Appendix 1. Radiomics score (Rad-score) calculation formula and list of radiomics features

Rad-score = -20.12+6.31*GLSZM_SZE_0.5_1.5_Lloyd_32
+0.01*GLSZM_SZHGE_0.5_2_Lloyd_8
+87.10*GLCM_Variance_1_1.2_Equal_16
+3.24*GLCM_Variance_1_2_Lloyd_32
+0.16*GLRLM_HGRE_1.5_0.67_Lloyd_8
+50.66*GLCM_Variance_1.5_1.2_Equal_16
+3.22*Solidity
-1.97*Eccentricity

Appendix 2

Stratification analysis of the combined model. Stratification analysis were performed based on patient characteristics (sex and age) and CT protocols (CT device and slice thickness) to evaluate robustness of the radiomics model. The results showed that the characteristics of patients and CT protocols had less impact on the stability and robustness of the proposed model.

Stratified analysis of sex. The patients were divided into two groups: Male (n=90) and female (n=89). The AUCs were 0.9315 and 0.8921, respectively. The P-values were 0.5254 and 0.6613, respectively, when the two groups and the overall cohort were compared using the Delong test. The ROC curves are shown in Fig. S2A.

Stratified analysis of age. The patients were divided into two groups: Age ≥60 years (n=63) and age <60 years (n=116). The AUCs were 0.9321 and 0.9001, respectively. The P-values were 0.4590 and 0.7860, respectively, when the two groups and overall cohort were compared using the Delong test. The ROC curves are shown in Fig. S2B.

Stratified analysis of the CT device. The patients were divided into four groups: Toshiba Aquilion One (n=100), Siemens Emotion 16 (n=23), GE Discovery 64 (n=16), and Siemens Force (n=40). The AUCs were 0.9301, 0.9924, 0.7833 and 0.8725, respectively. The P-values were 0.5258, 0.3245, 0.4056 and 0.5540, respectively, when the four groups and overall cohort were compared using the Delong test. The ROC curves are shown in Fig. S2C.

Stratified analysis of the CT slice thickness. The patients were divided into two groups: 2.5 mm (n=15) and 3.0 mm (n=164). The AUCs were 0.9096 and 0.8800, respectively. The P-values were 0.9957 and 0.8136, respectively, when the two groups and overall cohort were compared using the Delong test. The ROC curves are shown in Fig. S2D.

Figure S1. Calibration curves of the subjective finding, radiomics signature and combined models in the training (A, C and E), internal and external validation cohorts (B, D and F).

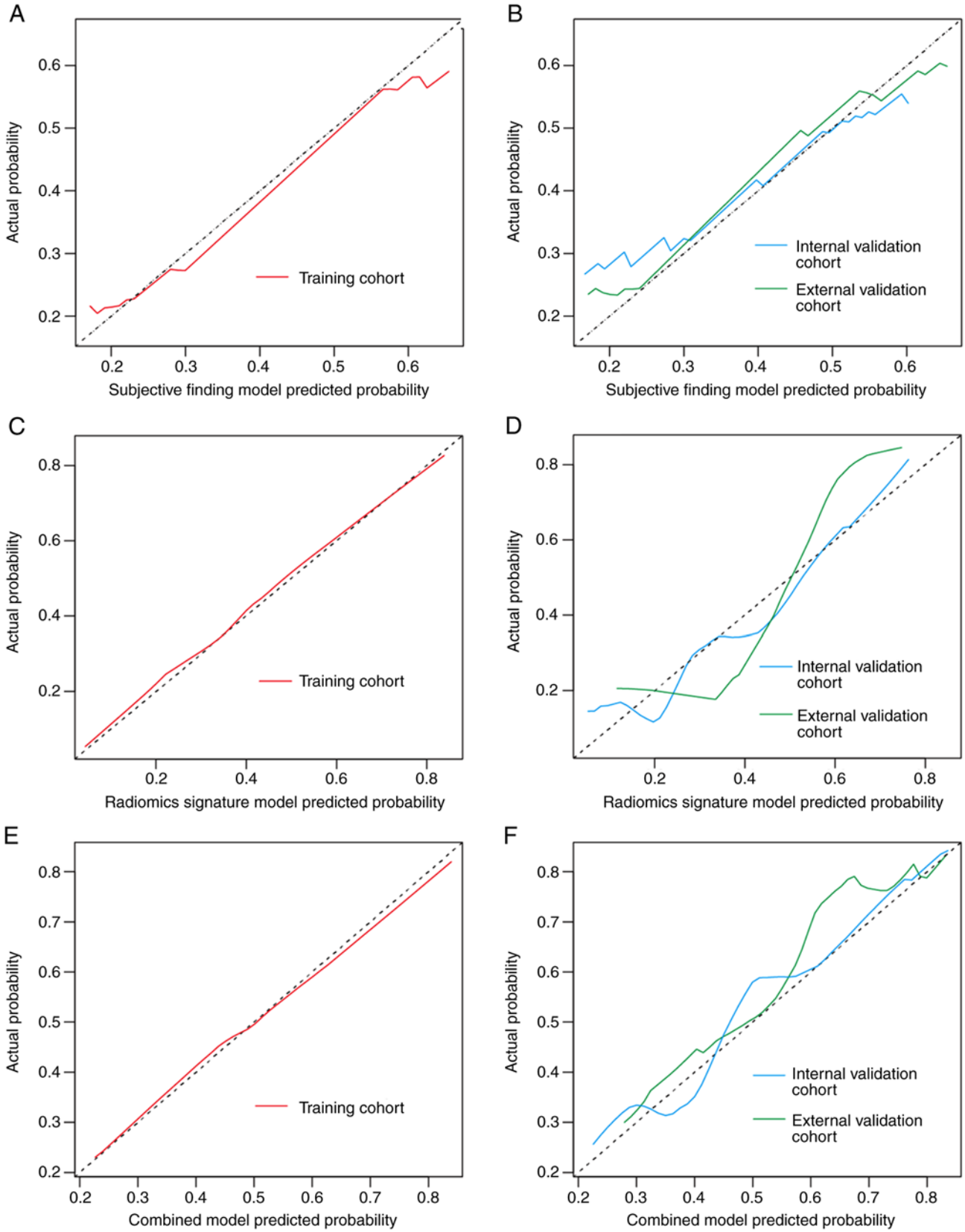

Figure S2. Stratified analysis of (A) sex, (B) age, (C) CT device, and (D) CT slice thickness. CT, computed tomography.

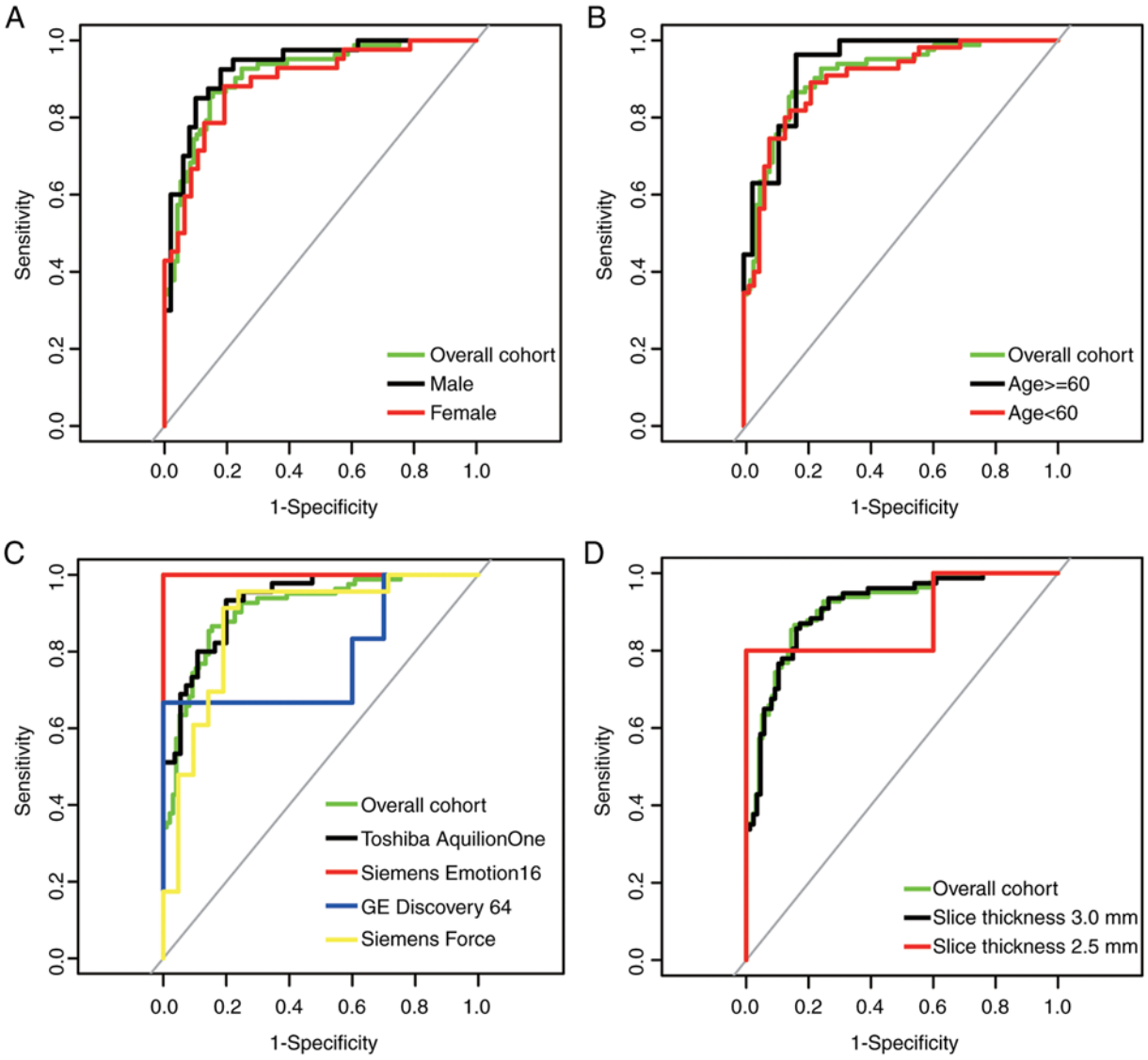

Table SI. Radiometric features and formulas.

Feature	Group	Formula
GLSZM	Small zone emphasis (SZE)	$SZE = \sum_{i=1}^{N_g}\sum_{j=1}^{L_z}\frac{p(i,j)}{j^2}$
	Small zone high gray-level emphasis (SZHGE)	$SZHGE = \sum_{i=1}^{N_g}\sum_{j=1}^{L_z}\frac{i^2 p(i,j)}{j^2}$
GLCM	Variance	$variance = \frac{1}{N_g \times N_g}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}[(i-\mu_i)^2 p(i,j) + (j-\mu_j)^2 p(i,j)]$
GLRLM	High gray-level run emphasis (HGRE)	$HGRE = \sum_{i=1}^{N_g}\sum_{j=1}^{L_r} i^2 p(i,j)$
Solidity		$Solidity = \frac{S}{S_{convex}}$
Eccentricity		$Eccentricity = \frac{\sqrt{\left[(A+B)-\sqrt{(A-B)^2+4H^2}\right]}}{\sqrt{\left[(A+B)+\sqrt{(A-B)^2+4H^2}\right]}}$

GLSZM, gray-level size zone matrix; GLCM, gray-level co-occurence matrix; GLRLM, gray-level run-length matrix.

Part 1: Gray-level size zone matrix (GLSZM)
Small Zone Emphasis (SZE):

$$SZE = \sum_{i=1}^{N_g}\sum_{j=1}^{L_z}\frac{p(i,j)}{j^2}$$

Small Zone High Gray-Level Emphasis (SZHGE):

$$SZHGE = \sum_{i=1}^{N_g}\sum_{j=1}^{L_z}\frac{i^2 p(i,j)}{j^2}$$

N_g represents the pre-defined number of quantized gray-levels set in volume.
L_z represents the size of the largest zone (of any gray-level) in volume.
$p(i,j)$ represents the normalized size zone matrix, defined as $p(i,j) = \frac{P(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{L_z} P(i,j)}$. $P(i,j)$ represents the number of 3D zones of gray-levels i and of size j in volume.

Part 2: Gray-level co-occurence matrix (GLCM)
Variance

$$variance = \frac{1}{N_g \times N_g}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}[(i-\mu_i)^2 p(i,j) + (j-\mu_j)^2 p(i,j)]$$

N_g represents the pre-defined number of quantized gray-levels set in volume.

$p(i,j)$ represents the normalized co-occurence matrix and equal to $\frac{P(i,j)}{\sum P(i,j)}$,

$P(i,j)$ represents the number of times voxels of gray-levels i were neighbors with voxels of gray-level j in volume.

μ_i represents the mean gray level intensity of pi and defined as $\mu_i = \sum_{i=1}^{N_g} p_i(i)i$,

μ_j represents the mean gray level intensity of pj and defined as $\mu_j = \sum_{i=1}^{N_g} p_j(j)j$.

Part 3: Gray-level run-length matrix (GLRLM):

$$HGRE = \sum_{i=1}^{N_g}\sum_{j=1}^{L_r} i^2 p(i,j)$$

N_g represents the pre-defined number of quantized gray-levels set in volume.
L_r represents the length of the longest run (of any gray-level) in volume.

$p(i,j)$ represents the normalized run length matrix, defined as $p(i,j) = \frac{P(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{L_r} P(i,j)}$, $P(i,j)$ represents the number of runs of gray-level i and of length j in volume.

Part 4: Solidity:

$$Solidity = \frac{S}{S_{convex}}$$

S is the area of the region and S_{convex} is the area of the convex hull for the region.

Part 5: Eccentricity:

$$Eccentricity = \frac{\sqrt{\left[(A+B)-\sqrt{(A-B)^2+4H^2}\right]}}{\sqrt{\left[(A+B)+\sqrt{(A-B)^2+4H^2}\right]}}$$

where $A = \sum m_i(y_i^2 + z_i^2)$, $B = \sum m_i(x_i^2 + z_i^2)$, $H = \sum m_i x_i y_i$.

(x_i, y_i, z_i) represents the ith pixel. m_i represents the grayscale of the ith pixel.

Table SII. Pearson's correlation coefficient analysis of the selected radiomics features.

Features	GLSZM_SZE_ 0.5_1.5_Lloyd_32	GLSZM_SZHGE_ 0.5_2_Lloyd_8	GLCM_Variance_ 1_1.2_Equal_16	GLCM_Variance_ 1_2_Lloyd_32	GLCM_Variance_ 1.5_1.2_Equal_16	GLRLM_HGRE_ 1.5_0.67_Lloyd_8	Eccentricity	Solidity
GLSZM_SZE_0.5_1.5_Lloyd_32	\	0.116	-0.329	0.528	0.315	-0.409	0.288	0.172
GLSZM_SZHGE_0.5_2_Lloyd_8	-	\	0.379	-0.035	0.249	0.302	0.288	-0.070
GLCM_Variance_1_1.2_Equal_16	-	-	\	0.061	-0.108	0.881	0.336	-0.230
GLCM_Variance_1_2_Lloyd_32	-	-	-	\	0.200	-0.017	0.264	0.137
GLCM_Variance_1.5_1.2_Equal_16	-	-	-	-	\	-0.126	0.242	-0.114
GLRLM_HGRE_1.5_0.67_Lloyd_8	-	-	-	-	-	\	0.248	-0.221
Eccentricity	-	-	-	-	-	-	\	-0.224
Solidity	-	-	-	-	-	-	-	\

‘-’ indicates that the Pearson correlation coefficient analysis was conducted. ‘\’ indicates the Pearson correlation analysis of the same feature. Pearson correlation analysis on selected features were performed using LASSO logistic regression, and then we selected the radiomics features with Pearson correlation coefficients of <0.9 as independent features for building the radiomics signature.

Table SIII. Hosmer-Lemeshow test of the three models.

Model	Hosmer-Lemeshow test for each model		
	Training cohort	Internal validation cohort	External validation cohort
Subjective finding model	P=0.312	P=0.414	P=0.354
Radiomics signature model	P=0.954	P=0.741	P=0.435
Combined model	P=0.413	P=0.334	P=0.198

The calibration curve of each model showed good agreement in the training, internal and external validation cohorts. The Hosmer-Lemeshow test was not significant (all P>0.05), demonstrating a good fit.