Supplementary

Table S1 Performance of five DL models in different datasets

	AUC (95% CI)	SEN	SPE	ACC	PPV	NPV	F1-score
Training dataset							
MobileNet-V2	0.868 (0.848–0.888)*	0.794	0.779	0.786	0.765	0.807	0.779
DesNet201	0.830 (0.806–0.855)	0.699	0.845	0.776	0.804	0.756	0.749
ResNet50	0.854 (0.830–0.876)	0.747	0.812	0.781	0.783	0.779	0.765
VGG19	0.793 (0.765–0.819)	0.822	0.646	0.730	0.679	0.799	0.743
Xception	0.864 (0.843–0.885)	0.695	0.878	0.791	0.838	0.761	0.760
Validation dataset							
MobileNet-V2	0.919 (0.897–0.941)*	0.868	0.797	0.829	0.782	0.879	0.821
DesNet201	0.815 (0.775–0.854)	0.638	0.858	0.758	0.789	0.739	0.704
ResNet50	0.845 (0.806–0.878)	0.780	0.795	0.788	0.762	0.812	0.771
VGG19	0.779 (0.737–0.819)	0.692	0.738	0.717	0.689	0.741	0.689
Xception	0.850 (0.813–0.883)	0.673	0.886	0.789	0.832	0.764	0.743
Test dataset							
MobileNet-V2	0.875 (0.845–0.901)*	0.687	0.914	0.800	0.890	0.742	0.776
DesNet201	0.771 (0.732–0.807)	0.776	0.624	0.700	0.675	0.734	0.722
ResNet50	0.796 (0.757–0.832)	0.736	0.703	0.720	0.714	0.726	0.725
VGG19	0.656 (0.610–0.706)	0.743	0.519	0.632	0.609	0.667	0.668
Xception	0.849 (0.814–0.880)	0.782	0.780	0.781	0.78	0.781	0.783

*, the AUC value is the largest. ACC, accuracy; AUC, area under the receiver operating characteristic curve; CI, confidence interval; DL, deep learning; NPV, negative predictive value; PPV, positive predictive value; SEN, sensitivity; SPE, specificity.

Table S2 Comparison of AUCs between different models

	Training cohort		Validation cohort		Test cohort	
	AUC	P value	AUC	P value	AUC	P value
US *vs.* DL	0.802 *vs.* 0.868	<0.001	0.799 *vs.* 0.919	<0.001	0.787 *vs.* 0.875	<0.001
US *vs.* USDL	0.802 *vs.* 0.922	<0.001	0.799 *vs.* 0.947	<0.001	0.787 *vs.* 0.907	<0.001
DL *vs.* USDL	0.868 *vs.* 0.922	<0.001	0.919 *vs.* 0.947	<0.001	0.875 *vs.* 0.907	<0.001

The DeLong test is used to compare the models from the training cohort, the validation cohort, and the test cohort. AUC, area under the receiver operating characteristic curve; DL, deep learning; US, ultrasound; USDL, ultrasound combined with deep learning.

https://dx.doi.org/10.21037/qims-2025-159

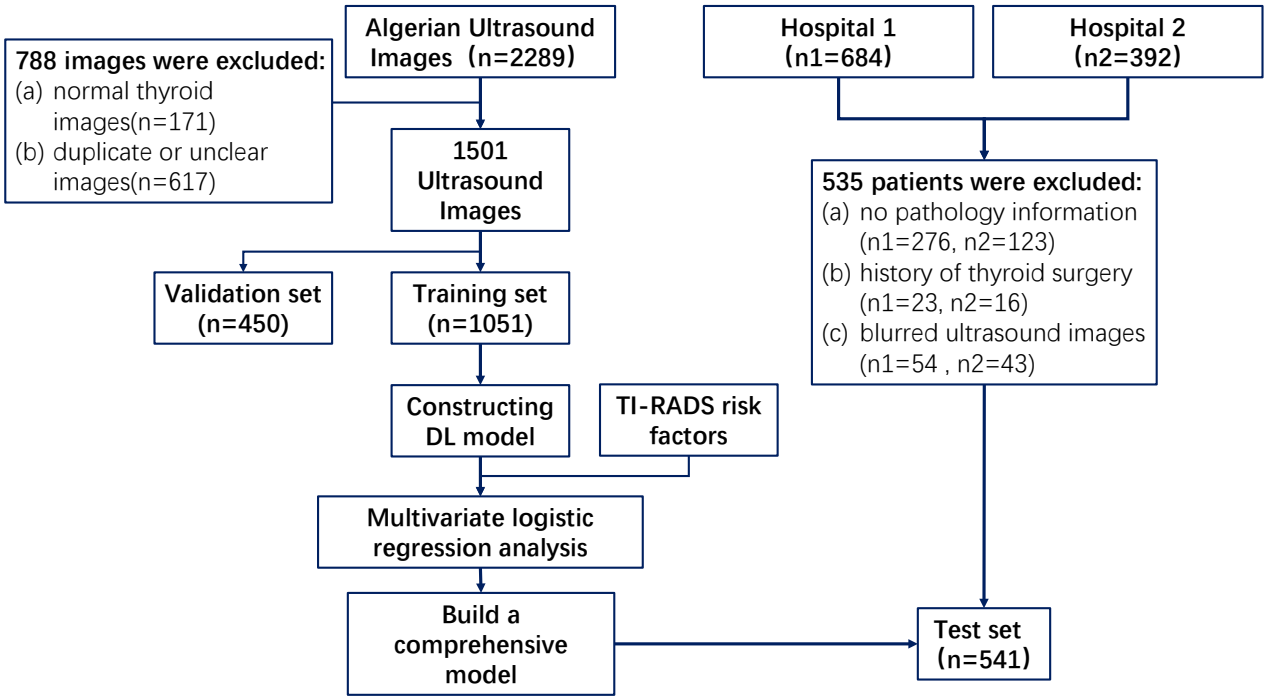

Figure S1 Flow chart of the research methodology. Hospital 1 refers to the First Affiliated Hospital of Anhui Medical University, and Hospital 2 refers to the Affiliated Hospital of Integration Chinese and Western Medicine with Nanjing University of Chinese Medicine. DL, deep learning; TI-RADS, Thyroid Imaging Reporting and Data System.

https://dx.doi.org/10.21037/qims-2025-159