Li *et al. Respiratory Research* (2024) 25:389
https://doi.org/10.1186/s12931-024-03004-9

Respiratory Research

RESEARCH Open Access

Quantitative texture analysis using machine learning for predicting interpretable pulmonary perfusion from non-contrast computed tomography in pulmonary embolism patients

Zihan Li[1†], Meixin Zhao[2†], Zhichun Li[1], Yu-Hua Huang[1], Zhi Chen[1], Yao Pu[1], Mayang Zhao[1], Xi Liu[3,4], Meng Wang[2], Kun Wang[3], Martin Ho Yin Yeung[1], Lisheng Geng[4], Jing Cai[1,5], Weifang Zhang[2*], Ruijie Yang[3*] and Ge Ren[1,5*]

Abstract

Background Pulmonary embolism (PE) is life-threatening and requires timely and accurate diagnosis, yet current imaging methods, like computed tomography pulmonary angiography, present limitations, particularly for patients with contraindications to iodinated contrast agents. We aimed to develop a quantitative texture analysis pipeline using machine learning (ML) based on non-contrast thoracic computed tomography (CT) scans to discover intensity and textural features correlated with regional lung perfusion (Q) physiology and pathology and synthesize voxel-wise Q surrogates to assist in PE diagnosis.

Methods We retrospectively collected ^{99m}Tc-labeled macroaggregated albumin Q-SPECT/CT scans from patients suspected of PE, including an internal dataset of 76 patients (64 for training, 12 for testing) and an external testing dataset of 49 patients. Quantitative CT features were extracted from segmented lung subregions and underwent a two-stage feature selection pipeline. The prior-knowledge-driven preselection stage screened for robust and non-redundant perfusion-correlated features, while the data-driven selection stage further filtered features by fitting ML models for classification. The final classification model, trained with the highest-performing PE-associated feature combination, was evaluated in the testing cohorts based on the Area Under the Curve (AUC) for subregion-level predictability. The voxel-wise Q surrogate was then synthesized using the final selected feature maps (FMs) and model

†Zihan Li, Meixin Zhao contributed equally.

*Correspondence:
Weifang Zhang
tsy1997@126.com
Ruijie Yang
Ruijyang@yahoo.com
Ge Ren
gary-ge.ren@polyu.edu.hk

Full list of author information is available at the end of the article

score maps (MSMs) to investigate spatial distributions. The Spearman correlation coefficient (SCC) and Dice similarity coefficient (DSC) were used to assess the spatial consistency between FMs or MSMs and Q-SPECT scans.

Results The optimal model performance achieved an AUC of 0.863 during internal testing and 0.828 on the external testing cohort. The model identified a combination containing 14 intensity and textural features that were non-redundant, robust, and capable of distinguishing between high- and low-functional lung regions. Spatial consistency assessment in the internal testing cohort showed moderate-to-high agreement between MSMs and reference Q-SPECT scans, with median SCC of 0.66, median DSCs of 0.86 and 0.64 for high- and low-functional regions, respectively.

Conclusions This study validated the feasibility of using quantitative texture analysis and a data-driven ML pipeline to generate voxel-wise lung perfusion surrogates, providing a radiation-free, widely accessible alternative to functional lung imaging in managing pulmonary vascular diseases.

Clinical trial number Not applicable.

Keywords Non-contrast computed tomography, Perfusion, Lung functional imaging, Pulmonary embolism, Radiomics

Introduction

Pulmonary embolism (PE) is recognized as the third leading cause of cardiovascular death, following acute myocardial infarction and stroke [1]. In PE, the obstruction of the pulmonary artery is typically caused by the accumulation of blood clots (thrombi), air bubbles, fat tissue, or amniotic fluid. This often occurs in individuals with pre-existing conditions such as heart disease, cancer, severe fractures, or during pregnancy [2, 3]. PE is associated with significant mortality, particularly in high-risk patients presenting with shock or cardiac arrest, where the 30-day mortality rate can reach 52–65% [4]. It is estimated that PE affects 300,000 to 600,000 people annually in the United States, resulting in at least 100,000 fatalities [5, 6]. Given the complex etiology and rapid progression of PE, timely and accurate diagnosis, as well as prompt treatment, are crucial for clinicians to improve patient outcomes.

Diagnosing PE cannot rely solely on clinical evaluation, as the supporting symptoms, signs, and laboratory data are often deceptively nonspecific [7]. Although computed tomography pulmonary angiography (CTPA) is the gold standard for diagnosing PE, its use of iodinated contrast agents (ICAs) poses risks for patients with contraindications such as ICA allergies or renal failure, limiting its universality. Studies have shown that repeated use of ICAs increases the risk of acute adverse reactions [8]. According to the 2019 guidelines for the diagnosis and management of pulmonary embolism, developed in collaboration with the European Respiratory Society, ventilation/perfusion single-photon emission computed tomography (V/Q-SPECT) is one of the most widely used imaging techniques in PE diagnosis, with a sensitivity and specificity of 97% and 91%, respectively [9, 10]. However, V/Q-SPECT is invasive, expensive, and less widely available than other imaging modalities like CT, often leading to longer waiting times for testing, especially in resource-limited hospitals or regions without nuclear medicine (NM) facilities [11].

In addition to mainstream pulmonary function examinations based on contrast agents, several studies have attempted to obtain perfusion information indirectly from more commonly available clinical thoracic scans, with computed tomography perfusion imaging (CTPI) gaining the most attention. A category of CTPI techniques employs deformable image registration (DIR) to physically model surrogates of pulmonary blood supply and circulation from non-contrast respiratory-related CT scans (i.e., four-dimensional CT, 4DCT), for restoring static perfusion distribution at the voxel level [12]. These methods typically have a strict and interpretable modeling process, but the prediction is severely affected by image quality (e.g. CT noise and motion artifacts) and DIR results. These limitations can lead to significant variability in perfusion assessments, potentially compromising diagnostic accuracy. Additionally, some deep-learning (DL) approaches have been proposed to directly synthesize pulmonary perfusion images from a single CT image via deep neural networks [13–15]. The application of end-to-end DL techniques, although innovative and efficient, faces challenges associated with their complex decision-making processes, and lack of consideration of clinically relevant pathological mechanisms. Overall, although still in the early stages of development, the emergence of CTPI has the potential to reduce the misuse of contrast agents and improve clinical technicians' efficiency.

To tackle these challenges, this study extends CTPI explorations by harnessing the capabilities of quantitative texture analysis. Texture analysis is capable of extracting high-dimensional features from routine non-contrast CT imaging that are typically not discernible by the human eyes. These features can then be quantitatively transformed into imaging biomarkers with explicit

mathematical definitions [16, 17]. By focusing on feature extraction and interpretability, our approach mitigates the "black box" issues seen in DL methods, while also addressing image quality limitations through robust feature selection and model optimization. We aimed to conduct the quantitative texture analysis and data-driven machine learning (ML) pipeline on non-contrast thoracic CT scans to discover both intensity and textural features that are correlated to the underlying regional lung perfusion physiology and pathology. Furthermore, by studying the spatial distribution of the perfusion-correlated features and fitted-model outputs, we will develop a reliable and interpretable method for predicting voxel-wise lung perfusion surrogates. The proposed method represents a novel perspective on CTPI, offering potentially improved diagnostic reference on PE and other perfusion-related lung diseases.

Materials and methods

Workflow overview

Figure 1 depicts the overall workflow of this study. In the feature selection stage, lung regions in CT images from all patients were segmented into multiple subregions. Based on a subject-specific threshold, each CT lung subregion was binarily labeled as either high- or low-functioning according to its spatial-averaged Q-SPECT signal. Within each high or low functional subregion, features were extracted for intensity and texture information and underwent a two-stage feature selection pipeline. The prior-knowledge-driven preselection stage initially screened for robust and non-redundant perfusion-correlated features, while the data-driven selection stage further filtered features by fitting ML models for a classification task. The final classification model was trained with the highest-performing PE-associated feature combination and evaluated for the subregion-level predictability. Feature maps (FMs) and model score maps (MSMs) of the final selected features were generated to study their spatial distributions and as potential surrogate perfusion maps. The voxel-wise Spearman correlation coefficient (SCC) and Dice similarity coefficient (DSC) for functional lung volumes were used to assess the spatial consistency between FMs or MSMs and Q-SPECT scans. The key stages in this workflow are further specified below.

Patient data and preprocessing

The internal discovery dataset was retrospectively collected from a database previously utilized in our research [14], with approval from the Institutional Review Board (IRB) of the affiliated institution. The raw database includes 173 patients who underwent pulmonary ^{99m}Tc-labeled macroaggregated albumin (^{99m}Tc-MAA) Q-SPECT/CT scanning at Queen Mary Hospital, from 2019 to 2023 for suspected lung diseases. Patient characteristics are summarized in Appendix A. Subject exclusion criteria for this study consisted of other lung conditions, such as pulmonary hypertension, lung cancer, or congestive heart failure, as well as incomplete pulmonary imaging. Additionally, an external testing dataset of Q-SPECT/CT scans collected from 2020 to 2023 at Peking University Third Hospital was included (Fig. 2) and was approved by the Medical Science Research Ethics Committee of the affiliated institution.

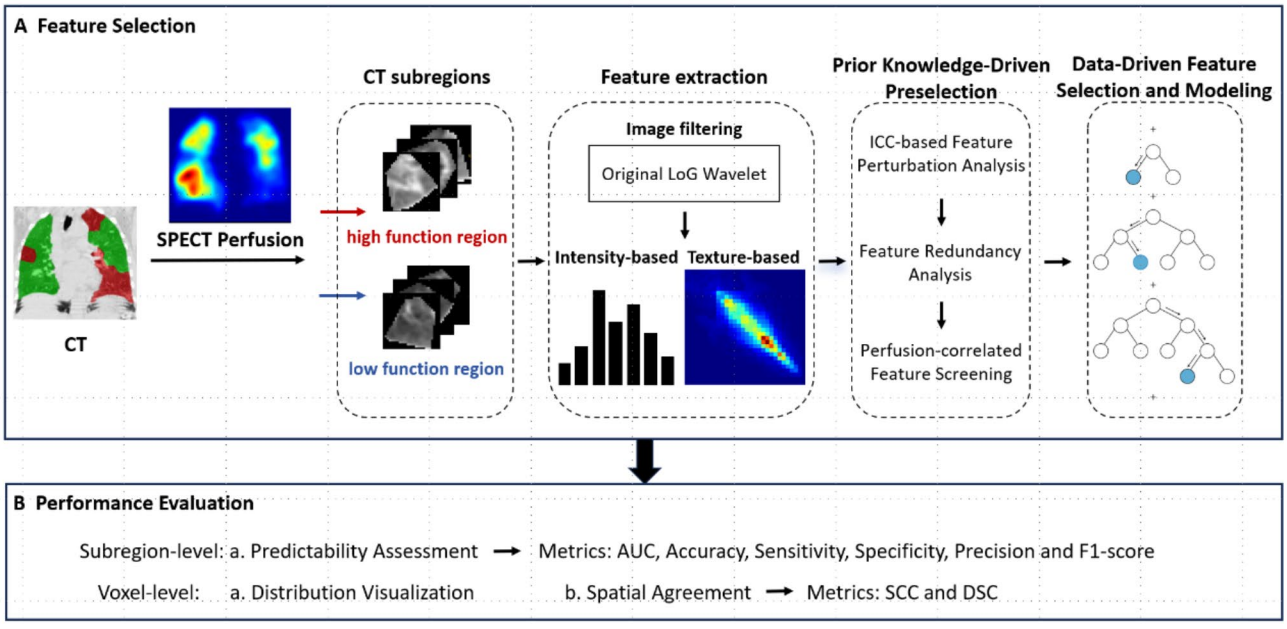

Fig. 1 The overall workflow of the study

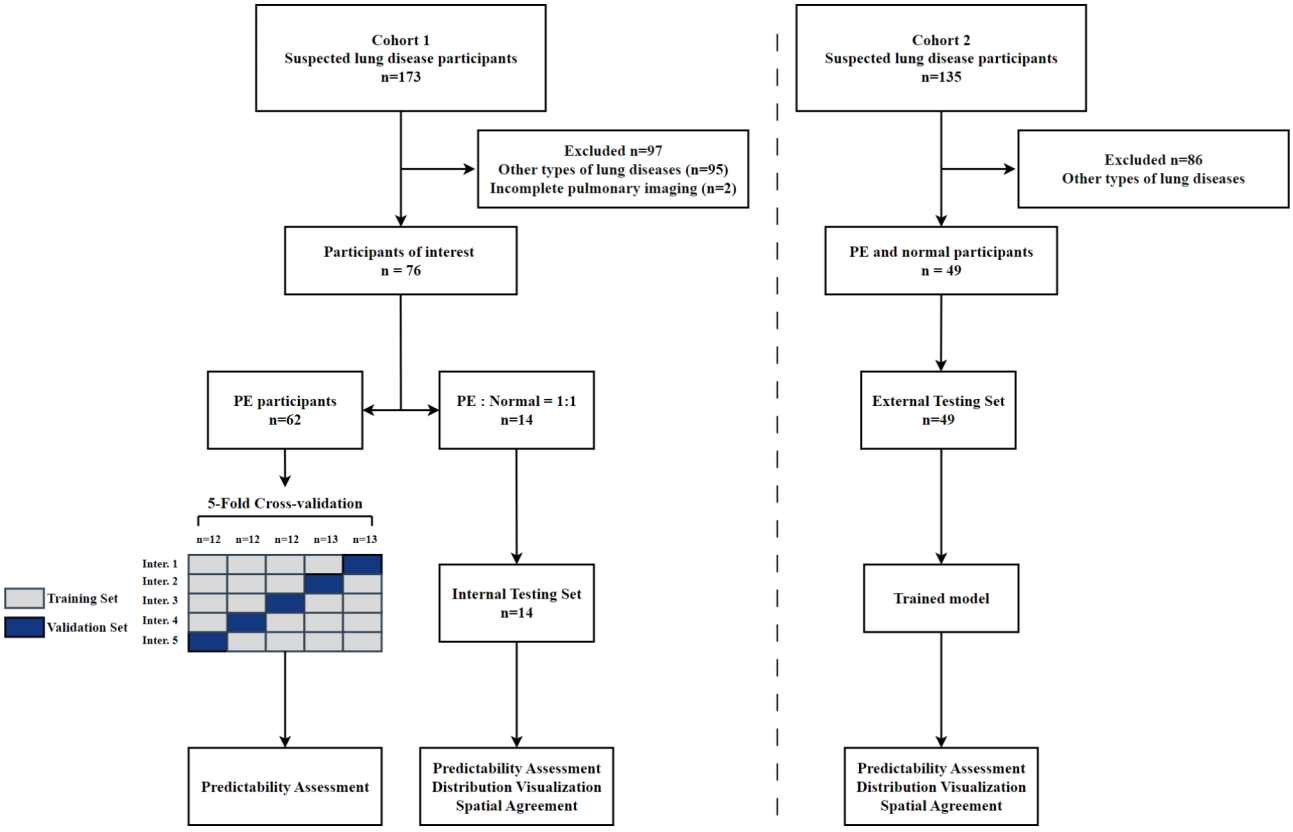

Fig. 2 The flowchart of Q-SPECT/CT scans included in the study. It shows the number of participants of interest in the pipeline development, the number of participants excluded from the study due to other types of lung diseases or incomplete lung imaging, and the number of participants used for external testing. Among them, the training cohort was divided into five groups. In each fold, the classifier was trained in four groups and validated in the remaining group

Patients' Q-SPECT/CT images for the internal discovery cohort were acquired using a GE Discovery 670 SPECT/CT scanner at a frame rate of 30 s/frame, totaling 60 frames. Prior to imaging, each patient was injected with 3 mCi of ^{99m}Tc-MAA in a supine position to prevent sedimentation of macroaggregates into the lung bases. The CT scans were acquired at 120 kVp and 80 to 120 mAs, and reconstructed into a 512×512 matrix with a resolution of 0.97×0.97×1.25 mm^3. Q-SPECT scans were acquired in a standard 128×128 matrix with a voxel size of 4.42×4.42×4.42 mm^3. On the other hand, Q-SPECT/CT images for the external testing cohort were collected using the Symbia Intevo SPECT/CT scanner by Siemens Healthineers. The CT scans were acquired at 130 kVp and 70 to 200 mAs, and reconstructed into a 512×512 matrix with a resolution of 0.97×0.97×5.00 mm^3. Q-SPECT scans were acquired in a standard 128×128 matrix with a voxel size of 4.80×4.80×4.80 mm^3. Q-SPECT images were rigidly registered with CT images to align their spatial positions.

All images, including CT scans and SPECT-based perfusion scans, were resampled to an isotropic resolution of 1.0×1.0×1.0 mm^3 using linear interpolation to ensure geometric consistency across modalities and subjects. To isolate the lung regions for focused analysis and reduce noise from surrounding tissues, lung masks were generated from the CT images for each patient using an open-source U-net-based model [18]. Each generated lung mask was manually checked, and the trachea regions were excluded from the mask. To reduce subsequent computational costs, all images were cropped to include only the lung regions. To suppress non-lung regions and anatomical structures irrelevant to PE while enhancing contrast, only lung voxels within the intensity range of -1000 Hounsfield unit (HU) to 200 HU were utilized from the CT images.

Subregion generation and functional label identification

For each patient, the lung mask was segmented into roughly equal-sized subregions using the masked simple linear iterative clustering (maskSLIC) algorithm [19], with each subregion set to a volume of 21×21×21 mm^3. These subregions were highly compact, balancing both intensity homogeneity and geometric proximity. Subsequently, based on the subject-specific signal threshold calculated in corresponding SPECT images, these

subregions were categorized into high-functional subregions and low-functional subregions. Specifically, the signal threshold, employed in previous NM-based and CT-based lung function studies, was defined as deviating by 15% from a normal lung function distribution, with careful mitigation of hotspots' impact [20, 21].

Subregion-wise feature extraction

The gray-level intensity of CT images was discretized into 64 bins ranging from −1000 to 200 HU for feature calculation. A total of 1116 features were extracted from each subregion using a radiomics software package (PyRadiomics; Harvard Medical School, Boston, Massachusetts, USA), which transforms images into quantitative data and is compliant with the Imaging Biomarker Standardization Initiative (IBSI) [22]. There were six types of features: first-order features, grey level co-occurrence matrix (GLCM) features, grey level size zone matrix (GLSZM) features, grey level run length matrix (GLRLM) features, grey level dependence matrix (GLDM) features, and neighboring gray-tone difference matrix (NGTDM) features. The latter five types are collectively referred to as textural features. In addition to the original images, features were also calculated from images processed through Laplacian of Gaussian (LoG) filtering and wavelet filtering. All features were extracted repeatedly from LoG-filtered images with sigma values of 0.5 mm, 1 mm, and 2 mm, and from wavelet-filtered images at eight different frequencies. Appendix B presents the radiomics features used in this study.

Prior knowledge-driven feature preselection

Feature robustness analysis

We conducted a perturbation analysis to evaluate the robustness of the extracted radiomics features. To simulate patient movement during imaging, variations in noise levels across different imaging devices, and uncertainties in region-of-interest (ROI) contouring, we randomly applied four fundamental image perturbation techniques: rotation (R), translation (T), noise addition (N), and contour randomization (C) [23]. The specific parameters used for each perturbation method are detailed in Appendix C. For each subregion, 10 images were independently generated with random perturbations, and radiomics features were re-extracted from them. Intraclass correlation coefficient (ICC) [24] was calculated for each feature to quantify its robustness against perturbations. The ICC threshold of 0.75 for identifying features with relatively high robustness was commonly used in previous radiomics literature [23]. Features that met or exceeded the threshold were considered reliable and retained for further analysis.

Feature redundancy analysis

To analyze feature redundancy within the dataset, we utilized the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [25]. The Pearson correlation coefficient (PCC) was employed as the metric to measure pairwise distances between features. A threshold of 0.95 for the PCC was set to identify and group highly correlated features into distinct redundant clusters. The threshold was chosen based on a published study [26]. In that paper, a PCC threshold of 0.9 was set to identify redundant feature clusters from 79 radiomic features. As for this study, we slightly increase the PCC threshold to 0.95, to retain more diverse features from different categories in a larger initial feature pool and to balance the number of features.

Perfusion-correlated feature screening

In this step, the features that best represented the function differences were filtered out. For all feature clusters, a non-parametric two-sided Mann-Whitney U test was conducted on the feature values extracted from high-function subregions and low-function subregions to calculate the effect size (ES), which measures the degree of rank differences between the two groups [27]. For each redundant cluster, we retained the feature with the highest absolute ES value, provided that the absolute ES value was greater than 0.33, to ensure the capability to distinguish between high-functional and low-functional regions of the lung. The ES threshold of 0.33 was set based on the statistical interpretations of the nonparametric test-derived effect size (rank-biserial correlation). A rank-biserial correlation ES > 0.33 indicates that the feature might have a medium-to-large ability to differentiate perfusion regions in lung CT. The filtered features served as the candidate feature set for subsequent data-driven feature selection.

Data-driven feature selection and modeling

We randomly selected an equal number of participants from the eligible PE participants to match the normal participants, together forming an internal testing cohort, while the remaining patients constituted the training cohort. Z-score normalization was applied to ensure that feature values were within a comparable range, preventing any particular feature from dominating the modeling process. Specifically, the feature values in the candidate feature set were scaled to have a mean of 0 and a standard deviation of 1. As shown in Fig. 2, the training cohort was divided into five groups, and a five-fold cross-validation with the CatBoost classifier was employed to avoid overfitting [28]. In each fold, one group was selected for validation, while the remaining four groups were used for classifier training. Feature selection in each fold was performed using the recursive feature elimination (RFE)

algorithm based on importance, eliminating the features with the lowest importance ranking until the number of remaining features reached the predetermined count. As a result, five fold-specific feature combinations with perfusion-correlated predictive capabilities were obtained. These five feature combinations were combined to create an integrated feature set, where each feature was ranked based on its frequency of occurrence during cross-validation. We retained the features that appeared more than three times as the final selected feature combination for training a final model, which was applied to the testing cohorts for evaluating subregion-level predictability. The predictabilities of the fold-wise models and the final model were evaluated by the receiver operating characteristic (ROC) curves, with the fold-averaged AUC for the internal validation cohort. 1000 bootstrap resamples were performed on all cohorts to calculate the 95% confidence intervals (CI) to assess the stability of the model performance. Additionally, the final model with the final feature combination was then applied to the external testing cohort to assess the accuracy and generalizability of the model. As performance comparisons, once the final feature combination was selected, logistic regression (LR) and support vector machine (SVM) were also employed to fit the training cohort and validate AUCs with the internal testing cohort.

Spatial distribution study

For the features within the final combination, we calculated their FMs for visualizing the spatial distribution of features on the CT image. In order to enhance computational efficiency and obtain relatively smooth FMs that retain the trend of the spatial distribution of features and eliminate artifacts, we employed a coarse-to-fine approximation method that transitioned from subregion-wise feature measurement to voxel-wise dense FM, instead of using the sliding-window filtering method [26]. For each patient, we first constructed a weighted matrix for each voxel with respect to all subregions based on Shepard's class of moving least squares approximation [29, 30]. Using the constructed weighted matrix, the feature values measured from each subregion can be transferred to each lung voxel that forms the FM.

We quantitatively evaluated the spatial similarities between FMs and reference Q-SPECT scan images using the SCC, as well as the Dice Similarity Coefficient for low functional region (DSC_{LO}) and high functional region (DSC_{HI}). The same approach as used for subregion functional label identification in Sect. 2.3 was applied to define subject-specific signal thresholds for dividing high- and low-functional regions from FMs and corresponding reference Q-SPECT images. Similarly, we generated MSMs based on the model's predicted probabilities for subregions and assessed the spatial consistency with the reference Q-SPECT scans.

Results

Study population and feature preselection

Of the 173 participants from QMH enrolled in the study, 97 were excluded due to other types of lung diseases ($n=95$) or incomplete pulmonary imaging ($n=2$). To ensure the homogeneity of the disease discovery and to simulate a realistic population, 49 patients diagnosed with PE and normal conditions out of 135 participants from the Institution B cohort were involved. The selected participants of interest ($n=76$) had a mean age of 67 ± 17 years, with a female proportion of 61.4%. Patient demographic data are presented in Table 1, which indicates that there was no statistically significant difference between the patients from the two centers ($p>0.05$).

In the process of prior knowledge-driven preselection, Fig. 3 illustrates the results of the feature robustness analysis. By setting the threshold of ICC value to 0.75, a total of 587 robust features were filtered out from all of the considered features. As shown in Appendix D, feature redundancy analysis on the robust features identified 329 independent non-redundant clusters. After removing features with ES values lower than 0.33, the candidate feature set included a total of 151 features, each selected as the highest ES value feature from each redundant cluster, for data-driven feature selection.

Subregion-level classification performance

Figure 4 depicts the change in the AUC of the CatBoost classifier with the number of selected features. The optimal model performance was determined when the

Table 1 Study Population

Characteristics	QMH							PUTH		
	Training Cohort ($n=62$)		Testing Cohort ($n=14$)		P Value	All Participants ($n=76$)		External Validation Cohort ($n=49$)		P Value
Age (y)	67 ± 17		66 ± 18		0.91	67 ± 17		61 ± 18		0.15
Gender					0.92					0.3
Male (Value, %)	23	37.1	6	42.9		46	38.7	11	22	
Female (Value, %)	39	62.9	8	57.1		29	61.4	39	78	

Note: Unless otherwise stated, the data are presented as mean ± standard deviation

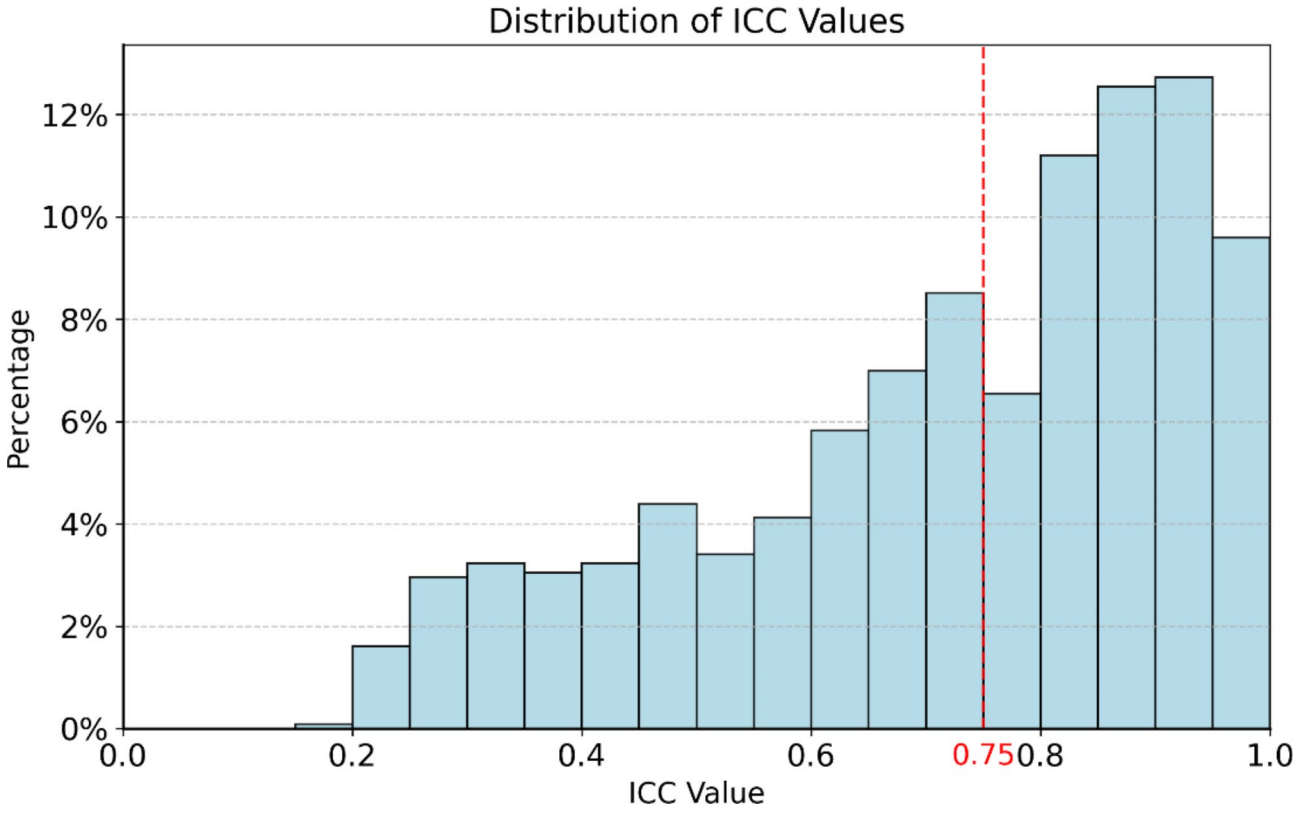

Fig. 3 Histogram of the distribution of the intraclass correlation coefficient (ICC) values for all considered features. The vertical axis represents the percentage of features within the corresponding ICC value range out of the total number of features. The red line indicates the threshold of 0.75, above which robust features are retained

number of selected features reached 16, achieving the highest predictability during the cross-validation. At this point, the AUC for the training cohort was 0.928 (95% CI: 0.922, 0.935), and the fold-averaged AUC for the internal validation cohort was 0.823 (95% CI: 0.814, 0.831). The evaluation of the classifier on predictability assessment, including AUC, accuracy, sensitivity, specificity, precision and F1-score, in the internal discovery and external testing dataset are summarized in Table 2. Figure 5A presents the ROC curves for the evaluation of subregion-level predictability in the training, internal validation, and internal testing cohorts. Fourteen features, which were selected more than three times in the five-fold cross-validation, were retained as the final feature combination, including features (1) original First-order 10Percentile, (2) original First-order Skewness, (3) log-sigma-2-0-mm NGTDM Busyness, (4) log-sigma-1-0-mm First-order Skewness, (5) log-sigma-2-0-mm First-order 90Percentile, (6) wavelet-LLH First-order Median, (7) wavelet-LLH First-order Mean, (8) wavelet-HLH First-order 90Percentile, (9) log-sigma-2-0-mm First-order 10Percentile, (10) wavelet-LHL First-order Minimum, (11) wavelet-LHH GLCM Imc1, (12) wavelet-LLH First-order 10Percentile, (13) wavelet-LLH First-order Minimum, (14) wavelet-HLH First-order RootMeanSquared.

As shown in Fig. 5B, we compared the performance of different classifiers with the final feature combination. The CatBoost model achieved an AUC of 0.863 (95% CI: 0.849, 0.877) during internal testing. Using the identical set of features, the LR and SVM classifier yielded AUCs of 0.797 (95% CI: 0.779, 0.814) and 0.847 (95% CI: 0.832, 0.861), respectively, indicating that the CatBoost classifier model outperformed the other classifiers in predictability. Additionally, the final CatBoost model also had a generalizable subregion-level predictability on the external testing cohort with an AUC of 0.828 (95% CI: 0.820, 0.838). Figure 6 displays the SHAP beeswarm summary plot with features and their contributions color-coded, where each dot corresponds to a subregion as a sample. Features in the final combination are ranked in descending order based on their contribution to the model's predictability, specifically by the mean absolute SHAP value. High feature values aligned with high contribution levels indicate a positive contribution to the model's predictability.

Voxel-wise perfusion surrogate analysis

Figure 7 visualizes the FMs and MSMs of a representative patient with PE, using the Q-SPECT scan as a reference. Major defects of perfusion can be identified with the

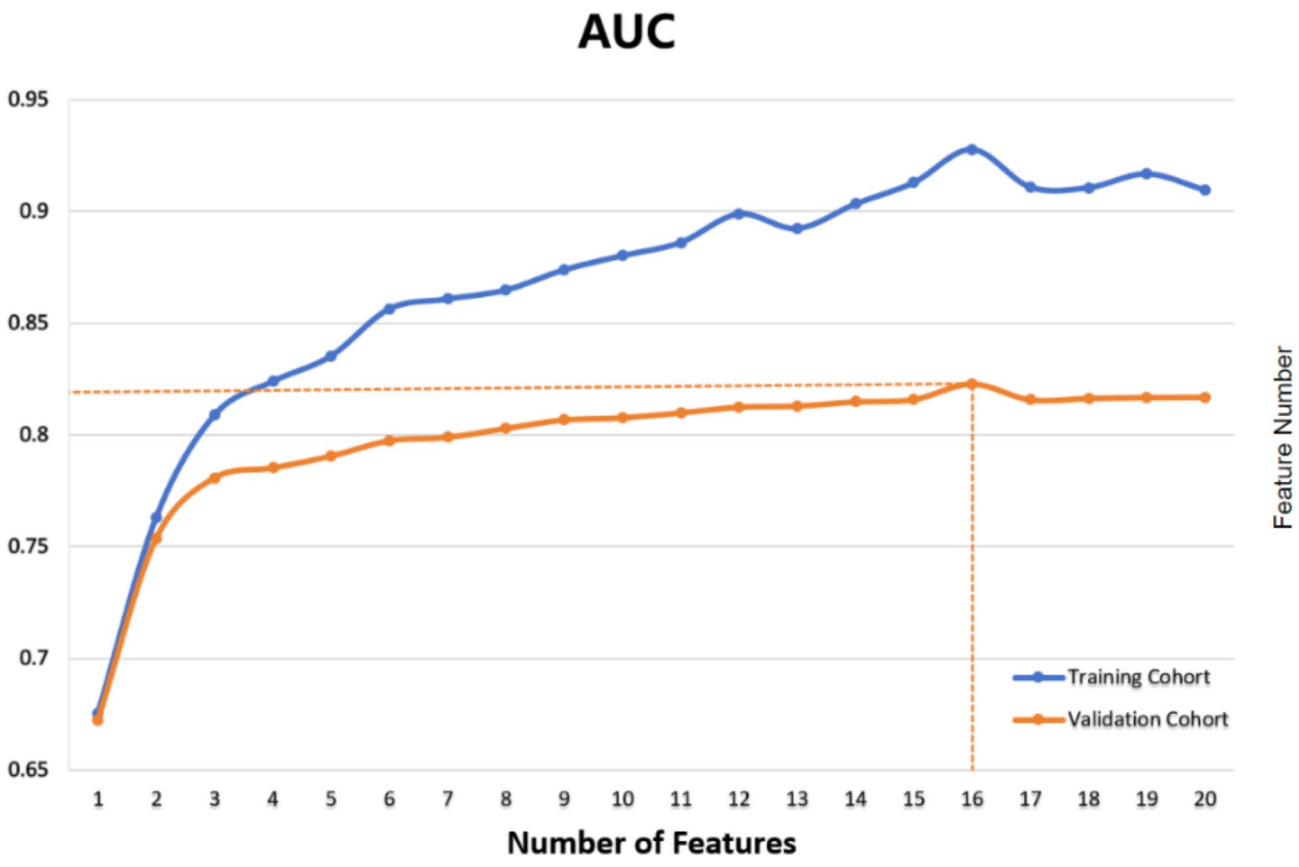

Fig. 4 The change in area under the receiver operating characteristic (ROC) curve (AUC) of the CatBoost model in training and internal validation cohorts from the internal discovery dataset with the number of selected features. The optimal model performance was achieved when 16 features were selected, resulting in the highest predictability (the fold-averaged AUC) in the internal validation cohort

Table 2 Subregion-level agreement in predictability metrics across the training, internal validation, and internal testing cohorts

	Training	Internal Testing			External Testing
		CatBoost	LR	SVM	
AUC	0.93 (0.92, 0.93)	0.86 (0.86, 0.87)	0.80 (0.80, 0.81)	0.85 (0.85, 0.86)	0.82 (0.81, 0.83)
Accuracy (%)	85 (84, 86)	78 (78, 79)	71 (71, 72)	75 (75, 76)	75 (74, 76)
Sensitivity (%)	92 (92, 93)	67 (65, 69)	70 (68, 72)	77 (76, 78)	71 (70, 72)
Specificity (%)	85 (84, 86)	84 (84, 85)	77 (76, 77)	82 (82, 83)	82 (81, 83)
Precision	0.84 (0.83, 0.85)	0.86 (0.85, 0.87)	0.84 (0.84, 0.85)	0.87 (0.85, 0.89)	0.94 (0.92, 0.96)
F1-score	0.88 (0.87, 0.89)	0.82 (0.82, 0.83)	0.74 (0.73, 0.75)	0.78 (0.76, 0.80)	0.83 (0.83, 0.84)

Note: All Data in parentheses are 95% confidence intervals

Abbreviation: LR, Logistic Regression; SVM, Support Vector Machine; AUC, Area under the curve

Q-SPECT scan in the left lower and right middle lobes of the lungs. Notably, the relationship between these FMs and the perfusion defects aligns with their previously calculated ES. Several positively correlated FMs (with positive ES) show low signal regions that broadly correspond to the perfusion defects. Conversely, negatively correlated FMs (with negative ES) display high signal regions in areas of reduced perfusion. The MSM demonstrates a positive correlation with perfusion distribution and exhibits greater visual similarity to the Q-SPECT scan than any single FM.

Figure 8; Table 3 summarize the quantitative evaluation results on spatial consistency among internal and external testing cohorts. In box plots, the central points

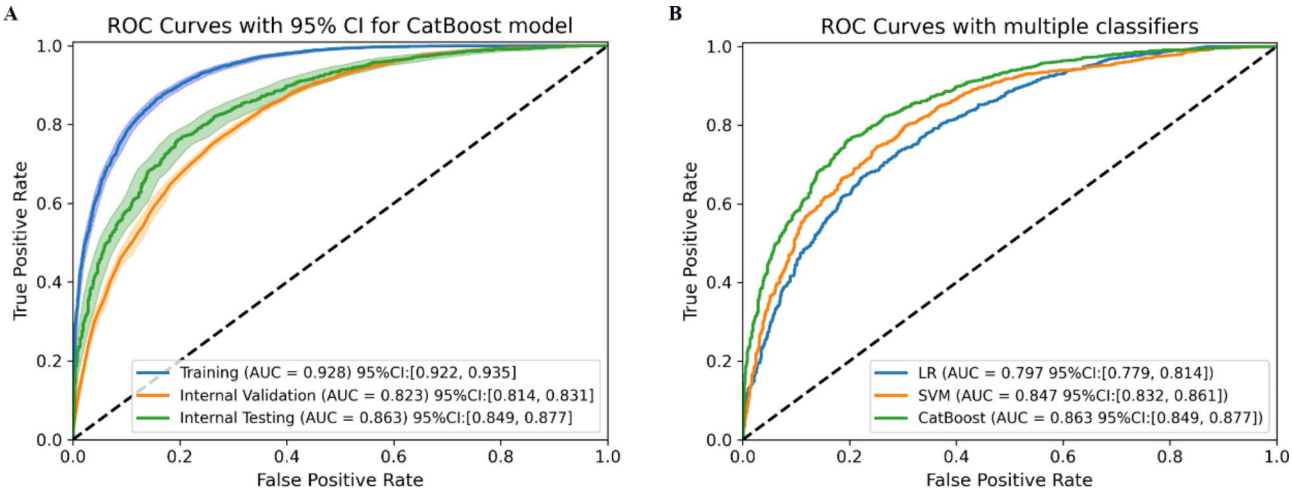

Fig. 5 Receiver operating characteristic (ROC) curves show the result of subregion-level evaluation on agreement in the training, internal cross-validation and internal testing cohorts (**A**) and for multiple classifiers in the internal testing cohort (**B**). The shaded areas represent the 95% confidence intervals (CI) for the ROC curves, indicating the range within which the model's true performance is expected to fall with 95% confidence

represent the mean, reflecting the central tendency of the dataset. The central horizontal line denotes the median. The top and bottom edges of the box indicate the first quartile (Q1) and the third quartile (Q3), respectively, illustrating the interquartile range (IQR). The whiskers extend to the smallest and largest values within 1.5 times the IQR from Q1 and Q3, respectively, with values beyond this range, marked by blue points, are considered outliers. Evaluations in the internal testing cohort are shown MSMs reached the median SCC of 0.66 and the median DSC of 0.64 and 0.86 for high and low functional regions, respectively. For the external testing cohort, MSMs achieved the median SCC of 0.60 as well as the median DSC of 0.49 and 0.85 for high and low functional regions, respectively. At the same time, we conducted subgroup analyses to evaluate the model performance across different clinical risk profiles in Appendix E.

Discussion

In this study, a cohort of 125 research participants underwent robustness analysis, redundancy analysis, and subregion-level perfusion-correlated feature screening on a set containing 1116 features. Subsequently, a data-driven feature selection and modeling pipeline was established. The model we developed identified 14 density-based and texture-based features, all of which were non-redundant, robust and capable of distinguishing between high and low functional regions of the lung (ES>0.33). The optimal model, which performed best during the cross-validation, was applied to the internal and external testing cohorts, yielding an AUC of 0.86 and 0.82, respectively, initially indicating effective predictability and cross-institutional generalizability.

The optimal model developed using the CatBoost algorithm retained 14 key features reflecting spatial heterogeneity of lung perfusion in PE cases. Most features were first-order features, including 10Percentile, 90Percentile, Minimum, Median, and Mean, which quantitatively assess image intensity values. Skewness and Root mean squared measure the asymmetry and variability of intensity distribution, respectively, indicating the density distribution and internal heterogeneity of defect subregions. The remaining features included two high-dimensional textural features: NGTDM Busyness, indicating the degree of gray-level variation between image voxels and their neighborhoods, suggesting higher irregularity in local textures, potentially containing edges; and GLCM-Imc1, indicating gray-level correlation within the image. The latter was extracted using a wavelet-LHH filter, applying high-pass filters in the y and z dimensions to enhance spatial heterogeneity and detect solid components of defect subregions.

Within the selected feature combination, the median SCC of the FMs for 7 features indicated moderate-to-high spatial similarity with reference Q-SPECT images and the MSMs showed significant spatial consistency with reference Q-SPECT images. Both visual comparisons and quantitative evaluations in the spatial distribution study of the selected features' FMs and MSMs suggest that while individual FMs capture various aspects of perfusion distribution, the fitted MSM has the potential to provide more comprehensive and accurate representations of lung perfusion. The improved spatial correlation between the MSM and Q-SPECT scans underscores the value of the proposed ML approach in integrating multiple intensity and texture features to predict pulmonary function.

In terms of methodology, this study proposes a new perspective on CTPI for assisting PE diagnosis, distinguishing itself from DIR-based and purely deep

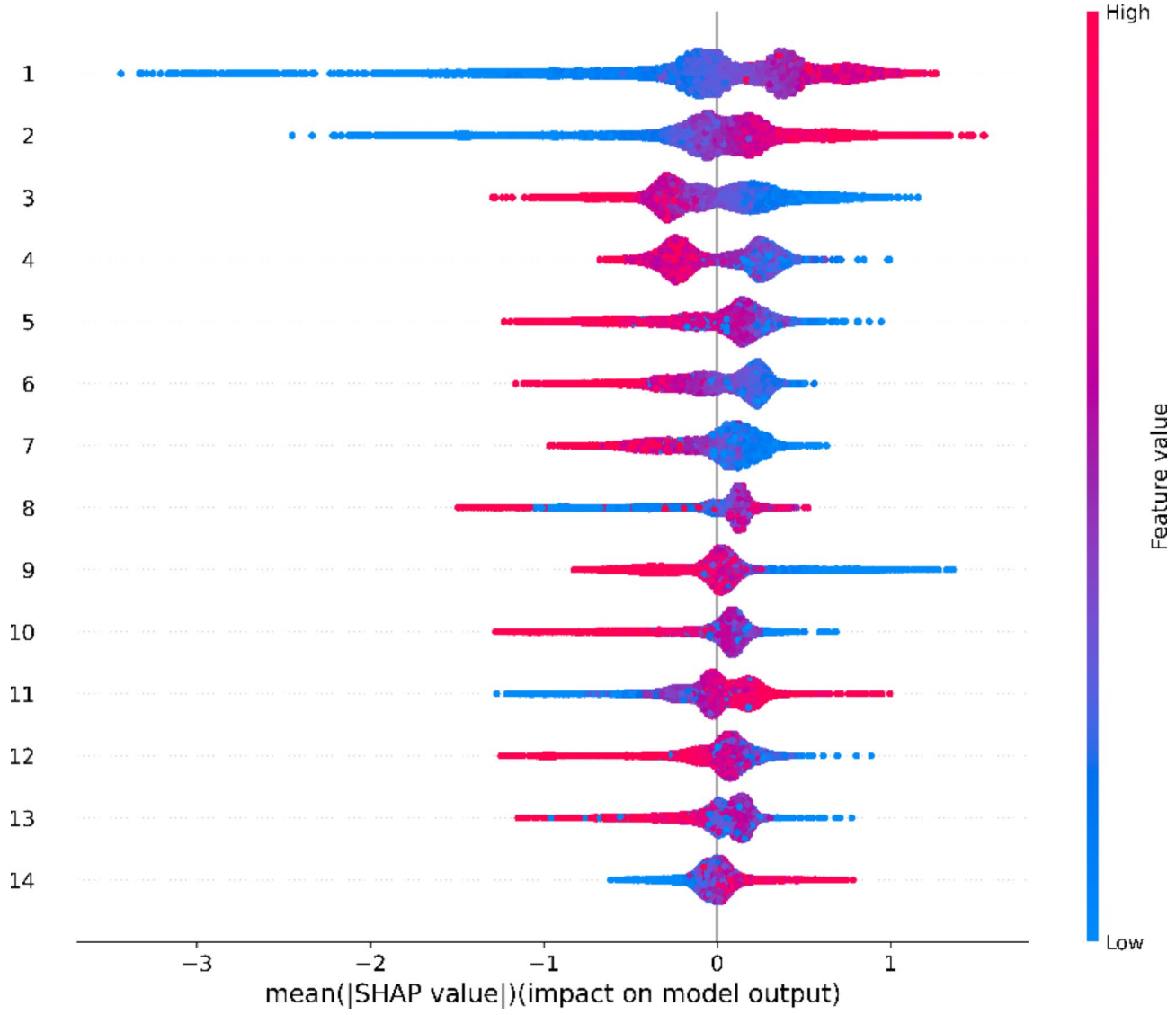

Fig. 6 SHAP (SHapley Additive exPlanations) summary plot for the final CatBoost classifier

learning-based methods, which are susceptible to image quality issues or lack interpretability [31, 32].In our framework, alongside intensity and texture features from raw images, we utilized high-dimensional features extracted from LoG- and wavelet-filtered images. Although high-dimensional features are more susceptible to motion artifacts [33, 34], our study conducted perturbation analysis to screen for robust features. Furthermore, acknowledges the nonlinear relationship between texture features and local lung function, while LR has a relatively simple parametric structure and provides a linear model to predict outcomes [35], CatBoost and SVM are more efficient and robust in addressing nonlinear problems. Prior studies have shown CatBoost's superior effectiveness in handling nonlinear data relationships compared to other ML techniques [36]. From a clinical standpoint, providing an alternative method of assessing lung perfusion through routine CT scanning may reduce the need for more invasive or radiation-intensive procedures. Our method generates voxel-level perfusion maps that help localize perfusion defects precisely, thereby enhancing the diagnosis and management of PE and other lung diseases associated with pulmonary perfusion. To minimize any bias that the algorithm might introduce, we applied a five-fold cross-validation for the training dataset from the prior experience [37]. In addition to this, feature robustness analysis selected features that appeared more than three times during cross-validation, avoiding reliance on a single feature selection method. When dealing with a relatively small sample size, we adjusted the model parameters based on prior experience to ensure that the AUC difference between the training

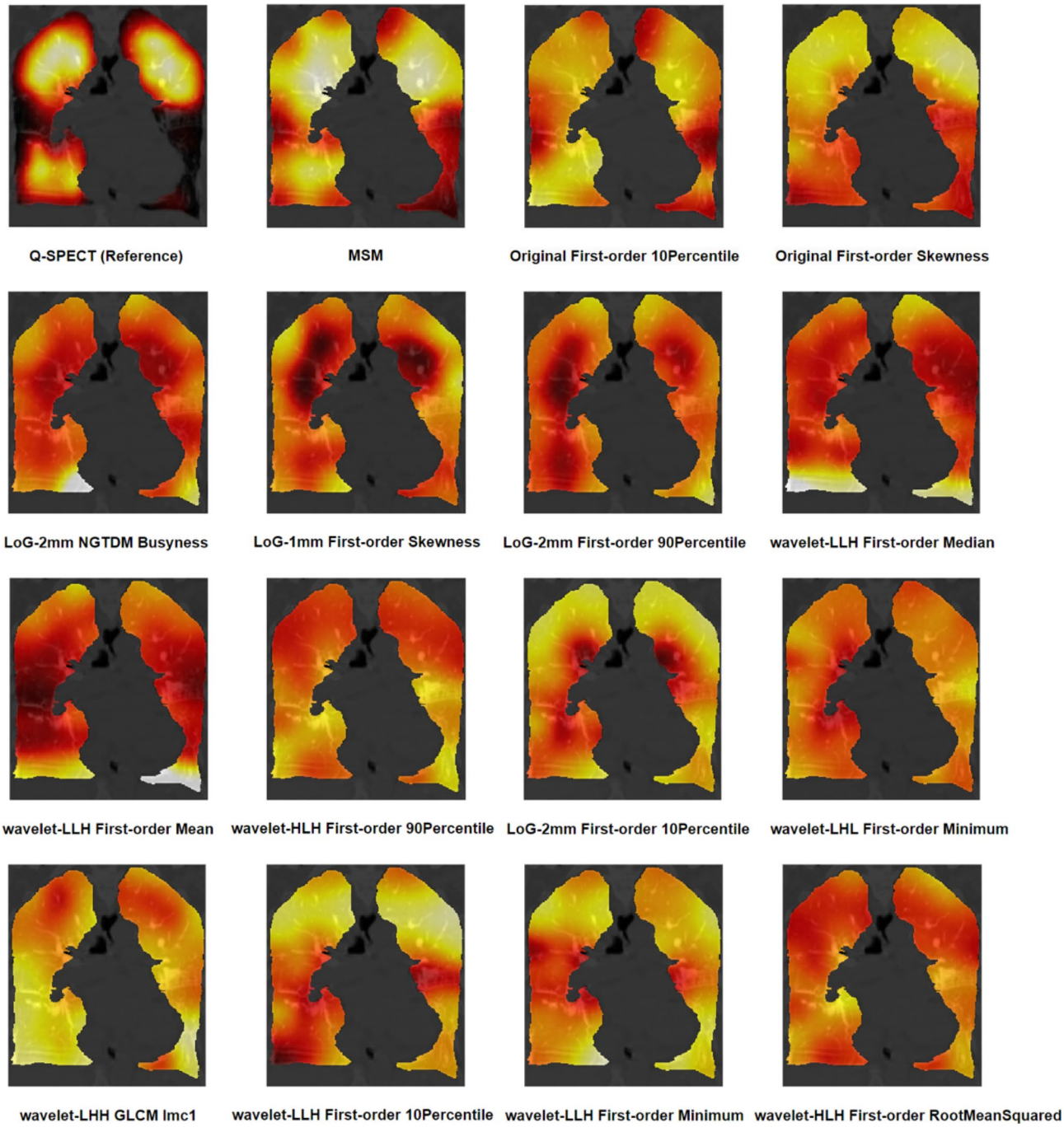

Fig. 7 Reference Q-SPECT scan, feature maps of the 14 selected features from the final feature combination and model score map (MSM) of a representative subject with pulmonary embolism disease

and testing datasets remained below 0.1, which helps to prevent overfitting.

Despite the promising results, our study has several limitations and points for improvement that are worth considering. The dataset's relatively homogeneous characteristics may limit the model's generalizability to more varied populations. The relatively small sample size, while mitigated by cross-validation and external testing, suggests that evaluation on larger, diverse (i.e., across different scanning devices and protocols) cohorts would be beneficial. Variations in CT scanning equipment, scanning protocols, and operator skill during image acquisition can lead to differences in image quality, resolution, and segmentation accuracy, which in turn may impact the study results. The age distribution of the participants in this study is limited, with an average of 65 ± 18 years, as

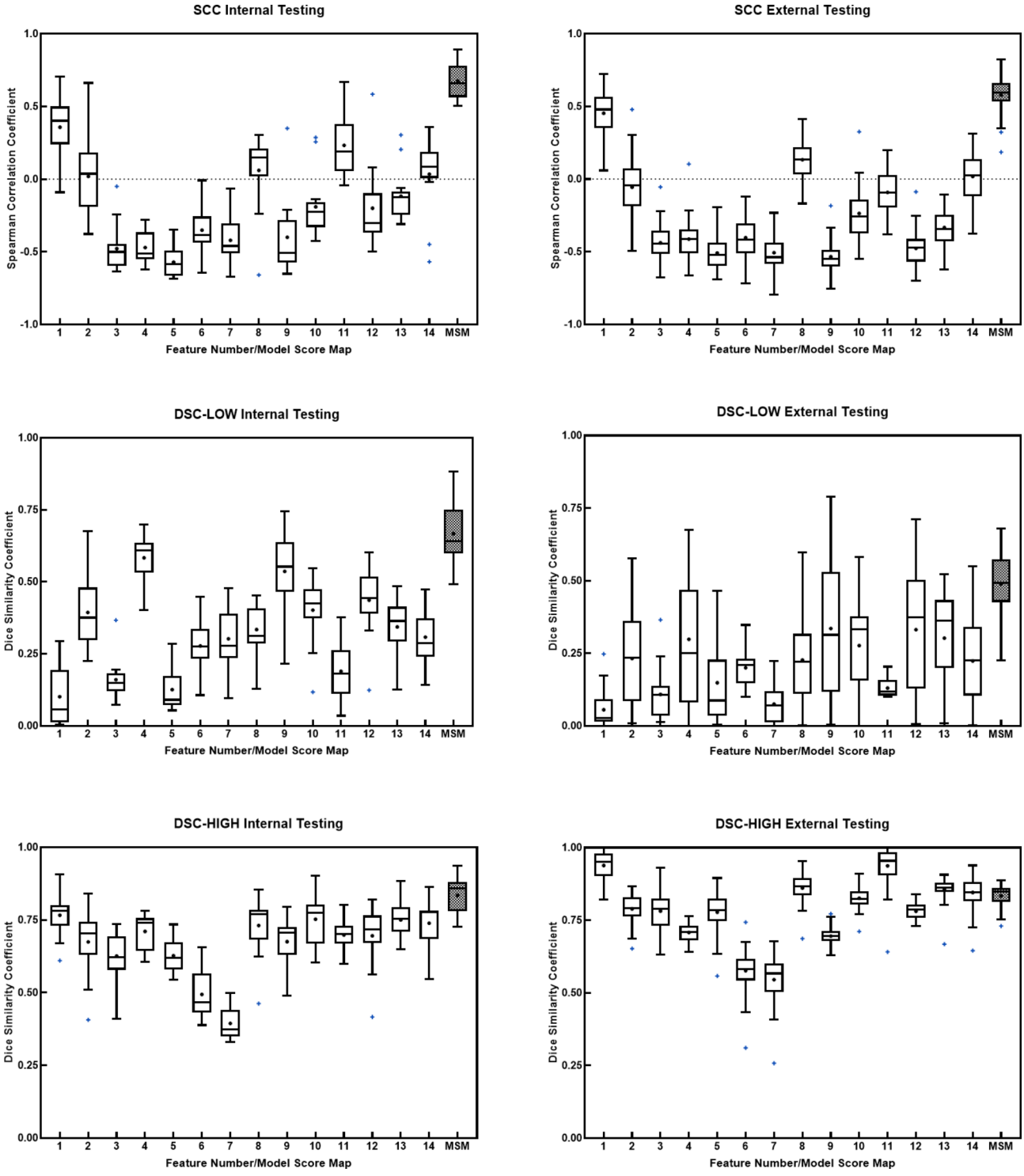

Fig. 8 Box plots of Spearman Correlation Coefficient and Dice Similarity Coefficient for internal and external testing cohorts

age-related physiological changes and the higher prevalence of lung disease in the elderly may affect the model's performance. Different racial and ethnic groups may vary in lung structure, function, and disease susceptibility, which could influence the generalizability of the model. The participants in this study lacked sufficient racial and ethnic diversity, and future studies should consider a wider range of groups to ensure the applicability of the model. In addition to the demographic limitations, geographic diversity was not adequately addressed in this

Table 3 Evaluation of feature and model-based spatial consistency across internal and external testing cohorts

Feature Number/Model	Internal Testing Cohort			External Testing Cohort		
	Median SCC	Median DSC_{LO}	Median DSC_{HI}	Median SCC	Median DSC_{LO}	Median DSC_{HI}
Model	0.66	0.64	0.86	0.60	0.49	0.85
1	0.40	0.06	0.78	0.48	0.03	0.95
2	0.04	0.38	0.70	-0.04	0.23	0.79
3	-0.50	0.15	0.62	-0.44	0.11	0.79
4	-0.51	0.61	0.74	-0.41	0.25	0.71
5	-0.59	0.09	0.62	-0.52	0.09	0.78
6	-0.38	0.28	0.47	-0.42	0.21	0.58
7	-0.46	0.28	0.37	-0.54	0.07	0.57
8	0.15	0.31	0.77	0.13	0.22	0.87
9	-0.51	0.55	0.71	-0.55	0.31	0.70
10	-0.23	0.43	0.77	-0.26	0.33	0.82
11	0.19	0.18	0.70	-0.09	0.12	0.95
12	-0.30	0.44	0.72	-0.47	0.37	0.79
13	-0.13	0.36	0.75	-0.35	0.36	0.86
14	0.08	0.29	0.78	0.03	0.23	0.85

Abbreviation: SCC, Spearman correlation coefficient; DSC, Dice similarity coefficient

study. Environmental factors such as air pollution, altitude, and other location-specific elements can affect lung health to varying degrees. Conducting evaluations across a wider range of geographic locations would enhance the model's robustness. Overall, future research should focus on these aspects of diversity to ensure the model's effectiveness and generalizability across diverse populations.

Additionally, when we attempted to replace the volume parameter settings for subregions generation with $26\times26\times26$ mm^3 (originally $21\times21\times21$ mm^3), the predictability on the internal testing cohort decreased to 0.79. The impact of the parameters on model performance indicates a need for further optimization of these settings. Future research directions also involve exploring the applicability of the method to pulmonary vascular diseases that affect perfusion other than PE discussed in this study.

Regarding the practical challenges of applying our classifier tool in clinical practice, we recognize that for any machine learning tool to be effectively implemented, it must be compatible with existing hospital information systems. This often requires relying on robust and secure APIs to interact with the clinical environment, ensuring seamless data transmission and system integration [38]. It also emphasizes the importance of designing user-friendly interfaces and providing comprehensive training to clinical staff, including mastering the technical operation of the tool and interpreting the model's outputs in a clinical setting to ensure the tool's effective use. Finally, the model needs to be validated on new datasets and continuously updated. External validation on data from diverse patient populations or institutions can further enhance the model's generalizability.

Furthermore, investigating the potential of the proposed method for longitudinal monitoring of changes in perfusion over multiple CT scans could provide valuable information for characterizing disease progression and response to therapy.

Conclusions

In conclusion, our study validated the feasibility of using quantitative texture analysis and a data-driven ML pipeline to generate voxel-wise lung perfusion surrogates across different institutions. We selected potential perfusion-correlated features to integrate into the model and found that the spatial distribution of the features and model output visualizations showed moderate to strong consistency with pulmonary functional imaging (i.e., perfusion SPECT imaging). Future collaboration within multidisciplinary teams could facilitate the non-invasive screening of PE and other perfusion-related lung diseases. Integrating our classifier into routine clinical practice could accelerate and optimize clinical decision-making.

Abbreviations

API	Application Programming Interface
AUC	Area Under the Curve
CT	Computed tomography
CTPA	Computed tomography pulmonary angiography
CTPI	Computed tomography perfusion imaging
4DCT	Four-dimensional computed tomography
CI	Confidence intervals
DIR	Deformable image registration
DL	Deep learning
DSC	Dice similarity coefficient
DSC_{HI}	Dice Similarity Coefficient for high functional region
DSC_{LO}	Dice Similarity Coefficient for low functional region
ES	Effect size
FM	Feature map
GLCM	Grey level co-occurrence matrix

GLDM	Grey level dependence matrix
GLRLM	Grey level run length matrix
GLSZM	Grey level size zone matrix
HU	Hounsfield unit
ICA	iodinated contrast agent
ICC	Intraclass correlation coefficient
LoG	Laplacian of Gaussian
LR	Logistic regression
ML	Machine learning
MSM	Model score map
NGTDM	Neighboring gray-tone difference matrix
NM	Nuclear medicine
PCC	Pearson correlation coefficient
ROI	Region-of-interest
ROC	Receiver operating characteristic
Q	Perfusion
SCC	Spearman correlation coefficient
SVM	Support vector machines
V/Q-SPECT	Ventilation/Perfusion single-photon emission computed tomography

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12931-024-03004-9.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
GR, ZL and YH conceived and designed research. MZ, XL, MW and KW collected data. ZL, YH and ZC interpreted results. ZL and YP prepared figures. ZL and YH drafted manuscript. GR, MY, JC, YP and MZ edited and revised manuscript. GR, LG, JC, WZ and RY provided supervision. All authors have contributed to the manuscript and approved the final version.

Funding
This work was supported in part by (1) the Health and Medical Research Fund (09200576) from the Health Bureau, the Government of the Hong Kong Special Administrative Region, (2) Shenzhen Science and Technology Program (JCYJ20230807140403007), (3) the Pneumoconiosis Compensation Fund Board in Hong Kong Special Administrative Region and (4) the Shenzhen Basic Research Program (JCYJ20210324130209023).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
The internal dataset was approved by the Institutional Review Board (IRB) of the University of Hong Kong/Hospital Authority Hong Kong West Cluster. Written informed consent was not required because of the retrospective nature of the study. The external dataset was approved by the Peking University Third Hospital Medical Science Research Ethics Committee (No. IRB00006761-M2024673). Informed written consent was obtained from all participants who agreed to take part in the study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR
[2]Department of Nuclear Medicine, Peking University Third Hospital, Beijing, China
[3]Department of Radiation Oncology, Peking University Third Hospital, Beijing, China
[4]School of Physics, Beihang University, Beijing, China
[5]The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China

Received: 6 August 2024 / Accepted: 4 October 2024
Published online: 28 October 2024

References

1. Goldhaber SZ, Bounameaux H. Pulmonary embolism and deep vein thrombosis. Lancet. 2012;379(9828):1835–46.
2. Di Nisio M, van Es N, Büller HR. Deep vein thrombosis and pulmonary embolism. Lancet. 2016;388(10063):3060–73.
3. Pantaleo G, et al. Amniotic fluid embolism: review. Curr Pharm Biotechnol. 2013;14(14):1163–7.
4. Kucher N, et al. Massive pulmonary embolism. Circulation. 2006;113(4):577–82.
5. Rahimtoola A, Bergin JD. Acute pulmonary embolism: an update on diagnosis and management. Curr Probl Cardiol. 2005;30(2):61–114.
6. Sung YK, Kline JA. Unchanging mortality from Pulmonary Embolism in the United States. Ann Am Thorac Soc. 2023;20(11):1554–6.
7. Squizzato A, Galli L, Gerdes VEA. Point-of-care ultrasound in the diagnosis of pulmonary embolism. Crit Ultrasound J. 2015;7(1):7.
8. Beckett KR, Moriarity AK, Langer JM. Safe use of contrast media: what the Radiologist needs to know. Radiographics. 2015;35(6):1738–50.
9. Moore AJE, et al. Imaging of acute pulmonary embolism: an update. Cardiovasc Diagnosis Therapy. 2017;8(3):225–43.
10. Konstantinides SV, et al. 2019 ESC guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS). Eur Heart J. 2020;41(4):543–603.
11. Hansen SL, et al. Ventilation–perfusion SPECT < em > versus CTPA in young adult females with suspected pulmonary embolism. Eur Respir J. 2020;55(6):2000448.
12. Castillo E, et al. Quantifying pulmonary perfusion from noncontrast computed tomography. Med Phys. 2021;48(4):1804–14.
13. Porter EM, et al. Synthetic pulmonary perfusion images from 4DCT for functional avoidance using deep learning. Phys Med Biol. 2021;66(17):175005.
14. Ren G, et al. Investigation of a Novel Deep Learning-based computed Tomography Perfusion Mapping Framework for Functional Lung Avoidance Radiotherapy. Front Oncol. 2021;11:644703.
15. Ren G, et al. Deep learning-based computed Tomography Perfusion Mapping (DL-CTPM) for pulmonary CT-to-perfusion translation. Int J Radiat Oncol Biol Phys. 2021;110(5):1508–18.
16. Mayerhoefer ME, et al. Introduction to Radiomics. J Nucl Med. 2020;61(4):488–95.
17. Kocak B, et al. Trends and statistics of artificial intelligence and radiomics research in Radiology, Nuclear Medicine, and Medical Imaging: bibliometric analysis. Eur Radiol. 2023;33(11):7542–55.
18. Hofmanninger J, et al. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp. 2020;4(1):50.
19. Achanta R, et al. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell. 2012;34(11):2274–82.
20. Parker JA, et al. SNM practice guideline for lung scintigraphy 4.0. J Nucl Med Technol. 2012;40(1):57–65.
21. Faught AM, et al. Evaluating the toxicity reduction with computed Tomographic Ventilation Functional Avoidance Radiation Therapy. Int J Radiat Oncol Biol Phys. 2017;99(2):325–33.
22. Zwanenburg A, et al. The image Biomarker Standardization Initiative: standardized quantitative Radiomics for High-Throughput Image-based phenotyping. Radiology. 2020;295(2):328–38.
23. Teng X, et al. Building reliable radiomic models using image perturbation. Sci Rep. 2022;12(1):10035.
24. Zwanenburg A, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep. 2019;9(1):614.
25. Ram A, et al. A density based Algorithm for discovering density varied clusters in large spatial databases. Int J Comput Appl. 2010;3(6):1–4.

26. Huang YH, et al. Respiratory invariant textures from static computed tomography scans for explainable lung function characterization. J Thorac Imaging. 2023;38(5):286–96.
27. Kerby DS. The simple difference formula: an Approach to Teaching nonparametric correlation. Compr Psychol, 2014. 3.
28. Ostroumova L et al. CatBoost: unbiased boosting with categorical features. in Neural Inform Process Syst. 2017.
29. Shepard DS. *A two-dimensional interpolation function for irregularly-spaced data.* Proceedings of the 1968 23rd ACM national conference, 1968: pp. 517–524.
30. Levin D. The approximation power of moving least-squares. Math Comput. 1998;67(224):1517–31.
31. Kiessling F. The changing face of cancer diagnosis: from computational image analysis to systems biology. Eur Radiol. 2018;28(8):3160–4.
32. Capobianco E, Deng J. Radiomics at a glance: a few lessons learned from learning approaches. Cancers. 2020;12(9):2453.
33. Huang Y, et al. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non—small cell lung cancer. Radiology. 2016;281(3):947–57.
34. Ji G-W, et al. Biliary tract cancer at CT: a radiomics-based model to predict lymph node metastasis and survival outcomes. Radiology. 2019;290(1):90–8.
35. Cui Y, et al. Radiomics analysis of multiparametric MRI for prediction of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. Eur Radiol. 2019;29:1211–20.
36. Zhang Z, Ho KM, Hong Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. Crit Care. 2019;23:1–10.
37. Torbati HM. Machine learning and texture analysis of [18F] FDG PET/CT images for the prediction of distant metastases in Non-small-cell Lung Cancer patients. Biomedicines. 2024;12(3):472.
38. Balch JA, et al. Machine learning–enabled clinical information systems using fast healthcare interoperability resources data standards: scoping review. JMIR Med Inf. 2023;11:e48297.

Publisher's note