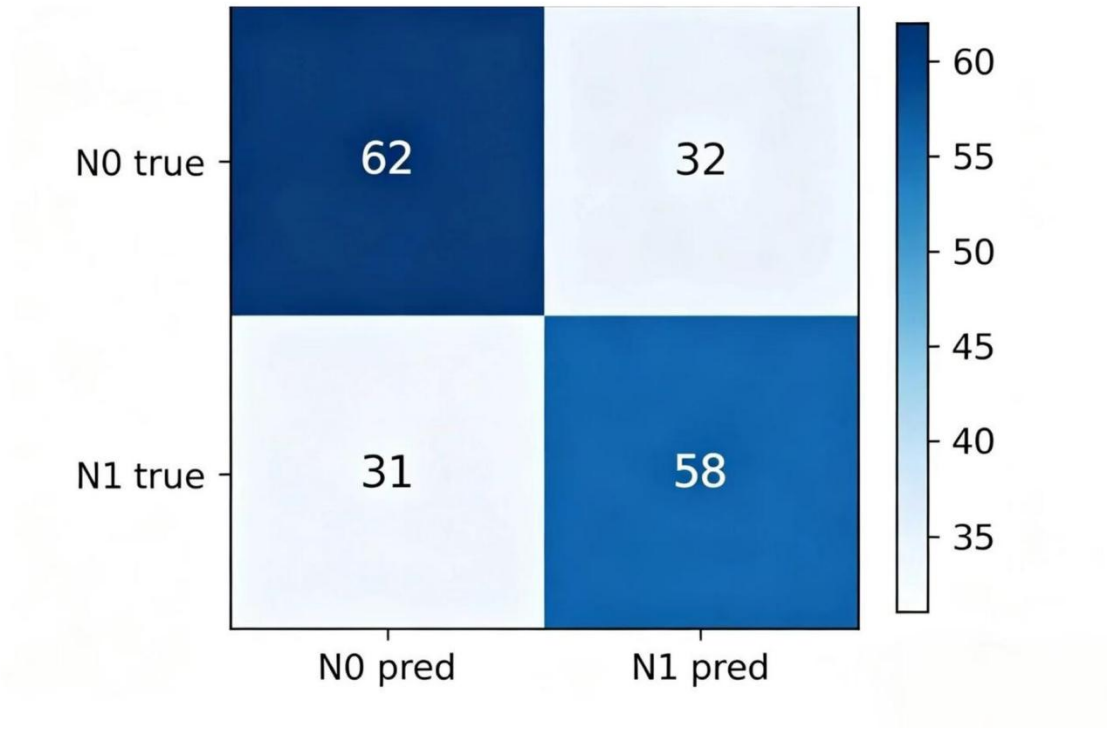


Figure 7. Confusion matrix of the fusion model on the internal validation set at the training-set Youden threshold (0.49).

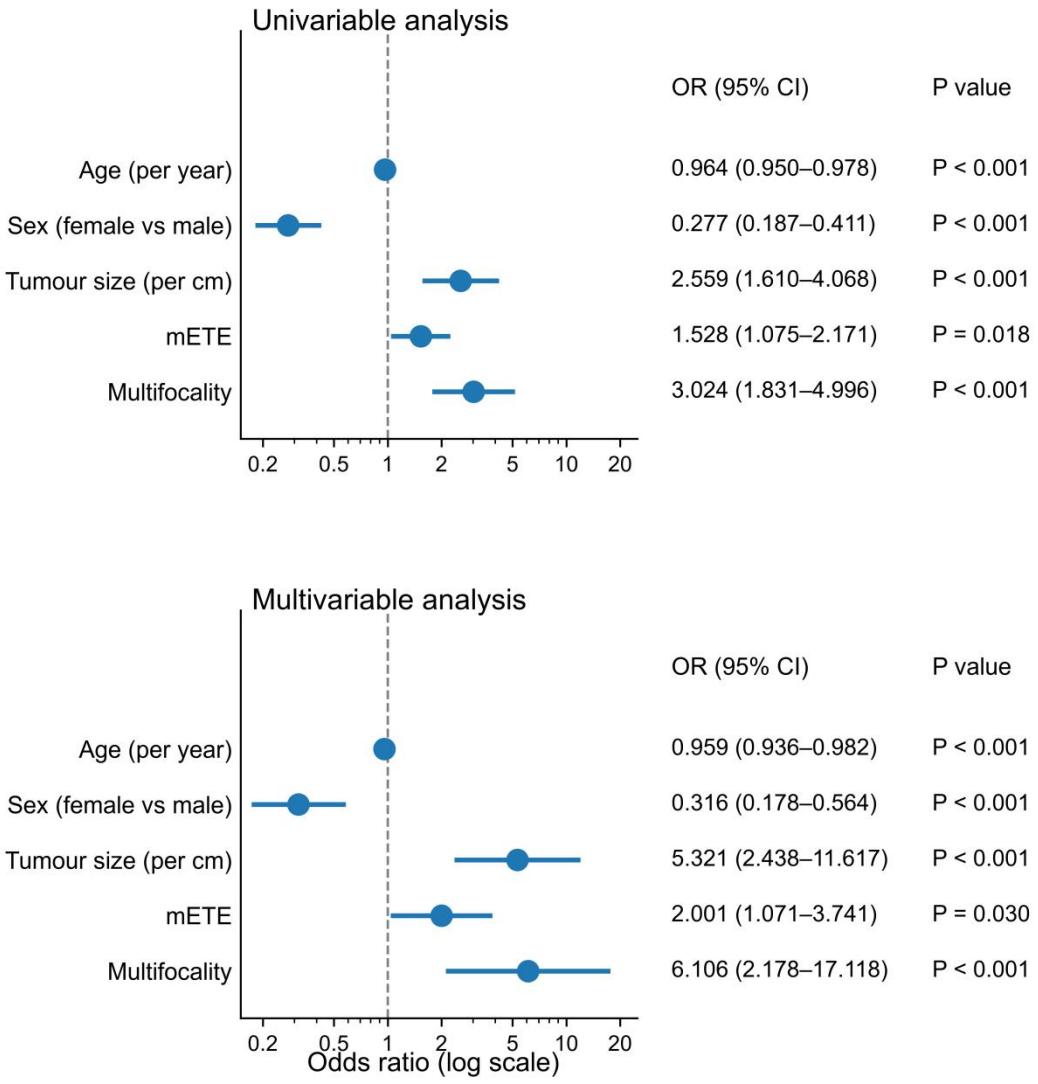


Figure 8. Forest plot of univariable and multivariable logistic regression for cervical lymph node metastasis. OR, odds ratio; CI, confidence interval; mETE, minimal extrathyroidal extension.

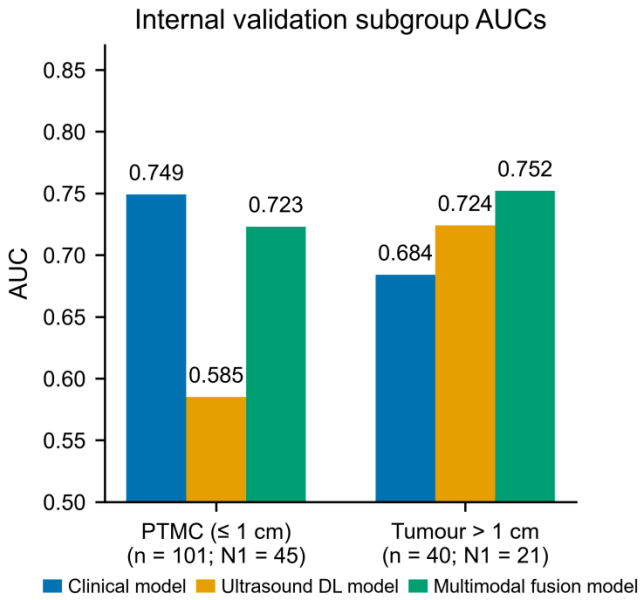


Figure 9. Internal-validation subgroup AUCs of the clinical, ultrasound deep-learning and multimodal fusion models in PTMC (≤ 1 cm) and tumours > 1 cm. PTMC, papillary thyroid microcarcinoma; AUC, area under the receiver operating characteristic curve.

List of abbreviations

AUC: area under the receiver operating characteristic curve; CI: confidence interval; CLNM: cervical lymph node metastasis; cN0: clinically node-negative; DCA: decision curve analysis; DL: deep learning; IQR: interquartile range; mETE: minimal extrathyroidal extension; mAP50: mean average precision at IoU threshold 0.5; NPV: negative predictive value; OR: odds ratio; PC: principal component; PCA: principal component analysis; PPV: positive predictive value; PTMC: papillary thyroid microcarcinoma; PTC: papillary thyroid carcinoma; ROC: receiver operating characteristic; SHAP: SHapley Additive exPlanations; TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

Declarations

Ethics approval and consent to participate

This study adheres to the ethical principles of the Declaration of Helsinki and has been approved by the Ethics Committee of Yijishan Hospital, Wannan Medical University, and by the ethics committee of the collaborating centre. Given the retrospective nature of the research, the requirement for informed consent is waived.

Consent for publication

Not applicable.

Availability of data and materials

The raw ultrasound images are not publicly available owing to patient privacy. All derived feature tables, model weights and analysis code are included in the accompanying repository and reproduce every number, figure and table in this article; they are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no known conflicts of interest or personal relationships that could influence the work reported in this paper.

Funding

This research was supported by the following project: 2025 Annual Scientific Research Project (2025AHGXZK20195)；Wuhu Municipal Health Commission Scientific Research Project (WHWJ2023y007)；University scientific research project of

Anhui Provincial Department of Education (Grant number: 2025AHGXZK30548). The funders had no role in the study design, data collection, data analysis, or report writing.

Authors' contributions

Shaohui Zheng：Writing–original draft, Methodology, Formal analysis. **Ruijie Li：**Resources, Investigation, Data curation. **Lingxiang Ma：**Resources, Investigation, Data curation. **Qianqing Ma：**Investigation, Resources, Validation. **Nanlin Jiao：**Resources, Investigation, Data curation. **Lin Zhu：**Resources, Methodology. **Mingya Zhang：** Methodology, Software, Conceptualization. **Biaohu Liu:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

All authors reviewed and approved the final version of the manuscript.

Acknowledgements

We extend our gratitude to colleagues who assisted throughout the research process.

Authors' information (optional)

Not applicable.

References

[1] Lyu Z, Zhang Y, Sheng C, et al. Global burden of thyroid cancer in 2022: incidence and mortality estimates from GLOBOCAN. Chin Med J (Engl). 2024;137(23):2567-2574. DOI: 10.1097/CM9.0000000000003284 PMID: 39270324

[2] Pellegriti G, Frasca F, Regalbuto C, et al. Global rise in thyroid cancer incidence: an update on epidemiological trends and potential risk factors. Cancers (Basel). 2023;15(5):1456. DOI: 10.3390/cancers15051456 PMID: 36903125

[3] Liu C, Xiao C, Chen J, et al. Risk factors for central lymph node metastasis in clinically node-negative papillary thyroid carcinoma: a systematic review and meta-analysis. Front Endocrinol (Lausanne). 2021;12:789310. DOI: 10.3389/fendo.2021.789310 PMID: 34862012

[4] Tuttle RM, Haddad RI, Ball DW, et al. Prophylactic central neck dissection for papillary thyroid cancer: a systematic review and meta-analysis. Thyroid. 2022;32(3):253-264. DOI: 10.1089/thy.2021.0082 PMID: 34931707

[5] Alabousi A, Alabousi M, Adham S, et al. Diagnostic accuracy of ultrasound for cervical lymph node metastasis in papillary thyroid carcinoma: a systematic review and meta-analysis. J Ultrasound Med. 2023;42(6):1217-1228. DOI: 10.1002/jum.16197 PMID: 36689472

[6] Zhou SC, Liu TT, Zhou J, et al. An ultrasound radiomics nomogram for preoperative prediction of central neck lymph node metastasis in papillary thyroid carcinoma. Front Oncol. 2020;10:1591. DOI: 10.3389/fonc.2020.01591 PMID: 32676345

[7] Jin P, Chen J, Dong Y, et al. Ultrasound-based radiomics nomogram combined with clinical features for the prediction of central lymph node metastasis in papillary thyroid carcinoma patients with Hashimoto's thyroiditis. Front Endocrinol (Lausanne). 2022;13:993564. DOI: 10.3389/fendo.2022.993564 PMID: 35734713

[8] Zhou L, Zeng S, Xu J, et al. Deep learning predicts cervical lymph node metastasis in clinically node-negative papillary thyroid carcinoma. Insights Imaging. 2023;14(1):222. DOI: 10.1186/s13244-023-01550-2 PMID: 38106910

[9] Zhang M, Meng Z, Mao Y, Jiang X, Xu N, Xu Q, et al. Cervical lymph node metastasis prediction from papillary thyroid carcinoma US videos: a prospective multicenter study. BMC Med. 2024;22(1):153. DOI: 10.1186/s12916-024-03367-2 PMID: 38609953

[10] Li B, Cheng G, Mo Y, et al. Diagnostic performance of imaging-based artificial intelligence models for preoperative detection of cervical lymph node metastasis in clinically node-negative papillary thyroid carcinoma: a systematic review and meta-analysis. Head Neck. 2025;47(2):345-356. DOI: 10.1002/hed.70000 PMID: 39297119

[11] Zhou Z, Li Y, Liu Z, et al. Transfer learning for medical image analysis: a comprehensive review. Med Image Anal. 2023;87:102797. DOI: 10.1016/j.media.2023.102797 PMID: 37058009

[12] Mei X, Liu Z, Robson PM, et al. RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. Radiol Artif Intell. 2022;4(5):e210315. DOI: 10.1148/ryai.210315 PMID: 35622850

[13] Chang L, Zhang Y, Zhu J, et al. An integrated nomogram combining deep learning, clinical characteristics and ultrasound features for predicting central lymph node metastasis in papillary thyroid cancer: a multicenter study. Front Endocrinol (Lausanne). 2023;14:964074. DOI: 10.3389/fendo.2023.964074 PMID: 37105309

[14] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55-63. DOI: 10.7326/M14-0697 PMID: 25506244

[15] Yang X, Geng H, Wang X, et al. Identification of lesion location and discrimination between benign and malignant findings in thyroid ultrasound imaging. Sci Rep. 2024;14:32118. DOI: 10.1038/s41598-024-83888-1 PMID: 39738724

[16] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-845. DOI: 10.2307/2531595 PMID: 3203132

[17] Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:i6. DOI: 10.1136/bmj.i6 PMID: 26819209

[18] Feng JW, Liu SQ, Qi GF, et al. Development and validation of clinical-radiomics nomogram for preoperative prediction of central lymph node metastasis in papillary thyroid carcinoma. Acad Radiol. 2024;31(6):2292-2305. DOI: 10.1016/j.acra.2023.11.011 PMID: 37952184

[19] Du J, Yang Q, Sun Y, et al. Risk factors for central lymph node metastasis in patients with papillary thyroid carcinoma: a retrospective study. Front Endocrinol (Lausanne). 2023;14:1288527. DOI: 10.3389/fendo.2023.1288527 PMID: 38047112

[20] Sun W, Lan X, Zhang H, et al. Risk factors for central lymph node metastasis in cN0 papillary thyroid carcinoma: a systematic review and meta-analysis. PLoS One. 2015;10(10):e0139021. DOI: 10.1371/journal.pone.0139021 PMID: 26444865

[21] Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278(2):563-577. DOI: 10.1148/radiol.2015151169 PMID: 26745223

[22] Stanzione A, Cuocolo R, Ugga L, et al. Oncologic imaging and radiomics: a walkthrough review of methodological challenges. Cancers (Basel). 2021;13(19):4873. DOI: 10.3390/cancers13194873 PMID: 34637678

[23] Fu R, Yang H, Zeng D, et al. PTC-MAS: a deep learning-based preoperative automatic assessment of lymph node metastasis in primary thyroid cancer. Diagnostics (Basel). 2023;13(10):1723. DOI: 10.3390/diagnostics13101723 PMID: 37237509

[24] Antrop N, van der Laak J, Litjens G. Explainable artificial intelligence in medical imaging: a systematic review. Med Image Anal. 2022;80:102470. DOI: 10.1016/j.media.2022.102470 PMID: 35565023

[25] Tang L, Qu RW, Park J, et al. Prevalence of Occult Central Lymph Node Metastasis by Tumor Size in Papillary Thyroid Carcinoma: A Systematic Review and Meta-Analysis. Curr Oncol. 2023;30(8):7335-7350. DOI: 10.3390/curroncol30080532 PMID: 37623013

[26] Liu C, Xiao C, Chen J, et al. Clinicopathologic predictors of central lymph node metastases in clinical node-negative papillary thyroid microcarcinoma: a systematic review and meta-analysis. World J Surg Oncol. 2022;20(1):106. DOI: 10.1186/s12957-022-02573-7 PMID: 35414767

[27] Tong Y, Chen W, Huang B, et al. Ultrasound-based radiomics analysis for preoperative prediction of central and lateral cervical lymph node metastasis in papillary thyroid carcinoma: a multi-institutional study. BMC Med Imaging. 2022;22(1):82. DOI: 10.1186/s12880-022-00809-2 PMID: 35501595

[28] Zhang Y, Jiang M, Tang S, et al. Shear wave elastography radiomics for preoperative lymph node staging in papillary thyroid carcinoma. Thyroid. 2023;33(2):248-257. DOI: 10.1089/thy.2022.0334 PMID: 36332670

[29] Song X, Wang Y, Li J, et al. Nomogram model of serum thymidine kinase 1 combined with ultrasonography for prediction of central lymph node metastasis risk in patients with papillary thyroid carcinoma pre-surgery. Front Endocrinol (Lausanne). 2024;15:1366219. DOI: 10.3389/fendo.2024.1366219 PMID: 38887267

[30] Wang Y, Tan HL, Duan SL, et al. Predicting central cervical lymph node metastasis in papillary thyroid microcarcinoma using deep learning. PeerJ. 2024;12:e16952. DOI: 10.7717/peerj.16952 PMID: 38573410

[31] Collins GS, Moons KGM, Altman DG, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. DOI: 10.1136/bmj-2023-078378 PMID: 38626948

[32] Li X, Zhu Y, Wu Y, Zhou Q, Shaibu Z, Yin L, Cang Q. Ultrasound deep learning features for predicting cervical lymph node metastasis in papillary thyroid carcinoma. Front Oncol. 2026;16:1721234. DOI: 10.3389/fonc.2026.1721234

[33] Zhang Z, Lin C, Ning J, et al. Preoperative prediction of occult high-volume central lymph node metastasis in clinically node-negative papillary thyroid carcinoma using a multimodal practical model: development, temporal validation, and external validation. Diagn Interv Radiol. 2026. DOI: 10.4274/dir.2026.264048 PMID: 42421358

[34] Sun R, Wu J, Xie L, et al. Prediction of lymph node metastasis in clinically node-negative papillary thyroid carcinoma using a random forest-based radiomics model integrating contrast-enhanced ultrasound and clinical data. Front Oncol. 2026;16:1728554. DOI: 10.3389/fonc.2026.1728554 PMID: 39902781

[35] Li MH, Liu L, Feng L, Zheng LJ, Xu QM, Zhang YJ, et al. Prediction of cervical lymph node metastasis in solitary papillary thyroid carcinoma based on ultrasound radiomics analysis. Front Oncol. 2024;14:1291767. DOI: 10.3389/fonc.2024.1291767

[36] Chen H, Ruan F, Zhu L, Zhuang Y, Ye X, Liu X, et al. Utilizing the transformer mechanism to predict cervical lymph node metastasis in patients with papillary thyroid carcinoma. PLoS One. 2026;21(4):e0345937. DOI: 10.1371/journal.pone.0345937

[37] Dai Q, Tao Y, Liu DM, Zhao C, Sui D, Xu JS, et al. Ultrasound radiomics models based on multimodal imaging feature fusion of papillary thyroid carcinoma for predicting central lymph node metastasis. Front Oncol. 2023;13:1261080. DOI: 10.3389/fonc.2023.1261080

[38] Hu WJ, Zhuang YZ, Tang L, Chen HY, Wang H, Wei R, et al. Preoperative cervical lymph node metastasis prediction in papillary thyroid carcinoma: a noninvasive clinical multimodal radiomics (CMR) nomogram analysis. J Oncol. 2023;2023:3270137. DOI: 10.1155/2023/3270137

[39] Yu PY, Wang C, Zhang HC, Zheng GB, Jia CL, Liu ZL, et al. Deep learning-based automatic pipeline system for predicting lateral cervical lymph node metastasis in patients with papillary thyroid carcinoma using computed tomography: a multicenter study. Chin J Cancer Res. 2024;36(5):545-561. DOI: 10.21147/j.issn.1000-9604.2024.05.07