Cancer Imaging

RESEARCH

Open Access

# CT-based radiomics as a non-invasive virtual biopsy for high PD-L1 expression prediction in non-small cell lung cancer

Michela Destito[1†], Caterina Battaglia[1,2†], Paolo Zaffino[1*], Giulio Caridà[1], Maria Cucè[3†], Alessandro Pullano[1,2], Martina Frangipane[1,2], Maria Francesca Spadea[4], Domenico Laganà[1,2], Pierfrancesco Tassone[1], Pierosandro Tagliaferri[1] and Carlo Cosentino[1]

## Abstract

**Purpose** Non-small cell lung cancer (NSCLC) remains a major clinical challenge, with Programmed death-ligand 1 (PD-L1) expression serving as a crucial biomarker to guide immunotherapy. However, its current assessment through invasive biopsies may not capture tumor heterogeneity. This study explores the feasibility of a CT-based radiomics approach, combined with machine learning (ML), as a potential non-invasive virtual biopsy to predict high PD-L1 expression (≥50%) in NSCLC patients.

**Methods** Contrast-enhanced CT scans from 55 patients with histologically confirmed NSCLC were retrospectively analyzed. Radiomic features were extracted from tumor volumes, and multiple ML classifiers were trained and evaluated through repeated stratified k-fold cross-validation.

**Results** Among the models evaluated, the Support Vector Machine (SVM) classifier demonstrated the best performance, achieving a median accuracy of 0.77 (quartiles: 0.66–0.82) and an area under the curve (AUC) of 0.83 (0.63–0.92). Feature importance analysis using SHAP (Shapley Additive Explanations) revealed that texture features were the most informative in predicting PD-L1 expression levels. Notably, the integration of clinical data did not improve model performance, highlighting the dominant predictive value of radiomic features alone.

**Conclusion** Our findings support the feasibility of CT-based radiomics as a potential tool for virtual biopsy to identify NSCLC patients with high PD-L1 expression (≥50%), potentially serving as a complementary or alternative tool to tissue biopsy, especially in cases where biopsy is contraindicated or insufficient.

**Keywords** Non-small cell lung cancer, Radiomics, Virtual Biopsy, Machine Learning, Feature importance

†Michela Destito and Caterina Battaglia contributed equally to this work.

*Correspondence:
Paolo Zaffino
p.zaffino@unicz.it
[1]Department of Experimental and Clinical Medicine, University of Catanzaro, Viale Europa, 88100 Catanzaro, Italy
[2]Radiology Unit, Department of Experimental and Clinical Medicine, University of Catanzaro, Viale Europa, 88100 Catanzaro, Italy
[3]Medical Oncology Unit, R. Dulbecco Hospital, University of Catanzaro, Viale Europa, 88100 Catanzaro, Italy
[4]Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany

## Introduction

Non-small cell lung cancer (NSCLC) represents approximately 85% of all newly diagnosed cases of lung cancer [1]. This tumor is characterized by considerable heterogeneity [2], encompassing multiple histological subtypes that differ in their molecular profiles and clinical behaviors [3]. Despite notable advances in diagnostic imaging and therapeutic strategies, the overall prognosis for NSCLC patients remains poor, particularly when the disease is diagnosed in an advanced stage [4].

Programmed death-ligand 1 (PD-L1) has emerged as a critical biomarker in NSCLC [5–7], playing a crucial role in guiding treatment decisions involving immune checkpoint inhibitors. However, conventional biopsy-based techniques used to obtain such information are inherently invasive and often fail to capture the spatial and molecular heterogeneity of the tumor. This limitation underscores the need for non-invasive, comprehensive methods capable of assessing PD-L1 expression across the entire tumor volume.

Radiomics analysis [8, 9] extracts quantitative features from medical images and can offer a non-invasive alternative to predict PD-L1 expression, potentially overcoming the limitations of biopsies. Furthermore, this approach enables capturing tumor heterogeneity and spatial variations that might be missed by localized sampling, providing a more comprehensive assessment of the tumor microenvironment. In oncology, Artificial Intelligence (AI) [10], and particularly machine learning (ML) [11] algorithms, can process radiomic features extracted from regions of interest defined in the image (in this case the tumor) to identify complex patterns associated with specific biological signatures [12–14]. Once trained on large annotated image datasets, these models can effectively stratify patients, predict therapeutic response, and support personalized treatment planning.

Recent studies have demonstrated the feasibility and promise of ML-based radiomics in the prediction of molecular and immunological biomarkers, including PD-L1 expression.

In particular [15–17], have examined the use of FDG-PET/CT imaging (fluorodeoxyglucose positron emission tomography alongside computed tomography) to predict PD-L1 expression in NSCLC patients. Although these approaches are promising, their reliance on dual imaging modalities may limit clinical applicability. Other recent works [18–22], more in line with the present study, have explored CT image-only methods as a potentially more accessible alternative. However, all these studies do not clearly specify the ML algorithms employed, making it difficult to assess or replicate their findings.

In contrast, the proposed study systematically evaluates different classification algorithms, explores different strategies for feature selection, and incorporates diverse performance metrics to provide a more robust and comprehensive overview of the use of ML-based radiomics to predict PD-L1 expression $\geq$50% in NSCLC patients, using only CT imaging. The final aim is to improve immunotherapy stratification through a non-invasive and reproducible method to offer a viable alternative to biopsy-based PD-L1 evaluation, contributing to more effective and personalized patient management.

## Materials and methods

### Study population

CT images and clinical data were retrospectively collected from patients referred to the Oncology Unit of Magna Graecia University in Catanzaro, Italy, who had been diagnosed with NSCLC as of 2018. From an initial cohort of 182 lung tumor cases, patients diagnosed with a non-NSCLC tumor or staged exclusively on the basis of PET were excluded.

Inclusion criteria were: confirmed diagnosis of NSCLC (regardless of histologic subtype), availability of pre-treatment chest CT with contrast agent, and biomarker immunohistochemistry assessed PD-L1 status.

CTs were performed using multidetector scanners from different manufacturers. Acquisition parameters included a tube voltage of 100–120 kVp and automatic tube current modulation (reference mAs adapted to patient body habitus). Images were reconstructed using both soft-tissue and high-spatial-resolution kernels for mediastinal and lung parenchymal assessment. The field of view was tailored to the patient's size to include the entire thorax. Slice thickness ranged from 0.7 to 5.0 mm (median = 0.8 mm) with overlapping reconstructions. Pitch values ranged between 0.9 and 1.2.

Figure 1 reports the flowchart that led to selecting 55 patients. The final cohort was stratified according to PD-L1 expression levels: 26 patients had PD-L1 expression $\geq$50%, while 29 patients had PD-L1 expression $<$ 50%.

PD-L1 expression data was obtained from tissue biopsy or surgical samples, with a 50% threshold used to classify patients as PD-L1 positive. All data were anonymized and handled according to ethical guidelines. In addition to imaging information, the main clinical characteristics of the patients were collected. Those features are summarized in Table 1. Approval for this observational study was granted by the Ethics Committee of the Calabria Region – Central Area Section (protocol no. 79, dated March 16, 2023). All study steps were performed in accordance with the Declaration of Helsinki.

### Image pre-processing

The CT volumes used in this work were collected with different scanners. Considering this, voxel resampling ($1 \times 1^{*}1$ mm) [23] was applied to standardize the spatial

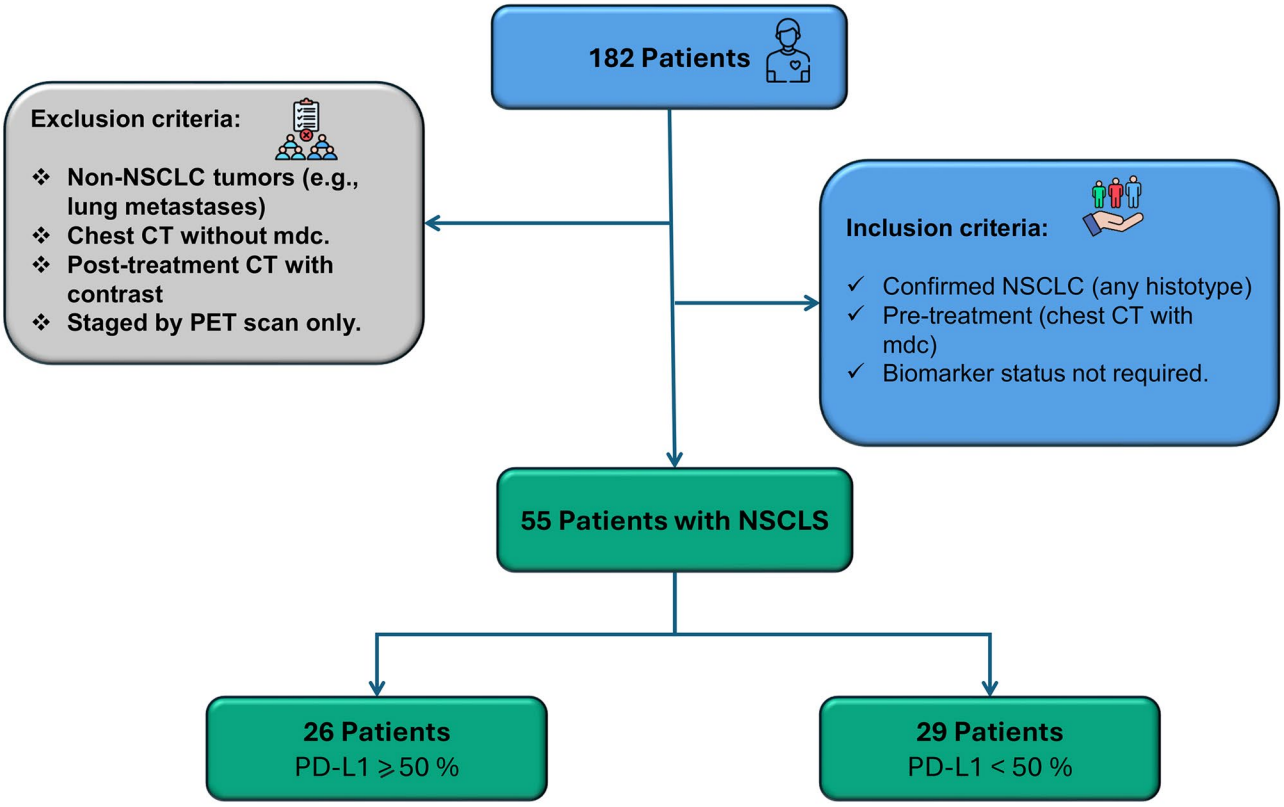


**Fig. 1** Patient selection flowchart summarizing the screening process, inclusion and exclusion criteria, and final cohort composition used for radiomics analysis of PD-L1 expression in NSCLC patients

resolution of the images, ensuring a more homogeneous analysis.

#### Segmentation volume of interest and feature extraction

Tumor segmentation was performed by an expert radiologist (C.B.) manually delineating tumor regions in the CT images using the 3D Slicer software tool [24]. This segmentation enabled the extraction of precise radiomic features from the tumor regions, which were then used in the subsequent analysis. To assess inter-rater variability, a random subset of 10 patients was also contoured by two independent radiologists (A.P. and M.F.). A comprehensive set of 120 radiomic features was extracted from the segmented tumor. Features were extracted from each patient using PyRadiomics [25], an open-source Python library [26] for extracting high-resolution features from medical images. These features were categorized as follows: First-order statistics, which describe the distribution of voxel intensities (e.g., mean, standard deviation, skewness); Shape-based features, which quantify the three-dimensional geometry of the tumor (e.g., volume, surface area, sphericity); Gray Level Co-occurrence Matrix (GLCM) features, capturing spatial relationships and texture patterns between neighboring voxels (e.g., contrast, correlation, Imc1); Gray Level Run Length Matrix (GLRLM) features, measuring the length of consecutive voxels with the same intensity; Gray Level Size Zone Matrix (GLSZM) features, assessing the size of homogeneous zones within the lesion; Gray Level Dependence Matrix (GLDM) features, evaluating the degree of dependence between gray levels in a given neighborhood (e.g., Dependence Entropy). This diverse feature set provided a detailed and non-invasive characterization of tumor heterogeneity, supporting the development of ML models for PD-L1 expression prediction.

#### Feature selection and machine learning model training

As a preliminary step, the correlated features (threshold = 0.7) were identified and removed to reduce redundancy and mitigate multicollinearity, which can negatively affect model interpretability and performance. This was followed by data normalization using the Standard Scaler method from scikit-learn [27], which standardized each feature by removing the mean and scaling to unit variance. This pre-processing step was essential to ensure that all features contributed equally to the model training process. Subsequently, a combination of feature selection methods was employed to identify the most predictive radiomic features within each training fold using three approaches: (i) Recursive Feature Elimination

**Table 1** Clinical characteristics of the eligible patient cohort. Values are reported as counts unless otherwise specified. SD: standard deviation

| Clinical feature | Eligible patients (n = 55) |
|---|---|
| Sex | |
| Male | 43 |
| Female | 12 |
| Median age, years (SD) | 70 (9.74) |
| Average tumor volume, $mm^3$ (SD) | 115.8 (167.7) |
| Tumor location | |
| Left lung | 18 |
| Right lung | 37 |
| Histological type | |
| Squamous cell carcinoma | 13 |
| Adenocarcinoma | 42 |
| Smoking history | |
| Smoker | 48 |
| Never smoked | 7 |
| Stage | |
| I | 0 |
| II | 3 |
| III | 6 |
| IV | 46 |
| Treatment | |
| Immunotherapy only | 14 |
| Chemotherapy + immunotherapy | 24 |
| Chemotherapy only | 12 |
| Target therapy | 5 |
| Pleural invasion | |
| Present | 18 |
| Absent | 37 |
| PD-L1 expression | |
| ≥50% | 26 |
| <50% | 29 |

(RFE) [28] with a logistic regression base estimator (solver=liblinear); (ii) SelectFromModel with L2-regularized logistic regression (penalty='l2'); and (iii) SelectFromModel with Random Forest (n_estimators=100). Based on a heuristic analysis, the optimal number of features for each method was set to 18. The final feature set for the subsequent classification step was determined by retaining only those variables selected by at least two out of the three strategies. This ensemble-like approach aimed to improve feature robustness and reduce model variance by focusing on features with stable selection across multiple algorithms. The complete list of selected features and their selection frequency across folds is reported in the supplementary material.

Various ML models [29] were trained on the selected features to predict PD-L1 expression levels. The following six ML classifiers were tested in this study: Extra Trees Classifier (ETC), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), and XGB (XGBoost). All classifiers were trained using fixed hyperparameters. Extensive hyperparameter optimization was intentionally avoided to reduce the risk of overfitting, given the limited cohort size. Indeed, conducting the same experiments with hyperparameter tuning led to a decrease in performance, confirming the robustness of the initial configuration. This design prioritizes stability and reproducibility over aggressive tuning and reflects the exploratory nature of the study.

To ensure robustness, reduce overfitting, and obtain a reliable estimate of generalization performance, the entire modeling pipeline was implemented using repeated stratified k-fold cross-validation (5 folds, 2 repetitions) [30]. In each fold, 80% of the data were used for training and 20% for validation. The stratified design ensures that class distribution is preserved across folds, generating balanced training and validation subsets and reducing sampling bias. Both feature normalization and selection steps were performed independently within the training portion of each cross-validation fold.

All analyses were conducted in Python 3.12.12 using PyRadiomics version 3.0.1, scikit-learn version 1.6.1, and SHAP version 0.50.0.

#### *Model evaluation*

The performance of each model was evaluated using the following metrics [31]: Accuracy, F1-Score, Precision, Recall, AUC (Area Under the Curve). To interpret the models and understand which radiomics features were most influential in predicting PD-L1 expression, a feature importance analyses using SHAP (Shapley Additive Explanations) [32] was conducted. This method provides a way to interpret ML models by evaluating the contribution of each feature to the model's output. SHAP values were used to quantify the impact of each radiomics feature on the prediction of PD-L1 expression.

Statistical differences between classifiers were assessed using the Wilcoxon signed-rank test [33] across multiple cross-validation folds. The workflow of the study is shown in Fig. 2.

### Contribution of clinical features

The included clinical variables were chosen as they are commonly used in clinical practice to provide a comprehensive description of the patient's clinical history.

To assess the importance of the clinical features described in Table 1, the same ML models were trained using: (i) clinical features alone and (ii) the combination of clinical and radiomic features. The clinical variables considered as model inputs were sex, age, tumor location, histological type, and smoking history. Stage and pleural invasion were excluded from the training process.

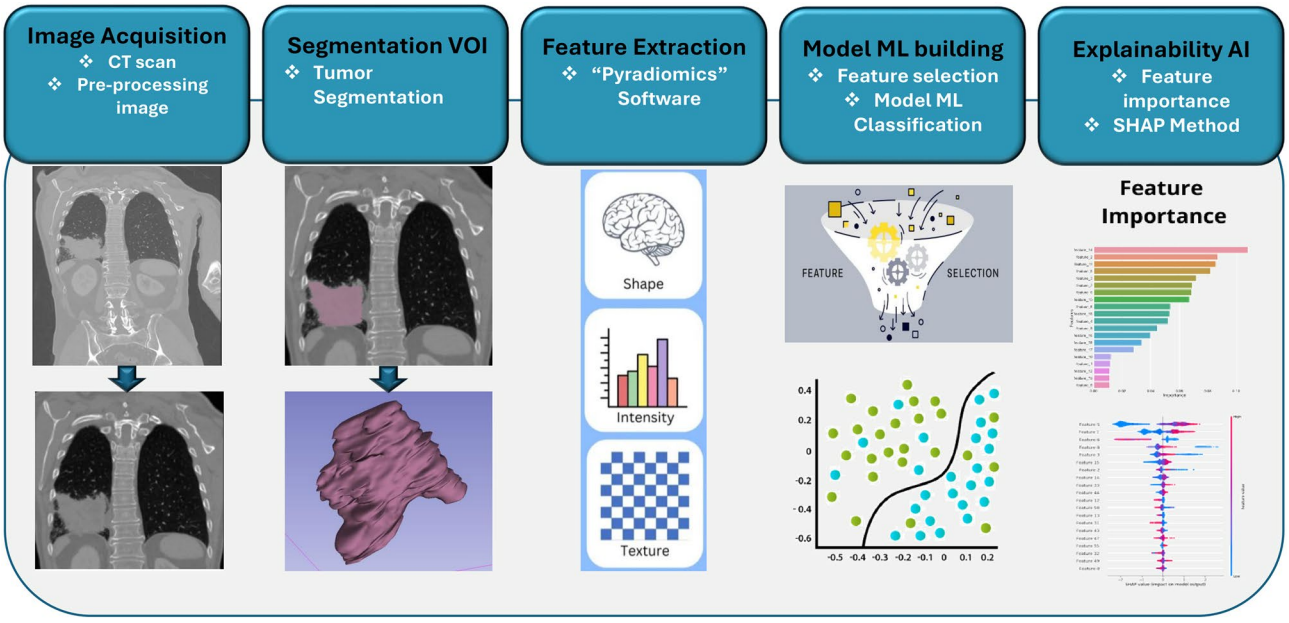


**Fig. 2** Workflow of the proposed radiomics pipeline. The diagram illustrates CT image acquisition, manual tumor segmentation, radiomic feature extraction, feature selection embedded within cross-validation, machine learning model training, performance evaluation, and SHAP-based interpretability analysis

**Table 2** Classification performance across repeated 5-fold cross-validation. Values are reported as median (interquartile range). *M* = machine learning model; F = feature set (R: radiomic; R+C: radiomic + clinical; C: clinical)

| M | F | Classification Metrics | | | | |
|---|---|---|---|---|---|---|
| | | **Accuracy** | **F1-score** | **Precision** | **Recall** | **AUC** |
| ETC | R | 0.50 (0.45–0.64) | 0.50 (0.41–0.60) | 0.50 (0.43–0.60) | 0.60 (0.35–0.60) | 0.53 (0.40–0.63) |
| | R+C | 0.50 (0.45–0.61) | 0.50 (0.40–0.53) | 0.46 (0.40–0.60) | 0.40 (0.40–0.57) | 0.44 (0.32–0.64) |
| | C | 0.50 (0.39–0.54) | 0.53 (0.31–0.54) | 0.46 (0.34–0.50) | 0.60 (0.28–0.60) | 0.49 (0.32–0.55) |
| LR | R | 0.68 (0.57–0.73) | 0.70 (0.63–0.73) | 0.67 (0.59–0.67) | 0.80 (0.70–0.80) | 0.75 (0.53–0.77) |
| | R+C | 0.64 (0.48–0.64) | 0.57 (0.50–0.70) | 0.56 (0.45–0.68) | 0.60 (0.60–0.80) | 0.67 (0.53–0.73) |
| | C | 0.54 (0.48–0.54) | 0.54 (0.51–0.62) | 0.50 (0.45–0.55) | 0.60 (0.52–0.75) | 0.58 (0.45–0.64) |
| SVM | R | **0.77 (0.66–0.82)** | **0.76 (0.68–0.82)** | **0.71 (0.67–0.80)** | **0.80 (0.67–0.80)** | **0.83 (0.63–0.92)** |
| | R+C | 0.54 (0.45–0.64) | 0.61 (0.43–0.65) | 0.50 (0.43–0.59) | 0.60 (0.43–0.80) | 0.62 (0.57–0.75) |
| | C | 0.45 (0.45–0.61) | 0.48 (0.35–0.62) | 0.58 (0.41–0.59) | 0.50 (0.35–0.80) | 0.62 (0.40–0.72) |
| RF | R | 0.64 (0.54–0.73) | 0.59 (0.46–0.65) | 0.63 (0.53–0.79) | 0.60 (0.40–0.60) | 0.65 (0.58–0.72) |
| | R+C | 0.45 (0.39–0.61) | 0.45 (0.29–0.57) | 0.41 (0.34–0.69) | 0.45 (0.25–0.60) | 0.48 (0.33–0.65) |
| | C | 0.45 (0.39–0.54) | 0.49 (0.41–0.53) | 0.46 (0.38–0.50) | 0.55 (0.40–0.60) | 0.50 (0.35–0.61) |
| KNN | R | 0.59 (0.48–0.64) | 0.44 (0.41–0.58) | 0.53 (0.50–0.65) | 0.40 (0.35–0.60) | 0.60 (0.47–0.76) |
| | R+C | 0.50 (0.45–0.54) | 0.47 (0.41–0.52) | 0.50 (0.41–0.55) | 0.40 (0.35–0.75) | 0.53 (0.42–0.61) |
| | C | 0.50 (0.45–0.54) | 0.49 (0.24–0.56) | 0.42 (0.21–0.50) | 0.55 (0.25–0.60) | 0.56 (0.44–0.63) |
| XGB | R | 0.64 (0.56–0.73) | 0.61 (0.58–0.71) | 0.63 (0.53–0.73) | 0.63 (0.60–0.80) | 0.66 (0.63–0.79) |
| | R+C | 0.68 (0.54–0.73) | 0.64 (0.58–0.71) | 0.63 (0.52–0.75) | 0.63 (0.60–0.80) | 0.65 (0.46–0.79) |
| | C | 0.45 (0.37–0.54) | 0.46 (0.41–0.50) | 0.43 (0.38–0.50) | 0.45 (0.40–0.60) | 0.41 (0.30–0.55) |

## Results

### PD-L1 expression classification performance

The Table 2 summarizes the performance (in terms of median and quartiles) of the six ML models (ETC, LR, SVM, RF, KNN, XGB),evaluated using five test folds with two repetitions of cross validation across three different feature sets: radiomics features, radiomics and clinical features, and only clinical features.

Among all models, SVM achieved the best overall performance using the identified radiomic feature set, considering medians and quartiles equal to: 0.77 (0.66–0.82) for accuracy, 0.76 (0.68–0.82) for F1-score, 0.71 (0.67–0.80) regarding precision, 0.80 (0.67–0.80) for recall, and 0.83 (0.63–0.92) for AUC.

The LR model trained only on radiomic features provided good results as well, reporting a median/quartiles

F1-score of 0.70 (0.63–0.73) and a recall of 0.80 (0.70–0.80). Similarly, XGB trained on radiomic features yielded competitive results, with an AUC of 0.66 (0.63–0.79) and precision of 0.63 (0.53–0.73).

Pairwise Wilcoxon signed-rank tests showed significant differences between SVM and ETC ($p = 0.0098$) and between SVM and KNN ($p = 0.0371$). Other comparisons were not statistically significant (RF: $p = 0.0625$; LR: $p = 0.2324$; XGB: $p = 0.1680$).

Integrating clinical features led to a drop in performance across models.

Overall, models trained solely on clinical features generally exhibited the lowest performance, with median accuracies rarely exceeding 0.54, regardless of the classifier.

Figure 3 illustrates the ROC curves for the best-performing model across the three feature sets: radiomic features, radiomic with clinical features, and clinical features only. The model trained on radiomic features demonstrated superior discriminative ability, with an AUC of 0.83 (0.63–0.92), clearly outperforming the other configurations.

The combination of radiomic and clinical features resulted in a moderate AUC of 0.67 (0.53–0.73), while the model trained solely on clinical features achieved a lower AUC of 0.62 (0.40–0.72). These results confirm that radiomic features contribute significantly to the model's predictive performance, and that the addition of clinical features does not necessarily enhance classification capability.

### SHAP analysis

SHAP analysis was performed to assess the importance of features in predicting PD-L1 expression in NSCLC patients. The analysis was carried out on models trained in multiple folds, first using only radiomic features and then including clinical features. For each independent training run, the selection frequency of each feature was computed. Subsequently, a selection frequency analysis was performed; based on this distribution, the 15 most frequently selected features were identified and retained for the subsequent SHAP analysis.

In the model with only radiomic features (Fig. 4) the most relevant features identified were: Textural features like GLCM Imc1, GLDM DependenceEntropy, and GLSZM SmallAreaEmphasis, which showed the highest impact on the prediction. Specifically, GLCM Imc1 (Informational Measure of Correlation 1) quantifies the degree of linear dependence between the intensities of neighboring pixels, reflecting the local correlation of gray levels; GLDM Dependence Entropy captures the complexity and randomness in the spatial arrangement of

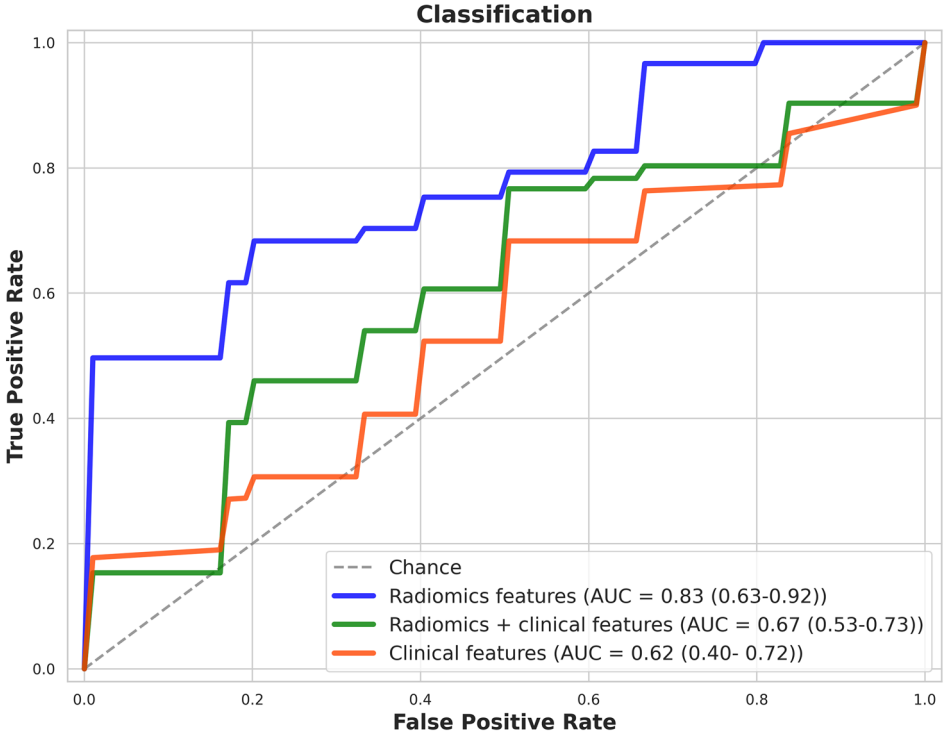


**Fig. 3** Receiver operating characteristic (ROC) curves obtained from cross-validation validation folds for the best-performing classifier across the three feature configurations on test set: radiomics only, radiomics combined with clinical features, and clinical features only. The figure highlights the superior discriminative ability of the radiomics-based model

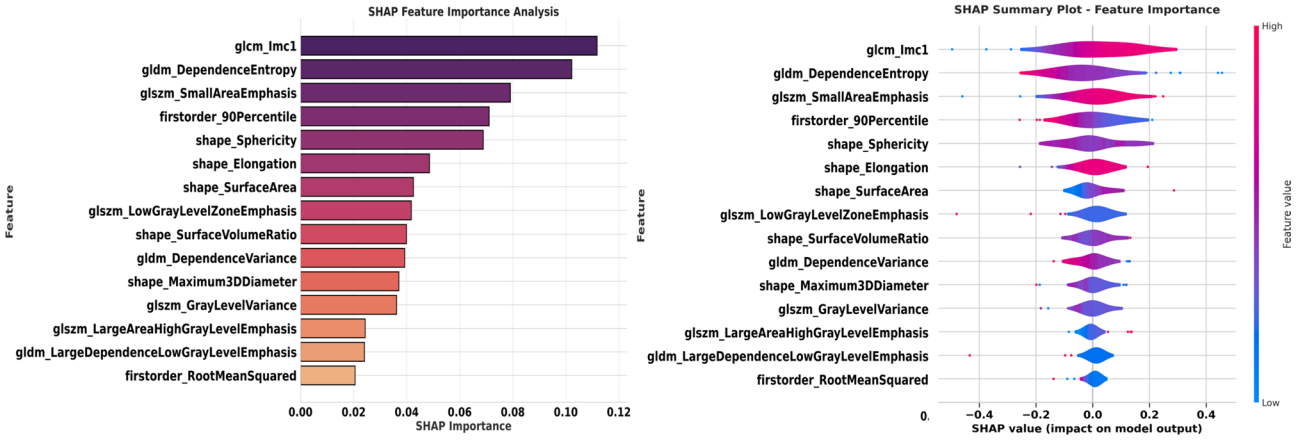


**Fig. 4** SHAP summary plot illustrating the relative contribution of radiomic features to PD-L1 prediction. Each point represents an individual patient, and features are ranked according to their mean absolute SHAP value, indicating their overall importance in the model

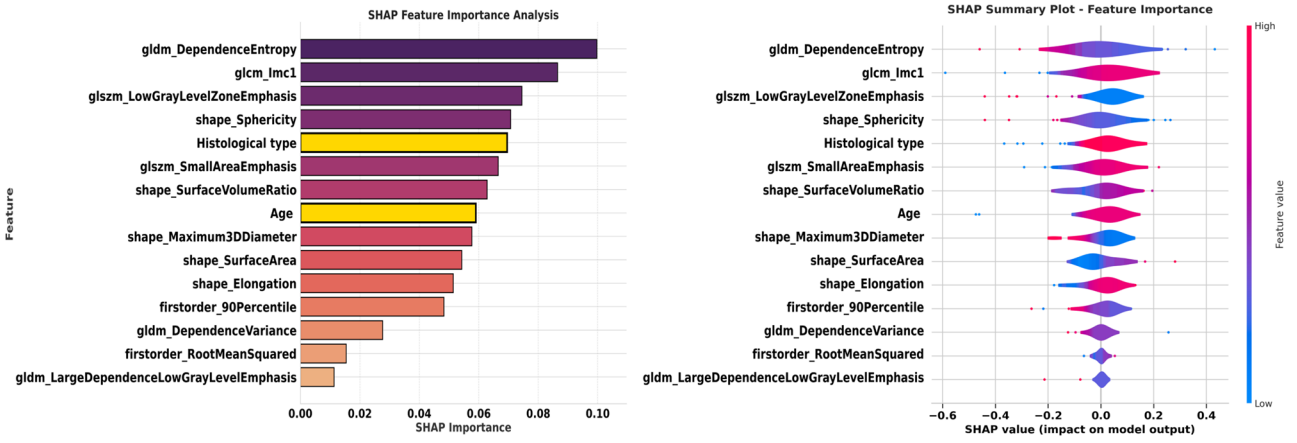


**Fig. 5** SHAP summary plot for the combined radiomic and clinical model. Radiomic features remain the dominant predictors, while age and histological subtype emerge as the most relevant clinical contributors

voxels with similar gray levels, and GLSZM Small Area Emphasis highlights the presence of fine textures by measuring the prevalence of small, homogeneous regions. Morphological features such as shape Sphericity, shape Elongation, and shape SurfaceVolumeRatio were also among the top contributors, and First-order statistics like firstorder 90Percentile played a significant role as well.

In the model with radiomic plus clinical features (Fig. 5) the clinical features were added to the model, radiomics remained dominant in terms of predictive power. However, two clinical variables emerged among the top features: Age and Histological type were selected in the top 10 most impactful features. Despite their inclusion, textural radiomic features such as gldm DependenceEntropy, glcm Imc1, glszm LowGrayLevelZoneEmphasis, and shape Sphericity continued to show the highest SHAP importance values. This analysis confirms that radiomic features are more informative than clinical variables.

## Discussion

This work investigates the potential of radiomic features as a predictive tool for PD-L1 expression in NSCLC patients. PD-L1 is an important biomarker used for predicting the efficacy of immunotherapies [34] and, as a consequence, accurately capturing its presence can significantly impact treatment strategies and outcomes [7, 35–38]. In line with recent advances in medical imaging and ML [39], our study suggests that radiomic features extracted from CT images offer promising results for prediction of PD-L1 expression, providing a potential alternative to more invasive biopsies and offering clinicians a noninvasive tool to support treatment decisions in cases where tissue sampling is limited or risky.

Pairwise Wilcoxon signed-rank tests revealed that the SVM significantly outperformed both the ETC ($p = 0.0098$) and KNN ($p = 0.0371$), whereas differences with other classifiers did not reach statistical significance. Given the limited dataset size and the superior metrics achieved by the SVM—with accuracy and AUC scores

of 0.77 (0.66–0.82) and 0.83 (0.63–0.92), respectively—it emerges as the most robust and competitive choice for this task.

The inclusion of clinical features resulted in diminished performance; furthermore, classifiers trained on clinical variables alone failed to achieve acceptable results. This may be attributed to the insufficient predictive power of these variables for this specific task. When combined with the limited dataset size, their inclusion increases model complexity, leading to overfitting and, consequently, degraded classification outcomes.

It should be noted that two clinical variables listed in Table 1—stage and pleural invasion—were excluded from the analysis: pleural invasion was omitted because it is confirmed by biopsy; stage was not included because most patients in the cohort were diagnosed with stage IV disease, providing little discriminative power for the models.

An important strength of this work is the feature selection process, which played a key role in improving the robustness and generalizability of the model [40, 41]. The same approach was used in [13] to investigate the overall survival of primary central nervous system lymphoma patients, combining multiple selection methods to identify the most relevant features while reducing redundancy and noise. By keeping only features selected by at least two methods, the process helped ensure that the final model was stable, robust, and interpretable. Careful feature selection was particularly important given the limited sample size and contributed significantly to the excellent performance of the best classifiers.

In addition, a small number of patients had CT images with relatively low spatial resolution. Specifically, three patients had a slice thickness of 5.0 mm and one patient had a slice thickness of 2.5 mm. Interestingly, these cases were still classified correctly by the model, suggesting that the predictive algorithm was able to extract robust features even from low-resolution scans. This result reinforces the potential clinical applicability of the model, especially in real-world scenarios where high-resolution imaging is not always available or feasible.

Compared to previous studies, the proposed workflow leads to better classification performance and offers deep insight into the most informative features for the prediction of PD-L1 expression based on CT scans. To provide a clearer comparison with previously published studies, a summary of the current state of the art in radiomics-based PD-L1 prediction in NSCLC (including imaging modality, machine learning approaches, PD-L1 classification thresholds, and performance metrics) is reported in Supplementary table S1.

The first and most important difference concerns the PD-L1 classification threshold. Several radiomics studies classify PD-L1 positivity using lower thresholds (e.g., $\geq 1\%$) [16, 42–45]. In contrast, this study specifically targets high PD-L1 expression ($\geq 50\%$), which represents a clinically meaningful cutoff used to guide treatment decisions. In particular, patients with PD-L1 expression $\geq 50\%$ may receive first-line pembrolizumab monotherapy, as demonstrated in the pivotal study [46], while patients having PD-L1 expression $\geq 50\%$ are candidates for chemotherapy plus immunotherapy. In contrast, $1\%$ is a threshold used to guide second-line treatment strategies, which affect fewer patients than the first line. Several studies based on FDG-PET/CT images and radiomic features alone reported AUC values of 0.74 [15], 0.76 [16] and 0.73 [17]. Among studies based solely on CT images, reported AUC values include 0.72 for Wen et al. [18], 0.81 for Li et al. [15], 0.78 for Bracci et al. [20], 0.66 for Yoon et al. [21], and 0.59 for Yolchuyeva et al. [22]. However, all of the above studies exclusively report AUC as a performance parameter, without reporting complementary indicators such as accuracy or confidence intervals. Importantly, none of the previous studies conducted an in-depth analysis of the importance of features, limiting the understanding of which radiomic features contribute most to the prediction task. In contrast, our study not only achieves comparable or better performance (median AUC = 0.83), but also provides complete methodological transparency by reporting results from multiple classifiers with multiple metrics (AUC, accuracy, f1-score, precision, and recall, including quartiles) and incorporating a robust feature selection process followed by feature importance analysis.

Recent work has highlighted the increasing potential of CT-based radiomics and radiomics-derived signatures for predicting immunotherapy response, including treatment benefit, survival outcomes, and biomarkers such as tumor mutational burden (TMB) in NSCLC patients treated with immune checkpoint inhibitors [47]. Furthermore, systematic reviews have confirmed the feasibility of radiomics models for estimating immunotherapy benefit beyond PD-L1 alone [48]. Emerging evidence also suggests that radiomic features can capture aspects of the tumor immune microenvironment linked to immunotherapy response [49].

To evaluate inter-observer segmentation reproducibility, a random subset of 10 patients was independently segmented by two additional radiologists using the same protocol. Radiomic features were re-extracted from these segmentations, and agreement was quantified using intraclass correlation coefficients (ICC3, two-way mixed-effects model). Segmentation reproducibility analysis showed a median ICC3 of 0.79. Overall, 58% of features demonstrated good-to-excellent agreement (ICC $\geq 0.75$), and 27% achieved excellent agreement (ICC $\geq 0.90$).

The main limitation of this work is the relatively small cohort size, which may affect the generalizability of the

results. The relatively wide interquartile range observed in AUC values reflects variability in model performance across cross-validation folds, which is expected given the limited sample size. In addition, the limited number of patients compared with the dimensionality of the radiomic feature space may increase the risk of model overfitting. Although feature selection strategies were implemented to reduce the number of predictors, the ratio of variables to cases remains relatively low. To mitigate this issue, feature selection was embedded within each cross-validation training fold and a consensus-based multi-method selection strategy was adopted to retain only the most stable features. To further minimize the impact of the small dataset on generalization performance, repeated stratified cross-validation was used. However, larger multicenter datasets will be necessary to confirm robustness and reduce uncertainty in performance estimates.

Although rigorous cross-validation and feature selection were used to mitigate overfitting and improve robustness, independent external validation remains essential. An external cohort would allow verification of predictive performance in a real-world setting.

## Conclusion

In conclusion, this proof-of-concept study highlights the promising role of radiomics for NSCLC, offering a non-invasive approach to infer PD-L1 expression directly from CT images. It also underscores the relevance of the radiomic features that contributed to the final prediction.

Future studies could expand this approach by implementing larger machine learning platforms based on federated learning, involving multicenter collaborations. This strategy may pave the way for further research on image-guided PD-L1 assessment, addressing the sampling limitations of physical biopsies, which may fail to capture the full intratumoral spatial heterogeneity of PD-L1 expression. Such efforts may contribute to enhancing inter-institutional standardization, reducing reliance on human resources, lowering diagnostic costs, and ultimately improving access to molecular predictive diagnostics, a key goal in the era of precision oncology.

### Abbreviations

| | |
|---|---|
| NSCLC | Non-small cell lung cancer |
| PD-L1 | Programmed death-ligand 1 |
| AI | Artificial Intelligence |
| ML | Machine Learning |
| FDG-PET | Fluorodeoxyglucose positron emission tomography |
| CT | Computed Tomography |
| GLCM | Gray Level Co-occurrence Matrix |
| GLRLM | Gray Level Run Length Matrix |
| GLSZM | Gray Level Size Zone Matrix |
| GLDM | Gray Level Dependence Matrix |
| RFE | Recursive Feature Elimination |
| ETC | Extra Trees Classifier |
| SVM | Support Vector Machine |
| LR | Logistic Regression |
| RF | Random Forest |
| KNN | K-Nearest Neighbors |
| XGB | XGBoost |
| AUC | Area Under the Curve |
| SHAP | Shapley Additive Explanations |
| Imc1 | Informational Measure of Correlation 1 |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40644-026-01079-9.

Supplementary Material 1

### Author contributions

M.D. Writing – original draft, Writing – review & editing, Visualization, Validation, Methodology. C.B. : Writing – review & editing, Data curation. G.C. : Writing – review & editing, Investigation, Data curation. M.C. : Writing – review & editing, Data curation. A.P. : Writing – review & editing, Data curation. M.P.: Writing – review & editing, Data curation. D.L.: Writing – review & editing, Resources. PierFrancesco T.: Writing – review & editing, Resources, Conceptualization. PieroSandro T.: Writing – review & editing, Resources, Conceptualization. M.F.S. : Writing – review & editing, Methodology, Conceptualization. P.Z.: Writing – review & editing, Supervision, Resources, Methodology, Investigation, Data curation, Conceptualization. C.C.: Writing – review & editing, Supervision, Resources, Conceptualization.

### Funding

This work was partially supported by the POS RADIOAMICA project, funded by the Italian Ministry of Health.

### Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Declarations

### Ethics approval and consent to participate

All data were anonymized and handled according to ethical guidelines. This observational study was approved by the Ethics Committee of the Calabria Region – Central Area Section (protocol no. 79, dated March 16, 2023). All study steps were performed in accordance with the Declaration of Helsinki.

### Competing interests

The authors declare no competing interests.

Received: 1 November 2025 / Accepted: 18 June 2026
Published online: 26 June 2026

### References

1. Gridelli C, Rossi A, Carbone DP, Guarize J, Karachaliou N, Mok T, et al. Non-small-cell lung cancer. Nat Rev Dis Primers. 2015;1(1):1–16. https://doi.org/10.1038/nrdp.2015.9.
2. Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. Nat Rev Cancer. 2014;14(8):535–46. https://doi.org/10.1038/nrc3775.
3. Thomas A, Liu SV, Subramaniam DS, Giaccone G. Refining the treatment of NSCLC according to histological and molecular subtypes. Nat Rev Clin Oncol. 2015;12(9):511–26. https://doi.org/10.1038/nrclinonc.2015.90.
4. Selvaggi G, Novello S, Torri V, Leonardo E, De Giuli P, Borasio P, et al. Epidermal growth factor receptor overexpression correlates with a poor prognosis in completely resected non-small-cell lung cancer. Ann Oncol. 2004;15(1):28–32. https://doi.org/10.1093/annonc/mdh011.
5. Sacher AG, Gandhi L. Biomarkers for the clinical use of PD-1/PD-L1 inhibitors in non–small-cell lung cancer: a review. JAMA Oncol. 2016;2(9):1217–22. https://doi.org/10.1001/jamaoncol.2016.0639.

6. Yu H, Boyle TA, Zhou C, Rimm DL, Hirsch FR. PD-L1 expression in lung cancer. J Thorac Oncol. 2016;11(7):964–75. https://doi.org/10.1016/j.jtho.2016.04.014.
7. Abdel-Rahman O. Correlation between PD-L1 expression and outcome of NSCLC patients treated with anti-PD-1/PD-L1 agents: a meta-analysis. Crit Rev Oncol/Hematol. 2016;101:75–85. https://doi.org/10.1016/j.critrevonc.2016.03.007.
8. Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, et al. Introduction to radiomics. J Nucl Med. 2020;61(4):488–95. https://doi.org/10.2967/jnumed.118.222893.
9. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imag. 2012;30(9):1234–48. https://doi.org/10.1016/j.mri.2012.06.010.
10. Shimizu H, Nakayama KI. Artificial intelligence in oncology. Cancer Sci. 2020;111(5):1452–60. https://doi.org/10.1111/cas.14377.
11. Cuocolo R, Caruso M, Perillo T, Ugga L, Petretta M. Machine learning in oncology: a clinical appraisal. Cancer Lett. 2020;481:55–62. https://doi.org/10.1016/j.canlet.2020.03.032.
12. van Timmeren Je, Leijenaar RTH, van Elmpt W, Reymen B, Oberije C, Monshouwer R, et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiother Oncol. 2017;123(3):363–69. https://doi.org/10.1016/j.radonc.2017.04.016.
13. Destito M, Marzullo A, Leone R, Zaffino P, Steffanoni S, Erbella F, et al. Radiomics-based machine learning model for predicting overall and progression-free survival in rare cancer: a case study for primary CNS lymphoma patients. Bioengineering. 2023;10(3):285. https://doi.org/10.3390/bioengineering10030285.
14. Zhang L, Du Q, Shen M, He X, Zhang D, Huang X. Interpretable model based on MRI radiomics to predict the expression of ki-67 in breast cancer. Sci Rep. 2025;15(1):13318. https://doi.org/10.1038/s41598-025-97247-1.
15. Li J, Ge S, Sang S, Hu C, Deng S. Evaluation of PD-L1 expression level in patients with non-small cell lung cancer by 18F-FDG PET/CT radiomics and clinicopathological characteristics. Front Oncol. 2021;11:789014. https://doi.org/10.3389/fonc.2021.789014.
16. Zhao X, Zhao Y, Zhang J, Zhang Z, Liu L, Zhao X. Predicting PD-L1 expression status in patients with non-small cell lung cancer using [18F] FDG PET/CT radiomics. EJNMMI Res. 2023;13(1):4. https://doi.org/10.1186/s13550-023-00956-9.
17. Monaco L, De Bernardi E, Bono F, Cortinovis D, Crivellaro C, Elisei F, et al. The "digital biopsy" in non-small cell lung cancer (NSCLC): a pilot study to predict the PD-L1 status from radiomics features of [18F] FDG PET/CT. Eur J Nucl Med Mol Imag. 2022;49(10):3401–11. https://doi.org/10.1007/s00259-022-05783-z.
18. Wen Q, Yang Z, Dai H, Feng A, Li Q. Radiomics study for predicting the expression of PD-L1 and tumor mutation burden in non-small cell lung cancer based on CT images and clinicopathological features. Front Oncol. 2021;11:620246. https://doi.org/10.3389/fonc.2021.620246.
19. Sun Z, Hu S, Ge Y, Wang J, Duan S, Song J, et al. Radiomics study for predicting the expression of PD-L1 in non-small cell lung cancer based on CT images and clinicopathologic features. J X-Ray Sci Technol: Clin Appl Diagnosis Ther. 2020;28(3):449–59. https://doi.org/10.3233/XST-200642.
20. Bracci S, Dolciami M, Trobiani C, Izzo A, Pernazza A, D'Amati G, et al. Quantitative CT texture analysis in predicting PD-L1 expression in locally advanced or metastatic NSCLC patients. Radiol Med. 2021;126(11):1425–33. https://doi.org/10.1007/s11547-021-01399-9.
21. Yoon J, Suh YJ, Han K, Cho H, Lee HJ, Hur J, et al. Utility of CT radiomics for prediction of PD-L1 expression in advanced lung adenocarcinomas. Thorac Cancer. 2020;11(4):993–1004. https://doi.org/10.1111/1759-7714.13352.
22. Yolchuyeva S, Giacomazzi E, Tonneau M, Lamaze F, Orain M, Coulombe F, et al. Radiomics approaches to predict PD-L1 and PFS in advanced non-small cell lung patients treated with immunotherapy: a multi-institutional study. Sci Rep. 2023;13(1):11065. https://doi.org/10.1038/s41598-023-38076-y.
23. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys. 2017;44(3):1050–62. https://doi.org/10.1002/mp.12123.
24. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D slicer as an image computing platform for the quantitative imaging network. Magn Reson Imag. 2012;30(9):1323–41. https://doi.org/10.1016/j.mri.2012.05.001.
25. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77(21):e104–07. https://doi.org/10.1158/0008-5472.CAN-17-0339.
26. PyRadiomics developers. PyRadiomics Documentation. 2025. Accessed: 026-02-05. Available from: https://pyradiomics.readthedocs.io/en/latest/.
27. Developers S-L. StandardScaler — scikit-learn documentation. 2025. Accessed: 2026-02-05. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.
28. Chen X, Jeong JC. Enhanced recursive feature elimination. Sixth international conference on machine learning and applications (ICMLA 2007). IEEE; 2007. p. 429–35.
29. F.Y O, J.E.T A, Awodele A, J. O H, Olakanmi O, Akinjobi A, et al. Supervised machine learning algorithms: classification and comparison. Int J Comput Trends Technol (IJCTT). 2017;48(3):128–38. https://doi.org/10.14445/22312803/IJCTT-V48P126.
30. Developers S-L. RepeatedStratifiedKFold — scikit-learn documentation. 2025. Available from: Accessed: 2026-02-05. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedStratifiedKFold.html.
31. Naidu G, Zuva T, Sibanda EM. A review of evaluation metrics in machine learning algorithms. Computer science on-line conference. Springer; 2023. p. 15–25.
32. Mangalathu S, Hwang SH, Jeon JS. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Eng Struct. 2020;219:110927. https://doi.org/10.1016/j.engstruct.2020.110927.
33. Taheri S, Hesamian G. A generalization of the Wilcoxon signed-rank test and its applications. Stat Papers. 2013;54(2):457–70. https://doi.org/10.1007/s00362-012-0443-4.
34. Sgambato A, Casaluce F, Sacco P, Palazzolo G, Maione P, Rossi A, et al. Anti PD-1 and PDL-1 immunotherapy in the treatment of advanced non-small cell lung cancer (NSCLC): a review on toxicity profile and its management. CDS. 2016;11(1):62–68. https://doi.org/10.2174/1574886311207040289.
35. Brody R, Zhang Y, Ballas M, Siddiqui MK, Gupta P, Barker C, et al. PD-L1 expression in advanced NSCLC: insights into risk stratification and treatment selection from a systematic literature review. Lung Cancer. 2017;112:200–15. https://doi.org/10.1016/j.lungcan.2017.08.005.
36. Lamberti G, Sisi M, Andrini E, Palladini A, Giunchi F, Lollini PL, et al. The mechanisms of PD-L1 regulation in non-small-cell lung cancer (NSCLC): Which are the involved players? Cancers. 2020;12(11):3129. https://doi.org/10.3390/cancers12113129.
37. Passiglia F, Bronte G, Bazan V, Natoli C, Rizzo S, Galvano A, et al. PD-L1 expression as predictive biomarker in patients with NSCLC: a pooled analysis. Oncotarget. 2016;7(15):19738. https://doi.org/10.18632/oncotarget.7582.
38. Ilie M, Hofman V, Dietel M, Soria JC, Hofman P. Assessment of the PD-L1 status by immunohistochemistry: challenges and perspectives for therapeutic strategies in lung cancer patients. Virchows Arch. 2016;468(5):511–25. https://doi.org/10.1007/s00428-016-1910-4.
39. Giger ML. Machine learning in medical imaging. J Am Coll Radiol. 2018;15(3):512–20. https://doi.org/10.1016/j.jacr.2017.12.028.
40. Zhang W, Guo Y, Jin Q. Radiomics and its feature selection: a review. Symmetry. 2023;15(10):1834. https://doi.org/10.3390/sym15101834.
41. Perniciano A, Loddo A, Di Ruberto C, Pes B. Insights into radiomics: impact of feature selection and classification. Multimed Tools Appl. 2024;84(26):31695–721. https://doi.org/10.1007/s11042-024-20388-4.
42. Zhang H, Meng X, Wang Z, Zhou X, Liu Y, Li N. Predicting PD-L1 in lung adenocarcinoma using 18F-FDG PET/CT radiomic features. Diagnostics. 2025;15(5):543. https://doi.org/10.3390/diagnostics15050543.
43. Peng M, Wang M, Yang X, Wang Y, Xie L, An W, et al. Prediction of PD-L1 expression in NSCLC patients using PET/CT radiomics and prognostic modelling for immunotherapy in PD-L1-positive NSCLC patients. Clin Radiol. 2025;86:106915. https://doi.org/10.1016/j.crad.2025.106915.
44. Durmaz M, Emec M, Yirgin IK, Ilgin C, Kaval G, Bunul I, et al. CT-based radiomics for predicting PD-L1 expression status in non-small cell lung cancer using a hybrid machine learning model. Clin Imag. 2026;132:110742. https://doi.org/10.1016/j.clinimag.2026.110742.
45. Li J, Yang Z, Liu L, Tang F, Zheng T, Zhang C, et al. Predicting PD-L1 expression in advanced EGFR-mutant lung adenocarcinoma patients using NECT, CECT radiomics and clinical features. BMC Cancer. 2026;26(1):178. https://doi.org/10.1186/s12885-025-15514-w.
46. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Pembrolizumab versus Chemotherapy for PD-L1–positive non–small-Cell lung cancer. N Engl J Med. 2016;375(19):1823–33. https://doi.org/10.1056/NEJMoa1606774.
47. Cousin F, Louis T, Dheur S, Aboubakar F, Ghaye B, Occhipinti M, et al. Radiomics and delta-radiomics signatures to predict response and survival

in patients with non-small-cell lung cancer treated with immune checkpoint inhibitors. Cancers. 2023;15(7):1968. https://doi.org/10.3390/cancers15071968.
48. Zheng J, Xu S, Wang G, Shi Y. Applications of CT-based radiomics for the prediction of immune checkpoint markers and immunotherapeutic outcomes in non-small cell lung cancer. Front Immunol. 2024;15:1434171. https://doi.org/10.3389/fimmu.2024.1434171.
49. Min J, Dong F, Chen Y, Li W, Wu Y, Tan Y, et al. The NSCLC immunotherapy response predicted by tumor-infiltrating T cells via a non-invasive radiomic approach. Front Immunol. 2024;15:1379812. https://doi.org/10.3389/fimmu.2024.1379812.

## Publisher's Note