International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

THYROX: A Unified Explainable Deep Learning System for Thyroid Nodule Classification and Lymph Node Metastasis Assessment in Ultrasound

Tanish P Jadhav
Dept of CSE(AI&ML)
Don Bosco Instituite of Technology
Bengaluru, India

Sarvesh Patil
Dept of CSE(AI&ML)
Don Bosco Institute of Technology
Bengaluru, India

Prof. Padmavathi B
Dept of CSE(AI&ML)
Don Bosco Institute of Technology
Bengaluru, India

***Abstract* - Thyroid nodule is common but rarely malignant, and ultrasound interpretation remains subject to substantial inter-observer variability. In papillary thyroid carcinoma (PTC), staging additionally requires assessment of the cervical lymph node basin, which is frequently under-detected preoperatively. Existing deep learning systems treat the nodule and the nodal basin separately, rarely quantify their explanation and are seldom validated on public data. We present THYROX, a unified decision support system performing nodule malignancy classification, lesion localization, risk stratification and cervical nodal assessment in one explainable pipeline. Two public datasets were used:TN5000(5,000 biopsy confirmed B-mode images with bounding box annotations) and LymphUs (338 PTC patients, two centers). Four ImageNet pretrained backbones ResNet50, DenseNet121, EfficientNetV2-S and ConvNeXt-Tiny were fine-tuned under an identical protocol and combined by soft vote ensembling with three view test time augmentation, oversampling only the training split. For the nodal task a DenseNet121 was warm started from the thyroid checkpoint rather than ImageNet, testing whether an in-domain ultrasound representation transfers across anatomical sites under label scarcity. On the TN5000 test set ($n = 750$) the ensemble achieved 91.87% accuracy (95% CI 89.69-93.62), 95.34% sensitivity, 83.18% specificity and ROC-AUC 0.9675. on LymphUs ($n = 51$) the warm-started model reached 88.24% accuracy with 100% specify from only 233 trained images. Three backbones performed equivalently (AUC range 0.0021) while ConvNeXt-Tiny failed to converge, and single-box regression proved inadequate for localization (mean IoU 0.286); both negative results are reported.**

Keywords: thyroid nodule, cervical lymph node metastasis, ultrasound, deep learning, transfer learning, explainable AI, Grad-Cam, ensemble learning

I. INTRODUCTION

Thyroid cancer is the most common endocrine malignancy, with an estimate over 820,000 new cases and 47,000 deaths worldwide in 2022 [1]. Recorded incidence has risen steeply for four decades while mortality has remained flat a divergence attribute to over detection of small, indolent papillary carcinomas rather than a true rise in disease burden [1]. The clinical problem is therefore not detection but discrimination.

Ultrasonography is the first-line modality. Structured reporting, principally ACR TI-RADS [25], translates sonographic features into risk categories and fine needle aspiration (FNA) thresholds, reducing biopsy rates while improving accuracy [21]. The features themselves, however are assessed with only moderate reproducibility: pooled meta-analysis report inter-rater agreement of $\kappa \approx 0.51 \approx 0.51$ [13]. Downstream, FNA cytology returns an indeterminate result in a substantial fraction of sampled nodules, driving diagnostic surgery in patients with benign disease [24].

A second problem concerns disease extent. Cervical lymph node metastasis is common in PTC and materially changes surgical planning [27], yet preoperative ultrasound detection is far less sensitive: a meta-analysis of 27 studies covering more than 23,000 patients reports pooled radiologist sensitivity of 0.51 against 0.80 for AI models [6].

Deep learning has addressed the first problem extensively [19], [9], [3]; nodal prediction has followed a smaller trajectory [5], [6]. Three limitations recur. First, the nodule and the nodal basin are treated separately despite being assessed in the same examination for the same decision. Second, explanations are displayed but not evaluated, although benchmarking in other modalities shows saliency methods localising markedly worse than human annotation [14]. Third, evaluation often uses small private cohorts, sometimes with resampling applied before partitioning, which inflates reported accuracy.

This work addresses all three. Our contributions are:

1) A **unified two-task pipeline** covering nodules malignancy, localisation and risk stratification together with a prallel cervical nodal module, mirroring clinical PTC staging.
2) **Cross anatomical in domain transfer:** the nodal classifier is warm started from a thyroid ultrasound that speckle statistics and margin priors are shared across ultrasound anatomies.
3) A **four architechture benchmark with ensembling** under one shared regularisation protocol, isolating back-bone choice from training recipe.
4) **Quanttitaive explainability,** evaluating Grad-CAM against TN5000's ground-truth boxes rather than pre senting it illustratively.
5) **Public data, reported negative results and released code.**

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

II. RELATED WORK

A. Noduel classification, nodal metastais and transfer

Multicohort convolutional studies established that networks approach expert performance and that the endpoint is better framed as a biopsy-management decision [19], and that assistance raises junior readers toward senior level accuracy prospectively [9]; the field has since moved toward transformer and hybrid designs [3], with Rashed *et al.* [4] the closest analogue to this work. Nodal prediction is younger progressing from an initial pilot [22] to explainable multi-modal learning [5]; meta-analysis identifies the same gap throughout most studies are single-centre and few release data [6]. On initialisation, Raghu *et al*. [20] showed that much of ImageNet transfer's benefit derives from weight scaling rather than feature reuse, and benchmarks confirm in-domain pretraining generally wins under label scarcity [12]. Most relevant here, USFM [10] demonstrated that ultrasound representations generalise across organs. Grad-CAM [17] is the standard explanation tool, yet across seven saliency methods on chest radiographs all localised significantly worse than a human benchmark [14], and quantitative protocols are rarely applied to thyroid ultrasound.

B. Positioning against immediate atecedents

Three studies from the direct antecedents of this work(Table I). Nugroho *et al*. [26] classify nodule margin from nine geometric descriptors with SMOTE balancing, reporting 94.79% accuracy on 165 private images. We adopt the descriptor family but differ in endpoint, data provenance, and importantly in applying oversampling only after partitioning: in [18] SMOTE expands 165 images to 480 instance before the 80/20 split, so synthetic samples interpolated from test set originals enter training. Pimpalkar and Gulame [27] fuse transverse and longitudinal views through two YOLOv5 detectors and an XGBoost head, reporting 99.64% accuracy on an 81 case test set; we adopt the multi-view and detector then classifier structure, while noting their component detectors reach mAP@0.5 of only ≈0.70-0.72. Dency Flora et al. [11] survey thyroid dysfunction classification from laboratory attributes rather than malignancy from images; we cite it for motivation and exclude it from performance comparison.

III. MATERIALS AND METHODS

The architecture is shown in Fig. 1; this section follows it stage by stage.

A. Datasets

TN5000 [7] compromises 5,000 B-mode thyroid ultra-sound images 3,572 malignant and 1,428 benign labelled by FNA biopsy with pathological confirmation and annotated by experienced radiologists. Each image carries a PASCAL-VOC XML annotation encoding a lesion box ($x_{\min}$, $y_{\min}$, $x_{\max}$, $y_{\max}$); images are JPEG standardised to 718 × 500 × 3.

LymphUs [2] comprises cervical lymph node ultrasound from 338 PTC patients, across two independent centres (Centre 1:180 patients, 99 benign / 81 malignant; Centre 2: 158 patients, 76 benign / 82 malignant), with segmentation masks alongside B-Mode images. Only B-mode images were used; mask, lesion and fusion renderings were excluded, yielding 338 images (175 benign, 163 malignant). Dataset properties and partitioning appear in Table II.

Partitioning was stratified by class. Because the TN5000 release does bot expose patient identifiers, thyroid partitioning is performed at the image level; no image appears in more that one split, but patient level disjointness cannot be claimed. Nodal partitioning is patient-level and verified disjoint.

B. Pre-processing and augmentation

An input frame $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ is bilinearly resized to 224×224 and normalized channel-wise as $\hat{x}^{(c)}=(x^{(c)}/255-\mu_c)/\sigma_c$ with the ImageNet statistics μ=(0.485, 0.456, 0.406), σ=(0.229, 0.224, 0.225). Augmentation (Table III) uses only ultrasound plausible transforms, with random erasing acting as a cutout regularizer simulating acoustic shadowing. A heavier variant was used for the nodal task, where the training set is an order of magnitude smaller and oversampled duplicates must be rendered visually distinct at each epoch.

Critically, the two branches take separate images: a thyroid image is not a valid input to the nodal classifier, and the interface reports "not assessed" for any branch whose input is absent.

C. Localisation and geometric features

Lesion localisation used a ResNet50 backbone with a four unit sigmoid head regressing the box in normalised coordinates b=($x_{\min}$ / W, $y_{\min}$ / H, $x_{\max}$ / W, $y_{\max}$ / H) ∈ $[0,1]^4$, trained with the Smooth-L1 objective

$$\mathcal{L}_{\text{box}} = \sum_{k=1}^{4} \rho\left(\hat{b}_k - b_k\right), \quad \rho(z) = \begin{cases} \frac{1}{2}z^2, & |z| < 1 \\ |z| - \frac{1}{2}, & \text{otherwise,} \end{cases} \tag{1}$$

Using AdamW at 1 × 10^{-4} for up to 20 epochs (patience 5) Localisation quality is measured by the Jaccard index

$$\text{IoU}\left(\hat{B}, B\right) = \frac{|\hat{B}\cap B|}{|\hat{B}\cup B|} \tag{2}$$

Reported as the mean over the test set and as the fraction exceeding 0.5.

Nine geometric margin descriptors follow the formulation of [18], computed from the largest external contour recovered by median filtering, CLAHE, Otsu thresholding and elliptical morphological closing and opening within the region of interest. Writing A and P for contour area and perimeter, A_h and P_h for those of the convex hull, and $d_i = \| p_i - \bar{p} \|_2$ for the radial distance from the centroid to boundary point i, the three most discriminative are

$$\kappa = \frac{4\pi A}{p^2}, \quad \chi = \frac{P_h}{P}, \quad \delta = \frac{\sigma_d}{\bar{d}}, \tag{3}$$

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

TABLE I

THROX COMPARISON WITH PREVIOUS AND RECENT STUDIES. REPORTED RESULTS ARE NOT DIRECTLY COMPARABLE DUE TO DIFFERENCES IN DATASETS, METRICS, AND TEST SIZES.

STUDY	YEAR	DATA	ENDPOINT	n TEST
Dency Flora et al [11]	2024	Tabular (yes)	Dysfunction	-
Nugroho et al. [18]	2020	165 img. (no)	Margin reg./irreg	96
Pimpalakar & Gulame [27]	2025	983 lumps (no)	Benign /Malignant	81
Rashed et al. [4]	2025	Public	Benign /Malignant	-
Zhang et al. [7]	2025	TN5000(yes)	Detection	1,000
AI Pooled [12]	-	>23k pts (mixed)	Nodal Metastasis	-
Radiologists pooled [12]	-	>23k pts (mixed)	Nodal Metastasis	-
THYROX thyroid (ours)	**2026**	**TN5000 (yes)**	**Benign/ Malignant**	**750**
THYROX nodal (ours)	**2026**	**LymphUs (yes)**	**Nodal metastasis**	**51**

TABLE II

DATASET CHARACTERISTICS AND SPLITS. OVERSAMPLING WAS LIMITED TO TRAINING DATA, KEEPING VALIDATION AND TEST SET UNCHANGED FOR UNBIASED EVALUATION.

METHOD	ACCURACY	NODAL
Survey: hybrid DCNN	93.85%	No
9 Geometric feat + MLP	94.79%	No
2 × YOLOv5 + NMS + XGBoost	99.64%	No
9 backbones benchmarked	ResNet50 best	No
DETR, Faster R-CNN, SSD	mAP ≈ 0.58	No
Meta-analysis, 27 studies	sens. 80%	Yes
Human readers	Sens. 51%	Yes
4-backbone ens +TTA	**91.87%**	**Yes**
Warm-start DenseNet121 + TTA	**88.24%**	**Yes**

TABLE III

AUGMENTATION PIPELINE, MODEL CONFIGURATIONS AND FOOTPRINT. MODEL SIZE DID NOT CORRELATE WITH ACCURACY: THE SMALLEST MODEL PERFORMED BEST, WHILE THE LARGEST FAILED TO CONVERGE.

PROPERTY	TN5000[19]	LYMPHUS [20]
Anatomical target	Thyroid nodule	Cervical lymph node
Clinical endpoint	Benign/malignant	Non-met. / metastatic
Images used	5,000	338
Benign / malignant	1,428 / 3,572	175 / 163
Class ratio	1 : 2.50	1 : 0.93
Patients	Not disclosed	338
Centres	Single release	2 (180 + 158)
Modality	B-mode	B-mode (+masks, unused)
Train (ben. / mal.) after balancing	3,500 (998 / 2,502)	233 (120 / 113)
Validation	750 (214 / 536)	54 (28 / 26)
Test	750 (214 / 536)	51 (27 / 24)
Partition level	Image	Patient

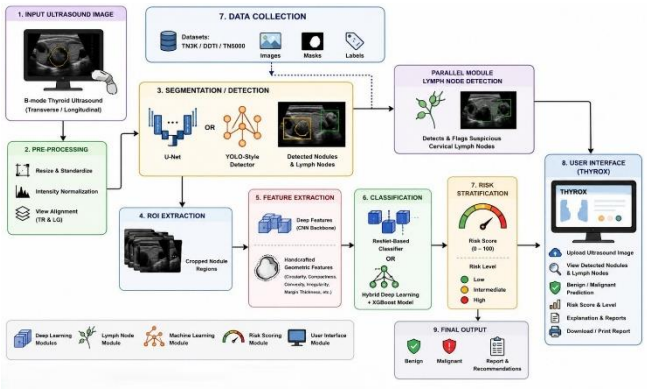


Fig. 1. The THYROX architecture.

Where compactness κ equals unity for a circle, convexity $\chi < 1$ signals an irregular margin, and dispersion δ quantifies radial irregularity directly.

TABLE IV

THYROID CLASSIFICATION ON TN5000 TEST SET ($n = 750$) using three-view TTA.

TRANSFORM	THYROID	NODAL

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Horizontal / Vertical flip	0.5 / 0.3	0.5 / 0.5
Rotation	±20°	±30°
Affine translate / scale	6% / ±10%	10% / ±15%
Brightness / contrast jitter	±0.20	±0.25
Random erasing	p=0.25	p=0.30
Mixup [25] rate π	0.5	0.3

circularity $\bar{d}$ / σ_d, solidity A / A_h , elongation b / a, roughness P / P_h, equivalent diameter ration $2\sqrt{(A/\pi)}/a$ and margin thickness $(d_{max} - d_{min})/\bar{d}$, with a, b the fitted ellipse axes follow analogously.

D. Architechutres and training

Four backbones spanning the residual, densely connected, compound scaled and modernized-convolutional families were selected: ResNet50[28], DenseNet121[26], EfficientNetV2-S[16] and ConvNeXt-Tiny[23], all instantiated from ImageNet-pretrained weights via `timm` with two-way head. Transformer backbones were excluded on the grounds that at 5,000 images the data regime favours convolutional inductive bias. Pre-architecture regularization (Table IV) was set after a preliminary run in which EfficientNetV2-S exhibited the largest train validation gap.

Because the thyroid classes are imbalanced at 1:2.5, each class c is weighted by inverse frequency, normalized to sum to the number of classes C:

$$w_c = C \cdot \frac{1/n_c}{\sum_{c'=1}^{C} 1/n_{c'}} \tag{4}$$

Targets are smoothed by $\epsilon = 0.1$ to discourage over-confident logits, $\tilde{y}_c = (1-\varepsilon)y_c + \varepsilon/C$, giving the training objective

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} w_c\, \tilde{y}_c \log p_c\,, \qquad p_c = \frac{e^{z_c}}{\sum_{c'} e^{z_{c'}}}. \tag{5}$$

Mixup [25] is applied with probability π perbatch: for a pair of examples $(x_i, y_i), (x_j, y_j)$ and $\lambda \sim \text{Beta}(\alpha, \alpha)$

$$\tilde{x} = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \tag{6}$$

$$\mathcal{L}_{\text{mix}} = \lambda \mathcal{L}_{\text{CE}}(f(\tilde{\mathbf{x}}), y_i) + (1-\lambda)\mathcal{L}_{\text{CE}}\big(f(\tilde{\mathbf{x}}), y_j\big), \tag{7}$$

With α = 0.2 and π = 0.5 (thyroid) or 0.3 (nodal). Optimization used AdamW with automatic mixed precision under a cosine-annealed learning rate,

$$\eta_t = \tfrac{1}{2}\eta_0(1 + \cos(\pi t/T_{\max})), \tag{8}$$

Where T_{max} = 25 (thyroid) or 30 (nodal). Model selection used best validation accuracy with early stopping (patience 7 thyroid, 10 nodal); batch size was 32 and all seeds were fixed at 42.

A note affecting Fig. 4: on mixup epochs, training accuracy is computed by comparing predictions on the blended batch $\tilde{x}$ against the original labels y_i, and is therefore systematically understated. This is why validation accuracy exceeds training throughout, and is a measurement artefact rather than underfitting

E. Cross-anatomical transfer

The nodal training set contains 233 images, a regime in which ImageNet transfer yields limited benefit [*20*] and indomain pretraiing repeadtedly helps [*12*]. Evidence that ultrasound represents generalise across organs *[10]* motivates our hypothesis: a backbone trained to discriminate benign from malignant thyroid nodules acquires speckle-texture and margin morphology features that transfer to cervical n*odal* assessment despite the change of organ. Operationally, a DenseNet121 is loaded from the thyroid checkpoint with `strict=False`, chosen over ResNet50 for parameter economy (27.1 vs 90.0 MB), and trianed with $\eta_0 = 5 \times 10^{-5}$, weight decay 2×10^{-4} and mixup rate 0.3 a gentler schedule that avoids ealry destruction of the trnasferred representation.

F. Inference, ensembling and risk stratification

At inference each image is presented under three transforms identity, horizontal flip and vertical flip and the softmax vectors averaged, so that for model m

$$\mathbf{p}_m^{\text{TTA}}(\mathbf{x}) = \tfrac{1}{3}\textstyle\sum_{k\in\{\text{I,H,V}\}} \text{softmax}\left(f_m\big(\mathcal{T}_k(\mathbf{x})\big)\right). \tag{9}$$

Architecture level soft voting over the M converged backbones then gives the ensemble posterior and decision

$$\bar{p}(\mathbf{x}) = \tfrac{1}{M}\textstyle\sum_{m=1}^{M} p_{m,\text{mal}}^{\text{TTA}}(\mathbf{x}), \qquad \hat{y} = \mathbb{1}[\bar{p}(\mathbf{x}) > \tau], \tag{10}$$

with $\tau = 0.5$. the posterior is mapped to an integer risk score $s \left\lfloor 100\bar{p}(\mathbf{x}) + \frac{1}{2} \right\rfloor \in \{0, \ldots, 100\}$ and thence to three tiers mirroring the structure of TI-RADS [25],

$$\text{tier}(s) = \begin{cases} \text{Low}, & s \le \tau_\ell \\ \text{Intermediate}, & \tau_\ell < s \le \tau_h \\ \text{High}, & s > \tau_h. \end{cases} \tag{11}$$

Boundaries are selected on the validation split only, τ_l as the largest score retaining $NPV \ge 0.95$ and τ_h as the smallest reaching $PPV \ge 0.90$. Tier level results are reserved for future work.

G. Explainability

Grad-CAM [17] is computed on the final convolutional block of the selected model. For target class c and feature maps A^k of that block, the channel weights and the localization map are

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \quad L^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right), \tag{12}$$

Where Z is the number of spatial positions. Because TN5000 supplies a ground truth box $\mathcal{B}$ for every image,

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

attention quality is quantifies rather than merely displayed. Over N test images we report the pointing game [14] and the energy-in-box fraction,

$$\mathrm{PG} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[\arg\max_{(i,j)} L_n^c(i,j) \in \mathcal{B}_n\right], \tag{13}$$

$$\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N} \frac{\sum_{(i,j)\in\mathcal{B}_n} L_n^c\,(i,j)^c}{L}\,_n\,(i,j), \tag{14}$$

Together with the IoU of (2) between $\mathcal{B}_n$ and L_n^c thresholder at half its maximum.

TABLE V
TN5000 CONFUSION MATRICES AND ERROR COSTS. ENSEMBLE ACHIEVED 91.87% ACCURACY, 95.34% SENSITIVITY AND 83.18 SPECIFICITY, SLIGHTLY OUTPERFORMING DENSENET121

MODEL	TN	FP	FN	TP	MISSED	UNNEC.
ResNet50	178	36	34	502	6.3%	16.8%
DenseNet121	177	37	25	511	4.7%	17.3%
EfficientNetV2-S	175	39	32	504	6.0%	18.2%
ConvNeXt-Tiny	0	214	0	536	0.0%	100%
Ensemble	178	36	25	511	4.7%	16.8%

H. Evaluation metrics

With malignant as the positive class and (TP, FP, FN, TN) the confusion-matrix counts, we report accuracy, precision TP/(TP+FP), specificity TN/(TN+FP), negative predictive value TN/(TN+FN), F_1 = 2TP/(2TP+FP+FN), Youden's J = Sens +Spec -1, and balanced accuracy $\frac{1}{2}$(Sens + Spec). ROC-AUC and PR-AUC are computed from the ensemble posteriors of (10). Because the nodal test set contains only 51 images, interval estimates use the Wilson score method rather than the normal approximation: for k successes in n trials with $\hat{p} = k/n$ and $z = 1.96$,

$$\mathrm{CI}_{95\%} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}. \tag{15}$$

Localization is evaluated by mean IoU and IoU@0.5. Experiments ran on a single NVIDIA Tesla T4 (16GB) with PyTorch 2.11 and CUDA 12.8.

IV. RESULTS

A. Thyroid nodule classifications

The performance of the four backbone models and the ensemble on the TN5000 test set with three view test-time augmentation (TTA) is evaluated using confusion matrices and Wilson confidence intervals. The corresponding results are visualised in the Figs. 2,3 and 6

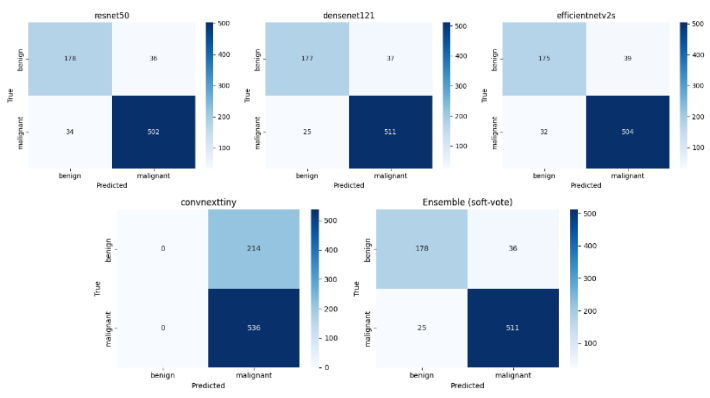
B. Two negative results

ConvNeXt-Tiny did not converge. Its training remained At 0.49-0.51 for all eight epochs before early stopping, validation accuracy oscillated between the class priors (0.7147 and 0.2853), and it assigned the malignant label to all 750 test images yielding accuracy exactly equal to majority class prevalence and ROC-AUC 0.5601 (Fig. 4). We attribute this to optimizer architecture mismatch rather than a deficiency of the architecture mismatch rather than a deficiency of the architecture: the shared protocol applies AdamW at $\eta_0 = 3 \times 10^{-4}$ with uniform weight decay, whereas ConNeXt-family models are conventionally trained with far lower peak rates, layer-wise decay, stochastic depth and much longer schedules [15]. At 5,000 images and 25 epochs none of those conditions held. Identical regularization protocols therefore do not transfer across architecture families directly relevant to any multi-backbone benchmark under a shared recipe.

Single-box regression proved inadequate for localization. Validation Smooth L1 loss fell monotonically from 0.0078 to 0.0049, so optimization succeeded, but localization quality did not: mean IoU 0.2859, with only 23.73% of predictions exceeding IoU 0.5. The failure is structural: regressing four coordinates form a global embedding presumes a single lesion of predictable scale and position, with no mechanism for scale invariance, competing candidate regions, or confidence ranked suppression. Detectors on this dataset reach mAP@0.5 ≈ 0.58 [7] and dedicated single stage detectors elsewhere exceed 0.70 [27]. The consequence is visible in Fig. 7, where the predicted box sits well within the true lesion extent. Because region-of-interest extraction, geometric features and the hybrid classifier all depend on a reliable lesion region, this stage is keystone of the pipeline, not a peripheral component.

C. Cervical nodal metastasis

The warm-started DenseNet121 reached best validation accuracy 0.8704 at epoch 8, with early stopping at epoch 17. Test performance appears in Table VI and Fig. 5. The model made no false-positive predictions: every benign node was correctly identified, giving 100% specificity and precision, at the cost of six missed metastases of 24(75.00% sensitivity). At 51 test images the intervals are necessarily wide sensitivity spans 55.10-88.00% and the point estimates should be read accordingly rather than as precise performance claims.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Fig. 2. Confusion matrices for the TN5000 test set.

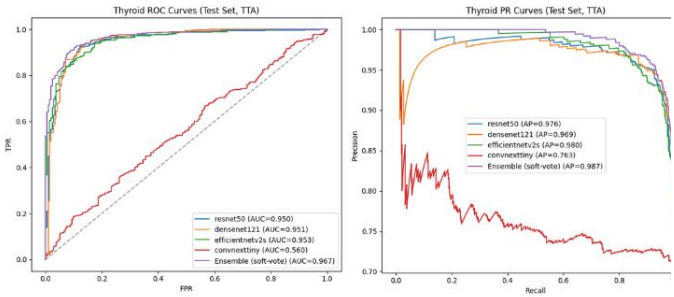


Fig. 3. Receiver operating characteristic (left) and precision recall (right) curves on the TN5000 test set

TABLE VI
LYMPHUS TEST SET CLASSIFICATION (n=51) USING WARM STARTED DENSENET121 WITH TTA. WILSON 95% CONFIDENCE INTERVALS

METRIC	VALUE	95% CI
Accuracy	88.24%	76.62-94.50
Precision	100.00%	-
Sensitivity	75.00%	55.10-88.00
Specificity	100.00%	87.54-100.00
Negative Predictive Value	81.82%	-
F1-Score	85.71%	-
Balanced accuracy	87.50%	-
Confusion Matrix	TN 27, FP	FN 6, TP 18

D. *Explainability and end-to-end output*

Fig. 7 shows Grad-CAM and the complete pipeline output for an unseen transverse acquisition. Attention concentrates on the isthmic nodule, which the model classifies as malignant with 90.8% confidence the desired behaviour. The same map, however, contains a second high-activation region at the inferior border, over an acoustic-shadow area posterior to the lesion that corresponds to no nodular structure. We report this explicitly: it shows that Grad-CAM attention here is not fully lesion-specific, consistent with benchmarking evidence from other modalities [14], and it is the direct motivation for quantifying attention

E. *Comparision with published methods*

Table I places these results beside published work. Against pooled radiologist sensitivity of 51% for cervical nodal metastasis across 27 studies [6], the nodal module's 75% from 233 training images is a meaning margin, though the 51 images test set precludes a strong claim. Against the pooled AI sensitivity of 80% and specificity of 83% in the same meta-analysis, THYROX sits below on sensitivity and above on specificity, reflecting the threshold asymmetry noted above.

V. DISCUSSION

A. *Why the smallest model won, and what that implies*

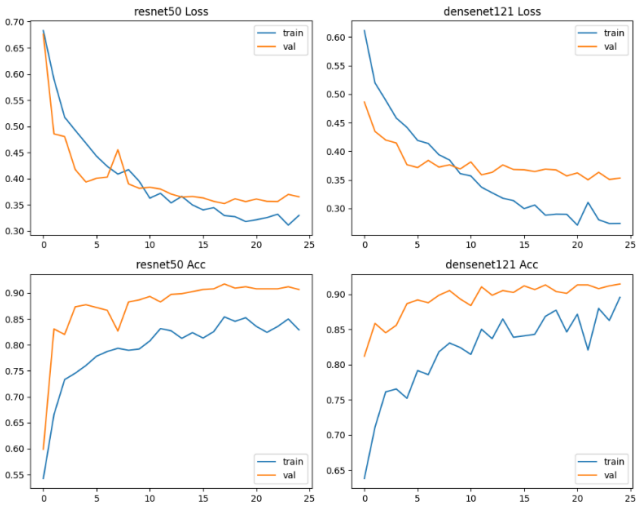
DenseNet121 achieved the highest accuracy, sensitivity and F1 with the smallest checkpoint by a factor of three to four. A plausible mechanism is architectural: dense connectivity keeps early-layer representations directly available at the head, and in ultrasound the discriminative features margin irregularity, microcalcification, subtle echogenicity gradients are fine-grained and high-frequency, precisely the information deep residual bottlenecks discard. Feature reuse rather than capacity appears to be the useful property, which also explains why the largest model performed worst. The broader implication is that at few thousand images the training recipe dominates architectural choice: three architectures with different inductive biases converged to within 0.002 AUC under one protocol, with the same protocol applied to a fourth failed completely.

This bounds what our results support. The 0.0021 ROC-AUC range across three backbones on 750 test images is far smaller than the sampling variability such a text set admits, and we performed neither DeLong testing nor repeated seed runs; we therefore do not claim any of three is superior. The caution applies with greater force to the nodal results, where a sensitivity interval spanning 55-88% is compatible both with clinically useful performance and with performance that would not be.

B. *The operating point matters more than the accuracy*

The clinacally weakest number here is thyroid specificity at 83.18% approximately one benign nodule in six is lfagged malignant, which in an FNA triage aid translates into un-necessary biopsies the burden that motivated weighted at 2.5:1 while screening populations are overwhelmingly benign, postive predicitve value would fall substantially in deployment. The 0.5 treshold is arbirary: tuning tier boundaries on validation data converts it into a deliberate operating point, and a tiered output that defers explicitly is more clinically honest than a binary label.

The nodal module shows the opposite asymmetry and need the opposite correction. Perfect specifity with 75% sensitivity eans it never over calls metastasis but misses one in four. For nodal staging that is the wrong trade-off a missed metastssis may lead to indequate initial suregery, whereas a false positve prompts further imaging so the theshold should move toward sensitivity.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

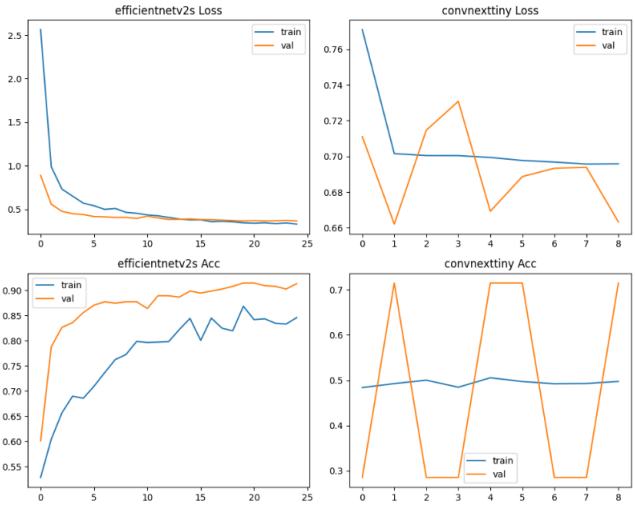


Fig. 4. Training, validation loss and accuracy

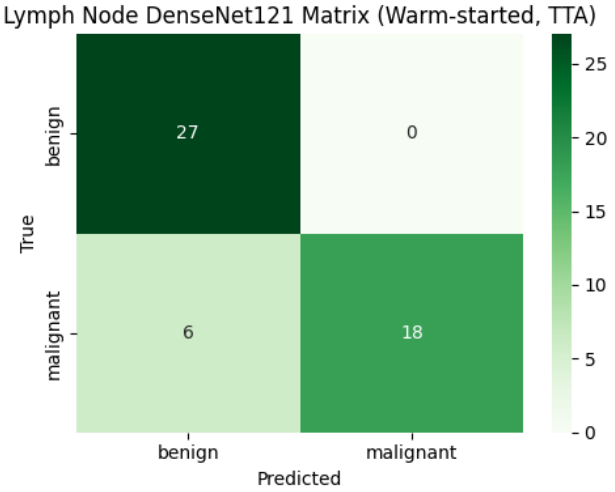


Fig. 5. Confusion matrix for lymph node classification

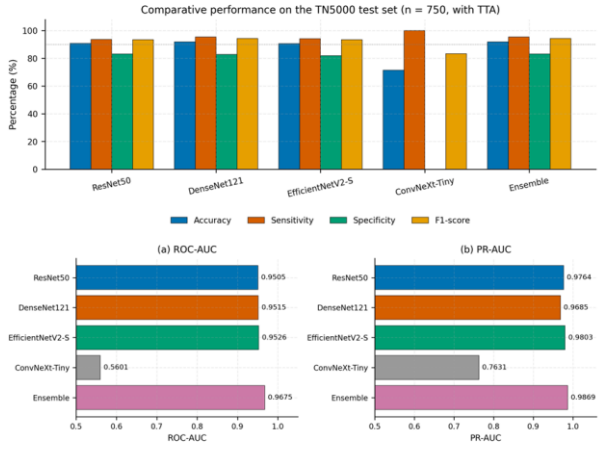


Fig. 6. Comparative performance. Top: accuracy, sensitivity, specificity and F1 by model. Bottom: ROC-AUC and PR-AUC by model

C. Positioning, limitations and future work

THYROX is a triage and second-reader aid, a framing with a stronger base than autonomus operation [19], [9]: the tiered output of (11), a visual explaination permitting a reader

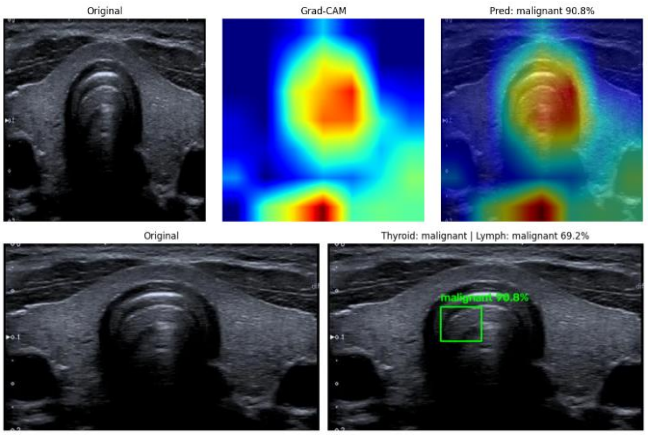


Fig. 7. Qualitative output. Top: original image, Grad-CAM heat map and overlay with predicted class. Bottom: end-to-end output.

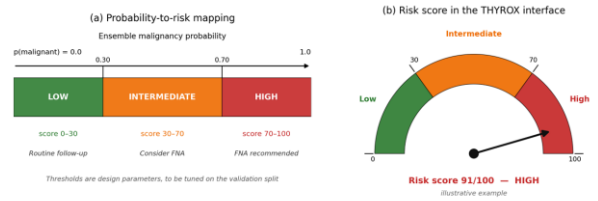


Fig. 8. Risk stratification. (a) Probability-to-risk mapping. (b) Risk score in the THYROX interface.

to reject an implausible prediction, and a footprint compatible with modest harware all fit that role.

The limitations are substantial. All results derive from retrospective secondary analysis of public data; thyroid partitioning is image-level because the release does not expose patient identifiers; and the thyroid model has no external validation. The nodal test set contains 51 images, giving intervals too wise for precise claims. Only B-mode imaging was used elastography, Doppler and the LymphUs masks were not exploited. Results come from a single partition and seed without cross-validation or statistical testing, and the partition used (3,500/750/750) differs from TN5000's official 3,500/500/1,000, limiting comparability. The endpoint is binary, and no comparison against radiologists on these images was performed.

Immediate priorities are therefore an ablation isolating ImageNet from thyroid-ultrasound initialization, repeated seed runs with DeLong testing, and re-evaluation on the official partition. Beyond these, LymphUs's two-centre structure permits genuine externally validation by training on Centre 1 and testing on Centre 2 the strongest credibility improvement available at low cost while the distributed masks would support a segmentation stage far more reliable than box regression, enabling the geometric feature and hybrid classifier stages. Multi-view fusion [27], multimodal fusion [5] and ultrasound foundation model backbones [10]

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

are natural extensions; a prospective reader study would be required before any deployment claim.

VI. CONCLUSION

We presented THYROX, a unified explainable system addressing both components of PTC ultrasound assessment nodule malignancy and cervical nodal status in one pipeline using exclusively public, biopsy confirmed data. A soft-vote ensemble of three convolutional backbones with test time augmentation achieved 91.87% accuracy, 95.34% sensitivity, 83.18% specificity and ROC-AUC 0.9675 on the TN5000 test set, and a nodal classifier warm started from the thyroid backbone reached 88.24% accuracy from only 233 training images indicating that in-domain ultrasound representations offer a practical route to usable performance under label scarcity. Backbone choice within the convolutional family proved immaterial and the smallest model performed best, wile ConvNeXt-Tiny under the shared recipe and single-box regression for localization both failed in ways we report rather than omit. THYROX is positioned as a triage and second reader aid; its specify, absence of external validation and small nodal evaluation preclude any stronger claim

This study is a secondary analysis of publicly released, de-identified datasets; no new patients data were collected and institutional review board approval was not required

REFERENCES

[1] D. W. Chen and M. R. Haymart, “Unravelling the rise in thyroid cancer incidence and adressing overfiagnosis,” Nat. Rev. Endocrinol., vol. 22, no.1, pp. 10-20, 2026.

[2] A. Mohammadi et al., “lyumphUs: a mulitcenter open-access database of lymph node ultrasound images in patients with papillary thyroid carcinoma,” Data Brief, vol. 66, art 112694, 2026.

[3] G. N. Sujini and S. Balakrishna, “automated thyroid nodule classification in ultrasound imaging using a hybrid vision transformer and Wasserstein GAN with gradient penalty,” Sci. Rep., vol 15, art. 40786, 2025.

[4] A. Rashed, T. Medhat, and A. Elgarayhi “enhancing automatic diagnosis of thyroid nodules from ultrasound scans leveraging deep learning models,” Sci. Rep., vol. 15, art. 4036, 2025.

[5] P. Shen et al., “Explainable multimodal deep learning for predicting thyroid cancer lateral lymph node metastasis using ultrasound imaging,” Nat. Commun., vol. 16, art. 7052, 2025.

[6] X. Wang et al., “Ultrasound based artificial intellignece for predicting cervical lymph node metastsis in papillary thyroid cancer: a systematic reviw and meta-analysis,” Front. Endocrinol., vol. 16, art. 1570811, 2025.

[7] H. Zhang, Q Liu, X. Han, L. Niu, and W.Sun, “TN5000: and ultrasound image dataset for thyroid nodule detection and classification,” Sci. Data, vol.12, art. 1437, 2025.

[8] A. Pimpalkar and M. Gulame, “Thyroid lumps detection in ultrasound images utilizing muli-view deep learning approach,” in Prox. IEEE Int. Conf. Recent Adv. Ciout. Syst. (REACS), 2025.

[9] D. Xu et al., “The clinical value of artifical intelligence in assisting junior radiologists in thyroid ultrasound: a multicenter prospective study,” BMC Med., vol. 22, art. 293, 2024.

[10] J. Jiao et al., “USFM: a universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis,” Med. Image Anal., vol. 96, art. 103202, 2024.

[11] G. Dency Flora et al., “Identification of thyroid disorder using deep learning techniques a survey,” in Proc. 10th Int. Conf. Adv. Comput. Commun. Syst.(ICACCS), 2024.

[12] S. Azizi et al., “Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging,” Nat. Biomed. Eng., vol. 7, no. 6, pp. 756-779, 2023.

[13] W. Li et al., “Systematic review and meta-analysis of ACR TI-RADS inter-reader reliabilty for risk stratification of thyroid nodules,” Front. Oncol., vol. 12, art.840516, 2022.

[14] A. Saporta et al., “Benchmaking saliency methods for chest X-ray interpretation,” Nat. Mach.Intell., vol. 4, no. 10, pp. 87=67-878, 2022.

[15] Z. Liu et al., “A ConvNet for the 2020s,” in Proc. IEEE/CVF CVPR, 2022, pp. 11966-11976.

[16] M. Tan and Q. V. Le, “EffiicientNetV2: smaller models and faster training,” in Proc. ICML, PMLR vol. 139,2021, pp. 10096-10106.

[17] R. R Selvaraju et al., “Grad-CAM: visual explanations from deep networks via gradient-based localization,” Int. J. Comput. Vis., vol. 128, no. 2, pp. 336-359, 2020.

[18] H. A Nugroho, E. l. Frannita, and A. H. T. Hutami, “Thyroid nodules catego rization based oargin features using deep learning,” in Proc. 3rd Int. Semin. Res. Inf. Technol. Intell. Syst. (ISRITI), 2020, pp. 499-504.

[19] M. Budha et al., “Management of thyroid nodules seen on US images deeo kearing may match performance of radiologists,” Radiology, vol. 292, no. 3, pp. 695-701, 2019.

[20] M. Raghy, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: understanding tranfer learning for medical imaging,” in Adv. Neural Inf. Process. Syst., vol. 32, 2019.

[21] J. K. Hoang et al., “Reduction in thyroid nodule biopsies and improved accuracy with ACR TI-RADS,” Radiology, vol.287, no. 1, pp. 185-193, 2018.

[22] J. H. Lee et al., “Deep learning-based computer-aided diagnosis system for localization and diagnosis of metastatic lymph nodes on ultrasound: a pilot study,” Thyroid vol. 28, no. 10 pp. 1332-1338, 2018.

[23] H. Zhang, M. Ciss’e, Y. N. Dauplin, and D. Lopez-Paz, “mixup: beyond epirical risk minimization,” in Proc. ICLR, 2018.

[24] E. S. Cibas and S. Z. Ali, “The 2017 Bethesda System for Reporting Thyroid Cytopathology,” Thyroid, vol. 27, no. 11, pp. 1341-1346, 2017.

[25] F. N. Tessler el al.,”ACR Thyroid Imaging, Reporting and Data System (TIRADS): white paper of the ACR TI-RADS Committee,” J. Am. Coll. Radiol., vol. 14, no. 5, pp. 587-595, 2017.

[26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Qeinberger, “Densely connected convolutional networks,” in Proc. IEEE CVPR, 2017, pp. 2261-2269.

[27] B. R. Haufen et al., “2015 American Thyroid Association management guidelies for adult patients with thyroid nodules and diffrentiated thyroid cancer,” Thyroid, vol. 26, no. 1, pp. 1-133, 2016.

[28] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE CVPR, 2016, pp.770-778.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)