Biomedical Physics &
Engineering Express

Endorsed by IPEM Institute of Physics and Engineering in Medicine
PURPOSE-LED PUBLISHING™

PAPER • **OPEN ACCESS**

Multisequence MRI-driven assessment of PD-L1 expression in non-small cell lung cancer: a pilot study

To cite this article: Agnese Robustelli Test *et al* 2026 *Biomed. Phys. Eng. Express* **12** 015019

View the article online for updates and enhancements.

You may also like

- Insights from a discrete generalized beta distribution analysis of heart rate and blood pressure variability: an integrated approach to study end-stage renal disease
Alejandro Aguado-García, Claudia Lerma, Juan C Echeverría et al.
- A simple yet effective microfluidic device for the *in-situ* formation of uniform-sized cell-laden microgels
Hajar Mohamadzade Sani, Seyed Mostafa Hosseinalipour, Sarah Salehi et al.
- Biomaterials to biofabrication: advanced scaffold technologies for regenerative endodontics
Arun Mayya, Akshatha Chatra, Vinita Dsouza et al.

This content was downloaded from IP address 2.36.99.217 on 13/04/2026 at 15:21

Biomedical Physics & Engineering Express

Check for updates

OPEN ACCESS

RECEIVED
15 July 2025

REVISED
13 October 2025

ACCEPTED FOR PUBLICATION
1 December 2025

PUBLISHED
11 December 2025

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

Multisequence MRI-driven assessment of PD-L1 expression in non-small cell lung cancer: a pilot study

Agnese Robustelli Test[1,2], **Chandra Bortolotto**[3,4], **Sithin Thulasi Seetha**[3,5,*], **Alessandra Marrocco**[3], **Carlotta Pairazzi**[3], **Gaia Messana**[3], **Leonardo Brizzi**[3], **Domenico Zacà**[6], **Robert Grimm**[7], **Francesca Brero**[1,2], **Manuel Mariani**[1,8], **Raffaella Fiamma Cabini**[9], **Giulia Maria Stella**[10,11], **Lorenzo Preda**[3,4,†] and **Alessandro Lascialfari**[1,2,†]

[1] Physics Department, University of Pavia, Pavia, Italy
[2] (INFN) National Institute for Nuclear Physics,Unit of Pavia, Pavia, Italy
[3] Diagnostic Imaging Unit, Department of Clinical, Surgical, Diagnostic, and Pediatric Sciences, University of Pavia, Pavia, Italy
[4] Radiology Institute, IRCCS Polyclinic San Matteo, Pavia, Italy
[5] Clinical Department, (CNAO) National Center for Oncological Hadrontherapy, Pavia, Italy
[6] Siemens Healthcare srl, Milan, Italy
[7] Siemens Healthineers AG, Erlangen, Germany
[8] (INFN) National Institute for Nuclear Physics, Unit of Milan, Milan, Italy
[9] (USI) University of Lugano, Lugano, Switzerland
[10] Department of Internal Medicine and Medical Therapeutics, University of Pavia, Pavia, Italy
[11] Unit of Respiratory Disease, Cardiothoracic and Vascular Department, IRCCS Polyclinic San Matteo, Pavia, Italy
[†] Preda L and Lascialfari A contributed equally to this work.
[*] Author to whom any correspondence should be addressed.

E-mail: sithin.thulasiseetha@unipv.it

Keywords: integrated radiomics–DL, diffusion MRI, immunotherapy biomarkers

Supplementary material for this article is available online

Abstract

Objective. Lung cancer remains the leading cause of cancer-related mortality worldwide, with Non-Small Cell Lung Cancer (NSCLC) accounting for approximately 85% of all cases. Programmed cell Death Ligand-1 (PD-L1) is a well-established biomarker that guides immunotherapy in advanced-stage NSCLC, currently evaluated via invasive biopsy procedures. This study aims to develop and validate a non-invasive pipeline for stratifying PD-L1 expression using quantitative analysis of IVIM parameter maps—diffusion (D), pseudo-diffusion (D^*), perfusion fraction (pf)—and T1-VIBE MRI acquisitions. *Approach.* MRI data from 43 NSCLC patients were analysed and labelled as PD-L1 positive ($\geqslant$1%) or negative ($<$1%) based on immunohistochemistry exam. After pre-processing, 1,171 radiomic features and 512 deep learning features were obtained. Three feature sets (radiomic, deep learning, and fusion) were tested with Logistic Regression, Random Forest, and XGBoost. Four discriminative features were selected using the Mann–Whitney U-test, and model performance was primarily assessed using the area under the receiver operating characteristic curve (AUC). Robustness was ensured through repeated stratified 5-fold cross-validation, bootstrap-derived confidence intervals, and permutation test. *Main Results.* Logistic Regression generally demonstrated the highest classification performance, with AUC values ranging from 0.78 to 0.92 across all feature sets. Fusion models outperformed or matched the performance of the best standalone radiomics or deep learning model. Among multisequence MRI, the IVIM-D fusion features yielded the best performance with an AUC of 0.92, followed by IVIM-D^* radiomic features that showed a similar AUC of 0.91. For IVIM-pf and T1-VIBE derived features, the fusion model yielded the best AUC values of 0.87 and 0.90, respectively. *Significance.* The obtained results highlight the potential of a combined radiomic-deep learning approach to effectively detect PD-L1 expression from MRI acquisitions, paving the way for a non-invasive PD-L1 evaluation procedure.

© 2025 The Author(s). Published by IOP Publishing Ltd

1. Introduction

Lung cancer accounts for the highest burden of cancer-related mortality [1, 2], with Non-Small Cell Lung Cancer (NSCLC) representing 85% of all cases [3]. In advanced-stage NSCLC patients, immunotherapies targeting Programmed cell Death Ligand-1 (PD-L1) have shown high therapeutic efficacy [4–6] in patients with positive PD-L1 expression [7]. Given the critical role of PD-L1 status in treatment response, an accurate immuno-molecular profile capable of predicting the tumour's response both at diagnosis and later when shifting between lines of therapy is of paramount importance [5]. In clinical practice, the Tumour Proportion Score (TPS) is the most used method to assess PD-L1 expression: if TPS $\geqslant$ 1%, patients are considered to have a higher likelihood of benefiting from immunotherapies. Currently, biopsy specimens or surgically resected tissue are typically employed for PD-L1 detection; such techniques are invasive and fail to capture the dynamic changes and the intra-tumour heterogeneity of PD-L1 expression [8]. Developing a non-invasive tool capable of dynamically assessing PD-L1 expression across the entire tumour could enhance patient survival outcomes by enabling the early selection of the most appropriate treatment. In this direction, in the last decades, the development of new data analysis methods, such as radiomics, Machine Learning (ML), and Deep Learning (DL), has paved the way for new approaches to quantify PD-L1 expression through medical images [9–11]. Radiomics and ML/DL have been mainly applied to Computed Tomography (CT) images, especially for lung cancer staging and grading [12–14]. However, while CT is the gold standard for lung investigations, it is not devoid of risks and limitations, the most significant being radiation exposure [15]. For these reasons, Magnetic Resonance Imaging (MRI) may represent a radiation-free alternative better suited to fully harness the potential of image analysis techniques.

Historically, lung MRI acquisitions showed poorer quality compared to CT, due to low proton density, susceptibility artifacts, cardiac and respiratory motion, and the lack of standard protocols [16]. Over the past decades, experience with the clinical use of lung MRI has grown, leading to standardized protocols and higher image quality [17–19], and demonstrating that specific MRI sequences, such as Volumetric-Interpolated-Breath-hold-Examination (VIBE) and Ultrashort-echo-TimE (UTE), can prospectively achieve performance comparable to CT [20–22]. Recently, MRI and Diffusion Weighted Imaging (DWI) have been proposed as potential tools for nodule characterization [23], and the correlation between PD-L1 expression and Intravoxel Incoherent Motion (IVIM)-DWI data has been statistically established [24]. IVIM is an advanced MRI technique that has the potential to differentiate true diffusion from microcirculation-related perfusion without requiring the use of contrast agents. IVIM sequences can be used to quantify the pure diffusion coefficient (IVIM-D), which reflects the true diffusion of water molecules in the extracellular space; the pseudo-diffusion coefficient (IVIM-D*), which captures the incoherent motion of water molecules within the capillary network; and the perfusion fraction coefficient (IVIM-pf), which indicates tissue perfusion [25–27].

Previous studies predicted PD-L1 expression using radiomic and DL models mainly from PET/CT or CT [28–31]. Contrast-enhanced MRI has been used in other anatomical districts, such as breast and liver, to pursue the same goal, achieving promising results with AUCs of 0.90 [32] in breast cancer and 0.81 [33] in hepatocellular carcinoma.

In the present study, we propose a non-invasive method to classify NSCLC patients as PD-L1 positive or negative by extracting radiomic and deep learning features from MRI images, considering IVIM parameter maps and pre-contrast T1-VIBE acquisitions. We aim to evaluate MRI's ability to characterize PD-L1 expressions in lung cancer, compare the performance of radiomic, deep learning, and fusion approaches, clarify the currently undefined role of non-enhanced MRI in this context, and develop a statistically reliable pipeline able to provide robust results in the context of a relatively small dataset.

2. Methods

2.1. Patient characteristics and clinical outcomes

We conducted a prospective study, enrolling a consecutive group of patients diagnosed with NSCLC (T $\geqslant$ 2, according to the TNM classification VIII edition). A total of 43 patients were included. Among them, 32 (74%) were males, and 11 (26%) were females. All the patients were of Caucasian ethnicity, aged between 50 and 86, with an average age of 69. There were 17 patients (39%) with lung adenocarcinoma, 14 (33%) with lung squamocellular carcinoma, and 12 patients (28%) with poorly differentiated NSCLC. Among these, 27 (63%) had a positive PD-L1 expression (TPS $\geqslant$ 1%), while 16 (37%) had no PD-L1 expression (TPS $<$ 1%).

Eligibility for inclusion required prior completion of tumour staging, pathological characterization, and immunohistochemical analysis to determine PD-L1 expression. Exclusion criteria for the study were established as follows: (i) insufficient compliance or presence of contraindications that prevented the execution of the MRI scan; (ii) any treatment prior to the MRI scan; (iii) lack of testing for PD-L1 expression. The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the hospital. Written informed consent was obtained from all participants.

PD-L1 expression was quantified through the 22C3 pharmDx assay (Agilent Technologies), and TPS was evaluated according to guidelines [34]. Based on immuno-histochemistry results, the tumours were categorized into two groups: PD-L1 positive (TPS $\geqslant$ 1%) and PD-L1 negative (TPS $<$ 1%).

2.2. MRI acquisition protocol

All patients underwent bilateral MRI lung examination, performed on a 1.5-Tesla scanner (MAGNETOM Aera; Siemens Healthineers, Forchheim, Germany) with an 18-channel surface coil. Scan sequences, on the axial plane, included pre-contrast T1-VIBE sequences and free-breathing IVIM-DWI. For the axial T1-VIBE scan, the parameters were: matrix 320 $\times$ 320, Repetition Time (TR) 2.12 ms, Echo Time (TE) 0.73 ms, Field Of View (FOV) 450 mm, slice thickness 2.5 mm, layer spacing 0 mm, number of slices 60. For axial IVIM-DWI, the parameters were: matrix 160 $\times$ 160, b values 0, 10, 20, 30, 50, 70, 100, 150, 200, 400, and 800 s mm^{-2}, TR 4800 ms, TE 64 ms, FOV 400 mm, slice thickness 6 mm, interval 1.2 mm, and number of slices 30.

2.3. Image analysis and segmentation

The IVIM-DWI sequence was analysed and post-processed using a prototype research software (MR Body Diffusion Toolbox V1.3.0., Siemens Healthineers) to extract parametric maps through bi-exponential analysis. MRI visible tumours were segmented manually by an experienced radiologist on the high-resolution pre-contrast T1-VIBE sequence. The resulting segmentation masks were propagated to the IVIM-DWI images using SimpleITK's deformable B-spline registration method. The toolbox-derived IVIM parametric maps were inherently aligned with the IVIM scan; hence, no further co-registration was necessary. Finally, any registration errors were corrected manually, if required.

Radiomic features are mainly affected by two image properties: voxel dimensions and grey-scale range. Images should first be resampled to a common voxel size, and the grey-level range needs to be normalized to offset the heterogeneity within the dataset. T1-VIBE images exhibit higher resolution with respect to the IVIM ones, with voxel dimensions of 1.4 $\times$ 1.4 $\times$ 2.5 mm^3 compared to 2.5 $\times$ 2.5 $\times$ 7.2 mm^3 for IVIM. Currently, there are no definitive guidelines on whether upsampling or downsampling is the preferred approach, as both have their advantages and drawbacks [35]. To ensure isotropic voxels for feature extraction, all images were resampled to 3.5 $\times$ 3.5 $\times$ 3.5 mm^3. This value was chosen as a compromise between minimizing information loss while downsampling the transverse plane of both sequences and reducing the impact of interpolation artifacts when upsampling the IVIM acquisition along the anterior-posterior axis. As medical images typically show in-plane resolution higher than the axial one, B-spline interpolation was employed for in-plane resampling and nearest neighbour interpolation for the axial plane [36]. As suggested in the literature, the Z-score method was chosen to standardize intensity values [37]. For IVIM, an absolute approach was applied, considering the mean and standard deviation calculated across all patients, as IVIM parametric maps quantify various physiological characteristics of the tissue, and such a normalization would preserve this global information. For T1-VIBE, a patient-specific approach was adopted, normalizing each image using the mean and standard deviation derived from the individual patient. Of note, in both cases, only grey-level intensities that fell within tumour voxels were considered to estimate the mean and standard deviation.

2.4. Handcrafted radiomic and deep learning feature extraction

From each of the considered MRI inputs—IVIM-D, IVIM-D*, IVIM-pf, and T1-VIBE—a total of 1,171 handcrafted radiomic features were extracted from the pre-processed images using the latest version of *PyRadiomics* (3.1.0) [38]. These features are primarily divided into four categories: (i) shape, describing the geometric properties of lesions; (ii) first order, based on the intensity histogram; (iii) texture, characterizing the relationship between a voxel and its neighbours; (iv) filtered-features, highlighting specific image characteristics enhanced through filters [39]. Specifically, 103 original features were analyzed (14 shape, 17 first order, 72 texture), mostly compliant with the Image Biomarker Standardization Initiative (IBSI) [35], 267 Laplacian of Gaussian-filtered features, 712 wavelet-filtered features, and 89 gradient-filtered features. All features were extracted using a fixed bin count, discretizing grey levels in 64 bins.

In the case of deep learning, 512 features were extracted from the global average pooling layer (before the classification head) of the pre-trained *ResNet18* model (see supporting information, section 1 for implementation details) for each input image. ResNet18 is a fully convolutional neural network (CNN), which was designed to address the vanishing gradient problem using residual connections (also known as skip connections) in the *ImageNet* competition [40, 41]. This model expects the inputs to be 2D RGB images [42]. However, lung MRI images are 3D volumes containing 2D axial slices of a target region in grey scale. For each MRI input, to obtain a representative 2D slice, the maximum intensity projection (MIP) was computed across the tumour-focalized slices. The MIP projects the highest intensity values along the axial direction onto a 2D plane, commonly used in medical imaging to enhance the visibility of high-intensity structures, such as tumours or blood vessels. The MIPs were normalized in a

workflow described in section 2.2 (i.e., global or local Z-score normalization based on the MRI sequence). Outliers were eliminated in an IBSI-compliant manner to fall between [−3, 3] [43]. To reduce the entropy, the non-tumour regions were masked. The normalized intensity values were then rescaled to be in the range of 0–255, which was duplicated in all three channels to construct an RGB image to be fed as input to the pre-trained ResNet18 model.

2.5. Feature selection

To prevent overfitting, four features were selected from both deep learning and radiomic features derived from the IVIM maps and the T1-VIBE acquisition. The features selection process involved two steps: (i) identification of the significant features through the Mann–Whitney U-test, a non-parametric statistical test, applied to detect the significance of the difference in features associated with PD-L1 expression; this test compares the distributions of each feature across two independent groups defined by the target variable, identifying those with statistically significant differences (p-value < 0.05); (ii) refinement of the features selected in the first step by picking the four features showing minimal inter-feature correlation and stronger association with the target variable. This refinement process utilized the Spearman correlation coefficient alongside the initial ranking to ensure that the selected features were both independent and relevant. Before building the classification models, feature values were normalized between zero and one.

2.6. Classification models

To predict the PD-L1 expression from the selected features, three different classification methods were employed: Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGB). LR is a linear, interpretable classifier that models the relationship between input features and the target class probability using a logistic function. RF is an ensemble method that combines multiple decision trees to improve predictive accuracy and robustness compared to individual trees. XGB is an advanced ensemble method based on boosting that sequentially builds decision trees to optimize performance by minimizing a specified loss function, known for its speed, scalability, and ability to capture complex patterns in data. For each of the mentioned classification methods, features selected from three different approaches were considered: radiomic features, deep learning features, and a mixed approach combining both deep learning and radiomic features.

2.7. Performance analysis

Models were evaluated through several criteria: area under the receiver-operating characteristic curve (AUC), accuracy, precision, recall, specificity, f1-score, and confusion matrix. Each metric provides specific insights on the model performance: (i) *AUC* is a measure of a classifier's ability to distinguish between classes, and represents the trade-off between True Positive (TP) rate and False Positive (FP) rate; (ii) *accuracy* is the ratio of correctly predicted instances to the total instances; (iii) *precision* quantifies how often a model properly estimates the positive class, i.e., dividing TP by the total number of positive predictions; (iv) *recall* is the ratio of TP to the total number of actual positive instances; (v) *specificity* relates with the models ability to identify negative cases, (vi) *f1-score* is the harmonic mean of precision and recall; (vii) the *confusion matrix* provides a visual representation of model's performance relating predicted and actual values for all classes [44]. Each parameter reflects the mean value obtained from 100 repetitions of stratified 5-fold cross-validation to ensure robustness against result variability; value is reported alongside the associated standard deviation. Moreover, as AUC is the most popular performance metric in classification tasks, the confidence intervals were estimated for each of the obtained AUCs through bootstrap analysis, and their statistical significance was assessed using a permutation test, where the null distribution of the AUCs was obtained by shuffling the labels in the target variable.

3. Results

The potential of MRI—including T1-VIBE and IVIM-derived parametric maps (D, D*, pf)—for predicting PD-L1 expression was assessed using three methodological approaches: radiomics, deep learning, and the fusion of both. Each feature set was evaluated using three machine learning classifiers: LR, RF, and XGB. Table 1 summarizes the mean cross-validated AUC (the primary metric of merit), along with their corresponding 95% confidence intervals and permutation-test-based p-values, for all classifier-feature combinations per sequence. The ROC curves are illustrated in figure 1.

The radiomic-only models demonstrated performance metrics within the following ranges: AUC [0.72, 0.91], accuracy [0.65, 0.80], precision [0.79, 0.94], recall [0.54, 0.75], specificity [0.75, 0.90], and f1-score [0.61, 0.82]. Models based on deep learning-derived features showed slightly improved results: AUC [0.76, 0.90], accuracy [0.68, 0.80], precision [0.83, 0.95], recall [0.59, 0.75], specificity [0.79, 0.95], and f1-score [0.67, 0.82]. Finally, the fusion features achieved the best overall performance, with metrics ranging as follows: AUC [0.82, 0.92], accuracy [0.72, 0.80], precision [0.89, 0.96], recall [0.64, 0.75], specificity [0.87, 0.94], and f1-score [0.72, 0.82]. Overall, these results reveal a consistent trend of higher specificity and precision compared to recall, suggesting a tendency toward more conservative yet correct

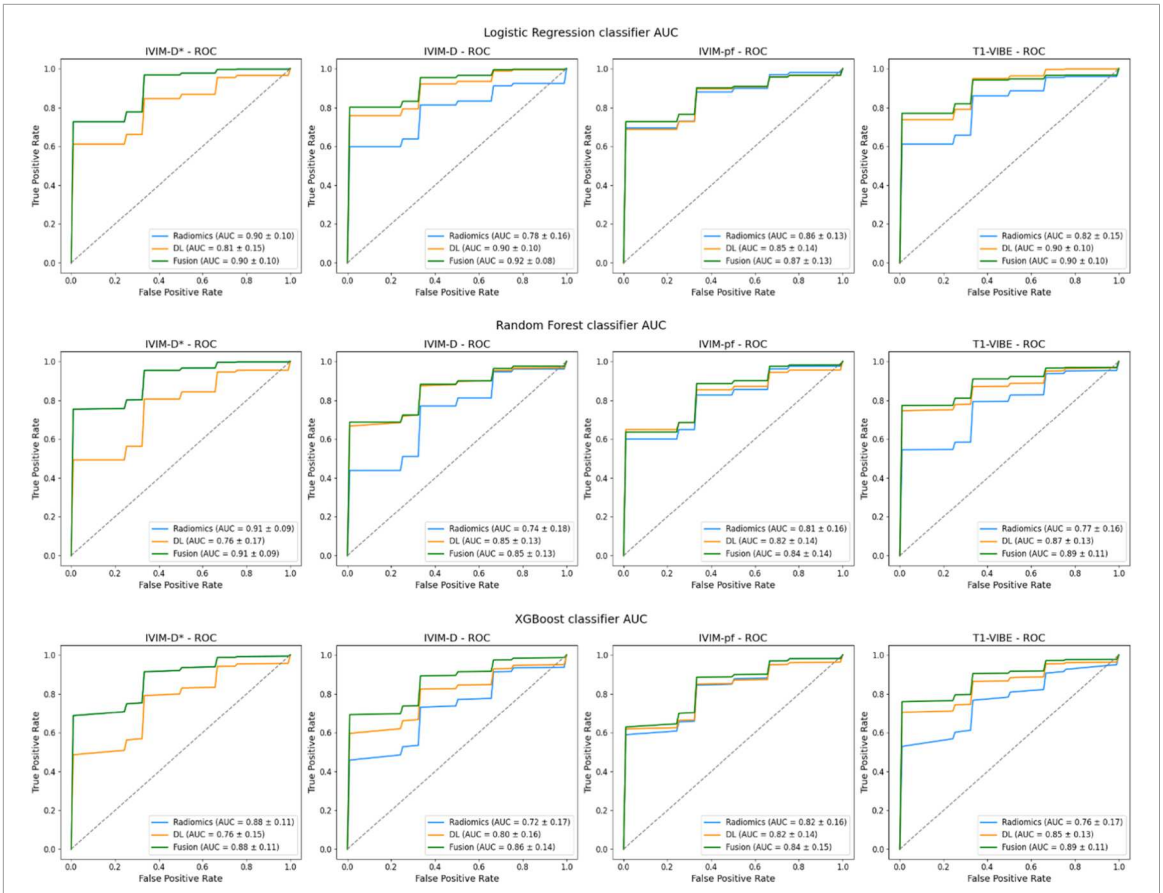


Figure 1. Mean cross-validation AUC ± standard deviation for all models, grouped by classifiers. Each of the four panels refers to one of the analysed MRI inputs, showing the AUC of the Radiomic-, Deep Learning-, and Fusion models.

Table 1. Mean cross-validation AUC, 95%confidence intervals, and p-value for all the analysed combinations of MRI input, classifiers, and features configuration approaches.

		Radiomics			Deep learning			Fusion		
MRI input	Model	AUC mean	AUC range	p-value	AUC mean	AUC range	p-value	AUC mean	AUC range	p-value
IVIM-D*	LR	0.90	[0.74; 0.98]	<0.0001	0.81	[0.70; 0.88]	0.0031	0.90	[0.74; 0.98]	<0.0001
	RF	0.91	[0.81; 0.98]	<0.0001	0.76	[0.63; 0.84]	0.0148	0.91	[0.82; 0.98]	<0.0001
	XGB	0.88	[0.78; 0.95]	<0.0001	0.76	[0.64; 0.84]	0.0128	0.88	[0.79; 0.96]	0.0001
IVIM-D	LR	0.78	[0.57; 0.90]	0.0043	0.90	[0.77; 0.98]	0.001	0.92	[0.83; 1.00]	<0.0001
	RF	0.74	[0.61; 0.85]	0.0212	0.85	[0.75; 0.88]	0.0036	0.85	[0.73; 0.90]	<0.0001
	XGB	0.72	[0.57; 0.79]	0.0403	0.80	[0.64; 0.84]	0.0115	0.86	[0.71; 0.91]	0.0013
IVIM-pf	LR	0.86	[0.68; 0.95]	0.0009	0.85	[0.61; 0.92]	0.0011	0.87	[0.70; 0.94]	0.0002
	RF	0.81	[0.71; 0.87]	0.0034	0.82	[0.77; 0.90]	0.0006	0.84	[0.76; 0.93]	0.0015
	XGB	0.82	[0.73; 0.89]	0.0023	0.82	[0.74; 0.90]	0.0015	0.84	[0.74; 0.91]	0.0015
T1-VIBE	LR	0.82	[0.63; 0.94]	0.0008	0.90	[0.78; 0.98]	<0.0001	0.90	[0.76; 0.98]	<0.0001
	RF	0.77	[0.69; 0.84]	0.0088	0.87	[0.79; 0.91]	0.0003	0.89	[0.83; 0.94]	0.0006
	XGB	0.76	[0.62; 0.86]	0.0114	0.85	[0.76; 0.91]	0.003	0.89	[0.82; 0.94]	0.0002

positive predictions. Moreover, IVIM-pf emerged as the most stable MRI map, as indicated by the relatively low fluctuations in evaluation scores shown in figure 2. Table S1 of the supporting information contains the values of the considered performance metrics for each combination of MRI input, classifier, and feature configuration.

Fusion models either outperformed or matched the AUC values of radiomic and deep learning standalone approaches across all MRI modalities and classifiers, as shown in figure 1s of the supporting information, presenting AUC values higher than 0.82. The performance gains obtained with the fusion approach were particularly notable compared with

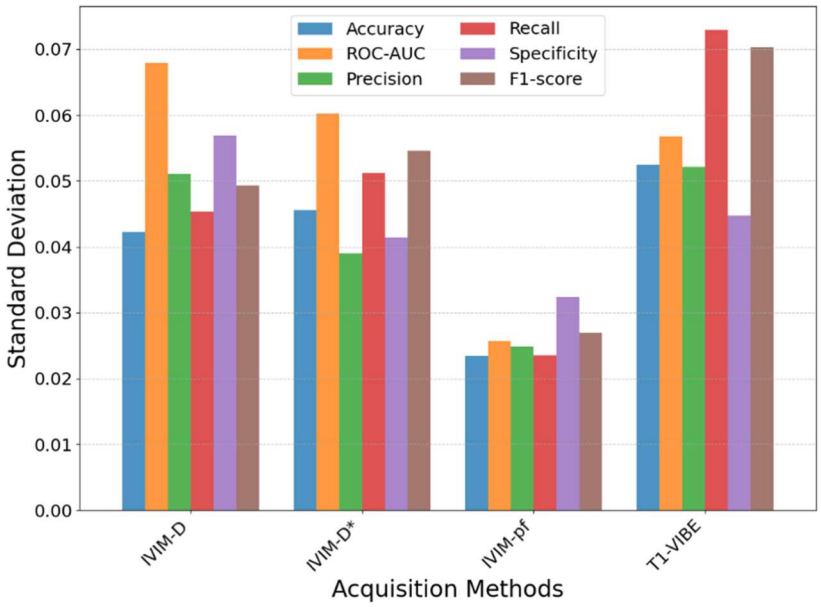


Figure 2. Stability of the model's evaluation metrics across the considered classifiers and approaches for each analysed MRI acquisition. The MRI metric showing more stability is the IVIM-pf.

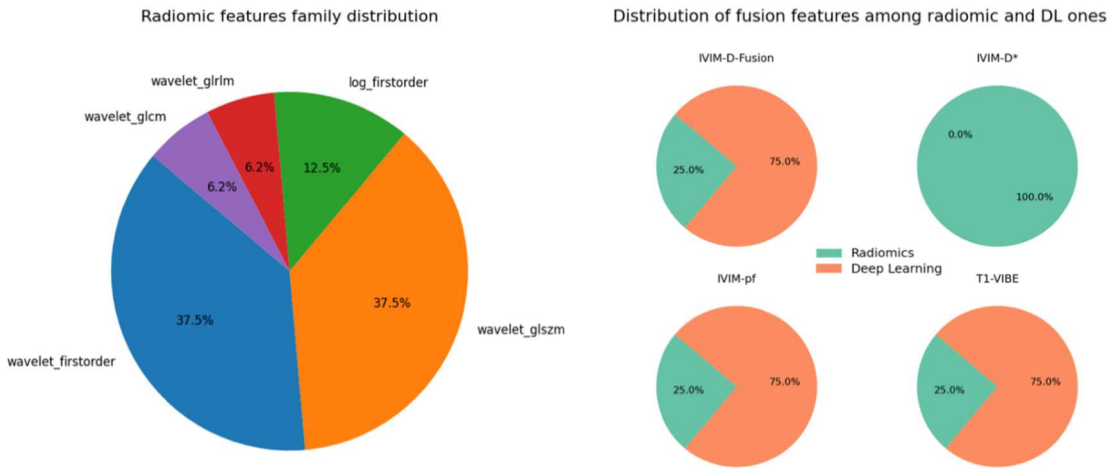


Figure 3. Selected features from the radiomic approach (left) and fusion approach (right). Radiomic features are exclusively filtered ones, derived from wavelet and Laplacian of Gaussian (LoG) transformations, first order and texture classes. In the fusion models, three out of four features are obtained from deep learning approach, with IVIM-D* being the only case composed solely of radiomic features.

radiomic ones in IVIM-D (15%–19%) and T1-VIBE (10%–17%), whereas improvements over deep learning were modest (<8% and <5%, respectively). For IVIM-pf, enhancements were limited across both approaches (<4%), reflecting IVIM-pf stability previously discussed and reported in figure 2. In contrast, for IVIM-D*, the radiomic model matched the fusion performance, while the fusion approach showed a considerable advantage over the deep learning one (11%–20%).

Across all feature types and image modalities, the Logistic Regression (LR) model generally provided the most consistent and highest AUC values, ranging from 0.78 to 0.92. Focusing on LR model, the highest AUC was observed for the fusion approach with IVIM-D features (AUC = 0.92, 95% CI: [0.83, 1.00], $p < 0.0001$), followed by T1-VIBE (0.90, [0.76, 0.98], $p < 0.0001$) and IVIM-D* (0.90, [0.74, 0.98], $p < 0.0001$). Among standalone features, radiomic outperformed deep learning features for IVIM-D* (0.90, [0.74, 0.98], $p < 0.0001$) and IVIM-pf (0.86, [0.68, 0.95], $p = 0.0009$), whereas deep learning features performed better for IVIM-D (0.90, [0.77, 0.98], $p = 0.0001$) and T1-VIBE (0.90, [0.78, 0.98], $p < 0.0001$). A visual representation of the AUC values obtained by LR for radiomic and deep learning approaches is depicted in figure 2s of the supporting information.

Figure 3 presents the signatures selected across different MRI modalities for the radiomic approach and the distribution among radiomic and deep learning features for the fusion approach. Within radiomic features, wavelet-derived first-order and texture features were the most frequently selected. Inside the texture families, GLSZM features were the most common, followed by GLRLM and GLCM. For IVIM-D*, the LoG-filtered first-order features were also among the selected features. In the fusion model, the feature composition varied by modality. In fact, for IVIM-D*, all selected features originated from the radiomic set, whereas for all other sequences, three out of the four selected features were deep learning features.

4. Discussion

Recently, immune checkpoint inhibitors targeting PD-L1 have emerged as a cornerstone of first-line therapy for advanced NSCLC [45]. Currently, PD-L1 expression is assessed through biopsies, which, while informative, are invasive and provide only a limited representation of the tumour heterogeneity. In this study, the role of lung MRI radiomics for non-invasive quantification of PD-L1 expression was explored, considering radiomics and deep learning features separately and their combination.

Previous studies assessed PD-L1 expression in lung cancer, primarily focusing on PET/CT images. Wei Mu *et al* used a CNN based on PET/CT images to predict PD-L1 expression in NSCLC patients using TPS $\geqslant$ 1% as the threshold [28], and reported an AUC of 0.82. Similarly, Panwen Tian *et al* designed a deep CNN, considering TPS cut-off $\geqslant$ 50%, which achieved an AUC of 0.76 [29]. Chengdi Wang *et al* focused on CT acquisitions to predict TPS level using radiomics , deep-learning, and their combination, with which they obtained AUCs $\geqslant$ 0.93 [30]. Similarly, our best-performing model (IVIM-D fusion approach) yielded an AUC of 0.92, aligning well with these previously reported results.

First-order features extracted from pharmacokinetic parameter maps–K^{trans} and K_{ep} derived from dynamic contrast-enhanced (DCE) MRI, have also been shown to discriminate between hypo- and hyper-expression groups (TPS $\geqslant$ 50%) with an AUC ranging from 0.71 to 0.76 [46]. However, DCE MRI requires the use of a contrast agent, which may not be suitable for all patients, whereas IVIM and T1-VIBE-derived features offer contrast-free alternatives. Although the role of advanced quantitative analysis using multisequence MRI radiomics for PD-L1 prediction largely remains unexplored, the mean ROI value extracted from IVIM-D* has previously demonstrated statistical significance ($p = 0.008$) in distinguishing PD-L1-negative from PD-L1-positive (TPS $\geqslant$ 1%) expression in NSCLC patients [24], while IVIM-D showed similar significance in cervical cancer ($p = 0.021$) [47]. Building upon this evidence, our study is, to the best of our knowledge, the first to investigate the predictive value of IVIM-D for PD-L1 expression in NSCLC.

Across modalities, the most frequently chosen radiomic features belonged to Laplacian of Gaussian (LoG)- and wavelet-filtered features, particularly histogram-based features and texture families such as GLSZM, GLRLM, and GLCM (see figure 3). These findings are in line with the literature, as filtered features are well known to significantly enhance predictive performance in radiomic studies [48], and histogram-based and texture features have shown strong predictive power in lung imaging tasks. These feature families, unlike tumour size and shape features, can capture the complex morphological, functional, and metabolic characteristics of tumours [49]. Moreover, prior studies have also identified GLSZM, GLRLM, and GLCM features as highly informative [50–52], with GLSZM features being significantly associated with semantic descriptors like tumour cavitation [52]. The advantages of the fusion approach are also consistent with findings from Chengdi Wang *et al*, who developed a similar pipeline using CT acquisitions [30]. Notably, the superiority of the fusion-based approach extends beyond PD-L1 prediction, with similar benefits observed in predicting EGFR mutation status in NSCLC [53], distinguishing benign from malignant ground-glass nodules [54], and assessing lung nodule malignancy [14].

The results suggest a bias towards conservative predictions, where the model shows high precision and specificity but low recall for the positive (majority) class. i.e., when the model makes a positive prediction, it is usually correct, but has an overall tendency to wrongly predict true positive cases as negative (false negatives). This may likely be influenced by the small dataset and the class imbalance present within our dataset, where, although the positive (majority) class dominates the training process, it lacks sufficient variability for robust generalization, sometimes misclassifying them across all considered classifiers, approaches, and acquisitions (see figure 3s of the supporting information, where the confusion matrices are reported). In fact, the small dataset size and the mono-centric nature of our study represent major limitations of this analysis. These challenges are common in medical imaging research, particularly in the context of lung MRI, which is still not part of routine clinical practice. For this reason, the development of models suitable for clinical environments, where data availability is often limited, was prioritized over the use of complex architectures that typically require large training datasets. Focus was directed to interpretable and accessible approaches—such as linear regression and ensemble-based methods—and ensured robustness through rigorous statistical

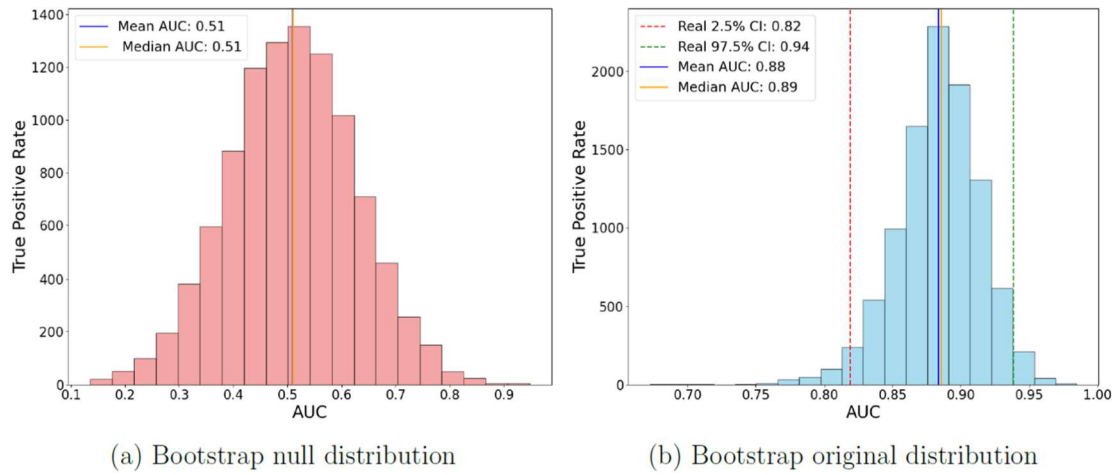


Figure 4. Bootstrap distribution for the T1-VIBE XGB fusion model. Both the null (a) and original (b) datasets are plotted; the mean and median of the original distribution correspond to the stratified 5-fold cross-validation AUC, while those of the null distribution align with random guessing.

validation. In fact, the entire pipeline was specifically designed to address the constraints of small datasets:

- Only four features were selected for each model to prevent overfitting, ensuring at least ten samples per feature.
- The pre-trained ResNet18 model was employed as a fixed feature extractor to obtain deep learning features. This approach has demonstrated strong performance in radiomic applications under limited data conditions [55–57].
- Model training was performed using 100-times repeated stratified 5-fold cross-validation to ensure stability and reproducibility.
- Model performance was evaluated through agreement across multiple metrics for a comprehensive assessment.

AUC confidence intervals and statistical significance were estimated using 10,000-iteration bootstrap simulations. The median and mean of the bootstrapped AUC distributions were closely aligned with the mean AUC value obtained through repeated stratified 5-fold cross-validation, further reinforcing the robustness of the results, as reported in table 1. As an illustrative example, figure 4 displays the bootstrap distribution of AUC values for the T1-VIBE XGB fusion model (AUC $= 0.89 \pm 0.11$) alongside the null distribution derived from the permutation test. The 95% confidence interval for the AUC ranges from 0.82 to 0.94. Moreover, the permutation test confirmed the statistical significance of the observed mean cross-validated AUC, yielding a p-value of 0.0002.

Additionally, the uncommon availability of multiple MRI acquisitions within the same cohort provided a unique opportunity to compare the predictive performance across different MRI sequences. Focusing on the stand-alone approaches, IVIM-D* showed the best performance when described with radiomic features, while IVIM-D performed better with deep features. A possible explanation is that IVIM-D* tends to be noisier, making radiomic aggregation within the ROI better suited to capture heterogeneity without overfitting, whereas IVIM-D has a higher signal-to-noise ratio and clearer edges that convolutional networks can exploit more effectively. Across all combinations, AUCs ranged from 0.72 to 0.92, indicating a learnable PD-L1–related signal in the analysed MRI acquisitions. In general, fusion approaches yielded the highest AUCs, likely by leveraging complementary information from radiomic and deep features—a trend also reported in other settings [30, 58–60].

Notably, these findings are in line with previous literature reporting a correlation between PD-L1 expression and IVIM-derived parameters in NSCLC [24], which reflect distinct aspects of tumour microstructure, especially if related to capillary perfusion (IVIM-D*). Given these findings, we can suppose that IVIM-D* may be particularly informative, since inflammatory presence is also linked to tumour neovascularization. IVIM-D derived features show predictive ability for PD-L1 expression (AUC 0.92), especially when described combining radiomic and deep learning features. IVIM-D has already been correlated with tumour cellularity and PD-L1 expression [61], since it reflects tumour cellularity and considers the contribution of the intralesional inflammatory cells. Finally, it is worth considering that IVIM-pf emerged as the most stable parameter across analyses (AUC [0.84–0.87]).

Lung tumour phenotypes can be divided into three main categories: immuno-inflamed, immuno-excluded, and immuno-desert, indicating decreasing levels of immune cell presence within the tumour stroma. As noted in the literature, the tumour

phenotype correlates with the ability to respond to immune checkpoint inhibitors, showing a correlation between phenotype and therapeutic efficacy, an aspect particularly valuable in clinical practice [62–65].

The baseline T1-VIBE sequence alone achieved an AUC up to 0.90 with fusion models. This highlights how morphological features, captured by both radiomics and deep learning, have predictive value. Signal heterogeneity, cavitations, and irregular margins can act as surrogates of the classical radiological qualitative features such as necrosis, stromal remodelling, and heterogeneous appearance. Histogram and texture features can, in fact, describe well the functional and morphological complexity of the tumour.

The multiparametric nature of MRI enables the capture of complementary information that, when analysed through radiomics and deep learning, can provide insights into the immune milieu of the tumour. Clinically, this positions MRI as a potential new tool for assessing the immune status of the tumour both at baseline and during therapy, which—as previously discussed—correlates with response to immunotherapy.

Moreover, the stability of IVIM-pf performances supports its role as a reliable parameter for longitudinal monitoring, offering a rapid, non-invasive tool for re-sampling during treatment, particularly in cases where resistance develops. Finally, considering that tissue biopsy samples only a portion of the tumour, these findings underscore the clinical value of whole-tumour, non-invasive imaging strategies to better capture the spatial heterogeneity of PD-L1 expression [66–68]. The consistency of the classification performance further supports the robustness of the approach across different classifiers and MRI inputs. However, to validate the proposed methodology and enable its integration into clinical practice, further studies involving multicentre cohorts and larger datasets will be necessary.

5. Conclusions

In this study, a non-invasive pipeline able to assess PD-L1 expression in NSCLC patients was developed by analysing non-enhanced MRI acquisitions, devoting considerable effort to achieve statistical robustness despite the limitation of small datasets. The fusion approach yielded the best performance for IVIM-D, IVIM-pf, and T1-VIBE, while IVIM-D* showed comparable results between the radiomic and fusion models. Across all MRI-derived features, the AUC ranged from 0.72 to 0.92, highlighting the potential of a combined radiomic and deep learning strategy to effectively detect PD-L1 expression from MRI in NSCLC patients. Our findings demonstrate that a fusion of radiomic and deep learning features can effectively detect PD-L1 expression, highlighting its potential as a non-invasive alternative for patient stratification in immunotherapy. However, as this is a pilot study based on a small dataset, further validation on a larger, multicentre cohort is required.

Acknowledgments

This research was funded by the grant 20190049910 of Foundation IRCCS Polyclinic San Matteo and by the INFN, NEXT_AIM project. The PNRR National Centre for HPC, Big Data and Quantum Computing, research project CN00000013, and MIUR Dipartimento di Eccellenza 2018–2022 project F11I18000680001 are also gratefully acknowledged.

Grant support

This research was funded by the grant 20190049910 of Foundation IRCCS Polyclinic San Matteo and by the INFN, NEXT AIM project. A. L. acknowledges the PNRR National Centre for HPC, Big Data and Quantum Computing, research project CN00000013.

Ethical statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Foundation IRCCS Polyclinic San Matteo. Written informed consent was obtained from all participants.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

References

[1] Sung H, Ferlay J, Siegel R L, Laversanne M, Soerjomataram I, Jemal A and Bray F 2021 Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries *CA Cancer J. Clin.* **71** 209–49

[2] Leiter A, Veluswamy R R and Wisnivesky J P 2023 The global burden of lung cancer: current status and future trends *Nat. Rev. Clin. Oncol.* **20** 624–39

[3] Chen R, Manochakian R and James L 2020 Emerging therapeutic agents for advanced non-small cell lung cancer *Journal of Hematology & Oncology* **13** 1–23

[4] Garon E B, Rizvi N A and Hui R 2015 Pembrolizumab for the treatment of non–small-cell lung cancer *New Engl. J. Med.* **372** 2018–28

[5] Herbst R S, Soria J C and Kowanetz M 2014 Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients *Nature* **515** 563–7

[6] Herbst R S, Baas P and Kim D W 2016 Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial *Lancet* **387** 1540–50

[7] Mok T S, Wu Y L and Kudaba I 2019 Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial *Lancet* **393** 1819–30
[8] Haragan A, Field J K and Davies M P 2019 Heterogeneity of PD-L1 expression in non-small cell lung cancer: Implications for specimen sampling in predicting treatment response *Lung Cancer* **134** 79–84
[9] Mayerhoefer M E, Materka A and Langs G 2020 Introduction to radiomics *J. Nucl. Med.* **61** 488–95
[10] Heilbroner S P and Miotto R 2023 Deep learning in medicine *Clinical Journal of the American Society of Nephrology* **18** 397–9
[11] Shen C, Nguyen D and Zhou Z 2020 An introduction to deep learning in medical physics: advantages, potential, and challenges *Phys. Med. Biol.* **65** 05TR01
[12] Binczyk F, Prazuch W and Bozek P 2021 Radiomics and artificial intelligence in lung cancer screening *Translational Lung Cancer Research* **10** 1186
[13] Tandungan S and Nurtanio I 2019 Comparison of accuracy in extreme learning machine based on hidden node structure variation for lung cancer classification *IOP Conf. Ser.: Mater. Sci. Eng.* (IOP Publishing) 012014
[14] Li S, Xu P and Li B 2019 Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features *Phys. Med. Biol.* **64** 175012
[15] Amicizia D *et al* 2023 Systematic review of lung cancer screening: advancements and strategies for implementation *Healthcare* **11** 2085
[16] Wielpütz M and Kauczor H U 2012 MRI of the lung: state of the art *Diagnostic and Interventional Radiology* **18** 344
[17] Biederer J N, Ohno Y and Hatabu H 2017 Screening for lung cancer: does MRI have a role? *Eur. J. Radiol.* **86** 353–60
[18] Raptis C A, Ludwig D R and Hammer M M 2019 Building blocks for thoracic MRI: Challenges, sequences, and protocol design *J. Magn. Reson. Imaging* **50** 682–701
[19] Voskrebenzev A and Vogel-Claussen J 2021 Proton MRI of the lung: how to tame scarce protons and fast signal decay *J. Magn. Reson. Imaging* **53** 1344–57
[20] Ohno Y, Takenaka D and Yoshikawa T 2022 Efficacy of ultrashort echo time pulmonary MRI for lung nodule detection and lung-RADS classification *Radiology* **302** 697–706
[21] Yang S, Zhang Y and Shen J 2020 Clinical potential of UTE-MRI for assessing COVID-19: patient- and lesion-based comparative analysis *J. Magn. Reson. Imaging* **52** 397–406
[22] Lee H, Choi E and Lee M K 2019 Morphologic evaluation of primary non-small cell lung cancer by 3 tesla MRI with free-breathing ultrashort echo time and radial T1-weighted gradient echo sequences: a comparison with CT analysis *Journal of the Korean Society of Radiology* **80** 466–76
[23] Owens C, Hindocha S and Lee R 2023 The lung cancers: staging and response, CT, 18F-FDG PET/CT, MRI, DWI: review and new perspectives *The British Journal of Radiology* **96** 20220339
[24] Bortolotto C, Stella G M and Messana G 2022 Correlation between PD-L1 expression of non-small cell lung cancer and data from IVIM-DWI acquired during magnetic resonance of the thorax: preliminary results *Cancers* **14** 5634
[25] Le Bihan D, Breton E and Lallemand D 1986 MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders *Radiology* **161** 401–7
[26] Le Bihan D and Turner R 1992 The capillary network: a link between IVIM and classical perfusion *Magn. Reson. Med.* **27** 171–8
[27] Iima M 2021 Perfusion-driven intravoxel incoherent motion (IVIM) MRI in oncology: applications, challenges, and future trends *Magnetic Resonance in Medical Sciences* **20** 125–38
[28] Mu W *et al* 2021 Non-invasive measurement of PD-L1 status and prediction of immunotherapy response using deep learning of PET/CT images *Journal for Immunotherapy of Cancer* **9** e002118
[29] Tian P, He B and Mu W 2021 Assessing PD-L1 expression in non-small cell lung cancer and predicting responses to immune checkpoint inhibitors using deep learning on computed tomography images *Theranostics* **11** 2098
[30] Wang C, Ma J and Shao J 2022 Non-invasive measurement using deep learning algorithm based on multi-source features fusion to predict PD-L1 expression and survival in NSCLC *Frontiers in Immunology* **13** 828560
[31] Shiinoki T, Fujimoto K and Kawazoe Y 2022 Predicting programmed death-ligand 1 expression level in non-small cell lung cancer using a combination of peritumoral and intratumoral radiomic features on computed tomography *Biomed. Phys. Eng. Express* **8** 025008
[32] Lo Gullo R, Wen H and Reiner J S 2021 Assessing PD-L1 expression status using radiomic features from contrast-enhanced breast MRI in breast cancer patients: Initial results *Cancers* **13** 6273
[33] Gong X Q, Liu N and Tao Y Y 2023 Radiomics models based on multisequence MRI for predicting PD-1/PD-L1 expression in hepatocellular carcinoma *Sci. Rep.* **13** 7710
[34] Marletta S, Fusco N and Munari E 2022 Atlas of PD-L1 for pathologists: indications, scores, diagnostic platforms and reporting systems *Journal of Personalized Medicine* **12** 1073
[35] Zwanenburg A, Vallières M and Abdalah M 2020 The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping *Radiology* **295** 328–38
[36] Lehmann T M, Gonner C and Spitzer K 1999 Survey: interpolation methods in medical image processing *IEEE Trans. Med. Imaging* **18** 1049–75
[37] Carré A, Klausner G and Edjlali M 2020 Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics *Sci. Rep.* **10** 12340
[38] Van Griethuysen J J, Fedorov A and Parmar C 2017 Computational radiomics system to decode the radiographic phenotype *Cancer Res.* **77** e104–7
[39] Bortolotto C, Pinto A and Brero F 2024 CT and MRI radiomic features of lung cancer (NSCLC): comparison and software consistency *European Radiology Experimental* **8** 71
[40] He K, Zhang X and Ren S 2016 Deep residual learning for image recognition *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–8
[41] Deng J, Dong W and Socher R 2009 Imagenet: a large-scale hierarchical image database *IEEE Conference on Computer Vision and Pattern Recognition* 248–55
[42] Long J, Shelhamer E and Darrell T 2015 Fully convolutional networks for semantic segmentation *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* 3431–40
[43] Zwanenburg A *et al* 2020 The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping *Radiology* **295** 328–38
[44] Vujović Ž 2021 Classification model evaluation metrics *International Journal of Advanced Computer Science and Applications* **12** 599–606
[45] Grant M J, Herbst R S and Goldberg S B 2021 Selecting the optimal immunotherapy regimen in driver-negative metastatic NSCLC *Nat. Rev. Clin. Oncol.* **18** 625–64
[46] Messana G *et al* 2025 Non-invasive PD-L1 stratification in non-small cell lung cancer using dynamic contrast-enhanced MRI *European Radiology* **35** 5569–80
[47] Liu K H, Yang W and Tian H P 2024 Relationships between intravoxel incoherent motion parameters and expressions of programmed cell death-1 (PD-1) and programmed cell death ligand-1 (PD-L1) in patients with cervical cancer *Clinical Radiology* **79** 264–e272
[48] Demircioğlu A 2022 The effect of preprocessing filters on predictive performance in radiomics *European Radiology Experimental* **6** 40
[49] Lee G, Lee H Y and Park H 2017 Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art *Eur. J. Radiol.* **86** 297–307

[50] Xu F, Zhu W and Shen Y 2020 Radiomic-based quantitative CT analysis of pure ground-glass nodules to predict the invasiveness of lung adenocarcinoma *Frontiers in Oncology* **10** 872
[51] Song F, Song X and Feng Y 2023 Radiomics feature analysis and model research for predicting histopathological subtypes of non-small cell lung cancer on CT images: a multi-dataset study *Med. Phys.* **50** 4351–65
[52] Yip S S, Liu Y and Parmar C 2017 Associations between radiologist-defined semantic and automatically computed radiomic features in non-small cell lung cancer *Sci. Rep.* **7** 3519
[53] Gong J, Fu F and Ma X 2023 Hybrid deep multi-task learning radiomics approach for predicting EGFR mutation status of non-small cell lung cancer in CT images *Phys. Med. Biol.* **68** 245021
[54] HU X, Gong J and Li H 2021 Computer-aided diagnosis of ground glass pulmonary nodule by fusing deep learning and radiomics features *Phys. Med. Biol.* **66** 065015
[55] Alkhatatbeh T, Alkhatatbeh A and Li X 2024 A single sequence MRI-based deep learning radiomics model in the diagnosis of early osteonecrosis of femoral head *Front. Bioeng. Biotechnol.* **12** 1471692
[56] Li L, Zhou X and Cui W 2023 Combining radiomics and deep learning features of intra-tumoral and peri-tumoral regions for the classification of breast cancer lung metastasis and primary lung cancer with low-dose CT *Journal of Cancer Research and Clinical Oncology* **149** 15469–78
[57] Wang X, Ma C and Jiang Q 2024 Performance of deep learning model and radiomics model for preoperative prediction of spread through air spaces in the surgically resected lung adenocarcinoma: a two-center comparative study *Translational Lung Cancer Research* **13** 3486
[58] Zhang Z, Zhang D, Yang Y, Liu Y and Zhang J 2024 Value of radiomics and deep learning feature fusion models based on dce-mri in distinguishing sinonasal squamous cell carcinoma from lymphoma *Frontiers in Oncology* **14** 1489973
[59] Astaraki M, Yang G, Zakko Y, Toma-Dasu I, Smedby O and Wang C 2021 A comparative study of radiomics and deep-learning based methods for pulmonary nodule malignancy prediction in low dose CT images *Frontiers in Oncology* **11** 737368
[60] Sun Q, Yu L and Song Z 2025 Deep learning and radiomics fusion for predicting the invasiveness of lung adenocarcinoma within ground glass nodules *Sci. Rep.* **15** 29285
[61] Meng N, Fu F and Sun J 2022 Sensitivity and specificity of amide proton transfer-weighted imaging for assessing programmed death-ligand 1 status in non-small cell lung cancer: a comparative study with intravoxel incoherent motion and 18F-FDG PET *Quantitative Imaging in Medicine and Surgery* **12** 4474
[62] Lizotte P H, Ivanova E V and Awad M M 2016 Multiparametric profiling of non-small-cell lung cancers reveals distinct immunophenotypes *JCI Insight* **1** 89014
[63] Chen D S and Mellman I 2017 Elements of cancer immunity and the cancer–immune set point *Nature* **541** 321–30
[64] Hegde P S, Karanikas V and Evers S 2016 The where, the when, and the how of immune monitoring for cancer immunotherapies in the era of checkpoint inhibition *Clinical Cancer Research* **22** 1865–74
[65] Thorsson V, Gibbs D L and Brown S D 2018 The immune landscape of cancer *Immunity* **48** 812–30
[66] Saito Y, Horiuchi S and Morooka H 2019 Inter-tumor heterogeneity of PD-L1 expression in non-small cell lung cancer *Journal of Thoracic Disease* **11** 4982–91
[67] Stella G M, Bortolotto C and Filippi A R 2018 Intrathoracic core needle biopsy and repeat biopsy for PD-L1 evaluation in non-small cell lung cancer *Journal of Thoracic Disease* **10** S4031–3
[68] Casale S, Bortolotto C and Stella G M 2021 Recent advancement on PD-L1 expression quantification: the radiologist perspective on CT-guided FNAC *Diagnostic and Interventional Radiology* **27** 214–8