Original Article

Multi-type classification of lung nodules based on CT radiomics and ensemble learning for diversity weighting

Guozhi Tang[1,2], Lingyan Du[1,2], Shihai Ling[1,2], Yue Che[1,2], Xin Chen[3]

[1]School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin, China; [2]Artificial Intelligence Key Laboratory of Sichuan Province, Yibin, China; [3]Department of Integrated Traditional Chinese and Western Medicine, Zigong First People's Hospital, Zigong, China

Contributions: (I) Conception and design: G Tang, L Du; (II) Administrative support: L Du; (III) Provision of study materials or patients: Y Che, S Ling; (IV) Collection and assembly of data: S Ling, G Tang; (V) Data analysis and interpretation: Y Che, G Tang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Lingyan Du, PhD. School of Automation and Information Engineering, Sichuan University of Science and Engineering, No. 1 Baita Road, Yibin 644000, China; Artificial Intelligence Key Laboratory of Sichuan Province, Yibin, China. Email: dulingyan@suse.edu.cn.

Background: The accurate classification of lung nodules is critical to achieving personalized lung cancer treatment and prognosis prediction. The treatment options for lung cancer and the prognosis of patients are closely related to the type of lung nodules, but there are many types of lung nodules, and the distinctions between certain types are subtle, making accurate classification based on traditional medical imaging technology and doctor experience challenging. This study adopts a novel approach, using computed tomography (CT) radiomics to analyze the quantitative features in CT images to reveal the characteristics of lung nodules, and then employs diversity-weighted ensemble learning to enhance the accuracy of classification by integrating the predictive results of multiple models.

Methods: We extracted lung nodules from the Lung Image Database Consortium image collection (LIDC-IDRI) dataset and derived radiomics features from the nodules. For the classification tasks of seven types of lung nodules, each was split into binary classifications. Two model-building methods were employed: M1 (equal-weighted voting ensemble classifier) and M2 (diversity-weighted voting ensemble classifier). Models were evaluated using 10-fold cross-validation with metrics including the area under the receiver operating characteristic curve (AUC), accuracy, specificity, and sensitivity.

Results: Both methods effectively completed classification tasks. The M2 method outperformed M1, particularly in classifying texture, calcification, and the benign and malignant nature of lung nodules. The AUC values of the M2 method in the four subclassifications of texture types of lung nodules were 0.9913, 0.8838, 0.9525, and 0.8845, with the corresponding accuracies of 0.9651, 0.8116, 0.9000, and 0.8284, respectively. In the classification of the degree of calcification of lung nodules, the AUC value of the M2 method was 0.9775 with an accuracy of 0.9642. In the classification of the benign and malignant nature of lung nodules, the AUC value of the M2 method was 0.8953 with an accuracy of 0.8168. The combination of CT radiomics and diversity-weighted ensemble learning effectively identifies lung nodule types, providing a novel method for the precise classification of lung nodules and aiding personalized lung cancer treatment and prognosis prediction.

Conclusions: The combination of CT radiomics and ensemble learning for diversity weighting can be well realized to identify the type of lung nodules.

Keywords: Radiomics; lung nodule classification; ensemble classifier; medical images; machine learning

Submitted Jun 28, 2024. Accepted for publication Oct 09, 2024. Published online Nov 29, 2024.
doi: 10.21037/qims-24-1315
View this article at: https://dx.doi.org/10.21037/qims-24-1315

Introduction

The global prevalence of lung cancer makes it one of the most lethal malignancies (1). It begins in the lungs and initially develops as a single or multiple nodules that can eventually spread to other organs and tissues in the body. Lung nodules can be categorized into different nodule types based on their shape, margins, and internal features, such as lobulated nodules, spiculated nodules, and ground-glass nodules (GGNs) (2). Determining the type of nodules is crucial for doctors to assess the risk of the nodule becoming cancerous and to choose the appropriate personalized treatment for the patient (3). In the past two decades, there has been a remarkable surge in artificial intelligence (AI) technology, leading to an increasing number of researchers focusing on investigating computer-assisted diagnostic (CAD) systems that integrate AI technology. For example, Ni *et al.* (4) developed an artificial neural network (ANN)-based model for the classification of eight types of lung nodules using computed tomography (CT) images. However, training deep learning models with CT images requires significant resources and costs, especially for three-dimensional (3D) medical images. Furthermore, the efficacy of deep learning methods heavily relies on large amounts of quality data, which are often limited by privacy and ethical concerns in medicine.

Radiomics is a technique for analyzing lesions by employing digital image processing methods to extract high-throughput features from medical images that are imperceptible to the human eyes (5,6). In contrast to image data, radiomics feature data are commonly stored in a tabular format, which offers a more straightforward and organized data structure. This format aligns well with the mature utilization of traditional machine learning algorithms, ensuring enhanced speed and efficiency in data processing. Compared to deep learning, traditional machine learning algorithms can be effectively trained on smaller datasets. For example, Rundo *et al.* (7) employed a combination of radiomics and machine learning techniques, which demonstrated relatively low training data requirements, training time, and computational power costs. Their approach successfully enabled the classification of solid versus sub-solid lung nodules as well as non-solid versus partially solid lung nodules. The mean area under the receiver operating characteristic (ROC) curve (AUC) for these two classifications in this study reached 0.89±0.02 and 0.80±0.18, respectively. This study confirms that radiomics combined with machine learning can achieve the classification of texture types of lung nodules. Based on this, we consider whether the combination of radiomics and machine learning can also achieve the classification of more lung nodule types, like the study by Ni *et al.* (4). Recent advancements have also explored the broad applications of radiomics in lung cancer management. Although the primary focus of this study is the classification of nodule morphology categories, it is worth noting that radiomics and the morphological characteristics of lung nodules have also shown potential in other related areas, such as predicting lung tumor growth intervals (8,9) and predicting lung adenocarcinoma and its subtypes (10,11). Additionally, accurately predicting the type of lung nodule (e.g., solid or subsolid) further contributes to the advancement of such studies. These related studies highlight the versatility of radiomics in offering deeper insights into lung cancer behavior, thereby supporting early detection and personalized treatment approaches.

Accurate classification of different nodule types requires the construction of multitasking models with high generalisability and robustness. Ensemble learning is a powerful strategy in machine learning that improves model performance by building and combining multiple learners. Ensemble learning is a powerful strategy in machine learning that improves model performance by building and combining multiple learners. It can effectively integrate the advantages of different models when dealing with multiple different tasks, thus demonstrating superior generalization ability and robustness than a single model.

In summary, the objective of this study is to integrate CT radiomics and ensemble learning methods for precise classification of seven distinct types of lung nodules, thereby providing more efficient and accurate computer-aided decision support for lung cancer diagnosis. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1315/rc).

Methods

Dataset

The Lung Image Database Consortium image collection (LIDC-IDRI) (12) is an international web-based resource containing images and lesion annotations for lung diagnosis, lung cancer screening, and chest CT scans, specifically designed for the development, training, and evaluation of CAD techniques for lung cancer detection and diagnosis.

The dataset includes 1,018 chest CT scans, with a peak voltage of 120 to 140 kV and a peak current of 40 to 624 mA during CT image acquisition. The CT scan images are in Digital Imaging and Communications in Medicine (DICOM) format, which is the standard image format in the medical field. The image size is 512×512 pixels, and the pixel values are expressed in Hounsfield units (HU). The CT scans that make up the dataset were acquired using the following slice thickness settings: 0.6, 0.75, 0.9, 1.0, 1.25, 1.5, 2.0, 2.5, 3.0, 4.0, and 5.0 mm. The lesions range from 3 to 30 mm in diameter, and each lesion was independently labeled by experienced radiologists. There is no shortage of multinodular cases in the CT data, and each nodule has been assessed in detail for the type of features (13), including malignant potential, sphericity, margin definition, spiculation, lobulation, texture, calcification, and internal structure. Considering the differences in nodule labeling by different physicians, we extracted nodules that obtained consensus from at least three physicians in the current study. After this screening, the final number of nodules identified was 1,426. Specific data on nodule types are shown in *Table 1*. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

To address the imbalance in the number of type categories among the same lung nodule types in the LIDC-IDRI dataset, we adopted a hierarchical stratification processing strategy. This approach first segments the type hierarchically based on the number of nodules. Then, it assigns binary classification labels based on the characteristics associated with the type itself. For example, based on the annotations provided in the original dataset for lobulated lung nodules, there are five lobulation grades: grade 1 (no lobulation), grade 2 (nearly no lobulation), grade 3 (medium lobulation), grade 4 (nearly marked lobulation), and grade 5 (marked lobulation). A hierarchical framework was constructed based on these grades, where grades 1 and 2 were categorized as "lowly lobulated nodules", grades 4 and 5 were categorized as "highly lobulated nodules", and grade 3 was categorized as "moderately lobulated nodules". Furthermore, to achieve precise binary classification, we established three sets of binary classification labels based on the aforementioned hierarchical stratification outcomes. The labels "3 [0]–12 [1]" represent the classification task between moderately lobulated and lowly lobulated lung nodules, "45 [0]–12 [1]" represent the classification task between highly lobulated and lowly lobulated lung nodules, and "45 [0]–3 [1]" represents the classification task between highly lobulated and moderately lobulated lung nodules. *Table 2* for further elaboration.

However, even after stratification, the problem of imbalance in the number of classification classes among classification labels remains prominent. For this reason, we further subdivide the classes with larger sample sizes into multiple small subsets and train on these subsets separately. These small subsets are collectively called the 'training subset'. For example, in the lung nodules sphericity classification task, the number of classification classes for binary classification is 411 [0] and 985 [1], respectively, with a ratio close to 1:2. Therefore, during training, we randomly divide the class labeled '1' in this classification dataset into two equal parts, forming two small training subsets together with the class labeled '0'. The number of classification classes for the two small training subsets is 441:493 and 441:491, respectively. The classification performance is finalized by calculating the average of the training results of each subset. This approach aims to mitigate the impact of imbalance in classification label categories on classification performance.

Feature processing

CT image pre-processing and region-of-interest segmentation

To eliminate batch effects in CT images from the LIDC-IDRI dataset due to differences in institutions and equipment, this study employed various standardization and correction techniques on the original CT images. First, intensity normalization (normalizing all CT image intensity values to the range of 0 to 1) and voxel resampling (resampling all slice thicknesses to 1 mm) were applied to ensure image uniformity. Next, the region of interest (ROI) masks required for radiomics feature extraction were obtained from the true nodule contour annotations in the dataset. Due to variations in contour annotations by different physicians, a 50% consistency criterion was used to extract the ROI masks. In practical applications, 3D Slicer (www.slicer.org) can be used for semi-automatic segmentation of ROI slices to achieve effective nodule ROI extraction.

Feature extraction

In this study, we used the PyRadiomics (5) tool to extract radiomics features from lung CT images provided by the LIDC-IDRI dataset. This process involves using nodule masks and actual nodule information from the dataset. We extracted a total of 1,064 features from each case of

Table 1 Distribution of nodule types in the LIDC-IDRI dataset

Annotation of nodule type in the dataset	Description	Nodule type grading	Quantities
Sphericity	3D shape of the nodule	1: linear	1
		2: ovoid/linear	67
		3: ovoid	373
		4: ovoid/round	598
		5: round	387
Lobulation	Degree of lobulation	1: no lobulation	730
		2: nearly no lobulation	363
		3: medium lobulation	201
		4: near marked lobulation	90
		5: marked lobulation	42
Spiculation	Degree of spiculation	1: no spiculation	843
		2: nearly no spiculation	334
		3: medium spiculation	127
		4: near marked spiculation	69
		5: marked spiculation	53
Texture	Nodule texture traits (solid, ground glass, or mixed)	1: non-solid/GGO	205
		2: non-solid/mixed	65
		3: part solid/mixed	115
		4: solid/mixed	214
		5: solid	827
Margin	Description of how well-defined the nodule margin is	1: poorly defined	115
		2: near poorly defined	194
		3: medium margin	186
		4: near sharp	399
		5: sharp	532
Calcification	Pattern of calcification	1: popcorn	0
		2: laminated	0
		3: solid	100
		4: non-central	4
		5: central	10
		6: absent	1,312
Malignancy	Subjective assessment of the likelihood of malignancy	1: highly unlikely	127
		2: moderately unlikely	164
		3: indeterminate	685
		4: moderately suspicious	276
		5: highly suspicious	174
Internal structure	Internal composition of the nodule	1: soft tissue	1,417
		2: fluid	3
		3: fat	0
		4: air	6

LIDC-IDRI, the lung image database consortium image collection; 3D, three-dimensional; GGO, ground glass opacity.

Table 2 Binary classification labels

Nodule's types	Nodule type grading [binary label]	Quantities	Code name
Sphericity	123 [0]–45 [1]	441–985	Sph1
Lobulation	3 [0]–12 [1]	201–1,093	Lob1
	45 [0]–12 [1]	132–1,093	Lob2
	45 [0]–3 [1]	132–201	Lob3
Spiculation	3 [0]–12 [1]	127–1,177	Spi1
	45 [0]–12 [1]	122–1,177	Spi2
	45 [0]–3 [1]	122–127	Spi3
Texture	1 [0]–5 [1]	205–827	Tex1
	1 [0]–234 [1]	205–394	Tex2
	23 [0]–5 [1]	180–827	Tex3
	4 [0]–5 [1]	214–827	Tex4
Margin	3 [0]–12 [1]	186–309	Mar1
	12 [0]–45 [1]	309–931	Mar2
	3 [0]–45 [1]	186–931	Mar3
Calcification	345 [0]–6 [1]	114–1,312	Cal1
Malignancy	123 [0]–45 [1]	976–450	Mal1
Internal structure	–	–	–

A code name is a surrogate name for that classification in this article.

nodule, which were categorized into three main groups based on the filtering status of the image: (I) unfiltered features of the original image, 124 in total; (II) features obtained by applying Laplacian of Gaussian filtering to the original image, 188 in total; and (III) features obtained by applying wavelet filtering to the original image, 752 in total. Further, these features are subdivided into first-order statistical features (222 in total), 3D shape features (17 in total), and higher-order texture features (825 in total). The higher-order texture features include gray level co-occurrence matrix (GLCM) features (14,15), gray level size zone matrix (GLSZM) features (16), gray level run length matrix (GLRLM) features (17), neighbouring gray tone difference matrix (NGTDM) features (18), gray level dependence matrix (GLDM) features (19). More detailed information about additional radiomics features and their extracted reproducibility can be found at the following link: https://pyradiomics.readthedocs.io/en/latest/.

For ease of presentation in this thesis, all feature names are abbreviated. For example: the feature name 'wavelet-LLL_gldm_DependenceNonUniformity' is abbreviated to 'wav-LLL_gldm_DN'. The feature name consists of three parts: (I) filtering status, (II) feature type, and (III) name.

Image filtering includes diagnostics (CT voxel statistics in the region of ROI), abbreviated as diag; original (unfiltered original data), abbreviated as org; wavelet-LLL (wavelet filtering. All possible combinations of applying either a high or a low pass filter in each of the three dimensions, respectively. Such as LLL, HHH, HHL...) abbreviated as wav-LLL; log-sigma-3-mm-3D (Laplacian of Gaussian filtering. Sigma is set to 3 to improve fine textures), abbreviated as log-sigma-3; log-sigma-5-mm-3D (Laplacian of Gaussian filtering. Sigma is set to 5 to improve rough textures), abbreviated as log-sigma-5.

Feature types include: image-original (CT image before resampling), mask-original (nodule mask before resampling), image-interpolated (CT image after resampling), mask-interpolated (nodule mask after resampling), first-order, shape, GLCM, GLSZM, NGTDM, GLDM.

Names include the name of each feature, including Energy, Entropy, etc. The abbreviations of the names refer to the official PyRadiomics documentation.

Feature pre-processing

The features need to be preprocessed before model training (20).

(I) Z-score normalization: after extracting radiomics features, Z-score normalization (subtracting the mean and dividing by the standard deviation) was performed to ensure that different features were on the same scale. This approach also effectively mitigates the differences introduced by varying equipment or imaging protocols.

(II) Near-zero variance analysis: near-zero variance analysis aims to screen out features that do not contribute to the model's prediction of the target category. The key to this approach lies in calculating the variances of features and eliminating those with minimal variances, thereby enhancing both the computational efficiency and predictive accuracy of the model.

(III) Redundant feature analysis: the purpose of performing redundant feature analysis is to eliminate redundant features to simplify the model. First, we construct the Spearman correlation coefficient matrix between features and filter out the feature pairs with a more than 90% correlation. Linear regression was then utilized to assess the predictive power of these features for the categorical variable. Finally, we retain the features with higher AUC values in the feature pairs. This step helps to improve the performance and predictive accuracy of the model.

Feature selection

To reduce the complexity of the model, improve the computational efficiency, and enhance the prediction accuracy of the model, this study applied the least absolute shrinkage and selection operator (LASSO) algorithm (21) to perform the dimensionality reduction and selection of the features. The core of the LASSO algorithm lies in introducing the L1 penalty term in the loss function, which enables the constraint and compression of the variable weights in the model. In the implementation process, we have chosen the LassoCV tool provided by the Sklearn library (22). The tool calculates the weight coefficients of each feature through the LASSO model with a value range between −1 and 1. The size of the absolute value of the feature weights directly reflects the importance of their contribution to the model; the larger the absolute value, the higher the contribution of the feature in the model. Finally, the downscaling and feature selection process is completed by retaining only the features with non-zero weights.

Classifier

The classification of the given data into specific classes necessitates the utilization of classifiers. Choosing the right classifier or the right combination of classifiers is an important part of building a classification model. The present study employed a weighted voting mechanism to integrate five distinct classifiers, thereby establishing an ensemble classifier. This ensemble approach aims to synthesize the unique advantages of each base classifier, and by weighing their prediction results, it improves the overall classification performance while allowing it to be better adapted to the seven different classification tasks in this paper.

Base classifiers

In the field of machine learning, support vector machine (SVM) (23) is a supervised learning algorithm for data classification that finds the optimal hyperplane. The core principle is to construct one or more hyperplanes to achieve linear differentiability of data in the feature space.

The K-nearest neighbors (KNN) algorithm (24), as an instance-based learning method, is based on the core principle of predicting the class of a sample by considering the classification information from its k nearest neighbors in the feature space.

Linear discriminant analysis (LDA) (25), also known as Fisher's discriminant analysis, aims to achieve maximum separability between classes. This method introduces a projection surface. Samples are mapped onto this surface, and their category is determined based on their resulting projection points.

Random forest (RF) (26), an ensemble learning algorithm, predicts by building multiple decision trees from bootstrapped dataset samples. Each tree splits on a subset of features selected randomly at each node, with the final prediction derived by aggregating individual tree outcomes through voting or averaging. This method extends and optimizes the Bagging approach (27).

eXtreme gradient boosting tree (XGBoost) (28), an advanced form of gradient boosting decision tree (GBDT) (29,30), enhances the traditional framework by adding regularization to prevent overfitting. Unlike GBDT, which uses a greedy algorithm to explore all split points, XGBoost employs an approximate greedy algorithm, improving

Table 3 Base classifier parameter settings

Classifiers	Parameter	Value
SVM	Regularization parameter: C	exp^{-3}–exp^{3}*
	Kernel	linear, rbf, sigmoid*
	Class_weight	balanced
	Gama	scale, auto*
KNN	N_neighbors	1–21*
	Power parameter for the Minkowski metric: P	1, 2*
	Weight	uniform, distance*
LDA	Solver	lsqr, eigen*
	Shrinkage	auto
RF	N_estimators	50–1,000*
	Max_depth	1–20*
	Criterion	gini, entropy*
	Max_features	sqrt, log2*
XGBoost	Colsample_bytree	0.6–1*
	Subsample	0.6–1*
	Gamma	0–0.5*
	Learning_rate	0.01–0.15*
	Max_depth	1–13*
	Min_child_weight	1–10*
	N_estimators	50–1,000*
	Objective	Binary: logistic

* the parameter was optimized using grid search in the experiment. SVM, support vector machine; KNN, k-nearest neighbors; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting tree.

efficiency by sorting eigenvalues and selecting splits based on quartiles. With parallel processing for faster split finding, XGBoost excels in speed and performance on large datasets.

Table 3 shows the parameters commonly used for the five base classifiers and the values at which these parameters were set in this paper.

Diversity-ensemble classifier

The fundamental concept of ensemble learning (31) is to construct and integrate multiple classifiers to accomplish diverse tasks. With ensemble learning, it is often possible to achieve superior performance over a single classifier. In ensemble learning, a good ensemble strategy is crucial, and different learning tasks may require different ensemble strategies. For numerical outputs, common ensemble methods include simple averaging and weighted averaging. On the other hand, for classification tasks, the voting method is typically employed for the ensemble. Voting methods can be further subdivided into absolute majority, relative majority, and weighted voting methods. Among them, the voting for class label classification is called hard voting, while the voting for class probability is called soft voting. In soft voting, the weighted voting method is often used. In this study, we adopted the weighted voting method as the ensemble strategy for the ensemble classifier.

In the weighted voting method, accurately determining the weighting factor is the most critical issue. In common ensemble learning algorithms, the confidence method and error rate weighting method are commonly employed to assign weights to the base classifiers, such as the RF algorithm, which is a representative example of an error rate weighting method. These methods can be collectively referred to as objective weighting methods. There is also a subjective weighting method, such as the hierarchical weighting method. Due to the differences in the mechanism and parameter settings of different base classifiers, it is more difficult for the subjective weighting method to assess the importance of each base classifier accurately. The objective weighting method mainly sets the weights based on the fluctuation of the performance index, but the uncertainty of the output of the base classifiers is significant, which leads to the lack of precision of the weights obtained by this type of method. Especially when the performance difference between base classifiers is slight, it is difficult for the confidence and error rate weighting methods to perform effective weighting. To compensate for these shortcomings, we used diversity weighting in this study. The diversity weighting method assigns weights by evaluating the uniqueness of individual base classifiers and their complementary roles in the overall decision. The diversity weighting method focuses on the performance differences between base classifiers and avoids simply favoring the best-performing base classifiers.

The effectiveness of the diversity weighting method in ensemble learning has been demonstrated with favorable outcomes. For example, Yang *et al.* (32) proposed an ensemble classification method based on accuracy and diversity. Experiments on the UC Irvine Machine Learning Repository show that the method can achieve good classification performance.

In this study, we have chosen five classifiers, namely

SVM, KNN, LDA, RF, and XGBoost, as the base classifiers and use the weighted voting method to integrate them and build the ensemble classifiers. In the weighting process, we use the diversity expressed by the disagreement measure (33) as the weights of the weighted voting method. The weights are calculated (34) as follows:

$$\omega(c_i) = \sum_{j=1}^{n} dval_{ij} / \sum_{i=1}^{n}\sum_{j=1}^{n} dval_{ij} \quad [1]$$

where c_i denotes the base classifier, n denotes the number of base classifiers, and $dval_{ij}$ denotes the number of c_j classified incorrectly and c_i classified correctly.

Model building and evaluation

Model building

This study aims to enhance the model's performance in classifying different types of lung nodules by integrating radiomics features with a diversity-weighted voting ensemble classifier. Based on this, we used two different model construction methods to build the corresponding nodule-type classification models. And analyzed and compared the modeling results of these models. The two modeling methods are notated as M1: the selected features are inputted into the equal-weighted voting ensemble classifier for classification (this method is used as a baseline to compare with improved methods); M2: the selected features are inputted into the diversity-weighted voting ensemble classifier for classification.

The overall flowchart of this study is shown in *Figure 1*. During the experimental process, we employed a 10-fold cross-validation method. Specifically, the training subset for each task was randomly divided into 10 equal subsets to ensure consistency in data distribution. In ten iterations, one subset was selected as the test set in each iteration, while the remaining nine subsets were combined to serve as the training set. Within the training set, 10% of the data was randomly chosen as the validation set, which was used to generate weighted voting weights. It should be noted that both the validation set and the test set were excluded from feature selection; their roles were confined to weight generation and model performance validation. The performance metrics of the model on the test set were recorded in each iteration. Finally, by calculating the average of the model performance metrics across all test sets, a comprehensive evaluation of the model performance was obtained.

Evaluation metrics

In this study, we used accuracy (ACC), AUC, specificity (SP), and sensitivity (SN) as evaluation metrics to assess the performance of the classification model quantitatively. Accuracy is the most intuitive performance metric, which is the ratio of the number of correctly classified samples to the total number of samples. The specificity metric evaluates the model's capacity to correctly identify negative samples (non-target classes), while sensitivity measures its ability to accurately recognize positive samples (target classes). These metrics constitute a comprehensive evaluation system to ensure a comprehensive quantitative assessment of the accuracy and effectiveness of the classification model. According to the confusion matrix (35), they are calculated as follows:

$$ACC = \frac{tp + tn}{tp + fp + tn + fn} \quad [2]$$

$$Specificity = \frac{tn}{tn + fp} \quad [3]$$

$$Sensitivity = \frac{tp}{tp + fn} \quad [4]$$

where *tp* (true positive) denotes an actual positive sample and a positive prediction; *fp* (false positive) denotes an actual negative sample but a positive prediction; *fn* (false negative) denotes an actual positive sample but a negative prediction; and *tn* (true negative) denotes an actual negative sample and a negative prediction as well.

AUC measures the overall ability of a classifier to discriminate between positive and negative samples (36), and its value ranges from 0.5 to 1. The closer the AUC approaches 0.5, the model exhibits limited discriminatory capacity among samples; conversely, as the AUC approaches 1, the model demonstrates a robust ability to discriminate between samples. It is calculated by the following formula:

$$AUC = \frac{s_p - n_p(n_p + 1)/2}{n_p n_n} \quad [5]$$

where s_p denotes the rank sum of positive samples, n_p and n_n denote the number of positive and negative samples.

In this study, to evaluate the statistical significance of the differences in the AUC between the two models, the DeLong test (37) was utilized. This test is specifically designed to compare the AUCs of correlated ROC curves, thereby providing a robust method for assessing the

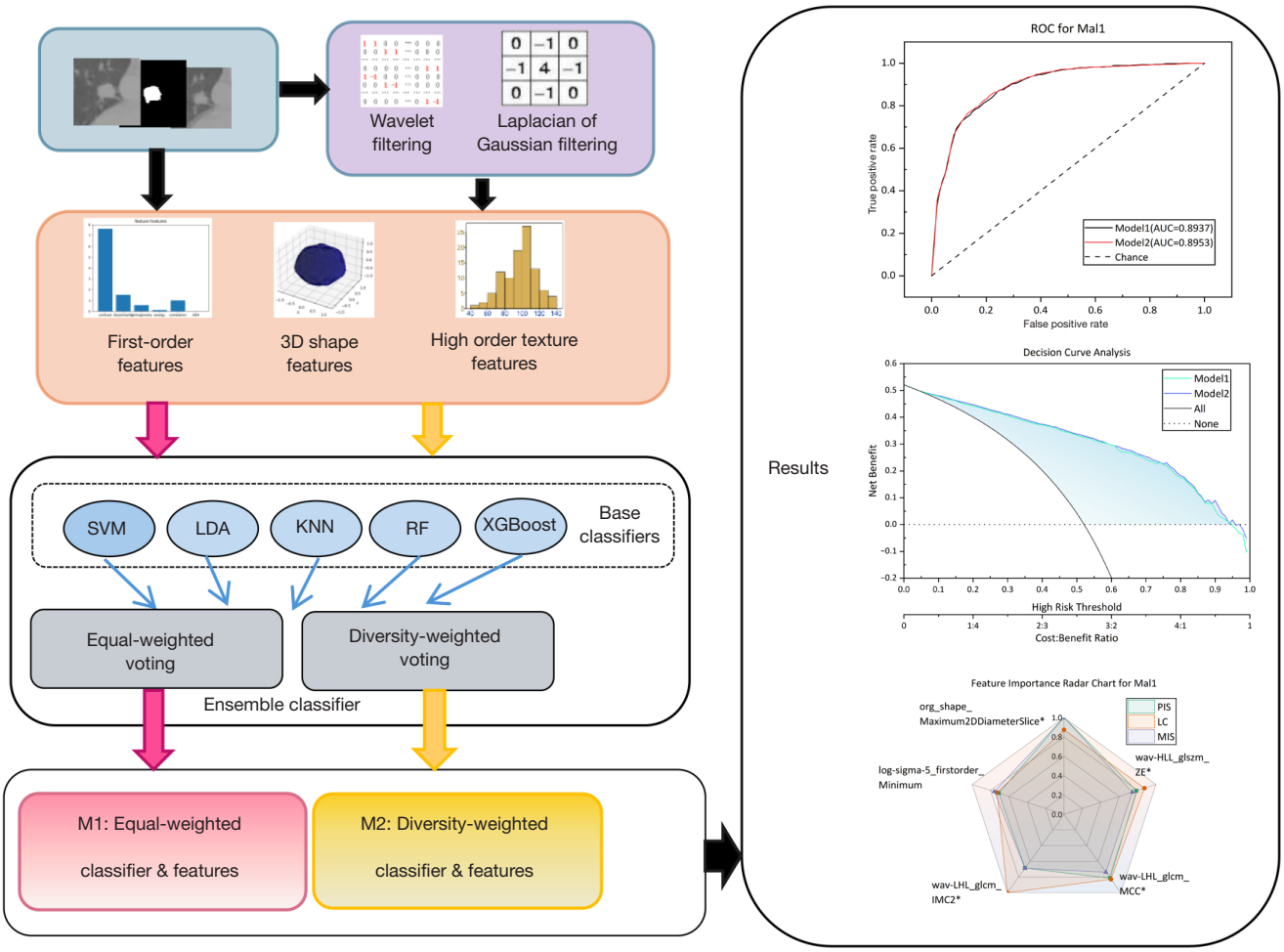

Figure 1 Experimental process. SVM, support vector machine; LDA, linear discriminant analysis; KNN, K-nearest neighbors; RF, random forest; XGBoost, extreme gradient boosting tree; ROC, receiver operating characteristic; AUC, area under the ROC curve; PIS, Permutation Importance score; LC, LASSO coefficient; MIS, mutual information score; L, low; H, high; LASSO, least absolute shrinkage and selection operator.

performance disparities in models concerning their ability to discriminate between binary outcomes. Furthermore, to assess the significance of differences in accuracy, sensitivity, and specificity between the two models, the Wilcoxon signed-rank test (38) was employed. This non-parametric test is favorable for paired data and is particularly useful in situations where the normality assumption may not hold. By implementing these statistical tests, the study ensures a rigorous evaluation of the models' comparative performance metrics. In the context of assessing the statistical significance between models, a P value threshold of less than 0.05 was adopted. Should the P value derived from the DeLong or Wilcoxon signed-rank tests fall below this threshold, it is interpreted as indicative of a statistically significant difference in the respective performance metrics under consideration between the two models.

To comprehensively assess the importance of input features for classifiers, we adopted three distinct evaluation methods: permutation importance, LASSO coefficient (LC), and mutual information approaches.

(I) Permutation importance: this method involves randomly shuffling the values of each feature and then calculating the impact of this perturbation on model performance (such as a decrease in accuracy) to assess the importance of features. If the model's performance significantly deteriorates

Table 4 Sphericity classification results

Class code	Method	AUC (95% CI)	ACC (95% CI)	SN (95% CI)	SP (95% CI)	P_{hl}	P value (AUC, ACC, SN, SP)
Sph1	M1	0.7109 (0.69, 0.73)	0.6545 (0.63, 0.67)	0.6861 (0.63, 0.74)	0.6193 (0.58, 0.65)	0.10	0.07, 0.002, 0.08, 0.37
	M2	0.7154 (0.69, 0.74)	0.6669 (0.63, 0.69)	0.7014 (0.64, 0.75)	0.6284 (0.58, 0.66)	0.11	

M1: equal-weighted ensemble; M2: diversity-weighted ensemble. AUC, area under the receiver operating characteristic curve; CI, confidence interval; P_{hl}, Hosmer-Lemeshow test statistic; ACC, accuracy; SN, sensitivity; SP, specificity.

Table 5 Lobulation classification results

Class code	Method	AUC (95% CI)	ACC (95% CI)	SN (95% CI)	SP (95% CI)	P_{hl}	P value (AUC, ACC, SN, SP)
Lob1	M1	0.7612 (0.74, 0.78)	0.6987 (0.65, 0.75)	0.7300 (0.65, 0.81)	0.6640 (0.60, 0.73)	0.13	0.02, 0.001, 0.04, 0.01
	M2	0.7662 (0.75, 0.79)	0.7135 (0.67, 0.76)	0.7484 (0.68, 0.82)	0.6750 (0.61, 0.74)	0.30	
Lob2	M1	0.8109 (0.79, 0.83)	0.7355 (0.68, 0.79)	0.7666 (0.69, 0.85)	0.7029 (0.63, 0.77)	0.39	0.0003, 0.0005, 0.006, 0.004
	M2	0.8191 (0.80, 0.84)	0.7631 (0.71, 0.82)	0.7864 (0.71, 0.87)	0.7385 (0.67, 0.81)	0.41	
Lob3	M1	0.5820 (0.53, 0.64)	0.5709 (0.50, 0.64)	0.3697 (0.26, 0.48)	0.7265 (0.62, 0.83)	0.13	0.01, 0.03, 0.01, 0.39
	M2	0.6004 (0.55, 0.66)	0.5926 (0.51, 0.68)	0.4146 (0.29, 0.54)	0.7308 (0.62, 0.84)	0.28	

M1: equal-weighted ensemble; M2: diversity-weighted ensemble. AUC, area under the receiver operating characteristic curve; CI, confidence interval; P_{hl}, Hosmer-Lemeshow test statistic; ACC, accuracy; SN, sensitivity; SP, specificity.

after shuffling a particular feature, then that feature is considered important.

(II) LC: in this study, we utilized the coefficients generated by the LASSO algorithm during the feature selection process. The larger the coefficient, the greater the contribution of the feature to the model; hence, the feature is deemed more important.

(III) Mutual information: mutual information measures the mutual dependence between two variables. In evaluating feature importance, the importance of a feature is assessed by calculating the mutual information value between each feature and the target variable. A high mutual information value indicates a strong mutual dependence between the feature and the target variable, thus making the feature very important for predicting the target variable.

In this paper, the normalized importance scores obtained by these three methods are referred to as Permutation Importance score (PIS), LC, and mutual information score (MIS), respectively.

Operations such as feature extraction, model establishment, and statistical comparison were all based on the Python (https://www.python.org/), Scikit-learn (https://scikit-learn.org/), and PyRadiomics libraries (https://pyradiomics.readthedocs.io/en/latest/index.html) of the Anaconda3 software platform (https://www.anaconda.com).

Results

Tables 4-10 demonstrate the evaluation results of the lung nodule type classification models constructed by the M1 and M2 methods. Additionally, we use the Hosmer-Lemeshow test statistic (P_{hl}) to show the model's calibration. If P_{hl} is >0.05, it indicates that the model is well-calibrated. Part A in *Figures 2-8* displays the ROC curves of M1 and M2 for each classification task. Part B in *Figures 2-8* displays the feature importance scores of the top five features ranked according to the average of the three values, PIS, LC, and MIS, in each classification task. If the average value of a feature is greater than 0.75, we consider it to have high feature importance and mark it with an asterisk (*). *Figure 8C* demonstrates the decision curve analysis in the

Table 6 Spiculation classification results

Class code	Method	AUC (95% CI)	ACC (95% CI)	SN (95% CI)	SP (95% CI)	P_{hl}	P value (AUC, ACC, SN, SP)
Spi1	M1	0.7723 (0.75, 0.79)	0.7050 (0.64, 0.77)	0.7239 (0.64, 0.80)	0.6853 (0.59, 0.78)	0.46	0.0001, 0.0006, 0.0009, 0.0007
	M2	0.7855 (0.77, 0.80)	0.7275 (0.66, 0.80)	0.7454 (0.66, 0.83)	0.7095 (0.61, 0.81)	0.62	
Spi2	M1	0.8079 (0.79, 0.83)	0.7398 (0.69, 0.79)	0.7781 (0.70, 0.86)	0.6981 (0.63, 0.77)	0.36	0.0001, 0.0001, 0.02, 0.0001
	M2	0.8185 (0.80, 0.84)	0.7610 (0.71, 0.81)	0.7943 (0.72, 0.87)	0.7250 (0.65, 0.80)	0.58	
Spi3	M1	0.5304 (0.44, 0.62)	0.5026 (0.39, 0.61)	0.5055 (0.39, 0.62)	0.5000 (0.36, 0.64)	0.06	0.03, 0.04, 0.1, 0.34
	M2	0.5627 (0.48, 0.65)	0.5721 (0.47, 0.67)	0.5888 (0.55, 0.62)	0.5577 (0.38, 0.74)	0.06	

M1: equal-weighted ensemble; M2: diversity-weighted ensemble. AUC, area under the receiver operating characteristic curve; CI, confidence interval; P_{hl}, Hosmer-Lemeshow test statistic; ACC, accuracy; SN, sensitivity; SP, specificity.

Table 7 Texture classification results

Class code	Method	AUC (95% CI)	ACC (95% CI)	SN (95% CI)	SP (95% CI)	P_{hl}	P value (AUC, ACC, SN, SP)
Tex1	M1	0.9920 (0.98, 0.99)	0.9596 (0.94, 0.98)	0.9660 (0.94, 0.99)	0.9530 (0.92, 0.98)	0.07	0.4, 0.006, 0.053, 0.05
	M2	0.9913 (0.98, 0.99)	0.9651 (0.95, 0.98)	0.9720 (0.95, 0.99)	0.9580 (0.93, 0.99)	0.15	
Tex2	M1	0.8784 (0.86, 0.89)	0.7929 (0.75, 0.84)	0.8123 (0.75, 0.87)	0.7742 (0.70, 0.85)	0.21	0.01, 0.001, 0.06, 0.005
	M2	0.8838 (0.87, 0.90)	0.8116 (0.77, 0.86)	0.8257 (0.77, 0.88)	0.7982 (0.73, 0.87)	0.31	
Tex3	M1	0.9501 (0.94, 0.96)	0.8844 (0.85, 0.92)	0.8963 (0.86, 0.94)	0.8710 (0.82, 0.92)	0.35	0.88, 0.001, 0.01, 0.018
	M2	0.9525 (0.94, 0.96)	0.9000 (0.87, 0.93)	0.9095 (0.87, 0.95)	0.8892 (0.84, 0.94)	0.36	
Tex4	M1	0.8799 (0.86, 0.90)	0.8034 (0.75, 0.85)	0.7871 (0.71, 0.87)	0.8194 (0.76, 0.88)	0.38	0.1, 0.0008, 0.002, 0.04
	M2	0.8845 (0.86, 0.91)	0.8284 (0.78, 0.87)	0.8251 (0.75, 0.90)	0.8318 (0.76, 0.90)	0.95	

M1: equal-weighted ensemble; M2: diversity-weighted ensemble. AUC, area under the receiver operating characteristic curve; CI, confidence interval; P_{hl}, Hosmer-Lemeshow test statistic; ACC, accuracy; SN, sensitivity; SP, specificity.

Table 8 Margin classification results

Class code	Method	AUC (95% CI)	ACC (95% CI)	SN (95% CI)	SP (95% CI)	P_{hl}	P value (AUC, ACC, SN, SP)
Mar1	M1	0.7433 (0.71, 0.78)	0.6990 (0.65, 0.75)	0.6198 (0.54, 0.70)	0.7642 (0.69, 0.84)	0.26	0.04, 0.49, 0.65, 0.9
	M2	0.7522 (0.72, 0.79)	0.7048 (0.65, 0.76)	0.6296 (0.54, 0.72)	0.7665 (0.69, 0.84)	0.51	
Mar2	M1	0.9335 (0.92, 0.94)	0.8607 (0.83, 0.89)	0.8691 (0.83, 0.91)	0.8524 (0.81, 0.89)	0.12	0.1, 0.01, 0.15, 0.05
	M2	0.9352 (0.92, 0.95)	0.8699 (0.84, 0.90)	0.8745 (0.83, 0.92)	0.8653 (0.82, 0.91)	0.61	
Mar3	M1	0.8300 (0.81, 0.85)	0.7560 (0.72, 0.79)	0.7600 (0.69, 0.83)	0.7515 (0.69, 0.81)	0.31	0.0009, 0.001, 0.16, 0.0013
	M2	0.8371 (0.82, 0.86)	0.7732 (0.73, 0.81)	0.7696 (0.71, 0.83)	0.7764 (0.71, 0.84)	0.39	

M1: equal-weighted ensemble; M2: diversity-weighted ensemble. AUC, area under the receiver operating characteristic curve; CI, confidence interval; P_{hl}, Hosmer-Lemeshow test statistic; ACC, accuracy; SN, sensitivity; SP, specificity.

Table 9 Calcification classification results

Class code	Method	AUC (95% CI)	ACC (95% CI)	SN (95% CI)	SP (95% CI)	P_{hl}	P value (AUC, ACC, SN, SP)
Cal1	M1	0.9764 (0.97, 0.98)	0.9500 (0.92, 0.98)	0.9618 (0.92, 1)	0.9375 (0.89, 0.98)	0.41	0.26, 0.0001, 0.0008, 0.0007
	M2	0.9775 (0.97, 0.98)	0.9642 (0.94, 0.99)	0.9763 (0.95, 1)	0.9512 (0.91, 0.99)	0.54	

M1: equal-weighted ensemble; M2: diversity-weighted ensemble. AUC, area under the receiver operating characteristic curve; CI, confidence interval; P_{hl}, Hosmer-Lemeshow test statistic; ACC, accuracy; SN, sensitivity; SP, specificity.

Table 10 Malignancy classification results

Class code	Method	AUC (95% CI)	ACC (95% CI)	SN (95% CI)	SP (95% CI)	P_{hl}	P value (AUC, ACC, SN, SP)
Mal1	M1	0.8937 (0.88, 0.91)	0.8088 (0.79, 0.83)	0.8133 (0.78, 0.85)	0.8040 (0.77, 0.84)	0.13	0.15, 0.01, 0.19, 0.07
	M2	0.8953 (0.88, 0.91)	0.8168 (0.80, 0.84)	0.8195 (0.79, 0.86)	0.8140 (0.78, 0.84)	0.25	

M1: equal-weighted ensemble; M2: diversity-weighted ensemble. AUC, area under the receiver operating characteristic curve; CI, confidence interval; P_{hl}, Hosmer-Lemeshow test statistic; ACC, accuracy; SN, sensitivity; SP, specificity.

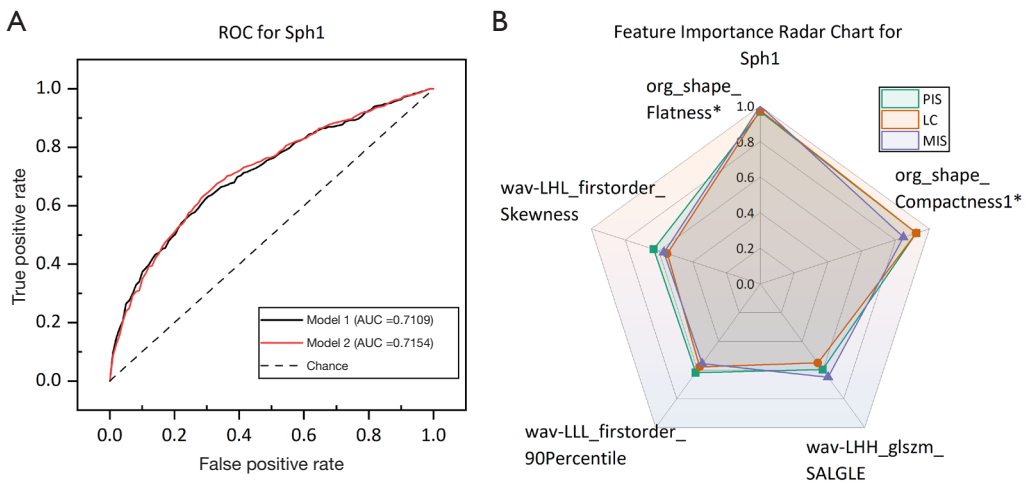

Figure 2 ROC curves and importance scores of the top five features ranked by average value in the sphericity classification task. (A) ROC for each model in the Sph1 classification. (B) The feature importance scores for the Sph1 classification. * denotes that the average values of PIS, LC, and MIS for that feature are greater than 0.75. ROC, receiver operating characteristic; AUC, area under the ROC curve; PIS, permutation importance score; LC, LASSO coefficient; MIS, mutual information score; L, low; H, high; LASSO, least absolute shrinkage and selection operator.

benign and malignant classification of lung nodules.

Sphericity

In the lung nodules sphericity classification task, we focus on a single classification problem, i.e., efficiently distinguishing between ellipsoid-biased and round-biased nodules. This task is denoted as Sph1. *Table 4* shows the evaluation results. *Figure 2B* shows the importance scores of the features. The ROC is shown in *Figure 2A*.

Lobulation

In this study, we divided the task of classifying the degree

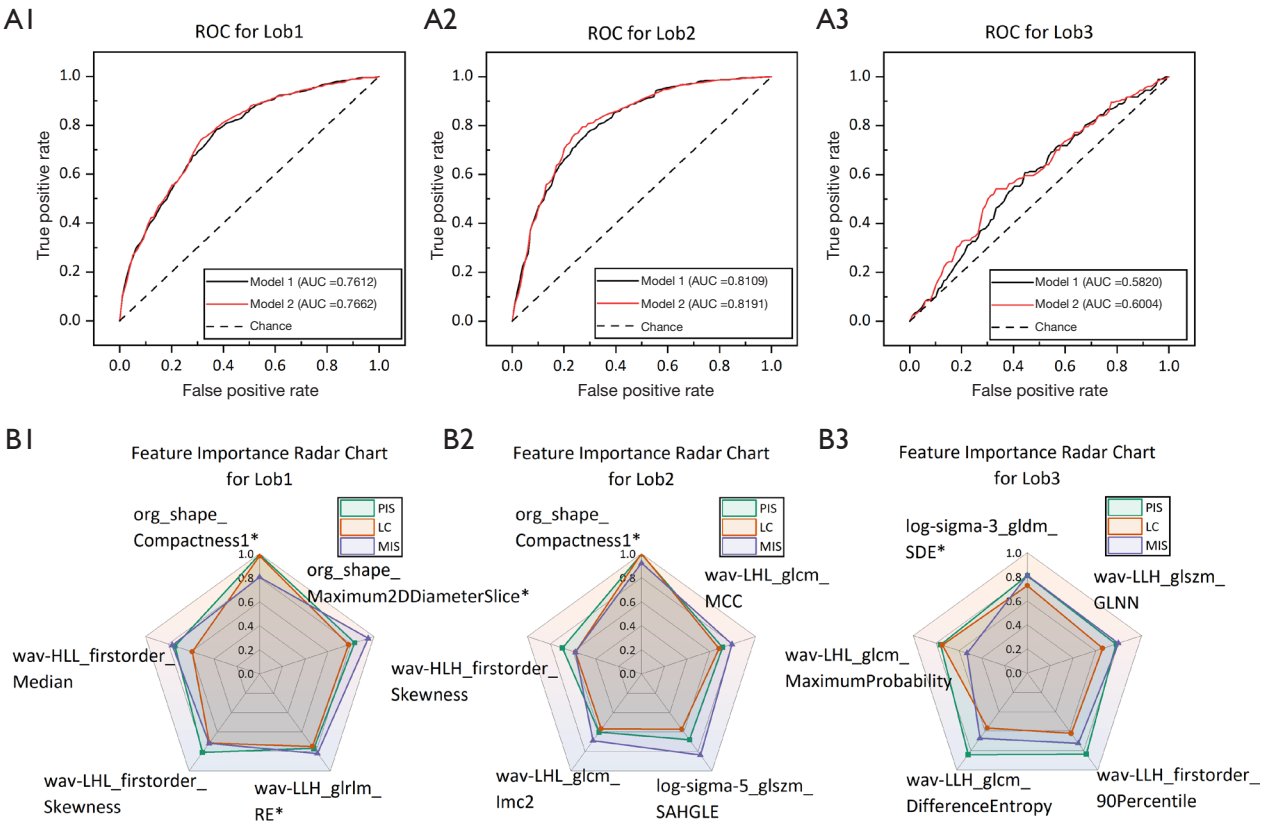

Figure 3 ROC curves and importance scores of the top five features ranked by average value in the lobulation classification task. (A1, A2, A3) ROC for each model in the Lob1, Lob2, and Lob3 classifications. (B1, B2, B3) The feature importance scores for the Lob1, Lob2, and Lob3 classifications. * denotes that the average values of PIS, LC, and MIS for that feature are greater than 0.75. ROC, receiver operating characteristic; AUC, area under the ROC curve; PIS, permutation importance score; LC, LASSO coefficient; MIS, mutual information score; L, low; H, high; MCC, maximal correlation coefficient; GLNN, gray level non-uniformity normalized; LASSO, least absolute shrinkage and selection operator.

of lobulation of lung nodules into three separate binary classification tasks named Lob1, Lob2, and Lob3, according to the degree of nodule lobulation given in the original dataset. Lob1 aims to distinguish moderately lobulated nodules from lowly lobulated nodules, Lob2 aims to distinguish highly lobulated nodules from lowly lobulated nodules, and Lob3 aims to distinguish highly lobulated nodules from moderately lobulated nodules. *Table 5* shows the evaluation results. *Figure 3B* shows the importance scores of the features. The ROC is shown in *Figure 3A*.

Spiculation

Similarly, we divided the task of classifying the degree of spiculation of lung nodules into three separate binary classification tasks named Spi1, Spi2, and Spi3, according to the degree of nodule spiculation given in the original dataset. Spi1 aims to distinguish moderately spiculated nodules from lowly spiculated nodules, Spi2 aims to distinguish highly spiculated nodules from lowly spiculated nodules, and Spi3 aims to distinguish highly spiculated nodules from moderately spiculated nodules. *Table 6* shows the evaluation results. *Figure 4B* shows the importance scores of the features. The ROC is shown in *Figure 4A*.

Texture

We divided the task of lung nodule texture classification into four separate binary classification tasks named Tex1, Tex2, Tex3, and Tex4, according to the nodule texture

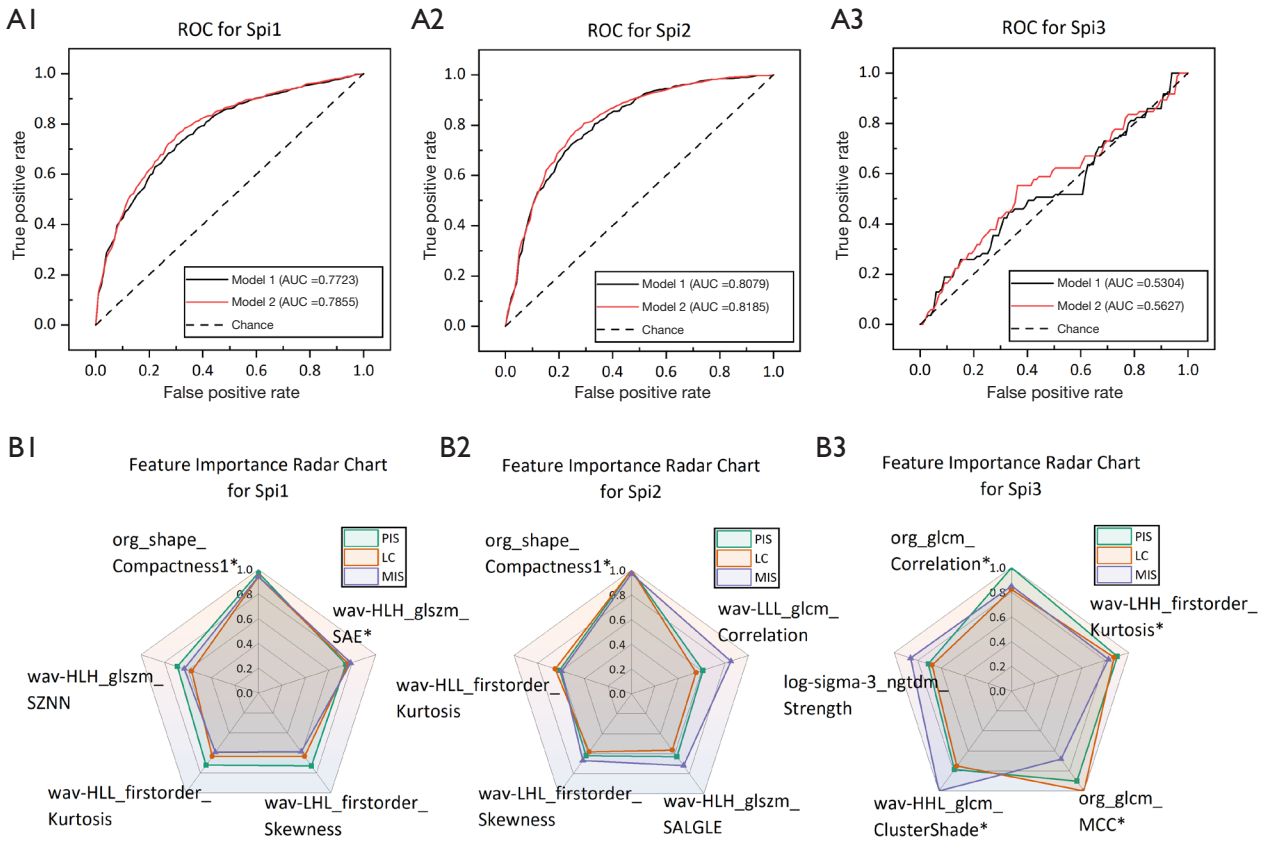

Figure 4 ROC curves and importance scores of the top five features ranked by average value in the spiculation classification task. (A1, A2, A3) ROC for each model in the Spi1, Spi2, and Spi3 classifications. (B1, B2, B3) The feature importance scores for the Spi1, Spi2, and Spi3 classifications. * denotes that the average values of PIS, LC, and MIS for that feature are greater than 0.75. ROC, receiver operating characteristic; AUC, area under the ROC curve; PIS, permutation importance score; LC, LASSO coefficient; MIS, mutual information score; L, low; H, high; SAE, small area emphasis; MCC, maximal correlation coefficient; LASSO, least absolute shrinkage and selection operator.

information given in the original dataset. Tex1 aims to distinguish pure GGNs from solid nodules, Tex2 aims to distinguish pure GGNs from mixed GGNs, Tex3 aims to distinguish mixed GGNs from solid nodules, and Tex4 focuses on the distinguishing of micro-mixed GGNs [the LIDC-IDRI dataset provides labels (solid/mixed)] from solid nodules. *Table* 7 shows the evaluation results. *Figure 5B* shows the importance scores of the features. The ROC is shown in *Figure 5A*.

Margin

For the classification of margin clarity in lung nodules, we divided this classification task into three separate binary classification tasks named Mar1, Mar2, and Mar3, according to the degree of nodule margin clarity given in the original dataset. Mar1 aims to distinguish moderately defined margin nodules from lowly defined margin nodules, Mar2 aims to distinguish lowly defined margin nodules from well-defined margin nodules, and Mar3 aims to distinguish moderately defined margin nodules from well-defined margin nodules. *Table 8* shows the evaluation results. *Figure 6B* shows the importance scores of the features. The ROC is shown in *Figure 6A*.

Calcification

In the task of classifying the degree of calcification in lung nodules, we focus on a single classification problem, i.e., effectively distinguishing calcified nodules from non-calcified nodules. And this task is denoted as Cal1. *Table 9* shows the evaluation results. *Figure 7B* shows the

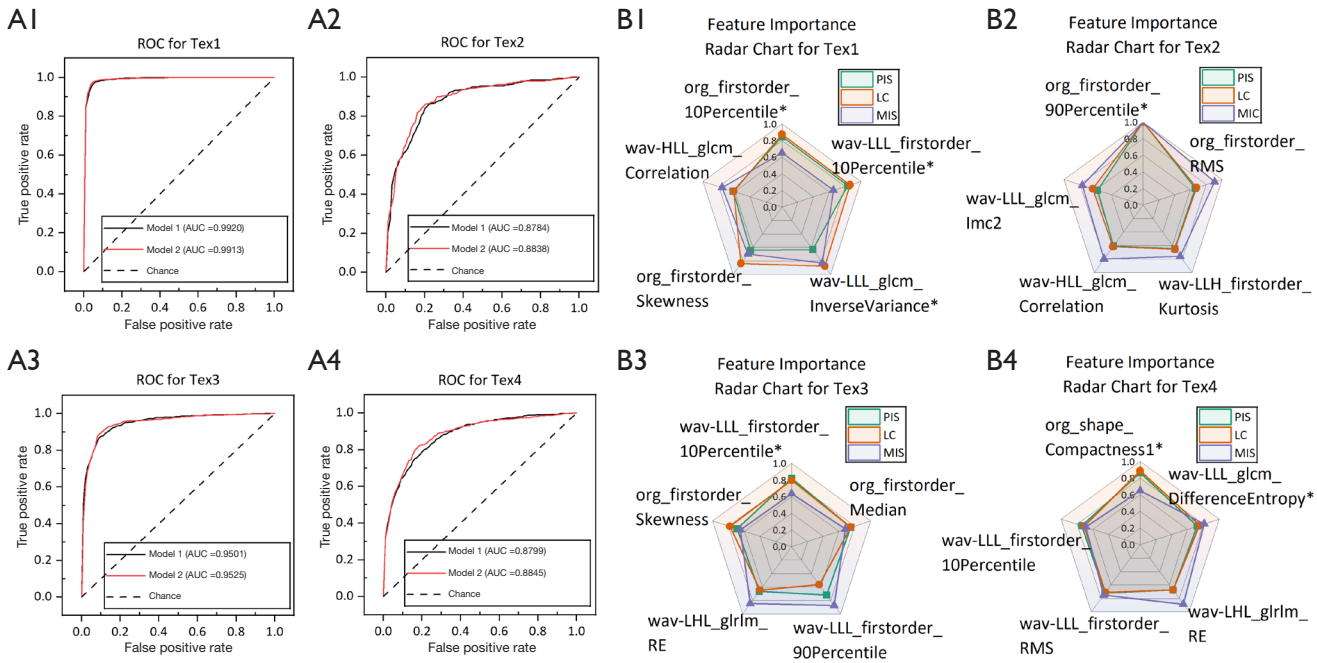

Figure 5 ROC curves and importance scores of the top five features ranked by average value in the texture classification task. (A1, A2, A3, A4) ROC for each model in the Tex1, Tex2, Tex3, and Tex4 classifications. (B1, B2, B3, B4) The feature importance scores for the Tex1, Tex2, Tex3, and Tex4 classifications. * denotes that the average values of PIS, LC, and MIS for that feature are greater than 0.75. ROC, receiver operating characteristic; AUC, area under the ROC curve; PIS, permutation importance score; LC, LASSO coefficient; MIS, mutual information score; L, low; H, high; RMS, root mean squared; LASSO, least absolute shrinkage and selection operator; RE, run entropy.

importance scores of the features. The ROC is shown in *Figure 7A*.

Malignancy

In this study, we denoted the classification of lung nodules' malignancy as Mal1. The binary classification labels were divided according to the malignant grades of nodules given by the original dataset. The first three grades were classified as benign nodules, and the last two grades were classified as malignant nodules. *Table 10* shows the evaluation results. *Figure 8B* shows the importance scores of the features. ROC is shown in *Figure 8A*. Decision curve analysis (DCA) is shown in *Figure 8C*.

Discussion

Low-dose CT (LDCT) screening can detect early lung cancer and reduce lung cancer mortality (39). In clinical practice, radiologists use superficial features (e.g., nodule size, margins, shape, etc.) on CT scans to make disease diagnoses. However, the information that these features can bring is limited, and it is difficult to make the diagnosis more accurate. CT radiomics research involves extracting high-throughput digital image features imperceptible to the human eye from CT scan results, followed by processing and analysis of these features to predict a reference outcome. This prediction can be utilized by doctors in conjunction with their own experience to enhance the accuracy of disease diagnosis. At present, it has been proved that radiomics has a good application prospect in the prediction of benign and malignant lung nodules, the differentiation of lung tumor subtypes, and the prognosis analysis of lung cancer (40).

Several recent studies have demonstrated the superiority of using radiomics to classify lung nodules. For example, Gupta *et al.* (41) extracted radiomics features from the LIDC-IDRI dataset and, after feature selection, utilized seven classifiers, such as decision tree, SVM, and Bayesian classifier, to build a classification model for distinguishing between benign and malignant lung nodules. The highest AUC of 0.96 and ACC of 0.908 were finally achieved. Xu

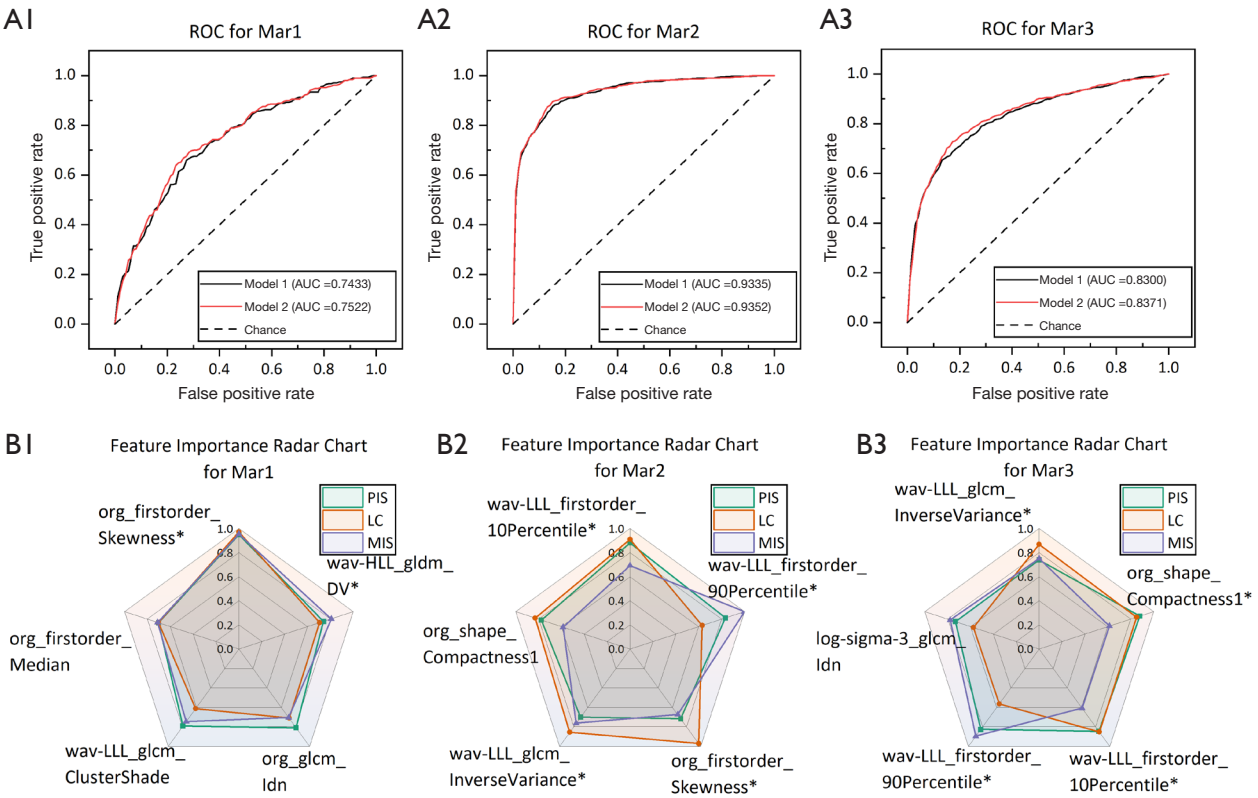

Figure 6 ROC curves and importance scores of the top five features ranked by average value in the margin classification task. (A1, A2, A3) ROC for each model in the Mar1, Mar2, and Mar3 classifications. (B1, B2, B3) The feature importance scores for the Mar1, Mar2, and Mar3 classifications. * denotes that the average values of PIS, LC, and MIS for that feature are greater than 0.75. ROC, receiver operating characteristic; AUC, area under the ROC curve; PIS, permutation importance score; LC, LASSO coefficient; MIS, mutual information score; L, low; H, high; DV, dependence variance; LASSO, least absolute shrinkage and selection operator.

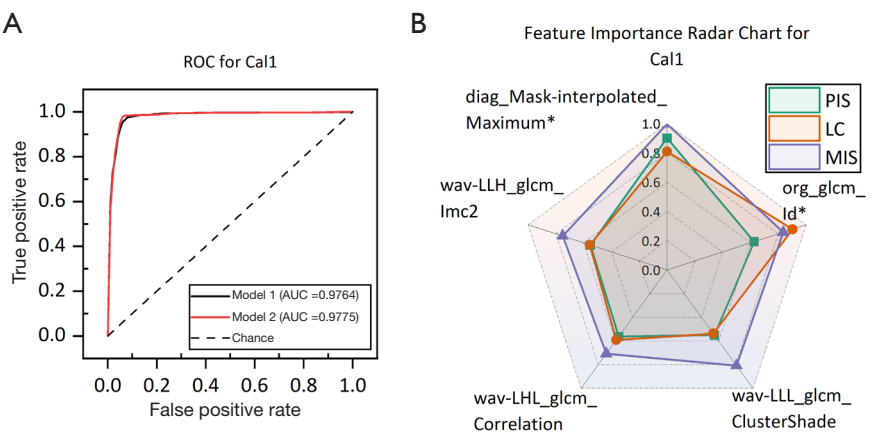

Figure 7 ROC curves and importance scores of the top five features ranked by average value in the calcification classification task. (A) ROC for each model in the Cal1 classification. (B) The feature importance scores for the Cal1 classification. * denotes that the average values of PIS, LC, and MIS for that feature are greater than 0.75. ROC, receiver operating characteristic; AUC, area under the ROC curve; PIS, permutation importance score; LC, LASSO coefficient; MIS, mutual information score; L, low; H, high; LASSO, least absolute shrinkage and selection operator.

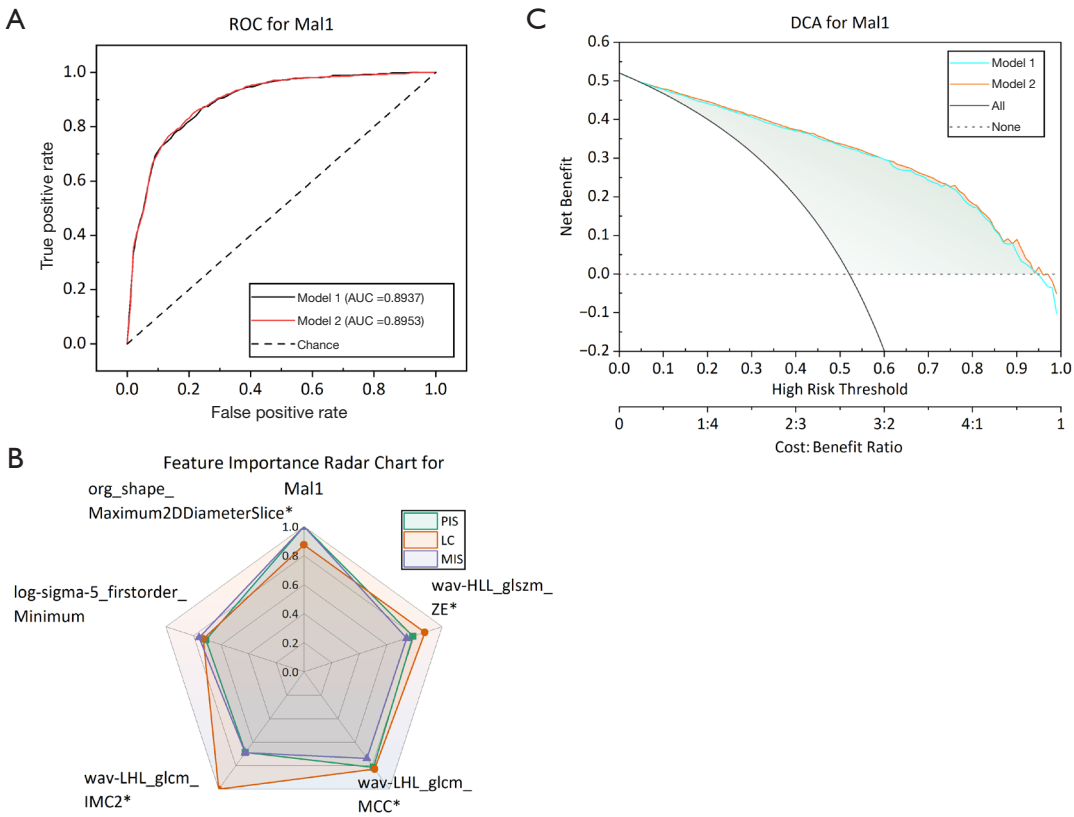

Figure 8 ROC curves and importance scores of the top five features ranked by average value in the malignancy classification task. (A) ROC for each model in the Mal1 classification. (B) The feature importance scores for the Mal1 classification. (C) DCA of classification models built by M1 and M2 methods in the Mal1 classification. * denotes that the average values of PIS, LC, and MIS for that feature are greater than 0.75. ROC, receiver operating characteristic; AUC, area under the ROC curve; DCA, decision curve analysis; PIS, permutation importance score; LC, LASSO coefficient; MIS, mutual information score; L, low; H, high; ZE, zone entropy; MCC, maximal correlation coefficient; LASSO, least absolute shrinkage and selection operator.

et al. (42) used radiomics features to establish three models to classify lung nodules' benign, malignant, and invasive nature. Their study achieved an AUC of up to 0.89. Jing *et al.* (43) developed a model to identify lung nodules' benign and malignant nature based on radiomics combined with LASSO and multivariate logistic regression. The AUC on the validation set achieved a score of 0.8454. Garau *et al.* (44) used radiomics to develop a LASSO-SVM model and an ANN model to predict the degree of malignancy of lung nodules. Both models were able to achieve an AUC of more than 0.89.

Although all of the above studies have used radiomics features to build models, there are obvious limitations in the choice of classification algorithms. Specifically, these studies usually use only a single classification algorithm for model training and do not take advantage of the potential advantages of the combination of multiple algorithms to improve model performance. In contrast, this study uses a diversity-weighted ensemble learning strategy, which integrates the advantages of multiple classifiers and aims to improve the overall performance of the model.

Specifically, we extracted 112 radiomics features totaling 1,064 from each nodule in the LIDC-IDRI dataset. These features were subjected to selection processes before being put into an ensemble classifier weighted by a diversity metric. Subsequently, we employed two model-building methods to construct a classification model for distinguishing between benign and malignant lung nodules. As shown in *Table 10*, M2 exhibited superior ACC (0.8168) compared to M1 (0.8088), and the difference is statistically significant (P=0.01<0.05), which indicates that M2 method may be more reliable in overall classification performance.

In addition, M2 was 0.16%, 0.62%, and 1% higher than M1 in AUC, sensitivity, and specificity, respectively, although these differences were not significant. In addition, as can be seen from the decision curves in *Figure 8C*, all two models in this study have high decision benefits in benign and malignant classification. The decision benefits of the diversity-weighted ensemble model (M2) are better than the equal-weighted ensemble classifier (M1). The specific numerical results are shown in *Table 10* and *Figure 8A*.

In medicine, lung nodules are classified into various types (e.g., lobulated nodules, spiculated nodules, GGNs, etc.), and different types of nodules have an impact on doctors' diagnosis of patients' conditions and formulation of diagnostic and therapeutic programs. For example, calcified nodules, particularly those with central or layered calcification, are usually indicative of benign conditions such as granulomas or sarcoidosis. The identification of these nodules often reduces the need for unnecessary invasive procedures (45). Lobulated and spiculated nodules often signify irregular tumor growth, indicating a higher risk of malignancy (46). GGNs, especially those with solid components, are strongly associated with early-stage lung adenocarcinoma. These nodules appear blurred on imaging and typically grow slowly, making them ideal for early intervention (47,48). Moreover, regular round nodules are generally benign, while irregular or blurred-margin nodules suggest potential malignancy, necessitating closer follow-up (49). Additionally, in recent years, some studies have utilized radiomics to explore the growth patterns of subsolid nodules, aiming to further reduce overtreatment and diagnostic delays (50). Accurately classifying the texture types and benign or malignant nature of lung nodules can contribute to these studies. Therefore, in addition to realizing the benign and malignant classification of lung nodules in this study, the classification of six other lung nodule types was modeled in the same way.

In the lung nodules sphericity classification task, the M2 method shows improvement in all metrics: AUC is increased by 0.45%, ACC is increased by 1.24%, sensitivity is increased by 1.53%, and specificity is increased by 0.91%. However, in terms of statistical significance, only the improvement in ACC was significant (P=0.002), but not in AUC, sensitivity, and specificity (P=0.07, 0.08, 0.37). This indicates that the M2 method has a slight advantage in overall performance compared to the M1 method. The specific numerical results are shown in *Table 4* and *Figure 2A*.

In the classification task of lung nodule lobulation degree, overall, the M2 method completely outperformed the M1 method on all metrics. In Lob1 classification, the M2 method was better than the M1 method in AUC, ACC, sensitivity, and specificity, and these differences were statistically significant (P=0.02, 0.001, 0.04, 0.01). For Lob2 classification, M2 also performed better, with all metrics higher than M1, and the improvement was also statistically significant (P=0.0003, 0.0005, 0.006, 0.004). As for the Lob3 classification, although M2 still outperformed M1 on all metrics, the overall performance was relatively low, with statistically significant improvements in AUC, ACC, and sensitivity (P=0.01, 0.03, 0.01). In general, Overall, the diversity-weighted M2 method performed better than the equal-weighted M1 method in the classification of lung nodule lobulation, especially in the Lob2 classification, demonstrating its effectiveness in handling the classification issues of high lobulation and low lobulation nodules. The specific numerical results are shown in *Table 5* and *Figure 3A*.

In the task of classifying the degree of spiculation in lung nodules, overall, the M2 method consistently outperforms the M1 method across all evaluation metrics. In the Spi1 classification, the M2 method surpasses the M1 method in AUC, ACC, sensitivity, and specificity, with these improvements being highly statistically significant (P<0.001). In the Spi2 classification, the M2 method also performs excellently, improving AUC, ACC, sensitivity, and specificity to 0.8185, 0.7610, 0.7943, and 0.7250, respectively, with significance tests showing these improvements are statistically significant (P<0.02). However, in the Spi3 classification, although the M2 method's performance (AUC of 0.5627, ACC of 0.5721, sensitivity of 0.5888, specificity of 0.5577) also exceeds that of the M1 method, except for the significant statistical improvements in AUC and ACC (P=0.03, 0.04), the increases in sensitivity and specificity are not significant (P=0.1, 0.34), and the overall performance is relatively low. In general, the diversity-weighted M2 method performs better than the equal-weighted M1 in the classification of lung nodule spiculation, particularly excelling in the Spi1 and Spi2 classifications, demonstrating its accuracy in addressing the classification issues of nodules with low and more pronounced spiculation. The specific numerical results are shown in *Table 6* and *Figure 4A*.

In the classification of lung nodule texture, the M2 method generally outperformed the M1 method in all metrics. In Tex1, the M2 method slightly outperformed M1 in ACC and specificity by 0.55% and 0.5%, respectively,

and the improvement in these two metrics was significant (p less than 0.05 for both). In Tex3, M2 outperformed M1 in all metrics, especially in ACC (0.9000, P=0.001), sensitivity (0.9095, P=0.01), and specificity (0.8892, P=0.018). The performance in the Tex2 task is similar, with M2 performing better on all metrics, particularly on AUC (0.8838, P=0.01), ACC (0.8116, P=0.001), and specificity (0.7982, P=0.005). In the Tex4 task, M2 similarly outperformed M1 on all assessment metrics than M1, and the boosts in ACC (0.8284, P=0.0008), specificity (0.8318, P=0.04), and sensitivity (0.8251, P=0.002) were statistically significant. These results indicate that the diversity-weighted M2 method exhibits higher accuracy in handling the more complex task of lung nodule texture classification. The specific numerical results are shown in *Table 7* and *Figure 5A*.

In the classification of margin clarity in lung nodules, the M2 method generally outperforms the M1 method across various metrics. In the Mar1 task, the M2 method shows higher AUC, ACC, sensitivity, and specificity than M1 by 0.89%, 0.58%, 0.98%, and 0.23%, respectively, with a significant difference in AUC (P=0.04). In the Mar2 task, M2 also performs better across all metrics, particularly showing significant improvements in ACC (0.8699, P=0.01) and specificity (0.8653, P=0.05). In the Mar3 task, the M2 method not only exceeds M1 across all metrics but also shows statistically significant improvements in AUC (0.8371, P=0.0009), ACC (0.7732, P=0.001), and specificity (0.7764, P=0.0013). These results indicate that the diversity-weighted M2 method performs better overall than the equal-weighted M1 method in classifying lung nodules of different clarities. The specific numerical results are shown in *Table 8* and *Figure 6A*.

In the Cal1 task, we focused on distinguishing between calcified and non-calcified nodules. The M2 method has shown improvements across multiple key performance metrics. Specifically, M2's AUC slightly increased by 0.11%, ACC improved by 1.42%, sensitivity increased by 1.45%, and specificity rose by 1.37%. Except for AUC, the improvements in other metrics have reached statistically significant levels (P<0.05). The specific numerical results are shown in *Table 9* and *Figure 7A*.

In each classification, the non-significant Hosmer-Lemeshow test statistic (P_{hl}) indicates that the models are well-calibrated. Moreover, the P_{hl} values of the M2 model are consistently greater than or equal to those of the M1 model, suggesting that the M2 model exhibits better calibration than the M1 model.

In summary, the lung nodule classification models in this study perform well in classifying various types of lung nodules. They particularly excel in classifying the texture types of lung nodules, the degree of calcification, and the benign or malignant nature of the nodules. Moreover, the innovation and advantages of combining CT radiomics with diversity-weighted ensemble learning the organic integration of the powerful expressive capabilities of high-dimensional imaging features with the robustness of diversity-weighted ensemble algorithms in complex classification tasks. This approach significantly improves the accuracy and robustness of multiple lung nodule classification tasks while demonstrating lower resource consumption compared to deep learning methods. Not only does this method excel in lung nodule classification, but it also has broad application potential. Beyond lung nodule type classification research, radiomics combined with machine learning can be extended to other medical imaging analysis fields, such as the study of lung adenocarcinoma interstitial growth (8), tumor volume doubling time analysis (9), and molecular subtyping of diffuse gliomas (51). Furthermore, in clinical practice, radiomics methods can be integrated into existing imaging diagnostic workflows, providing physicians with refined disease classification information to assist in formulating more precise diagnostic and treatment plans (40,50).

Table 11 shows the comparison of the results between the related methods and the proposed method for classifying multiple types of lung nodules. Among them, Ni *et al.* (4) used ANNs to develop a model for feature extraction of lung nodule types. Chen *et al.* (52) extracted convolutional neural network and stacked denoising autoencoder features from lung nodule CT data, integrating them into a hybrid feature set. An RF classifier was then used to classify different lung nodule types. As shown in the table, our method outperforms the other two studies in the classification of nodule calcification and texture. The absolute distance error (ADE) results also surpass them in the classification of nodule sphericity, margin, and benign-malignant distinction. However, in the classification of nodule lobulation and spiculation, the two studies based on CNN features perform better than ours, possibly due to CNN's superior ability to capture spatial information.

In this study, we analyzed the confidence levels of the original radiomics features for different classification tasks and found differences in the contribution of these features to each classification task. This suggests that the features differ in their ability to describe and differentiate between different types of lung nodules. Specifically:

Table 11 Comparison of the results of related method and proposed method for classifying multiple types of lung nodules

Nodule's types	Chen *et al.* (52)		Ni *et al.* (4)		Ours (M2)	
	ACC	ADE	ACC	ADE	ACC	ADE
Sphericity	–	0.8600	0.7121	0.7000	0.6669	0.6662
Lobulation	–	0.8000	0.9566	0.5000	0.6897	0.6205
Spiculation	–	0.6400	0.9431	0.5000	0.6869	0.6261
Texture	–	0.1800	0.7888	0.4800	0.8762	0.2473
Margin	–	0.9200	0.8111	0.7400	0.7826	0.4346
Calcification	–	0.8700	0.9221	0.3100	0.9642	0.0714
Malignancy	–	0.8700	0.8661	0.5900	0.8168	0.3662

The ADE is a metric that quantifies the disparity between the model's predicted value and the actual value. A smaller ADE indicates the model's superior predictive performance. In the classification of lobulation, spiculation, texture, and margin of nodules, this study performed a more detailed binary classification of nodules, while other studies were more general. To facilitate an approximate comparison, in this table, we averaged the metric results of the sub-tasks for each of these four classification tasks. ACC, accuracy; ADE, absolute distance error.

(I) In the lung nodules sphericity classification task, as shown in *Figure 2B*, the Flatness, which shows the relationship between the largest and smallest principal components in the ROI shape, and compactness1, which measures how dense the nodule is relative to the shape of the sphere, both have relatively high importance scores. Higher values of Flatness indicate that the nodule is closer to a sphere, and lower values indicate that the nodule is more flattened. The value of compactness1 ranges from 0 to $1/6\pi$. The closer its value is to $1/6\pi$, the more spherical the nodule shape is.

(II) In the lung nodule lobulation degree classification task, as shown in *Figure 3B*, compactness1, run entropy (RE), small dependence emphasis (SDE), and Maximum2DDiameterSlice were shown to have higher importance. Compactness1 focuses on the shape and compactness of the nodule, and RE indicates the complexity of the image texture. SDE highlights the grayscale dependence in small areas, and Maximum2DDiameterSlice measures the maximum diameter of the nodule in any two-dimensional (2D) slice.

(III) In the lung nodule spiculation degree classification task, as shown in *Figure 4B*, compactness1, kurtosis, cluster shade, correlation, small area emphasis (SAE), and maximal correlation coefficient (MCC) were shown to have higher importance. These features are closely related to the spiciness of lung nodules, including the shape irregularity of nodules (compactness1), the texture complexity on the surface and inside (Cluster Shade, MCC, correlation), and the characteristics of the pixel intensity distribution (Kurtosis, SAE).

(IV) In the lung nodule texture classification task, as shown in *Figure 5B*, 10th percentile, 90th percentile, inverse variance, compactness1, and difference entropy were shown to have higher importance. The 10th percentile and 90th percentile features help distinguish different types of nodules by reflecting the range of grayscale values, such as GGN typically having lower grayscale values. Inverse Variance measures texture uniformity, which is useful for identifying nodules like ground glass ones. Compactness1 shows shape compactness, differentiating nodules by shape. Difference Entropy indicates internal texture complexity.

(V) In the lung nodule margin clarity classification task, as shown in *Figure 6B*, skewness, 10th percentile, 90th percentile, compactness1, dependence variance (DV), and inverse variance were shown to have higher importance. These features reveal the asymmetry of margin pixel intensity distribution (skewness), the extreme distribution of pixel intensity (10th percentile, 90th percentile), the shape irregularity of nodules (compactness1), and

the consistency and complexity of texture structure (inverse variance, DV), providing key information for differentiating between lung nodules with clear margins and those with blurred margins.

(VI) In the task of classifying the degree of calcification of lung nodules, as shown in *Figure 7B*, Mask-interpolated maximum and inverse difference (Id) were shown to have higher importance. Mask-interpolated Maximum, by reflecting the highest density areas and the upper-density limits within the nodules, reveals the density characteristics of calcified nodules. In contrast, Id further differentiates between calcified and non-calcified nodules by measuring the smoothness of texture or the differences between pixels.

(VII) In the classification of benign and malignant lung nodules, as shown in *Figure 8B*, Maximum2DDiameterSlice, zone entropy (ZE), MCC, and informational measure of correlation2 (IMC2) were shown to have higher importance. They effectively capture key aspects like the size of the nodules (Maximum2DDiameterSlice) and the heterogeneity and complexity of their internal structure (MCC, ZE, IMC2). These characteristics are crucial for distinguishing between benign and malignant nodules.

Finally, there are some limitations to this study. First, to ensure more precise nodule contours, we directly used the nodule ROI masks provided by the LIDC-IDRI dataset during the ROI segmentation phase, which may not fully correspond to clinical practice procedures. Second, this study was retrospective, and all data were obtained from the LIDC-IDRI dataset, which may have introduced a bias in the distribution of different types of nodules compared to actual clinical conditions. Therefore, we need additional real-world samples for more in-depth research. Additionally, the imaging data in the LIDC-IDRI dataset come from different devices, introducing variability and randomness into the dataset. Although we have mitigated bias as much as possible through standardization and resampling methods, standardized imaging remains an issue to address in future research. In future studies, we plan to collaborate with various medical institutions to obtain real-world samples based on standardized imaging to enhance the dataset's diversity and the model's clinical applicability. Meanwhile, we will adopt a semi-automatic approach for ROI mask segmentation to improve the study's clinical realism.

Conclusions

In summary, the combination of CT radiomics and ensemble learning for diversity weighting provides a new, non-invasive, and efficient way for the diagnosis of lung diseases. In this study, we used this technique to successfully distinguish benign and malignant lung nodules and further accurately classified six different lung nodule types. And, of course, as they develop, they can be applied to many more areas than disease research.

Acknowledgments

Funding: This study was supported by the financial support from the National Natural Science Youth Foundation of China (No. 12304469).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-1315/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1315/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Leiter A, Veluswamy RR, Wisnivesky JP. The global

burden of lung cancer: current status and future trends. Nat Rev Clin Oncol 2023;20:624-39.
2. Loverdos K, Fotiadis A, Kontogianni C, Iliopoulou M, Gaga M. Lung nodules: A comprehensive review on current approach and management. Ann Thorac Med 2019;14:226-38.
3. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, Mehta AC, Ohno Y, Powell CA, Prokop M, Rubin GD, Schaefer-Prokop CM, Travis WD, Van Schil PE, Bankier AA. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. Radiology 2017;284:228-43.
4. Ni YF, Yang YY, Xie Z, Zheng DZ, Wang WD. Multi-Feature Extraction of Pulmonary Nodules Based on LSTM and Attention Structure. J Shanghai Jiao Tong Univ 2022;56:1078-88.
5. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts HJWL. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77:e104-7.
6. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441-6.
7. Rundo L, Ledda RE, di Noia C, Sala E, Mauri G, Milanese G, Sverzellati N, Apolone G, Gilardi MC, Messa MC, Castiglioni I, Pastorino U. A Low-Dose CT-Based Radiomic Model to Improve Characterization and Screening Recall Intervals of Indeterminate Prevalent Pulmonary Nodules. Diagnostics (Basel) 2021.
8. Tang EK, Wu YJ, Chen CS, Wu FZ. Prediction of the stage shift growth of early-stage lung adenocarcinomas by volume-doubling time. Quant Imaging Med Surg 2024;14:3983-96.
9. Wu FZ, Wu YJ, Chen CS, Tang EK. Prediction of Interval Growth of Lung Adenocarcinomas Manifesting as Persistent Subsolid Nodules ≤3 cm Based on Radiomic Features. Acad Radiol 2023;30:2856-69.
10. Wu FZ, Wu YJ, Tang EK. An integrated nomogram combined semantic-radiomic features to predict invasive pulmonary adenocarcinomas in subjects with persistent subsolid nodules. Quant Imaging Med Surg 2023;13:654-68.
11. Li R, Zhou L, Wang Y, Shan F, Chen X, Liu L. A graph neural network model for the diagnosis of lung adenocarcinoma based on multimodal features and an edge-generation network. Quant Imaging Med Surg 2023;13:5333-48.
12. Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 2011;38:915-31.
13. Hancock MC, Magnan JF. Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods. J Med Imaging (Bellingham) 2016;3:044504.
14. Nanni L, Brahnam S, Ghidoni S, Menegatti E, Barrier T. Different approaches for extracting information from the co-occurrence matrix. PLoS One 2013;8:e83554.
15. Iqbal N, Mumtaz R, Shafi U, Zaidi SMH. Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms. PeerJ Comput Sci 2021;7:e536.
16. Thibault G, Angulo J, Meyer F. Advanced statistical matrices for texture characterization: application to cell classification. IEEE Trans Biomed Eng 2014;61:630-7.
17. Dash S, Senapati MR. Gray level run length matrix based on various illumination normalization techniques for texture classification. Evol Intell 2021;14:217-26.
18. Chen S, Harmon S, Perk T, Li X, Chen M, Li Y, Jeraj R. Using neighborhood gray tone difference matrix texture features on dual time point PET/CT images to differentiate malignant from benign FDG-avid solitary pulmonary nodules. Cancer Imaging 2019;19:56.
19. Sassi OB, Sellami L, Slima MB, Chtourou K, Hamida AB. Improved spatial gray level dependence matrices for texture analysis. Int J Comput Sci Inf Technol 2012;4:209.
20. Papanikolaou N, Matos C, Koh DM. How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging 2020;20:33.
21. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58:267-88.
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. JMLR 2011;12:2825-30.
23. Widodo A, Yang BS. Support vector machine in machine condition monitoring and fault diagnosis. Mech Syst

Signal Process 2007;21:2560-74.
24. Uddin S, Haque I, Lu H, Moni MA, Gide E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep 2022;12:6256.
25. Zhu F, Gao J, Yang J, Ye N. Neighborhood linear discriminant analysis. Patt Recognit 2022;123:108422.
26. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9:e1301.
27. Breiman L. Bagging predictors. Mach Learn 1996;24:123-40.
28. Chen X, Huang L, Xie D, Zhao Q. EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. Cell Death Dis 2018;9:3.
29. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. Ann Stat 2000;28:337-407.
30. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189-232.
31. Dong X, Yu Z, Cao W, Shi Y, Ma Q. A survey on ensemble learning. Front Comput Sci 2020;14:241-58.
32. Yang LY. Classifiers selection for ensemble learning based on accuracy and diversity. Procedia Eng 2011;15:4266-70.
33. Tang EK, Suganthan PN, Yao X. An analysis of diversity measures. Machine Learning 2006;65:247-71.
34. Tang K, Li KL, Sun GW, Li HY, Zhang YZ, He W. New ensemble learning method for evidential reasoning based on diversity weighting. Appl Res Comput 2023;40:1012-8.
35. Hossin M, Sulaiman MN. A Review on Evaluation Metrics for Data Classification Evaluations. Int J Data Min Knowl Manag Process 2015;5:01-11.
36. Hand DJ, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Mach Learn 2001;45:171-86.
37. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.
38. Wilcoxon F. Individual comparisons by ranking methods. Breakthroughs in statistics: Methodology and distribution. Springer; 1992. p. 196-202.
39. Rankin NM, McWilliams A, Marshall HM. Lung cancer screening implementation: Complexities and priorities. Respirology 2020;25 Suppl 2:5-23.
40. Shi L, Sheng M, Wei Z, Liu L, Zhao J. CT-Based Radiomics Predicts the Malignancy of Pulmonary Nodules: A Systematic Review and Meta-Analysis. Acad Radiol 2023;30:3064-75.
41. Gupta H, Singh H, Kumar A. Texture and Radiomics inspired Data-Driven Cancerous Lung Nodules Severity Classification. Biomed Signal Process Control 2024;88:105543.
42. Xu QQ, Shan WL, Zhu Y, Huang CC, Bao SY, Guo LL. Prediction efficacy of feature classification of solitary pulmonary nodules based on CT radiomics. Eur J Radiol 2021;139:109667.
43. Jing R, Wang J, Li J, Wang X, Li B, Xue F, Shao G, Xue H. A wavelet features derived radiomics nomogram for prediction of malignant and benign early-stage lung nodules. Sci Rep 2021;11:22330.
44. Garau N, Paganelli C, Summers P, Choi W, Alam S, Lu W, Fanciullo C, Bellomi M, Baroni G, Rampinelli C. External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis. Med Phys 2020;47:4125-36.
45. Khan AN, Al-Jahdali HH, Allen CM, Irion KL, Al Ghanem S, Koteyar SS. The calcified lung nodule: What does it mean? Ann Thorac Med 2010;5:67-79.
46. Liu Y, Balagurunathan Y, Atwater T, Antic S, Li Q, Walker RC, Smith GT, Massion PP, Schabath MB, Gillies RJ. Radiological Image Traits Predictive of Cancer Status in Pulmonary Nodules. Clin Cancer Res 2017;23:1442-9.
47. Henschke CI, Yip R, Smith JP, Wolf AS, Flores RM, Liang M, Salvatore MM, Liu Y, Xu DM, Yankelevitz DF; . CT Screening for Lung Cancer: Part-Solid Nodules in Baseline and Annual Repeat Rounds. AJR Am J Roentgenol 2016;207:1176-84.
48. Yankelevitz DF, Yip R, Smith JP, Liang M, Liu Y, Xu DM, Salvatore MM, Wolf AS, Flores RM, Henschke CI; . CT Screening for Lung Cancer: Nonsolid Nodules in Baseline and Annual Repeat Rounds. Radiology 2015;277:555-64.
49. Erasmus JJ, Connolly JE, McAdams HP, Roggli VL. Solitary pulmonary nodules: Part I. Morphologic evaluation for differentiation of benign and malignant lesions. Radiographics 2000;20:43-58.
50. Liu YC, Liang CH, Wu YJ, Chen CS, Tang EK, Wu FZ. Managing Persistent Subsolid Nodules in Lung Cancer: Education, Decision Making, and Impact of Interval Growth Patterns. Diagnostics (Basel) 2023.
51. Li Y, Wei D, Liu X, Fan X, Wang K, Li S, Zhang Z, Ma K, Qian T, Jiang T, Zheng Y, Wang Y. Molecular subtyping of diffuse gliomas using magnetic resonance imaging: comparison and correlation between radiomics and deep

learning. Eur Radiol 2022;32:747-58.
52. Sihong Chen, Jing Qin, Xing Ji, Baiying Lei, Tianfu Wang, Dong Ni, Jie-Zhi Cheng. Automatic Scoring of Multiple Semantic Attributes With Multi-Task Feature Leverage: A Study on Pulmonary Nodules in CT Images. IEEE Trans Med Imaging 2017;36:802-14.

Cite this article as: Tang G, Du L, Ling S, Che Y, Chen X. Multi-type classification of lung nodules based on CT radiomics and ensemble learning for diversity weighting. Quant Imaging Med Surg 2024;14(12):8942-8965. doi: 10.21037/qims-24-1315