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.

Cross-cohort evaluation of thyroid ultrasound deep learning across different outcome definitions: a duplicate-controlled benchmark

Behnam Kiani Kalejahi

b.kiani@centralasian.uz

Central Asian University

Mohammad Javad Rajabi

INTI International University

Research Article

Keywords: Papillary thyroid carcinoma, Thyroid ultrasound, Deep learning, External validation, Reference standard, Probability calibration

Posted Date: September 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10733698/v1

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

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Medical Imaging on October 16th, 2026. See the published version at https://doi.org/10.1186/s12880-026-02944-6.

Abstract

Background.

Measured performance of thyroid ultrasound artificial intelligence depends on both cohort characteristics and the evaluation endpoint. We assessed frozen models across public cohorts with differing label provenance, separating evaluation against malignancy outcomes from agreement with ultrasound risk categories.

Methods.

Five backbones were trained on 387 thyroid ultrasound images from 385 public cases (222 papillary thyroid carcinoma, 163 benign) labelled by postoperative diagnosis; 1,332 fine-needle aspiration cytology blocks from the same archive formed a separate benchmark. Exact and perceptual near-duplicates were grouped before image-level partitioning; per-image patient identifiers were unavailable, so patient-level independence could not be confirmed. Frozen models were evaluated on a TN3K subset (n = 1,228; dataset-provided benign/malignant labels) and for concordance with an endpoint derived from radiologist Thyroid Imaging Reporting and Data System (TI-RADS) categories in DDTI (n = 637), without retraining or threshold adjustment.

Results.

On TN3K the ensemble reached an area under the receiver operating characteristic curve (AUROC) of 0.825 (95% confidence interval 0.800–0.849) and ResNet-50 reached 0.888 (0.870–0.904). Against the TI-RADS-derived endpoint, AUROC was 0.477 and 0.437, indicating little agreement with that categorisation. Because cohort, acquisition and endpoint changed together, this contrast cannot be attributed to label definition alone. The cytology benchmark reached AUROC 0.986 (0.970–0.998) against 0.733 for the ultrasound ensemble, a descriptive non-paired contrast.

Conclusions.

Frozen models performed differently across cohorts differing in acquisition and outcome definition. The findings support explicit reporting of label provenance, cautious interpretation of cross-cohort contrasts, and patient-level external evaluation against a clinically appropriate reference standard.

1 Introduction

Thyroid nodules are common, and their assessment typically begins with ultrasound risk stratification, followed by fine-needle aspiration (FNA) when clinical and sonographic criteria indicate sampling [1], [3]. The American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS)

standardises ultrasound descriptors and management-oriented risk categories [3], whereas the Bethesda System provides a separate framework for reporting thyroid FNA cytology [2]. These systems support related but distinct stages of the diagnostic pathway.

Artificial intelligence has been investigated extensively for thyroid ultrasound interpretation [4], [5] and cytology image classification [6]–[8]. Models that perform well in development data may perform differently in independent cohorts because of changes in population, equipment, acquisition, preprocessing, disease spectrum and label definition. Independent evaluation is therefore an important component of assessing potential clinical transportability [21]–[25]. In particular, postoperative disease outcomes and radiologist-assigned TI-RADS categories are not interchangeable endpoints: the former records a diagnostic outcome, whereas the latter expresses sonographic suspicion and guides management. Evaluation against either endpoint can be informative, but the resulting metrics answer different questions. Calibration should also be assessed separately from discrimination [19], [22].

This study aimed to benchmark duplicate-controlled thyroid ultrasound classifiers internally and across two independent public cohorts, while separating evaluation against dataset-provided benign/malignant labels from agreement with a TI-RADS-derived endpoint. A secondary aim was to report a descriptive, non-paired cytology benchmark from the development archive. Because each external cohort differs from the development data in several respects, the analysis was designed to characterise cross-cohort performance rather than to isolate a causal effect of label definition.

Relationship to prior work. This study follows earlier work by the authors on multimodal deep learning for papillary thyroid carcinoma using ultrasound and cytology [30]. The present manuscript differs in aim and design: it is a duplicate-controlled image-level benchmark whose subject is cross-cohort evaluation and endpoint provenance rather than multimodal diagnostic performance. Its distinct contributions are the exact- and perceptual-hash duplicate control described in Section 3.3, the evaluation of frozen models on two independent public cohorts without retraining or threshold adjustment, and the explicit separation of evaluation against malignancy outcomes from agreement with a TI-RADS-derived endpoint.

2 Related Work and Positioning

2.1 Thyroid ultrasound AI

Deep learning for thyroid ultrasound has progressed from single-image classifiers to multi-centre and multi-view systems. In a landmark multicohort study, Li et al. trained deep convolutional neural networks on sonographic images from three Chinese hospitals and reported sensitivity and specificity comparable to, or exceeding, skilled radiologists [4]. Peng et al. subsequently developed a multicentre deep-learning system to assist thyroid nodule diagnosis and management, again demonstrating radiologist-level discrimination across external cohorts [5]. Despite these advances, ultrasound interpretation remains sensitive to operator technique, scanner settings, nodule boundaries, interface overlays, and the absence of standardised ROI annotations, factors that depress image-only generalisation and motivate segmentation-guided and metadata-fused approaches [9], [10].

2.2 Thyroid cytopathology AI

Cytopathology AI is an equally active area. Lin et al. proposed a fast deep-learning screening framework operating on FNA smears and ThinPrep whole-slide images for PTC [6]; Duc et al. reported that AdaBoost-style ensembles of CNNs combined with stain normalisation achieved accuracies up to 0.997 on FNAC images [7]; and Guan et al. demonstrated an early VGG-16 pilot for differential diagnosis of PTC in cytological images [8]. The comprehensive review of Slabaugh et al. synthesises this body of work and emphasises that most studies are retrospective and that dataset curation, cytological interpretability, and workflow validation remain open problems [9]. Mechanistically, cytology exposes nuclear features, ground-glass nuclei, grooves, intranuclear inclusions, and chromatin clearing, that are more directly tied to the PTC diagnosis than gross sonographic texture, which is the biological rationale for our direct modality comparison.

2.3 Multimodal fusion

Multimodal models that combine clinical, radiological and deep ultrasound features may outperform single-modality baselines [10]. Valid ultrasound–cytology fusion, however, requires verified pairing of modalities within the same patients. Because the present archive does not provide usable per-image patient links, we analysed each modality separately and did not create artificial cross-modality pairs. In earlier work we examined multimodal deep learning for papillary thyroid carcinoma diagnosis combining ultrasound and cytology [30]; the present study is not a fusion model, and treats the two streams as separate benchmarks because the archive provides no verified cross-modality patient links.

2.4 Architectures, explainability, and calibration

ConvNeXt modernises convolutional networks using design choices inspired by vision transformers [11]; EfficientNet uses compound scaling to balance depth, width, and resolution [12]; the Swin Transformer applies shifted-window hierarchical attention [13]; ResNet introduces residual connections for stable deep training [14]; and DenseNet promotes feature reuse through dense connectivity [29]. For interpretability, Grad-CAM produces class-discriminative localisation maps from gradients flowing into convolutional features [18]. For probability reliability, temperature scaling is a simple, monotonic post-hoc calibration method that improves confidence estimates without changing rank-based metrics [19].

2.5 Study positioning

Table 1 contrasts representative studies with the present benchmark. The emphasis here is not a new architecture, but transparent evaluation across modern backbones, duplicate and near-duplicate control, probability calibration, and cautious cross-cohort interpretation with explicit attention to endpoint provenance.

Table 1
Positioning of the present study relative to representative prior thyroid imaging-AI work

Study	Year	Modality	Task	Validation design	Leakage / calibration safeguards	Headline result
Li et al. [4]	2019	Ultrasound	Cancer vs non-cancer	Retrospective multicohort	Not reported	≈ expert-radiologist Se/Sp
Peng et al. [5]	2021	Ultrasound	Benign vs malignant	Multicentre, prospective arm	Not reported	Radiologist-level, multicentre
Lin et al. [6]	2021	Cytology (WSI)	PTC screening	FNA + ThinPrep slides	Slide split	Fast automated screening
Duc et al. [7]	2022	Cytology (FNAC)	PTC vs benign	Single-source split	Not reported	Ensemble acc. up to 0.997
Guan et al. [8]	2019	Cytology	PTC differential dx	Pilot study	Not reported	VGG-16 proof-of-concept
This work	2026	Cytology + US (head-to-head)	Benign vs PTC	Duplicate-controlled image split; bootstrap; 5-fold CV; cross-cohort evaluation	Exact hash + pHash; ECE, NLL, Brier	Cytology AUROC 0.986; US ensemble 0.733 internal and 0.825 on TN3K

Note. WSI denotes whole-slide image; FNAC denotes fine-needle aspiration cytology; pHash denotes perceptual hash; Se/Sp denotes sensitivity/specificity; ECE denotes expected calibration error; NLL denotes negative log-likelihood. Headline results are summarised from the cited reports and the present experiments.

3 Methods

3.1 Study design and end-to-end analysis pipeline

The study used a fixed analysis pipeline executed in one computational environment (Python 3.10, PyTorch 2.5.1, CUDA 12.1, single NVIDIA RTX 4060 GPU; software manifest released with the public code). The stages were: (i) acquisition and harmonisation of the Mendeley Data archive [26]; (ii) image-level partitioning after exact-hash and perceptual-hash grouping; (iii) modality-specific training of five convolutional and transformer backbones; (iv) construction of a mean-probability ensemble; (v) internal image-level evaluation; and (vi) frozen-model evaluation on TN3K and DDTI. TN3K was treated as an external cohort benchmark against its distributed benign/malignant labels. DDTI was treated separately as an endpoint-concordance analysis against a binary grouping derived from radiologist TI-RADS categories, not as validation of malignancy diagnosis.

Modality-specific scope. Cytology and ultrasound were analysed as separate image-classification tasks. The archive does not provide verified per-image patient links across or within the two streams. Accordingly, the cytology-versus-ultrasound contrast is descriptive at the image-distribution level and does not estimate within-patient comparative diagnostic accuracy.

Backbone architectures. We trained five image-classification backbones spanning convolutional and transformer families to control for architectural inductive bias: (1) ConvNeXt-Small [11], a modernised convolutional design with depthwise convolutions, layer normalisation, and GELU activations; (2) EfficientNet-B3 [12], a compound-scaled convolutional architecture optimised for parameter-efficient performance; (3) Swin-Tiny [13], a hierarchical vision transformer with shifted-window self-attention; (4) ResNet-50 [14], the canonical deep residual architecture included as a stable comparator; and (5) DenseNet-121 [29], a densely connected convolutional network with strong feature reuse. All five models were initialised from publicly available ImageNet-pretrained weights via the timm library [16], with the original classification head replaced by a single linear logit head for binary classification. Pretrained-from-natural-image initialisation is justified by the small per-modality training set sizes and is consistent with current practice in medical image classification under similar data regimes.

Pre-processing and augmentation. Images were converted to three-channel RGB and resized to 224 × 224 pixels with bilinear interpolation. This resolution was selected because it is the native input scale of the ImageNet-pretrained backbones and allowed identical preprocessing across architectures under the available sample size and hardware constraints [15], [16]. Training augmentation comprised random horizontal flips, rotations of ± 8°, mild colour jitter and random crops from a slightly oversized resize; validation and test images received only deterministic resizing and ImageNet normalisation. Downsampling may reduce the visibility of punctate echogenic foci, thin halos and subtle margin irregularity. The results therefore apply to this input resolution; evaluation at higher resolution and with expert nodule segmentation is needed before clinical use.

Training protocol. Each model was trained for up to 40 epochs with AdamW (learning rate 3×10^{-5}; weight decay 1×10^{-4}) and a cosine learning-rate schedule. The loss was class-weighted binary cross-entropy with logits, and a WeightedRandomSampler was used to mitigate class imbalance. The best checkpoint was selected by validation AUROC and then evaluated once on the held-out image-level test partition. Random seeds (seed = 20251019) were fixed and cuDNN determinism enabled.

Ensemble construction. The duplicate-controlled ensemble averaged sigmoid probabilities from ConvNeXt-Small, EfficientNet-B3, Swin-Tiny and ResNet-50. Ensemble construction used no test or external labels. DenseNet-121 was retained as a separate comparator in accordance with the original four-backbone ensemble specification.

Evaluation and reporting. Discrimination metrics (AUROC and AUPRC) were complemented by accuracy, sensitivity, specificity and F1 at the Youden operating point selected on the validation partition, and by calibration metrics (Brier score, expected calibration error (ECE) and reliability diagrams). Ninety-five percent confidence intervals used 2,000 non-parametric bootstrap resamples of image-level predictions.

The non-paired cytology-versus-ultrasound contrast used 5,000 independent resamples and is reported descriptively; it does not identify a modality effect. Reporting was informed by STARD 2015 [21], CLAIM [22], [23], STARD-AI [24] and TRIPOD + AI [25].

3.2 Datasets, image counts and reference endpoints

Development archive. The public Mendeley Data release by Pang et al. [26] contains 385 source cases: 222 with a postoperative diagnosis of PTC and 163 with a postoperative diagnosis of a benign thyroid nodule. These postoperative diagnoses—not FNA cytology, core biopsy or TI-RADS—defined the binary case labels. The archive contains 387 B-mode ultrasound images and 1,332 ultrasound-guided FNA cytology H&E image blocks, giving 1,719 analysed images in total (Fig. 1). Image labels were inherited from the benign/PTC folder structure. Bethesda categories were not used to create the binary target; therefore no Bethesda categories were collapsed, and indeterminate cytology categories were not separately represented. The image counts imply that cases could contribute multiple cytology blocks and that more than one ultrasound image existed for at least some cases, but the released image manifest did not provide stable per-image case identifiers. Exact image-to-case multiplicity and cross-modality pairing could therefore not be reconstructed. Dataset provenance and per-modality characteristics are summarised in Table 2, and the duplicate-controlled image-level split distribution by modality and label is reported in Table 3.

Table 2
Development-dataset provenance and per-modality summary

Item	Cytology	Ultrasound
Number of images (this study)	1,332	387
Source case counts (whole dataset)	Shared source cohort: 222 postoperative PTC diagnoses + 163 postoperative benign diagnoses	Shared source cohort: 222 postoperative PTC diagnoses + 163 postoperative benign diagnoses
Image content	FNA cytology, H&E scanning blocks	B-mode thyroid ultrasound
Binary label reference	Folder-inherited postoperative diagnosis; Bethesda categories not used	Folder-inherited postoperative diagnosis; TI-RADS not used
Image-to-case identifiers	Unavailable in analysis manifest	Unavailable in analysis manifest
Resolution	Variable; 224×224 (384×384 for ROI)	Variable; 224×224 (384×384 for ROI)
Acquisition source	Mendeley Data, Pang et al. 2025 [26]	Mendeley Data, Pang et al. 2025 [26]
Licence	CC BY 4.0	CC BY 4.0
Patient-level split verification	Not possible	Not possible

Note. Source-case labels were based on postoperative diagnoses reported by Pang et al. [26]. Per-image case identifiers were unavailable in the analysis manifest; the 1,719-image total is 387 ultrasound images plus 1,332 cytology blocks and does not represent 1,719 independent patients.

Table 3
Duplicate-controlled image-level split distribution by modality and label

Split	Modality	Benign (0)	PTC (1)
Train	Cytology	444	488
Train	Ultrasound	114	154
Validation	Cytology	96	104
Validation	Ultrasound	25	34
Test	Cytology	95	105
Test	Ultrasound	26	34

Note. PTC denotes papillary thyroid carcinoma. Exact-hash and perceptual-hash groups did not cross partitions; different images from the same patient may nevertheless have crossed because patient identifiers were unavailable.

3.3 Duplicate and near-duplicate control and independence

Exact duplicates were identified by MD5 hash and visually near-identical images by perceptual hash. Images sharing either signature were assigned to one group and kept in a single partition. Preliminary analysis identified 29 exact-duplicate rows (three groups crossing the original split) and 33 same-perceptual-hash rows (four groups crossing the original split). After regrouping, no detected duplicate group crossed the train, validation and test partitions (Table 4). This procedure controls detected image-level duplication; it cannot establish patient-level independence because different images from one patient need not share a hash and per-image patient identifiers were unavailable. Internal estimates should therefore be interpreted as duplicate-controlled image-level performance and may remain optimistic if unrecognised within-patient overlap exists.

Table 4
Dataset composition and duplicate-control summary

Item	Value
Total images in final metadata	1,719
Cytology images	1,332
Ultrasound images	387
Benign labels (class 0)	800
PTC labels (class 1)	919
Detected duplicate groups crossing final splits	0
Original exact duplicate rows detected	29
Original cross-split exact duplicate groups (pre-safe-split)	3
Original same perceptual-hash rows detected	33
Original cross-split perceptual-hash groups (pre-safe-split)	4
Note. Perceptual hashing extends protection beyond byte-identical files but is not a substitute for patient identifiers.	

3.4 Model architectures and training configuration

Every single-image classifier used the encoder configured with global average pooling followed by a binary head: LayerNorm → Linear(features, 384) → GELU → Dropout(0.35) → Linear(384, 1), with the final logit mapped to a PTC probability via the sigmoid function. The cytology ensemble averaged predicted PTC probabilities using a by-row-order scheme, avoiding the duplicated case-ID Cartesian-merge inflation that affected earlier naive ensembles. Table 5 lists the backbones and rationale, and Table 6 the training configuration. Models were optimised with AdamW [17] under a cosine-annealing schedule with class-weighted binary cross-entropy and a WeightedRandomSampler; operating thresholds were selected on validation using the Youden index and applied unchanged to the held-out test set (and reselected within each fold for cross-validation).

Table 5
Model families and rationale

Backbone	Used for	Rationale
ConvNeXt-Tiny/Small	Cytology and ultrasound	Modern convolutional baselines; strong transfer learning.
EfficientNet-B3	Cytology and primary ultrasound	Compound-scaled CNN; strong efficiency–performance trade-off.
Swin-Tiny	Cytology and ultrasound	Hierarchical shifted-window transformer backbone.
ResNet-50	Comparator / strongest single	Canonical residual network; stable reference.
DenseNet-121	Comparator baseline	Densely connected CNN with strong feature reuse.
Duplicate-controlled mean-probability ensemble	Cytology and ultrasound	Averages by-order probabilities; avoids case-ID merge artefacts.

Note. All encoders were ImageNet-pretrained and fine-tuned with an identical classification head and training protocol.

Table 6
Training and optimisation configuration

Component	Configuration
Input size	224×224 for training; ROI crops resized to 384×384 before the final transform.
Normalisation	ImageNet mean/std [0.485, 0.456, 0.406] / [0.229, 0.224, 0.225].
Augmentation	Resize, random horizontal flip, random rotation (± 8°), mild colour jitter.
Loss	BCEWithLogitsLoss with positive-class weighting.
Sampling	WeightedRandomSampler to reduce class-imbalance effects.
Optimiser	AdamW; learning rate 3×10^{-5}; weight decay 1×10^{-4}.
Scheduler	CosineAnnealingLR over the configured epoch budget.
Regularisation	Dropout 0.35 in the head; gradient clipping at 1.0.
Hardware mode	CUDA when available; autocast mixed precision and GradScaler.
Threshold selection	Youden index on validation; final metrics on held-out test.

Note. Each model was trained on the train split with up to 40 epochs and a batch size of 8; the same recipe was applied across modalities for fairness.

3.5 Ultrasound ROI-cropping experiment

As an exploratory preprocessing study, ultrasound images were automatically cropped to a texture-based region of interest using grayscale conversion, Gaussian blurring, edge/texture response estimation, central masking, dilation, and bounding-box extraction of the largest textured region, with a central-crop fallback. The aim was to suppress peripheral interface regions, scanner overlays, and border artefacts and to focus learning on central sonographic tissue. This automatic crop is not equivalent to expert nodule segmentation; it tests whether reducing non-informative content improves ultrasound-only classification.

3.6 Calibration, explainability, and statistical testing

Post-hoc temperature scaling [19] was fitted by minimising negative log-likelihood on validation logits and assessed on the held-out image-level test partition. Calibration was quantified with Brier score, negative log-likelihood and ECE. Grad-CAM [18] was used for exploratory visualisation only and was not independently reviewed by specialists. Because cytology and ultrasound samples were neither patient-paired nor matched for sample size, acquisition or image generation, a paired DeLong comparison [20] was inappropriate. Independent bootstrap resampling was used only to describe the observed difference between the two image distributions; no causal or clinical modality-superiority inference was made.

3.7 Evaluation metrics

Performance was measured with AUROC, AUPRC, accuracy, precision, sensitivity, specificity, F1 and Brier score. Confidence intervals were estimated from 2,000 bootstrap resamples for selected models, and five-fold stratified, duplicate-controlled cross-validation was performed for the primary cytology ConvNeXt-Small model. Formal metric definitions are provided in Appendix A.

4 Results

4.1 Main held-out test performance

Selected duplicate-controlled image-level results are summarised in Table 7. On the cytology test partition (n = 200 images), ConvNeXt-Small achieved AUROC 0.988 and AUPRC 0.990, and the four-model ensemble achieved AUROC 0.986 and AUPRC 0.989. On the ultrasound test partition (n = 60 images), EfficientNet-B3 achieved AUROC 0.745 and AUPRC 0.802; the ultrasound ensemble used for external evaluation achieved AUROC 0.733. These are image-level estimates with duplicate control, not verified patient-level estimates; uncertainty is consequently greatest for ultrasound, and any residual within-patient correlation is unquantified.

Table 7
Selected duplicate-controlled held-out image-level performance

Model	Modality	AUROC	AUPRC	Acc.	Sens.	Spec.	F1	Brier	Thr.
Cytology ConvNeXt-Small	Cytology	0.988	0.990	0.975	0.981	0.968	0.976	0.030	0.925
Cytology Ensemble v2	Cytology	0.986	0.989	0.980	0.981	0.979	0.981	0.031	0.725
Cytology ConvNeXt-Tiny	Cytology	0.982	0.988	0.965	0.971	0.958	0.967	0.034	0.935
Cytology Swin-Tiny	Cytology	0.980	0.982	0.955	0.962	0.947	0.957	0.045	0.135
Cytology EfficientNet-B3	Cytology	0.976	0.972	0.935	0.933	0.937	0.938	0.062	0.515
Ultrasound EfficientNet-B3	Ultrasound	0.745	0.802	0.667	0.647	0.692	0.688	0.265	0.685

Note. Thr. is the validation-selected operating threshold. Acc., Sens. and Spec. denote accuracy, sensitivity and specificity. The four-model ensemble averages probabilities by image-row order and avoids duplicate case-ID merge artefacts.

4.2 Final-threshold confusion-matrix analysis

At the validation-selected threshold of 0.725, the cytology ensemble correctly classified 93 of 95 benign images and 103 of 105 PTC images. At threshold 0.925, cytology ConvNeXt-Small correctly classified 92 of 95 benign images and 103 of 105 PTC images. The ultrasound EfficientNet-B3 model at threshold 0.685 correctly classified 18 of 26 benign images and 22 of 34 PTC images (Table 8). These counts are image based and should not be interpreted as numbers of independent patients.

Table 8
Confusion matrices at the final selected operating thresholds

Model	Thr.	TN	FP	FN	TP	Acc.	Sens.	Spec.	Prec.
Cytology ConvNeXt-Small	0.925	92	3	2	103	0.975	0.981	0.968	0.972
Cytology Ensemble	0.725	93	2	2	103	0.980	0.981	0.979	0.981
Ultrasound EfficientNet-B3	0.685	18	8	12	22	0.667	0.647	0.692	0.733

Note. TN, FP, FN, and TP denote true negatives, false positives, false negatives, and true positives. Prec. denotes precision.

4.3 Bootstrap confidence intervals

Bootstrap resampling (2,000 replicates) yielded narrow intervals for the cytology image distribution and wider intervals for ultrasound (Table 9). The cytology ensemble achieved AUROC 0.986 (95% CI 0.970–0.998) and AUPRC 0.989 (95% CI 0.975–0.999); ultrasound EfficientNet-B3 achieved AUROC 0.745 (95% CI 0.612–0.865). Independent resampling estimated a descriptive cytology-minus-ultrasound AUROC difference of 0.247 (95% CI 0.120–0.384). Because the two image sets were not patient-paired and differ in size and image-generating process, this contrast does not establish relative clinical diagnostic performance.

Table 9
Bootstrap 95% confidence intervals for selected final models

Model	AUROC	AUPRC	Accuracy	Sensitivity	Specificity
Cytology ConvNeXt-Small	0.988 (0.976–0.998)	0.990 (0.979–0.998)	0.975 (0.950–0.995)	0.981 (0.952–1.000)	0.968 (0.929–1.000)
Cytology Ensemble v2	0.986 (0.970–0.998)	0.989 (0.975–0.999)	0.980 (0.960–0.995)	0.981 (0.952–1.000)	0.979 (0.947–1.000)
Ultrasound EfficientNet-B3	0.745 (0.612–0.865)	0.802 (0.664–0.907)	0.667 (0.550–0.783)	0.647 (0.485–0.800)	0.692 (0.500–0.870)

Note. Values are point estimate (95% confidence interval). Threshold-dependent metrics use the final validation-selected thresholds.

4.4 Five-fold cross-validation

Five-fold stratified, duplicate-controlled cross-validation produced mean cytology AUROC 0.982 ± 0.008 and mean AUPRC 0.985 ± 0.007 (Table 10). The low between-fold variance indicates stability with respect to the defined image groups, but patient-level stability cannot be assessed without identifiers.

Table 10
Five-fold cross-validation summary for cytology ConvNeXt-Small

Metric	Mean ± SD	Minimum	Maximum
AUROC	0.982 ± 0.008	0.974	0.994
AUPRC	0.985 ± 0.007	0.976	0.995
Accuracy	0.939 ± 0.020	0.921	0.974
Precision	0.939 ± 0.024	0.899	0.958
Sensitivity	0.945 ± 0.032	0.906	0.993
Specificity	0.932 ± 0.030	0.882	0.953
F1	0.942 ± 0.019	0.927	0.975
Brier	0.060 ± 0.020	0.026	0.079

4.5 Ultrasound ROI-cropping experiment

Automatic ROI cropping increased ultrasound AUROC from 0.745 to 0.817 and AUPRC from 0.802 to 0.850, with accuracy increasing from 66.7% to 70.0% and Brier score improving from 0.265 to 0.199 (Table 11). Sensitivity increased and specificity decreased slightly. This exploratory result suggests that reducing peripheral interface content may help at this resolution; it does not establish the benefit of expert segmentation.

Table 11
Exploratory ultrasound ROI-cropping experiment (EfficientNet-B3)

Preprocessing	AUROC	AUPRC	Accuracy	Sensitivity	Brier
Standard full-image resize	0.745	0.802	0.667	0.647	0.265
Automatic texture/central ROI crop	0.817	0.850	0.700	0.735	0.199
Note. Automatic ROI cropping is exploratory and is not equivalent to expert nodule segmentation.					

4.6 Calibration and temperature scaling

Temperature scaling with an optimal temperature of 2.414 left AUROC and AUPRC unchanged but improved probability reliability: ECE fell from 0.0317 to 0.0138, negative log-likelihood from 0.2150 to 0.1262, and Brier score from 0.0296 to 0.0284 (Table 12). The small decrease in accuracy at the fixed threshold of 0.925 reflects the changed probability scale; thresholds should be reselected after calibration if calibrated probabilities are used clinically.

Table 12
Post-hoc temperature scaling for cytology ConvNeXt-Small

Condition	Temp.	AUROC	AUPRC	Acc.@0.925	Brier	NLL	ECE
Uncalibrated	1.000	0.988	0.990	0.975	0.030	0.215	0.032
Temperature-scaled	2.414	0.988	0.990	0.960	0.028	0.126	0.014

Note. Temperature scaling improves probability calibration and does not change rank-based AUROC/AUPRC.

4.7 Descriptive cytology-versus-ultrasound contrast

Independent bootstrap resampling estimated a cytology-minus-ultrasound AUROC difference of 0.247 (95% CI 0.120–0.384) and AUPRC difference of 0.177 (95% CI 0.054–0.337) (Table 13). These quantities compare two non-paired image distributions with different sample sizes and image-generating processes. They are reported to describe this archive and protocol only; they do not show that cytology is intrinsically superior to ultrasound or estimate a within-patient difference.

Table 13
Bootstrap comparison between cytology and ultrasound performance

Comparison	Metric	Diff.	CI low	CI high	Scope
Cytology ConvNeXt-Small – Ultrasound EfficientNet-B3	AUROC	0.247	0.120	0.384	Non-paired
Cytology ConvNeXt-Small – Ultrasound EfficientNet-B3	AUPRC	0.177	0.054	0.337	Non-paired

Note. Differences used 5,000 independent bootstrap resamples. The analysis is descriptive because the modalities were not patient-paired and patient-level independence was not verifiable.

4.8 Grad-CAM explainability

Grad-CAM indicated that cytology ConvNeXt-Small attended primarily to cellular clusters and nuclear-rich regions (Fig. 2). In false-positive cytology examples, activation often concentrated on dense, crowded cellular clusters, a plausible source of malignant misprediction. Ultrasound Grad-CAM was more variable, with activation sometimes extending into superficial or interface regions (Fig. 3), consistent with the lower discriminative performance of ultrasound-only classification. These maps are used to surface failure modes, not as causal explanations of the diagnosis.

4.9 Cross-cohort evaluation and TI-RADS endpoint concordance

Frozen ultrasound models were applied to two public datasets without retraining or threshold refitting. The TN3K classification subset contained 1,228 images with dataset-provided binary labels (614 benign, 614 malignant). Public TN3K documentation establishes expert image annotation but does not document confirmatory pathology for every classification image; TN3K is therefore described here by its distributed

labels rather than as uniformly biopsy-confirmed. The processed DDTI set contained 637 images and a binary endpoint formed by grouping lower-suspicion TI-RADS categories (2–3) versus higher-suspicion categories (4–5, including 4a–4c), yielding 260 and 377 images. DDTI was originally annotated by radiologists using TI-RADS; its source report states that malignant diagnoses were biopsy-confirmed, but a TI-RADS threshold is not a malignancy reference standard for all nodules [28]. Accordingly, DDTI AUROC measures concordance with the grouped TI-RADS endpoint. The public datasets come from separate sources, with no known overlap with the development archive, but absence of patient overlap could not be verified from identifiers. Frozen-model performance on TN3K and concordance with the DDTI TI-RADS-derived endpoint are reported in Table 14.

Table 14
Frozen-model performance on TN3K and concordance with the DDTI TI-RADS-derived endpoint

Metric	Internal Ens.	TN3K ensemble	TN3K ResNet-50	DDTI ensemble (TI-RADS endpoint)	DDTI ResNet-50 (TI-RADS endpoint)
AUROC	0.733	0.825 (0.800–0.849)	0.888 (0.870–0.904)	0.477 (0.434–0.523)	0.437 (0.391–0.482)
AUPRC	0.791	0.866 (0.844–0.887)	0.883 (0.862–0.904)	0.573 (0.525–0.628)	0.537 (0.489–0.588)
Accuracy	0.672	0.763 (0.739–0.787)	0.770 (0.748–0.792)	0.444 (0.407–0.480)	0.408 (0.372–0.446)
Sensitivity	0.636	0.609 (0.570–0.649)	0.718 (0.683–0.754)	0.297 (0.255–0.342)	0.276 (0.234–0.320)
Specificity	0.720	0.917 (0.895–0.939)	0.823 (0.794–0.852)	0.658 (0.599–0.712)	0.600 (0.538–0.659)
F1	0.689	0.720 (0.689–0.751)	0.758 (0.729–0.784)	0.388 (0.339–0.435)	0.356 (0.308–0.403)
Brier (↓)	0.220	0.177 (0.167–0.187)	0.160 (0.147–0.173)	0.311 (0.296–0.326)	0.377 (0.355–0.399)
ECE (↓)	0.172	0.129 (0.109–0.152)	0.127 (0.111–0.151)	0.232 (0.197–0.271)	0.330 (0.292–0.368)

Note. Internal results are duplicate-controlled image-level estimates (n = 60). TN3K results use the dataset-provided benign/malignant label; universal pathology confirmation was not established in the public documentation reviewed. DDTI results use a TI-RADS-derived binary endpoint and therefore quantify agreement with suspicion categories, not diagnostic accuracy for malignancy. External intervals use 2,000 bootstrap resamples. The internal validation-selected threshold was held fixed.

On TN3K, the ultrasound ensemble achieved AUROC 0.825 and ResNet-50 achieved 0.888. Brier score and ECE were lower than on the internal ultrasound partition, although calibration comparisons across cohorts are influenced by prevalence and case mix. Against the DDTI TI-RADS-derived endpoint, AUROC was 0.477 and 0.437, respectively, indicating little agreement with that categorisation. This result is not evidence that the models fail to diagnose malignancy on DDTI because TI-RADS grouping was not used as a malignancy reference standard. Conversely, the higher TN3K estimates support transportability only to the TN3K subset and its distributed labels. Cohort, equipment, acquisition, preprocessing, disease spectrum and endpoint all changed together; the observed contrast therefore cannot be causally attributed to label semantics.

5 Discussion

Frozen thyroid ultrasound models performed differently across the TN3K and DDTI analyses. The TN3K AUROCs were higher than internal image-level estimates, whereas agreement with the DDTI TI-RADS-derived endpoint was near chance. These metrics answer different questions: TN3K evaluates discrimination against its distributed benign/malignant labels, while DDTI evaluates concordance with an ultrasound suspicion grouping. The DDTI result should therefore not be presented as diagnostic validation against malignancy.

The contrast nevertheless highlights a reporting issue. Training and evaluation endpoints should be named precisely, with their source and verification method, because postoperative diagnosis, FNA cytology and TI-RADS suspicion are not interchangeable. At the same time, this study cannot separate endpoint effects from distribution shift: the cohorts also differ in population, prevalence, scanners, acquisition, cropping, image quality and annotation. A same-cohort study containing both pathology outcomes and blinded TI-RADS assessments would be needed to estimate the effect of endpoint definition while holding images constant.

The cytology benchmark was more discriminative than the ultrasound benchmark within the development archive, but this was not a patient-paired comparison. The modalities differed in sample size (1,332 versus 387 images), image generation and spatial content, and the split could not be verified at patient level. The result is therefore a descriptive benchmark for these image distributions, not evidence that one modality is clinically superior.

Duplicate and perceptual-hash grouping reduced a documented source of image-level leakage. It does not exclude different views or fields from the same patient crossing partitions, so internal estimates may remain optimistic. The exploratory ROI result and calibration analyses should likewise be regarded as hypothesis-generating until repeated with patient-level splits, expert nodule localisation and independent outcome verification.

6 Limitations

- Unverified patient-level independence. Exact and perceptual duplicates were grouped, but per-image patient identifiers were unavailable. Different images from one patient may therefore have crossed partitions, and the internal estimates are image-level rather than patient-level.
- Reference-endpoint and cohort confounding. Development labels were based on postoperative diagnoses. TN3K used distributed binary labels whose universal pathology confirmation could not be established from the public documentation reviewed. The DDTI binary endpoint was derived from TI-RADS suspicion and is not a malignancy reference standard. Because endpoint, cohort and acquisition changed together, their separate effects cannot be estimated.
- Sample size and non-paired modalities. The internal test partitions contained 200 cytology images and 60 ultrasound images, not verified independent patients. Cytology and ultrasound were not paired within patients, so their performance contrast does not establish comparative clinical accuracy.
- Image resolution and ROI. Resizing to 224 × 224 may obscure small echogenic foci, halos and fine margins. Automatic cropping was approximate and not equivalent to expert nodule segmentation.
- Binary and retrospective scope. Only benign-versus-PTC classification was evaluated. Indeterminate Bethesda categories, non-PTC malignancies, prospective workflow performance and external cytology validation were not assessed.

7 Clinical and Research Implications

The results support three practical priorities for future thyroid imaging AI studies: use a clinically appropriate and explicitly documented reference standard; partition and analyse data at patient level; and distinguish diagnostic-outcome evaluation from agreement with risk-stratification systems. Higher-resolution inputs, expert or automated nodule segmentation, multi-view imaging and clinical variables warrant prospective evaluation. Model probabilities should be recalibrated and assessed for clinical utility only in the intended population and workflow.

8 Future Work

The most informative next study would assemble a patient-paired cohort containing ultrasound images, FNA cytology, postoperative pathology where available, clinical variables and blinded TI-RADS assessments. Evaluating identical images against both pathology outcomes and TI-RADS categories would help separate endpoint definition from cohort shift. Such data could then support segmentation-guided, multimodal modelling with patient-level external and prospective validation under current reporting guidance [24], [25].

9. Conclusions

In this duplicate-controlled image-level benchmark, frozen thyroid ultrasound models showed good discrimination against the TN3K subset's distributed binary labels but little agreement with a DDTI endpoint derived from TI-RADS categories. Because the cohorts differed in both endpoint and data distribution, this contrast does not isolate a label-semantic effect or validate malignancy diagnosis on

DDTI. Future studies should use verified patient-level splits, describe reference standards consistently and, where endpoint effects are of interest, compare pathology outcomes and TI-RADS assessments within the same cohort.

Declarations

Funding. No funding was received for this study.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval. This study is a secondary analysis of fully de-identified, publicly available datasets [26–28] and did not involve the collection of new human-subjects data by the authors. No additional ethics approval was required.

Consent to participate. Not applicable. The study used publicly available, de-identified datasets and involved no direct participation or recruitment of human subjects by the authors.

Consent for publication. Not applicable. The manuscript contains no identifiable individual person's data, images or case details requiring consent for publication.

Data availability. The development images are available from Mendeley Data (https://doi.org/10.17632/dp7zp2pb2v.1; CC BY 4.0) [26]. The TN3K [27] and DDTI [28] cohorts are available from their original distributors. Processing manifests, per-model probability files and figure source data are available at "https://github.com/b-kiani/Thyroid-Cancer-Classification-2026".

Code availability. All analysis scripts are available at https://github.com/b-kiani/Thyroid-Cancer-Classification-. Random seeds and software versions are documented in the repository.

Author contributions. Behnam Kiani Kalejahi: conceptualisation, methodology, software, formal analysis, investigation, data curation, visualisation, and writing of the original draft. Mohammad Javad Rajabi: conceptualisation, methodology, supervision, and review and editing. Both authors read and approved the final manuscript.

Acknowledgements. The authors thank Pang et al. for releasing the development cohort under an open licence, and the maintainers of the TN3K and DDTI datasets.

References

1. Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer. Thyroid. 2016;26(1):1–133. https://doi.org/10.1089/thy.2015.0020.
2. Ali SZ, Baloch ZW, Cochand-Priollet B, et al. The 2023 Bethesda System for Reporting Thyroid Cytopathology. Thyroid. 2023;33(9):1039–44. https://doi.org/10.1089/thy.2023.0141.

3. Tessler FN, Middleton WD, Grant EG, et al. J Am Coll Radiol. 2017;14(5):587–95. https://doi.org/10.1016/j.jacr.2017.01.046. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR TI-RADS Committee.
4. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol. 2019;20(2):193–201. https://doi.org/10.1016/S1470-2045(18)30762-9.
5. Peng S, Liu Y, Lv W, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. Lancet Digit Health. 2021;3(4):e250–9. https://doi.org/10.1016/S2589-7500(21)00041-8.
6. Lin YJ, Chao TC, Khalil MA, et al. Deep learning fast screening approach on cytological whole slides for thyroid cancer diagnosis. Cancers. 2021;13(15):3891. https://doi.org/10.3390/cancers13153891.
7. Duc NT, Lee YM, Park JH, Lee B. An ensemble deep learning for automatic prediction of papillary thyroid carcinoma using fine-needle aspiration cytology. Expert Syst Appl. 2022;188:115927. https://doi.org/10.1016/j.eswa.2021.115927.
8. Guan Q, Wang Y, Du J, et al. Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. J Cancer. 2019;10(20):4876–82. https://doi.org/10.7150/jca.28769.
9. Slabaugh G, Beltran L, Rizvi H, Deloukas P, Marouli E. Applications of machine and deep learning to thyroid cytology and histopathology: a review. Front Oncol. 2023;13:958310. https://doi.org/10.3389/fonc.2023.958310.
10. Li J, Bai C, Lin T, et al. Multi-modal feature integration of clinical data and ultrasound deep features for thyroid nodule malignancy prediction. J Radiat Res Appl Sci. 2024;17(4):101205. https://doi.org/10.1016/j.jrras.2024.101205.
11. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986 (2022).
12. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR, vol. 97, pp. 6105–6114 (2019).
13. Liu Z, Lin Y, Cao Y et al. Swin Transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021).
14. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016).
15. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009).
16. Wightman R. PyTorch image models (timm). GitHub repository. https://github.com/rwightman/pytorch-image-models (2019). Accessed 12 August 2026.

17. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019).
18. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017).
19. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR, vol. 70, pp. 1321–1330 (2017).
20. 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):837–45.
21. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527. https://doi.org/10.1136/bmj.h5527.
22. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2(2). https://doi.org/10.1148/ryai.2020200029. e200029.
23. Tejani AS, Mongan J, Kahn CE et al. Updating the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) for reporting AI research. Radiol Artif Intell 6(4), e240300 (2024).
24. Sounderajah V, Ashrafian H, Bossuyt PM, et al. Developing the STARD-AI reporting guideline for diagnostic accuracy studies of artificial intelligence. Nat Med. 2025;31(10):3283–9. https://doi.org/10.1038/s41591-025-03953-8.
25. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385. https://doi.org/10.1136/bmj-2023-078378. e078378.
26. Pang YY, Tang ZQ, Song C, Qu N, Chen JY, Xiong DD, Feng ZB, Chen G. Data Brief. 2025;60:111583. https://doi.org/10.17632/dp7zp2pb2v.1. https://doi.org/10.1016/j.dib. https://data.mendeley.com/datasets/dp7zp.2025.111583. Dataset: Mendeley Data, V1.
27. Gong H, Chen J, Chen G, Li H, Chen F, Li G. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Comput Biol Med. 2023;155:106389. https://doi.org/10.1016/j.compbiomed.2022.106389.
28. Pedraza L, Vargas C, Narváez F, Durán O, Muñoz E, Romero E. An open access thyroid ultrasound-image database (DDTI). In: Proceedings of the 10th International Symposium on Medical Information Processing and Analysis (SIPAIM), SPIE, vol. 9287, 92870W (2015). https://doi.org/10.1117/12.2073532
29. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708 (2017).
30. Kiani Kalejahi B, Rajabi MJ. Multimodal deep learning for papillary thyroid carcinoma diagnosis using ultrasound and cytology. BMC Med. Imaging (2026).

Figures

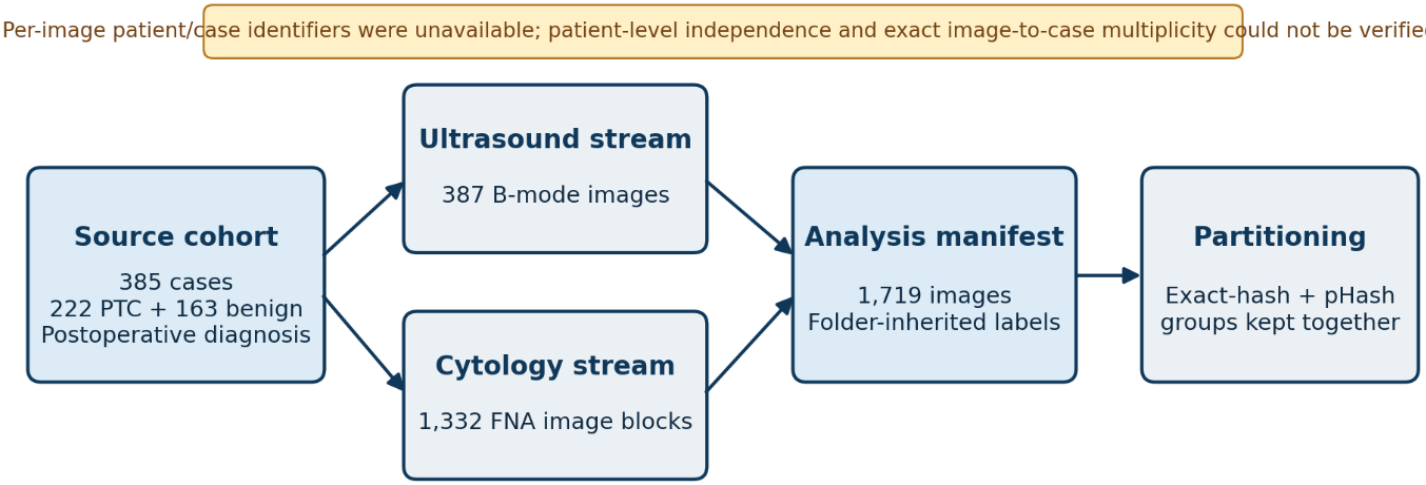


Figure 1

Development-dataset construction and partitioning. The 1,719 analysed images derive from 385 source cases; image-level patient identifiers were unavailable, so case multiplicity and patient-level independence could not be verified

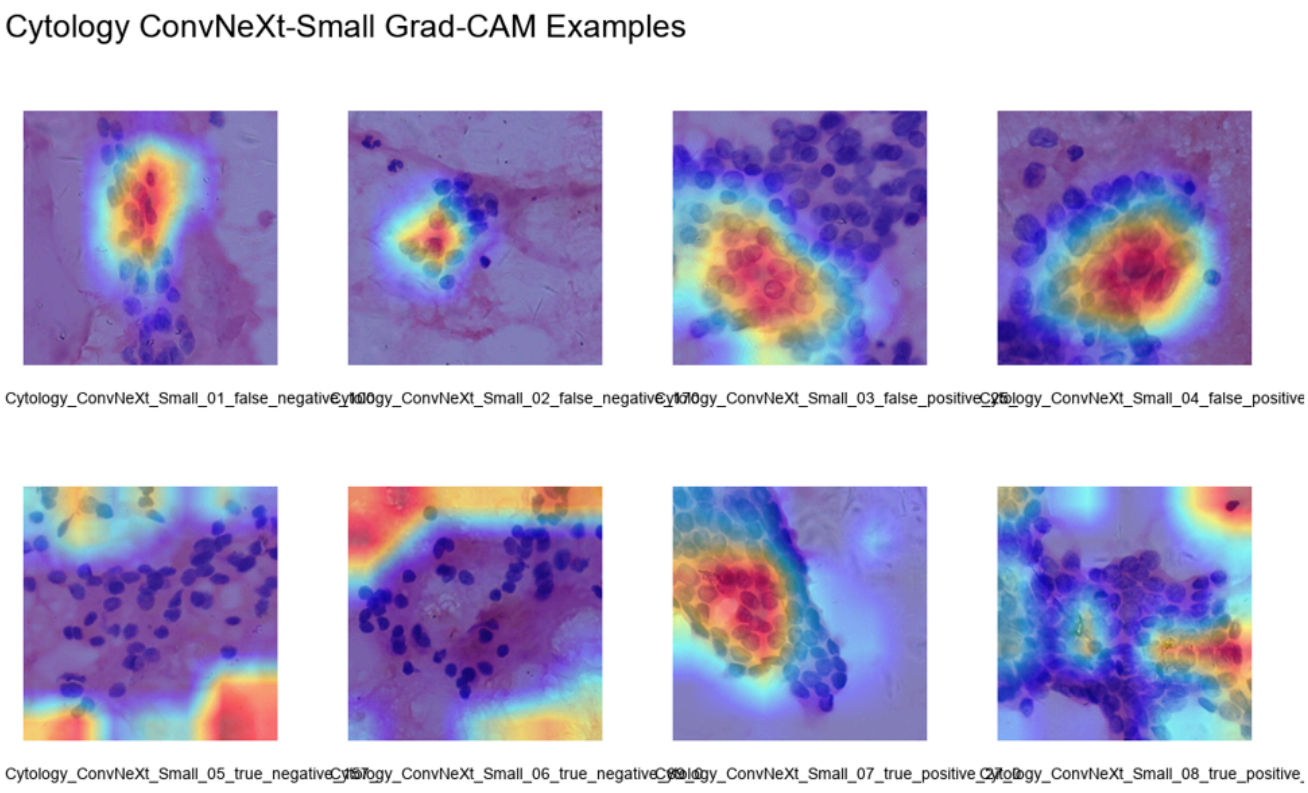


Figure 2

Cytology ConvNeXt-Small Grad-CAM examples (true-positive, true-negative, false-positive and false-negative images)

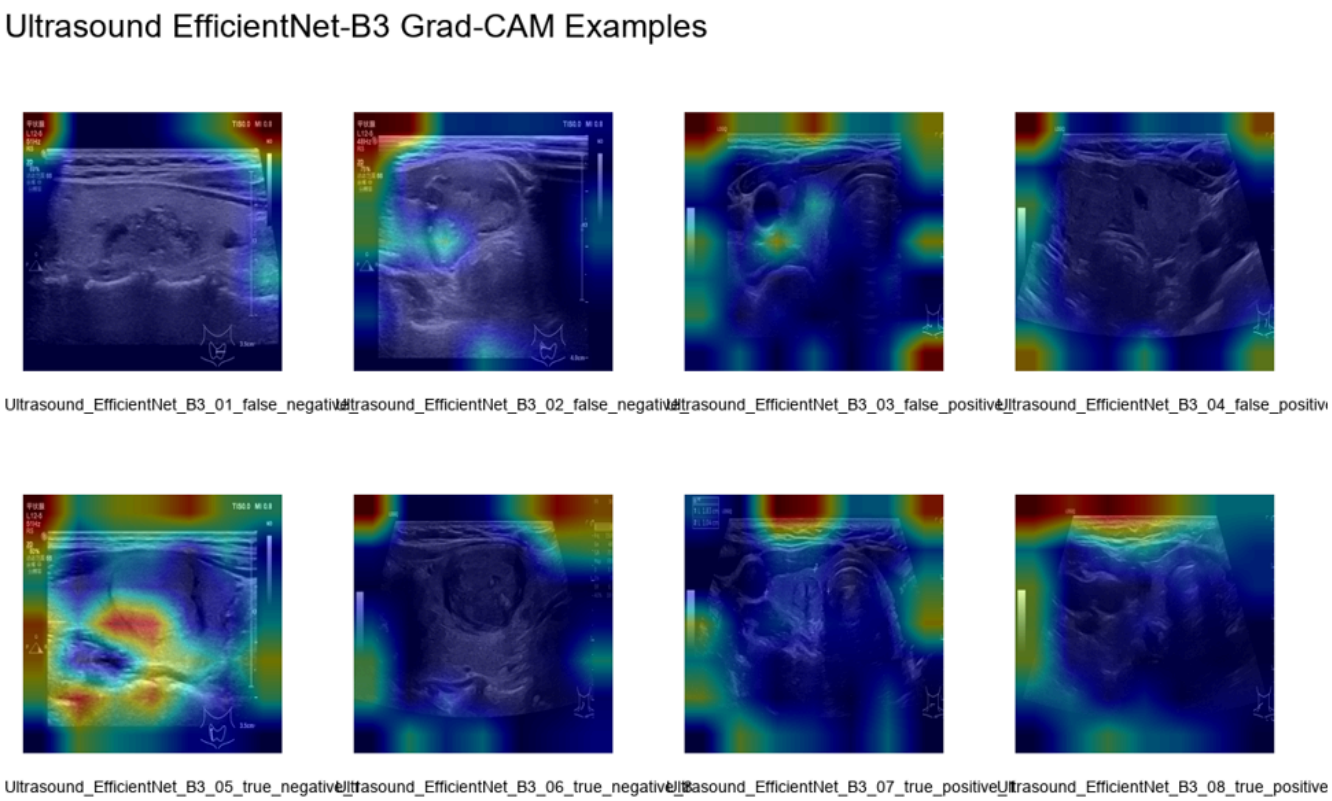


Figure 3

Ultrasound EfficientNet-B3 Grad-CAM examples showing diffuse and variable attention patterns