Appendix 1

The following steps were taken to obtain the ultrasound (US) images: the patient was placed in a supine position, raising the neck and back to fully expose the neck. Routine transverse and longitudinal scanning of the thyroid isthmus and left and right lobes was performed. Based on the characteristics of the high-frequency US images, we selected the clearest longitudinal section of the lesion, placed the probe on the patient's neck, and marked the lesion perpendicular to the probe, making the interface between the thyroid capsule and the sound beam perpendicular, increasing sound energy reflection, and ensuring the clear display of the US images. The longitudinal dimension, transverse dimension, morphology, internal echo, boundary, presence or absence of calcification, aspect ratio, and blood flow signal of each nodule were recorded. The US images were retrieved from the thyroid imaging database.

Appendix 2

Deep-learning (DL) procedure

Because of leak of image data, to better carry out the generalization, we carefully set the learning rate. We adapted the cosine decay learning rate algorithm in this study. Our learning rate is expressed as follows:

$$\eta_t = \eta_{\min}^i + \frac{1}{2}\left(\eta_{\max}^i - \eta_{\min}^i\right)\left[1+\cos\left(\frac{T_{cur}}{T_i}\pi\right)\right] \quad [1]$$

where $\eta_{\min}^i = 0$, $\eta_{\max}^i = 0.01$, and $T_i = 50$ represent the minimum learning rate, the maximum learning rate, and the number of iteration epochs, respectively. As the backbone part adopts pre-training parameters, to ensure the migration effect, on $T_{cur} = \frac{1}{2}T_i$, fine tune the parameters of the backbone part. Therefore, the learning rate of backbone part is expressed as follows:

$$\eta_t^{backbone} = \begin{cases} 0 & \text{if } T_{cur} > \frac{1}{2}T_i \\ \eta_{\min}^i + \frac{1}{2}\left(\eta_{\max}^i - \eta_{\min}^i\right)\left[1+\cos\left(\frac{T_{cur}}{T_i}\pi\right)\right] & \text{if } T_{cur} > \frac{1}{2}T_i \end{cases} \quad [2]$$

The other hyperparameter configurations are as follows: optimizer: Stochastic Gradient Descent (SGD), loss function: sigmoid cross entropy.

In the process of building deep-learning radiomics model, we combined the finally selected radiomics features with the result of the deep-learning model to form a new feature set, which is then input into a machine learning algorithm.

 https://dx.doi.org/10.21037/qims-23-1597

Table S1 Pathological classification of 1076 thyroid nodules

Pathological classification	Development set (n=719)	Validation set 1 (n=74)	Validation set 2 (n=283)
Benign nodule	411	8	175
Nodular goiter (%)	376 (91.5)	6 (75.0)	146 (83.4)
Adenoma (%)	18 (4.4)	0	8 (4.6)
Chronic lymphocytic thyroiditis (%)	12 (2.9)	2 (25.0)	20 (11.4)
subacute thyroiditis (%)	5 (1.2)	0	1 (0.6)
Malignant nodule	308	66	108
Papillary thyroid carcinoma (%)	306 (99.4)	65 (98.5)	107 (99.1)
Medullary thyroid carcinoma (%)	2 (0.6)	1 (1.5)	1 (0.9)

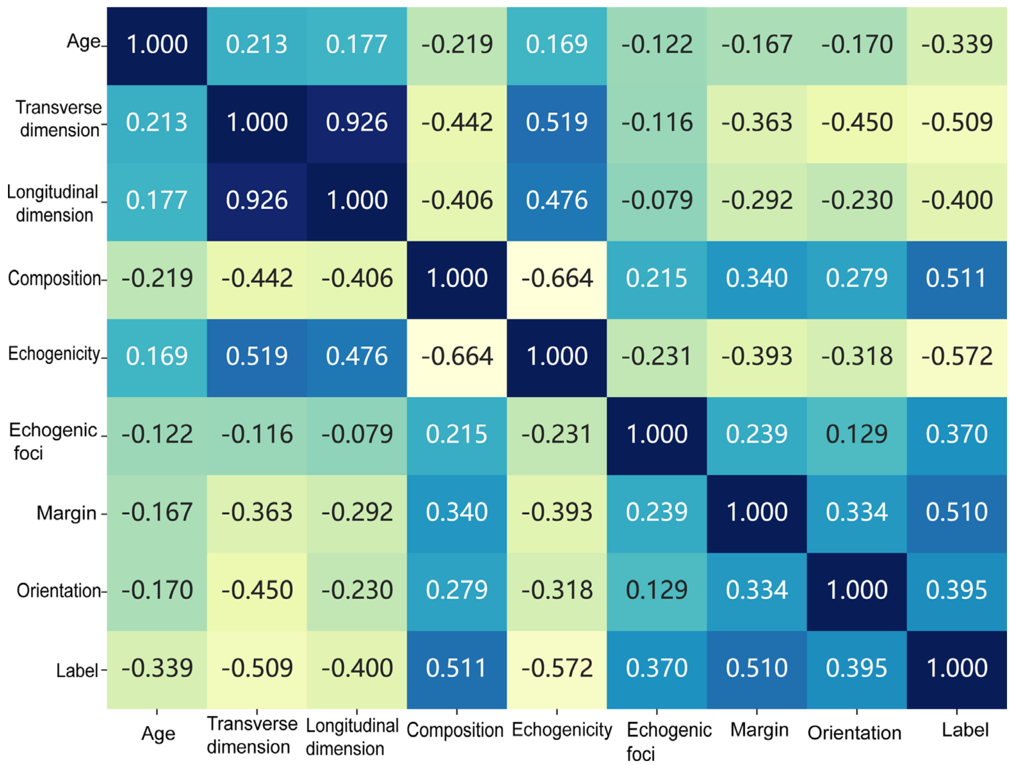

Figure S1 Clinical feature correlation coefficient heatmap. The closer the correlation coefficient was to 1 or −1, the stronger the correlation; The closer it was to 0, the weaker the correlation.

https://dx.doi.org/10.21037/qims-23-1597

Table S2 Logistic regression analysis of single and multiple factors

Variables	Univariate analysis				Multivariate analysis			
	OR	LCI	UCI	P value	OR	LCI	UCI	P value
Gender	0.922	0.856	0.993	0.073	–	–	–	–
Age	0.984	0.981	0.986	<0.001	0.992	0.99	0.994	<0.001
Transverse dimension	0.985	0.983	0.986	<0.001	0.994	0.99	0.999	0.035
Longitudinal dimension	0.982	0.98	0.984	<0.001	0.999	0.994	1.004	0.752
Orientation	1.689	1.571	1.815	<0.001	1.158	1.085	1.235	<0.001
Composition	1.328	1.29	1.368	<0.001	1.056	1.023	1.09	0.004
Echogenicity	0.819	0.806	0.833	<0.001	0.915	0.897	0.933	<0.001
Echogenic foci	1.271	1.214	1.331	<0.001	1.124	1.087	1.163	<0.001
Margin	1.195	1.169	1.223	<0.001	1.057	1.037	1.078	<0.001

OR, odds ratio; LCI, low-confidence interval; UCI, upper-confidence interval.

Table S3 The 45 radiomics features

Selected features
exponential_glcm_Correlation
exponential_glrlm_ShortRunLowGrayLevelEmphasis
exponential_glszm_SmallAreaHighGrayLevelEmphasis
exponential_ngtdm_Coarseness
gradient_glszm_LowGrayLevelZoneEmphasis
gradient_ngtdm_Busyness
lbp_3D_k_firstorder_10Percentile
lbp_3D_k_glcm_Imc1
lbp_3D_k_glszm_SizeZoneNonUniformity
lbp_3D_k_glszm_SmallAreaEmphasis
lbp_3D_m1_ngtdm_Busyness
lbp_3D_m2_glcm_Correlation
lbp_3D_m2_glszm_GrayLevelVariance
logarithm_glcm_DifferenceEntropy
original_shape_Elongation
square_glcm_JointAverage
square_glszm_LargeAreaLowGrayLevelEmphasis
square_ngtdm_Coarseness
squareroot_firstorder_InterquartileRange
squareroot_glszm_SmallAreaEmphasis
squareroot_glszm_SmallAreaLowGrayLevelEmphasis
wavelet_HHH_glrlm_GrayLevelVariance
wavelet_HHL_firstorder_Kurtosis
wavelet_HHL_glcm_Idn
wavelet_HLH_firstorder_Mean
wavelet_HLH_gldm_LargeDependenceHighGrayLevelEmphasis
wavelet_HLH_glrlm_RunLengthNonUniformityNormalized
wavelet_HLH_glrlm_ShortRunHighGrayLevelEmphasis
wavelet_HLH_glszm_LowGrayLevelZoneEmphasis
wavelet_HLH_glszm_SizeZoneNonUniformity
wavelet_HLH_glszm_SmallAreaEmphasis
wavelet_HLH_glszm_SmallAreaHighGrayLevelEmphasis
wavelet_HLL_firstorder_Median
wavelet_HLL_glcm_Idn
wavelet_LHH_firstorder_Maximum
wavelet_LHH_gldm_DependenceEntropy
wavelet_LHH_gldm_SmallDependenceHighGrayLevelEmphasis
wavelet_LHH_glszm_SizeZoneNonUniformityNormalized
wavelet_LHL_firstorder_Mean
wavelet_LHL_firstorder_Median
wavelet_LLH_ngtdm_Contrast
wavelet_LLL_gldm_LargeDependenceLowGrayLevelEmphasis
wavelet_LLL_glszm_SizeZoneNonUniformity
wavelet_LLL_glszm_SmallAreaHighGrayLevelEmphasis
wavelet_LLL_ngtdm_Coarseness

Table S3 (*continued*)

Table S3 (*continued*)

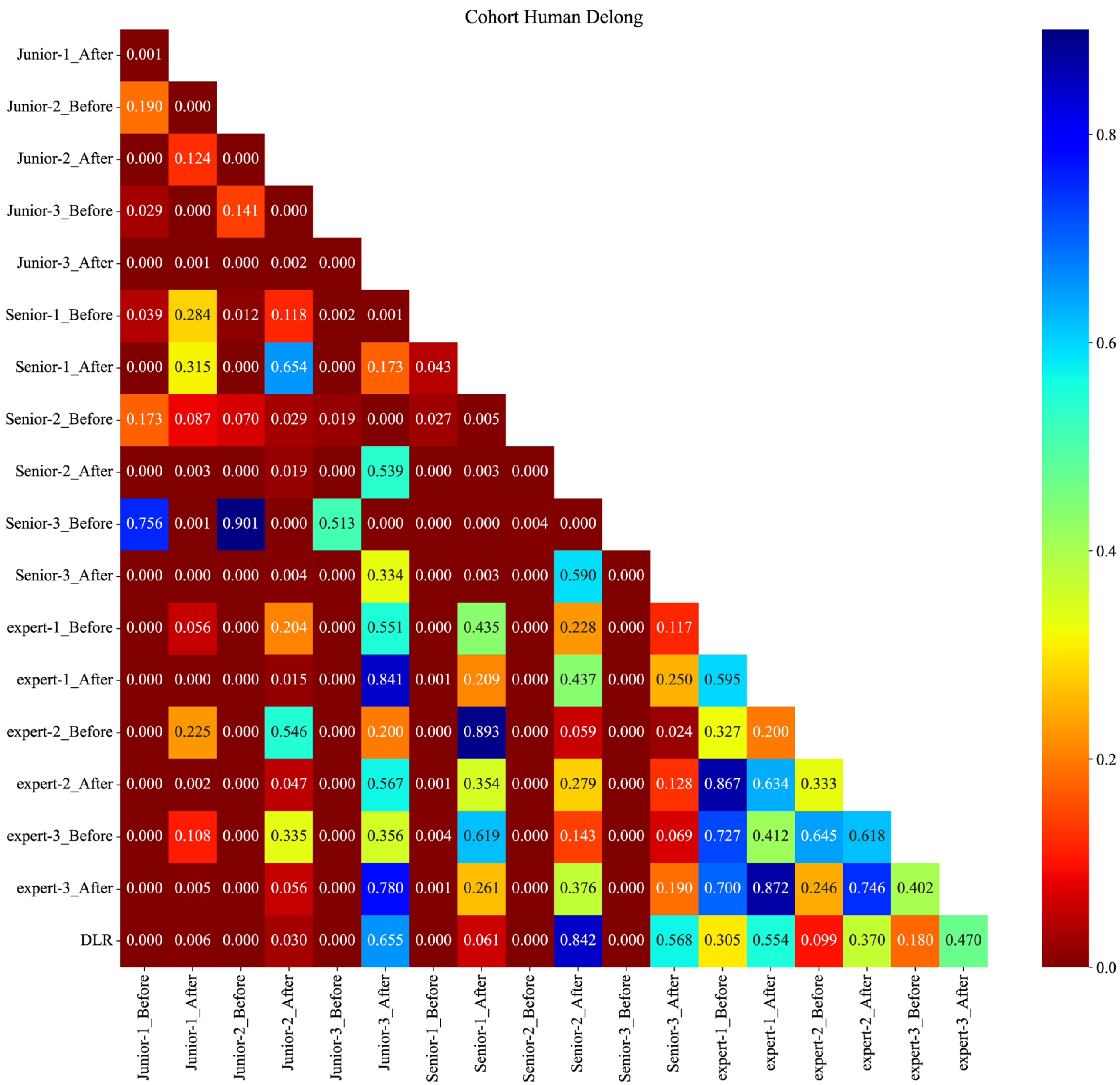

Figure S2 Delong test. The Delong test showed a statistically significant difference in the efficacy of the junior and senior physicians in diagnosing thyroid nodules before and after deep-learning radiomics assistance. However, there was no statistically significant difference in the diagnostic efficacy of the experts in diagnosing thyroid nodules. DLR, deep-learning radiomics.

https://dx.doi.org/10.21037/qims-23-1597

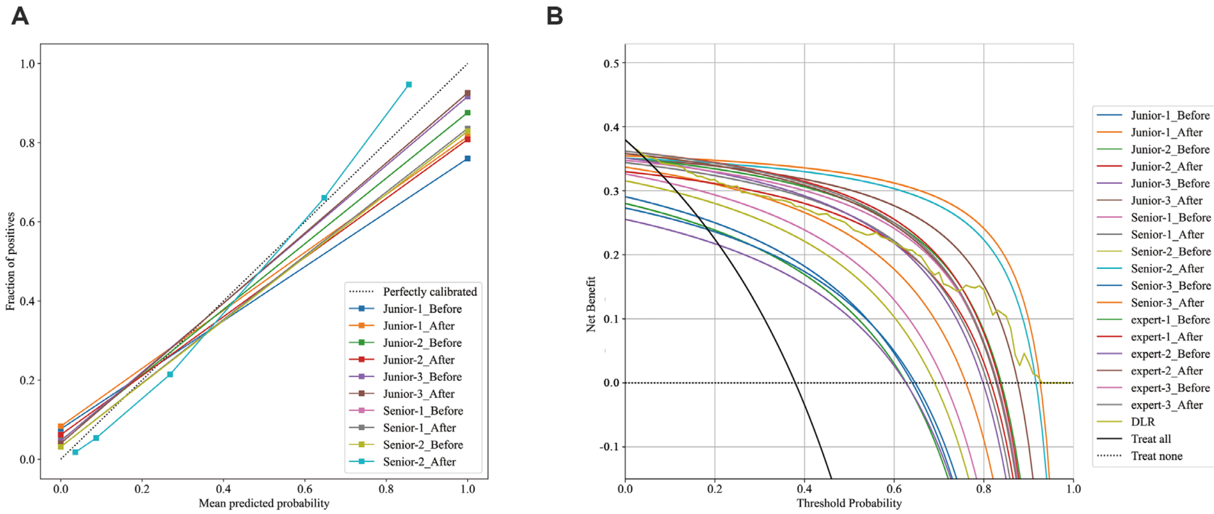

Figure S3 The calibration curve and DCA. (A) The calibration curve showed good calibration; (B) The DCA results showed that the models in which the junior physicians, senior physicians, and experts received deep-learning radiomics assistance had the best clinical utility. DLR, deep-learning radiomics; DCA, decision curve analysis.

https://dx.doi.org/10.21037/qims-23-1597