Appendix 1

1. Input Variables of This Study

The input variables encompass clinical, ultrasonographic, radiomic features, and deep learning features. Clinical features, such as patient age and gender, were obtained from electronic medical records. Ultrasonographic features consist of quantitative characteristics acquired through ultrasound imaging technology. Two radiologists with over five years of experience in thyroid ultrasound examination reviewed the dataset and interpreted the semantic features of all tumors based on the following criteria: tumor size, tumor location, and ultrasonic feature categories (position, orientation, margin, halo, composition, echogenicity, echotexture, shape, focal echogenic foci, and blood flow) according to the guidelines of Chinese -TIRADS (C-TIRADS) and the American College of Radiology TIRADS (ACR-TIRADS). Consensus was reached through discussion. The extracted ultrasound radiomics features included 18 first-order statistical features, 14 shape features, 24 gray-level co-occurrence matrix (GLCM) features, 14 gray-level dependence matrix (GLDM) features, 16 gray-level run length matrix (GLRLM) features, 16 gray-level size zone matrix (GLSZM) features, and 5 neighboring gray-tone difference matrix (NGTDM) features. A total of 107 radiomic features were extracted. Additionally, deep learning features were extracted using classic deep learning models such as ResNet18, ResNet50, DenseNet169, DenseNet201, and Inception_v3. A total of 2048 deep learning features were extracted.

2. Output Variables of This Study

The output variables of this study are radiomic features (Rad signature) and deep learning features (DTL signature), which are used to predict the malignancy risk of cystic-solid thyroid nodules (CSTN) and differentiate between benign and malignant cases.

The formula for calculating Rad-signature was illustrated as follows:
Rad signature= 0.10309278350515325 -
0.030996* original_gldm_GrayLevelNonUniformity –
0.020580 * original_glrlm_ShortRunEmphasis –
0.037967*original_glszm_SizeZoneNonUniformityNormalized –
0.024557 * original_ngtdm_Busyness –
0.036121 * original_ngtdm_Contrast +
0.086501 * original_shape_Elongation +
0.044142 * original_shape_SurfaceVolumeRatio

The formula for calculating DTL-signature was illustrated as follows:
DTL signature= 0.10309278350515462-0.022104 * DL_7 +0.006726 * DL_21 -0.001738 * DL_83 +0.006708 * DL_127 +0.019528 * DL_207 -0.002641 * DL_534 -0.001205 * DL_567 +0.000308 * DL_593 +0.005997 * DL_599 -0.009302 * DL_667 +0.003340 * DL_696 -0.002844 * DL_733 +0.003270 * DL_810 -0.023596 * DL_821 +0.011643 * DL_840 -0.006123 * DL_909 -0.002959 * DL_913 +0.015464 * DL_954 -0.013155 * DL_1030 +0.018269 * DL_1056 +0.006600 * DL_1068 +0.000661 * DL_1263 -0.003306 * DL_1356 -0.024122 * DL_1370 -0.005877 * DL_1513 -0.024585 * DL_1584 +0.004996 * DL_1620 -0.006139 * DL_141.1 -0.001575 * DL_283.1 -0.001104 * DL_410.1 +0.012490 * DL_421.1 +0.001094 * DL_502.1 +0.013885 * DL_613.1 -0.005574 * DL_959.1 -0.003843 * DL_1106.1 +0.010535 * DL_1118.1 -0.018431 * DL_1138.1 +0.006265 * DL_1180.1 +0.004419 * DL_1184.1 +0.006952 * DL_1201.1 -0.013578 * DL_1261.1 -0.001381 * DL_1285.1 -0.005365 * DL_1362.1 +0.008001 * DL_1478.1 +0.005565 * DL_1493.1 -0.004708 * DL_1509.1 -0.001743 * DL_1633.1 -0.011140 * DL_1767 +0.017847 * DL_1827 +0.002670 * DL_1848 -0.017633 * DL_1850 -0.002503 * DL_1872 +0.017027 * DL_1904

https://dx.doi.org/10.21037/gs-2024-551

3. Detailed description of the training process in this study:

3.1 Data preprocessing

Firstly, a physician with five years of experience in thyroid ultrasonography manually delineated the region of interest (ROI) of CSTN using ITK-SNAP software (version 3.8.0). Then, another thyroid ultrasonography expert with seven years of experience reviewed and reached a consensus on this ROI. Missing values in the original data were handled using multiple imputation methods. For imaging data, various processing techniques including segmentation, denoising, and normalization were employed, with grayscale values normalized to [0, 1].

3.2 Radiomics Signature

3.2.1. For image data, we utilized the open-source software PyRadiomics (https://pyradiomics.readthedocs.io/) to extract features, which were categorized into seven groups: (a) shape features; (b) first-order features; (c) gray-level co-occurrence matrix (GLCM) features; (d) gray-level dependence matrix (GLDM) features; (e) gray-level run-length matrix (GLRLM) features; (f) gray-level size zone matrix (GLSZM) features; and (g) neighboring gray level dependence matrix (NGTDM) features. These quantitative radiomic features were extracted from three types of images: (a) original images; (b) Laplacian of Gaussian (LoG) filtered images; and (c) wavelet-transformed images, yielding a comprehensive set of features denoted as Rad_feature.

3.2.2. These features were then normalized (Z-score) to transform the data to follow a normal distribution with a mean of 0 and a standard deviation of 1 (N~(0, 1)).

3.2.3. The Spearman correlation coefficient was used to calculate the correlation between the extracted radiomics features.

3.2.4. For features with a correlation coefficient greater than 0.9, only one of the two was retained.

3.2.5. Lasso regression was applied to perform cross-validation on the data and select the best penalty coefficient .

3.2.6. The data was randomly partitioned, and 5-fold cross-validation was employed, keeping the best model.

3.2.7. Various machine learning algorithms, including random forest (RF), k-nearest neighbors (KNN), logistic regression (LR), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and Extra Trees, were used to train the extracted features on the development set.

3.3 Deep Learning Signature

3.3.1. The largest ROI section in the imaging data was cropped to form roi_images. Based on transfer learning, a model was trained, and features were extracted to obtain deep learning features.

3.3.2. Ultrasound images in the training and testing sets were cropped according to the defined ROI, with a unified size adjustment to 224 × 224. The cropped ROI images were then randomly divided into a training set (70%) and a testing set (30%).

3.3.3. ResNet18, ResNet50, DenseNet169, DenseNet201, and Inception_v3 were used, loaded with ImageNet pre-trained parameters: learning rate of 0.01, loss function as cross-entropy loss, 30 epochs, and the model parameters with the highest accuracy on the test set were retained. The optimizer used was stochastic gradient descent (SGD).

3.3.4. Features were extracted from roi_images using the second-to-last layer of the model, namely the avgpool layer.

3.3.5. The avgpool layer contained 2048 features, which were reduced to 100 dimensions using principal components analysis (PCA).

3.3.6. The features were normalized (Z-score) to transform the data to follow a normal distribution N~(0, 1).

3.3.7. Various machine learning algorithms, including random forest (RF), k-nearest neighbors (KNN), logistic regression (LR), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and Extra Trees, were used to train the data on the development set.

3.4. Integrated Model Construction

First, we extracted features from the optimal radiomics model (Extra Trees) and the optimal deep transfer learning model (DenseNet169). Then, we performed feature fusion on the extracted radiomics and deep transfer learning features. Finally, we obtained the fused features of radiomics and deep transfer learning to construct the deep transfer learning and radiomics

https://dx.doi.org/10.21037/gs-2024-551

(DTLR) model (See *Figure S1*, *Figure S2*). The filtered clinical and ultrasound features were then combined with the handcrafted radiomics features and deep transfer learning features to create the final predictive model. The Stacking method was used to aggregate the model metrics.

4. The calculation formula of the performance metrics

1. Area Under the Curve (AUC): AUC represents the area beneath the receiver operating characteristic curve (ROC), which reflects the classification ability of a model across various thresholds. The AUC value ranges from 0 to 1, with higher values indicating better classification performance of the model. AUC comprehensively considers both the true positive rate (sensitivity) and false positive rate (1-specificity) of the model, providing an overall performance evaluation that is suitable for situations involving unbalanced datasets.

2. Accuracy: Accuracy measures the proportion of all correctly identified instances over the total instances. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP is the true positive instances, TN is the true negative instances, FP is the false positive instances, and FN is the false negative instances.

Accuracy serves as the most intuitive evaluation metric, suitable for scenarios where the distribution of sample categories is relatively balanced. However, in cases of unbalanced datasets, accuracy can be misleading, thus it should be used in conjunction with other metrics.

3. Sensitivity/Recall: Sensitivity, also known as recall, indicates the model's ability to correctly identify positive samples. Sensitivity measures the proportion of actual positives that are correctly identified by the model. It is calculated as:

$$Sensitivity = \frac{TP}{TP + FN}$$

In medical diagnosis, missed detections (false negatives) can lead to severe consequences, making sensitivity a crucial metric, especially in situations where minimizing missed detections is essential, such as tumor screening.

4. Specificity: Specificity measures the proportion of actual negatives that are correctly identified by the model. It is calculated as:

$$Specificity = \frac{TN}{TN + FP}$$

It measures the false positive rate, providing insights into the model's performance in recognizing negative samples. In certain clinical contexts, false positives can result in unnecessary treatments or anxiety, highlighting the importance of specificity.

5. Positive Predictive Value (PPV): The PPV represents the proportion of samples predicted as positive that are actually positive. It is calculated as:

$$PPV = \frac{TP}{TP + FP}$$

The PPV measures the accuracy of a model, particularly in clinical decision-making where doctors need to understand the reliability of positive results.

6. Negative Predictive Value (NPV): The NPV indicates the proportion of samples predicted as negative that are actually negative. It is calculated as:

$$NPV = \frac{TN}{TN + FN}$$

The NPV is equally important, providing critical information when evaluating model performance, especially when

https://dx.doi.org/10.21037/gs-2024-551

ensuring the reliability of negative results.

7. F1-score: The F1-score is the harmonic mean of precision and recall, It is calculated as:

$$F1\text{-}score = \frac{2*(Accuracy*Recall)}{Accuracy+\mathrm{Re}call}$$

The F1 score is particularly useful when dealing with unbalanced datasets, allowing a balance between precision and sensitivity. Through the F1-score, researchers can evaluate the overall performance of the model in classifying both positive and negative samples.

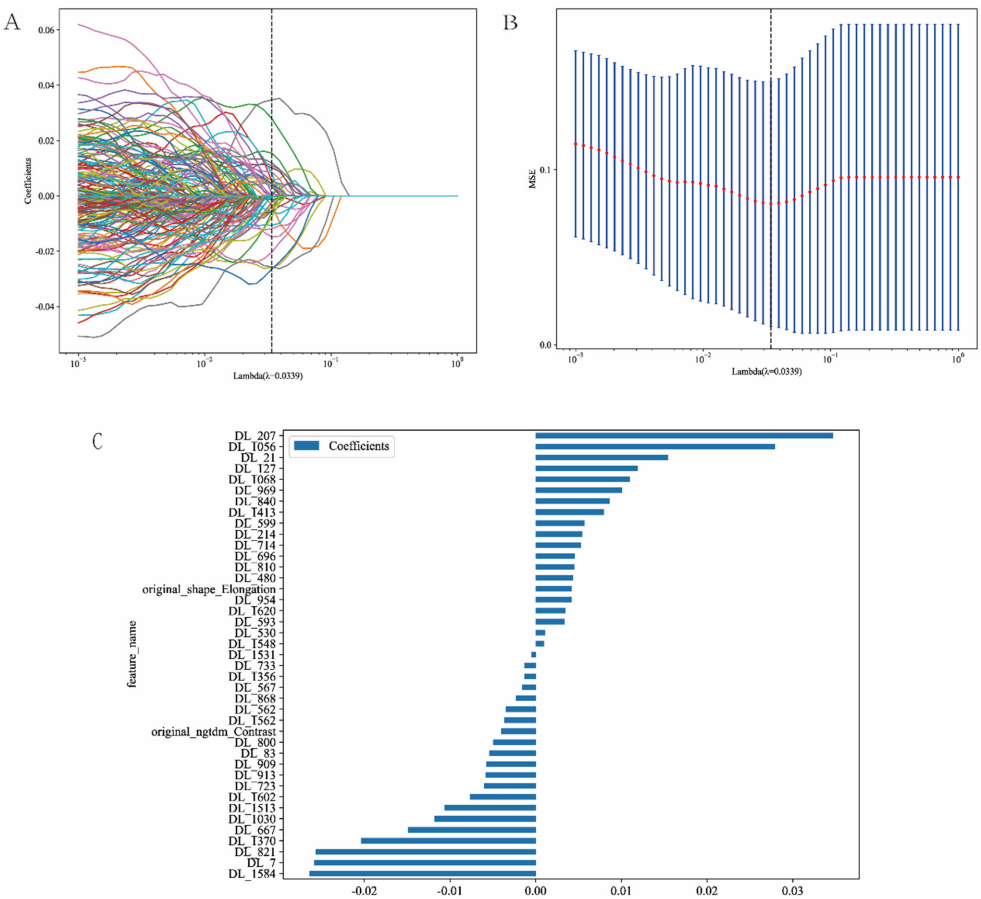

Figure S1 Deep transfer learning and radiomics (DTLR) feature selection using the least absolute shrinkage and selection operator (LASSO) logistic regression model in the training cohort and the histogram of the DLR-score based on the selected features. (A) LASSO regression. (B) Mean square error of 5-fold validation. (C) The histogram of the DTLR score is based on the selected features.

https://dx.doi.org/10.21037/gs-2024-551

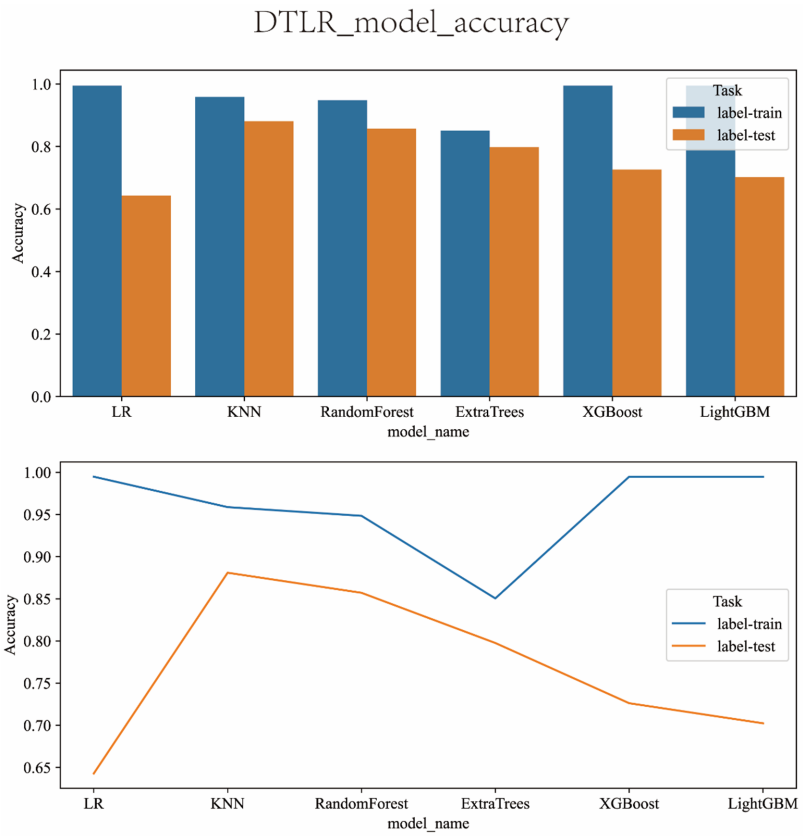

Figure S2 Deep transfer learning and radiomics (DTLR) model accuracy.

https://dx.doi.org/10.21037/gs-2024-551

Table S1 Demographic comparison between training and independent test cohorts

Characteristics and US features	Total (n=278)	Test cohort (n=84)	Train cohort (n=194)	P value
Age (years)	44.48±13.38	44.60±14.34	44.43±12.97	0.87
Size (mm)	31.55±11.81	30.54±12.30	31.98±11.59	0.37
Gender				0.55
Male	40 (14.39)	10 (11.90)	30 (15.46)	
Female	238 (85.61)	74 (88.10)	164 (84.54)	
Location				0.28
Upper	18 (6.47)	7 (8.33)	11 (5.67)	
Mid	167 (60.07)	45 (53.57)	122 (62.89)	
Lower	90 (32.37)	30 (35.71)	60 (30.93)	
Sthmus	3 (1.08)	2 (2.38)	1 (0.52)	
Composition				0.76
Predominately solid	227 (81.65)	70 (83.33)	157 (80.93)	
Predominately cystic	51 (18.35)	14 (16.67)	37 (19.07)	
Echogenicity				0.60
Hypoechoic	53 (19.06)	13 (15.48)	40 (20.62)	
Isoechoic	218 (78.42)	69 (82.14)	149 (76.80)	
Hyperechoic	7 (2.52)	2 (2.38)	5 (2.58)	
Echotexture				0.57
Homogeneous	12 (4.32)	5 (5.95)	7 (3.61)	
Heterogeneous	266 (95.68)	79 (94.05)	187 (96.39)	
Orientation				>0.99
Horizontal	272 (97.84)	82 (97.62)	190 (97.94)	
Vertical	6 (2.16)	2 (2.38)	4 (2.06)	
Echogenic_foc				0.96
No	234 (84.17)	70 (83.33)	164 (84.54)	
Microcalcifications	32 (11.51)	10 (11.90)	22 (11.34)	
Macrocalcifications	12 (4.32)	4 (4.76)	8 (4.12)	
Margin				0.65
Circumscribed	214 (76.98)	64 (76.19)	150 (77.32)	
Ill-defined	21 (7.55)	5 (5.95)	16 (8.25)	
Irregular margin	43 (15.47)	15 (17.86)	28 (14.43)	
Halo				0.29
Present halo	104 (37.41)	27 (32.14)	77 (39.69)	
Absent halo	174 (62.59)	57 (67.86)	117 (60.31)	
Acute_angle				0.94
No	246 (88.49)	75 (89.29)	171 (88.14)	
Yes	32 (11.51)	9 (10.71)	23 (11.86)	
CDFI				0.82
Peripheral vascularity	206 (74.10)	61 (72.62)	145 (74.74)	
Mixed vascularity	72 (25.90)	23 (27.38)	49 (25.26)	

Data are presented as mean ± standard deviation or n (%). CDFI, color Doppler flow imaging; US, ultrasound.

https://dx.doi.org/10.21037/gs-2024-551