Supplementary Materials

Content

Figure S1. Comparing AI models across different modalities

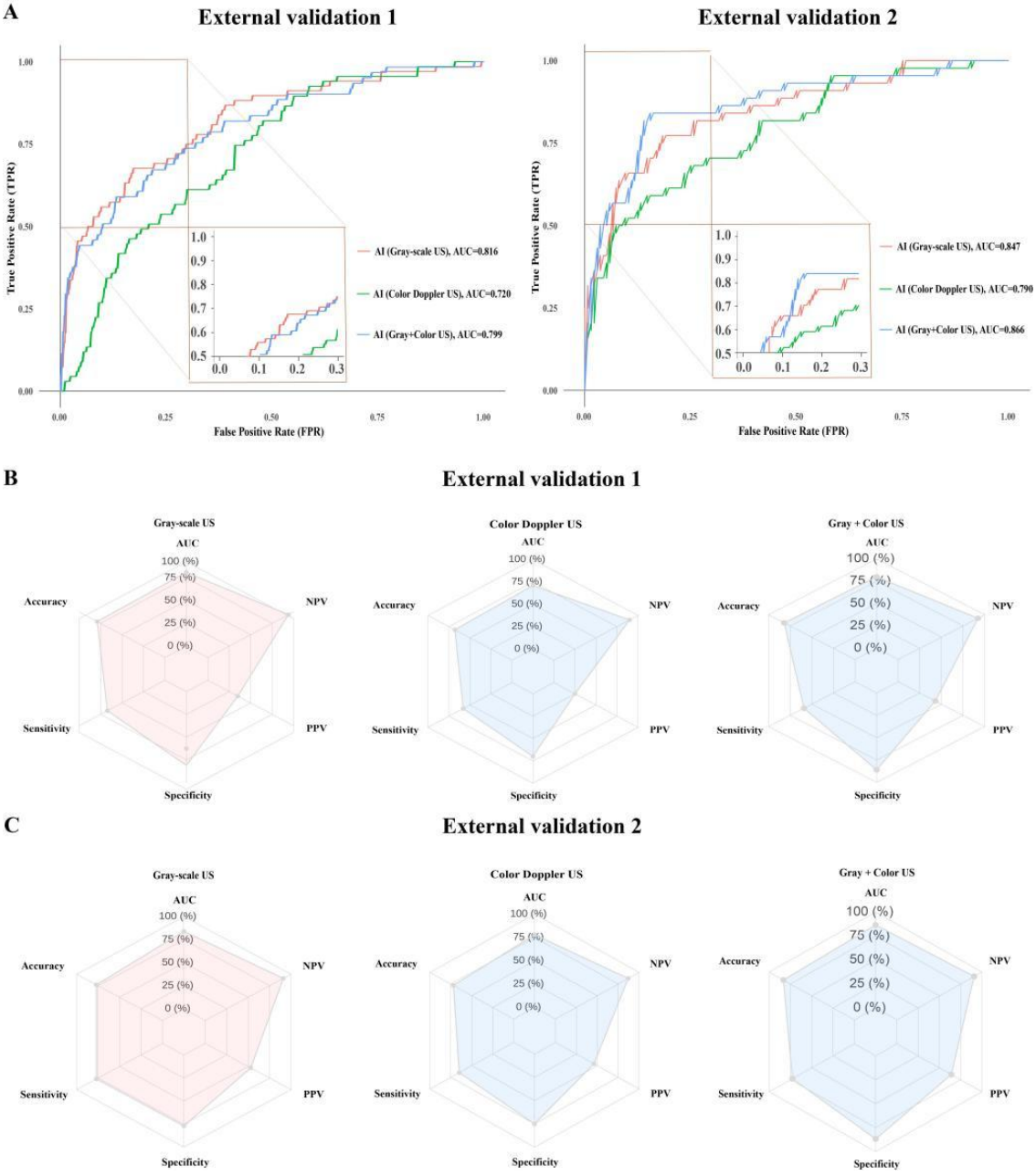


Figure S2. Model visualization and analysis of misdiagnosis cases

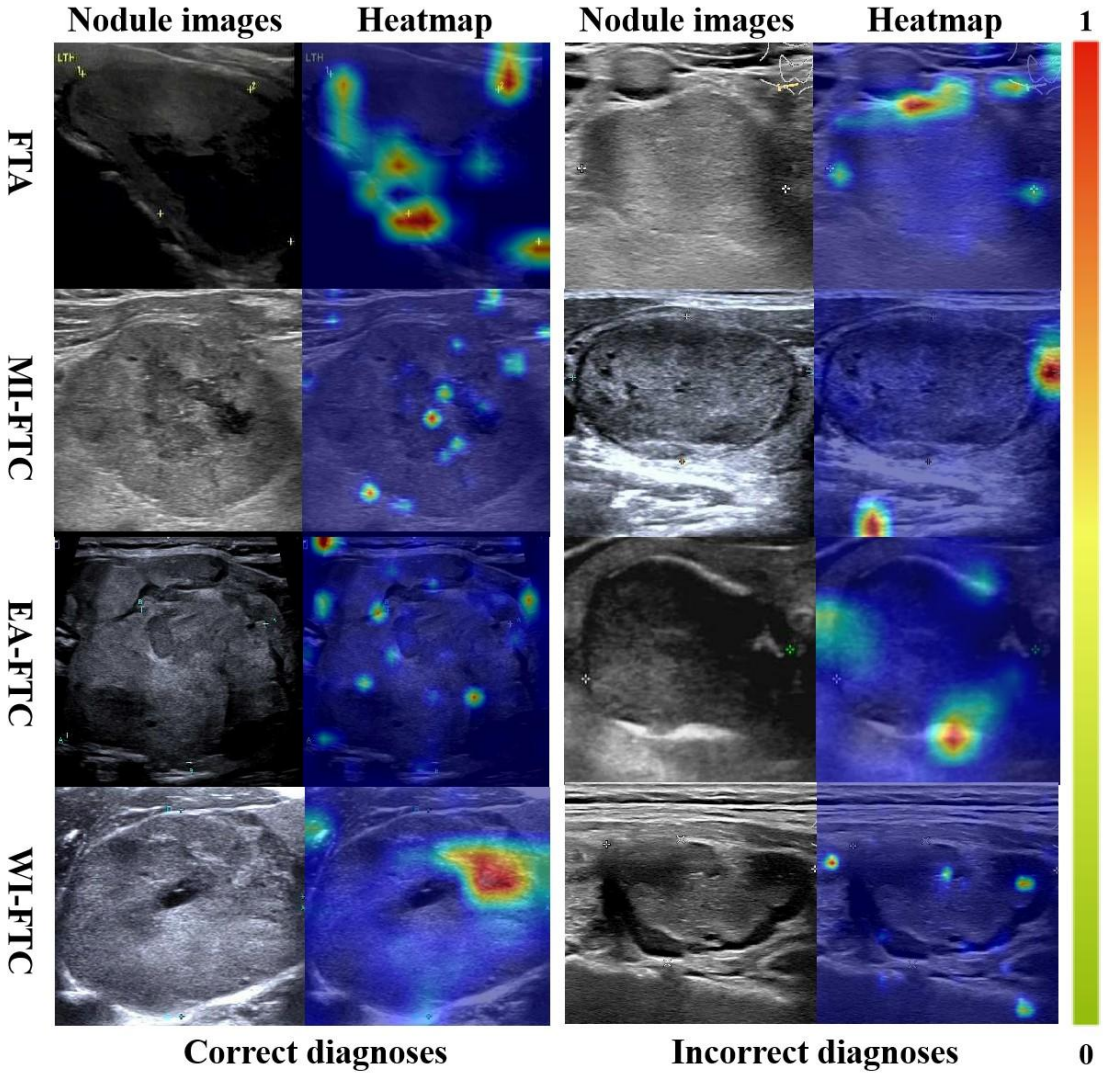
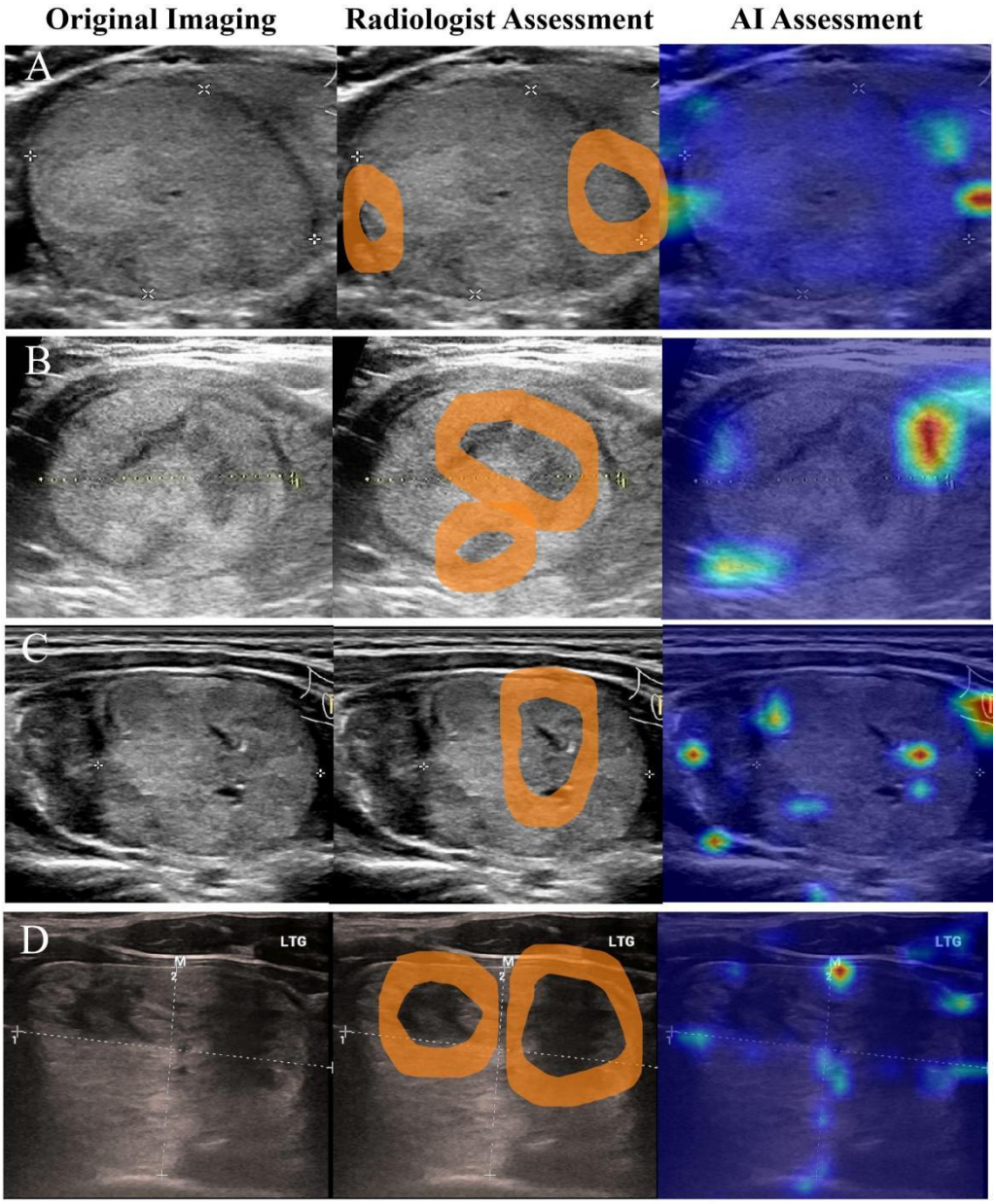


Figure S3. Case studies illustrating AI model performance in comparison with radiologist assessment



A–D present four independent thyroid nodule cases. For each case (row), the left panel shows the original B-mode ultrasound image with the lesion outlined; the middle panel summarizes the corresponding radiologist's assessment including key descriptive features; the right panel provides the AI model's prediction with the overlaid Grad-CAM attention map (red indicates regions of high attention). **A**, Case 1 (concordant diagnosis and feature localization); **B**, Case 2 (model correct, radiologists false negative); **C**, Case 3 (model correct subtyping, F-TIRADS underestimation); **D**, Case 4 (AI assistance correcting radiologist misdiagnosis). These examples highlight scenarios in which the AI model correctly identified malignancy or benignity, especially when radiologists initially misinterpreted subtle sonographic features, underscoring the model's potential to enhance conventional diagnostic evaluation.

Table S1. The diagnostic performance of the DL model in subgroup analysis

	No(%)	FTA(%)	FTC(%)	AUC(95%CI)	P value
Overall population	696(100%)	584(83.91%)	112(16.09%)	0.832(0.802-0.859)	-
Sex					
Male	149(21.41%)	111(19.01%)	38(33.93%)	0.829(0.744-0.913)	Reference
Female	547(78.59%)	473(80.99%)	74(66.07%)	0.831(0.779-0.883)	0.914
Lesion size, cm					
<3	353(50.72%)	307(52.57%)	46(41.07%)	0.785(0.706-0.864)	Reference
≥3	343(49.28%)	277(47.43%)	66(58.93%)	0.857(0.808-0.906)	0.085
Surgery method					
Total thyroidectomy	494(70.98%)	427(73.12%)	67(59.82%)	0.817(0.757-0.878)	Reference
Lateral thyroidectomy	202(29.02%)	157(26.88%)	45(40.18%)	0.857(0.796-0.918)	0.297
Nodule position					
Right	370(53.16%)	313(53.60%)	57(50.89%)	0.812(0.746-0.877)	Reference
Left	298(42.81%)	249(42.64%)	49(43.75%)	0.849(0.787-0.911)	0.347
Thyroid disease					
Yes	295(42.39%)	259(44.35%)	36(32.14%)	0.873(0.813-0.934)	Reference
No	401(57.61%)	325(55.65%)	76(67.86%)	0.812(0.755-0.870)	0.176
Location					
North District	407(58.48%)	338(57.88%)	69(61.61%)	0.864(0.811-0.912)	Reference

South District	192(27.59%)	179(30.65%)	13(11.61%)	0.804(0.671-0.937)	0.546
East District	57(8.19%)	30(5.14%)	27(24.11%)	0.754(0.624-0.885)	0.134
West District	40(5.75%)	37(6.34%)	3(2.68%)	0.910(0.789-0.999)	0.468

P-value: Comparison of AUC values among subgroups.

Table S2. Diagnostic performance of the DL model in cooperation with radiologists

	Accuracy(95%CI)	Sensitivity(95%CI)	Specificity(95%CI)	PPV(95%CI)	NPV(95%CI)	P value
AI	0.752(0.688-0.81)	0.8(0.644-0.909)	0.741(0.667-0.806)	0.427(0.313-0.546)	0.939(0.883-0.973)	Reference
Radiologists (without AI assistance)						
All radiologists	0.666(0.626-0.706)	0.438(0.337-0.534)	0.721(0.679-0.758)	0.289(0.205-0.384)	0.842(0.784-0.885)	<0.001
Mean senior radiologist	0.701(0.646-0.748)	0.412(0.291-0.54)	0.771(0.718-0.819)	0.319(0.223-0.437)	0.844(0.782-0.892)	0.002
Mean attending radiologist	0.655(0.597-0.703)	0.438(0.316-0.563)	0.708(0.651-0.762)	0.282(0.19-0.394)	0.84(0.777-0.886)	<0.001
Mean junior radiologist	0.641(0.587-0.687)	0.462(0.351-0.579)	0.684(0.63-0.738)	0.266(0.185-0.373)	0.843(0.781-0.896)	<0.001
Radiologists (with AI assistance)						
All radiologists	0.746(0.701-0.783)	0.567(0.454-0.667)	0.789(0.742-0.832)	0.393(0.286-0.499)	0.886(0.837-0.922)	0.223
Mean senior radiologist	0.718(0.653-0.762)	0.688(0.566-0.797)	0.726(0.67-0.781)	0.377(0.274-0.482)	0.907(0.852-0.945)	0.002
Mean attending radiologist	0.74(0.682-0.786)	0.575(0.438-0.709)	0.78(0.716-0.83)	0.384(0.274-0.502)	0.885(0.828-0.925)	0.021
Mean junior radiologist	0.779(0.723-0.82)	0.438(0.321-0.561)	0.861(0.809-0.904)	0.417(0.292-0.557)	0.865(0.814-0.906)	0.316
Radiologists (with AI assistance) one month post-initial assessment						
All radiologists	0.744(0.705-0.778)	0.575(0.457-0.677)	0.784(0.743-0.821)	0.41(0.303-0.51)	0.886(0.837-0.924)	0.261
Mean senior radiologist	0.767(0.721-0.801)	0.675(0.545-0.781)	0.789(0.742-0.829)	0.49(0.371-0.602)	0.911(0.863-0.946)	0.167
Mean attending radiologist	0.743(0.689-0.789)	0.412(0.273-0.544)	0.822(0.773-0.862)	0.367(0.234-0.497)	0.854(0.792-0.9)	0.047
Mean junior radiologist	0.721(0.665-0.774)	0.637(0.499-0.781)	0.741(0.676-0.801)	0.374(0.261-0.481)	0.895(0.836-0.937)	0.022

P value: Statistical comparison of diagnostic accuracy: Radiologists vs. AI models

Table S3. Comparative analysis of the proposed model and published models in differentiating FTC from FTA

	DL model	AUC(95%CI)	Accuracy(95%CI)	Sensitivity(95%CI)	Specificity(95%CI)	PPV(95%CI)	NPV(95%CI)	F1score	κcoefficient	P value
External test dataset-1	ConvNeXt	0.816(0.758-0.875)	0.795(0.757-0.829)	0.676(0.552-0.785)	0.812(0.773-0.848)	0.354(0.272-0.442)	0.943(0.915-0.964)	0.465	0.353	Reference
	OverLoCK	0.63(0.559-0.691)	0.609(0.572-0.641)	0.529(0.426-0.629)	0.621(0.581-0.656)	0.175(0.132-0.218)	0.897(0.866-0.925)	0.081	0.263	<0.001
External test dataset-2	ConvNeXt	0.847(0.777-0.917)	0.778(0.71-0.836)	0.773(0.622-0.885)	0.779(0.700-0.846)	0.531(0.402-0.657)	0.914(0.847-0.958)	0.630	0.479	Reference
	OverLoCK	0.679,(0.599-0.761)	0.644, (0.583-0.706)	0.614, (0.49-0.738)	0.654,(0.586-0.724)	0.365(0.28-0.466)	0.84(0.778-0.895)	0.218	0.458	<0.001
External test dataset-3	ConvNeXt	0.837(0.777-0.897)	0.752(0.688-0.81)	0.8(0.644-0.909)	0.741(0.667-0.806)	0.427(0.313-0.546)	0.939(0.883-0.973)	0.557	0.406	Reference
	OverLoCK	0.627(0.539-0.715)	0.767(0.718-0.816)	0.375(0.256-0.514)	0.861(0.816-0.906)	0.395(0.268-0.525)	0.851(0.805-0.896)	0.241	0.385	<0.001

P-value: AUC comparison (DL model) - ConvNeXt vs. OverLoCK