Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Construction of a Habitat Analysis-Integrated Radiomics Model to Predict KRAS Mutation Status and risk stratification in Non-Small Cell Lung Cancer

Huan Li
Fourth Affiliated Hospital of Harbin Medical University
Yang Ji
Fourth Affiliated Hospital of Harbin Medical University
Wei Meng
Third Affiliated Hospital of Harbin Medical University
Tong Zhang
yingxiang939@163.com

Fourth Affiliated Hospital of Harbin Medical University

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11444406/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Construction of a Habitat Analysis-Integrated Radiomics Model to Predict *KRAS* Mutation Status and risk stratification in Non-Small Cell Lung Cancer

Huan Li[1+],Yang Ji[1+],Wei Meng[2*+] and Tong Zhang[1,3*+]

[1]Department of Radiology, The Fourth Affiliated Hospital of Harbin Medical University, No. 37, YiYuan Street, NanGang District, Harbin, Heilongjiang, China

[2]Department of Radiology, Harbin Medical University, Harbin Medical University Cancer Hospital, 150 Haping Road, Harbin, Heilongjiang, China

[3]MRI clinical medical research center, MRI clinical medical research center of Heilongjiang, HeilongjiangHarbin, China

+ These authors contributed equally to this work.
*Corresponding authors.

Wei Meng,Department of Radiology, Harbin Medical University Cancer Hospital, 150 Haping Road, Harbin, Heilongjiang, 150081, China Tel: +86-188 4611 5742 E-mail address: articlemengwei@163.com ORCID: 0000-0002-3745-4036

Tong Zhang,Department of Radiology, Fourth Affiliated Hospital of Harbin Medical University, No. 37, YiYuan Street, NanGang District, Harbin, Heilongjiang, 150001, China Tel: 0451-82576831 E-mail address: yingxiang939@163.com ORCID: 0000-0003-1123-7874

Abstract

Objectives To explore the predictive performance of a habitat analysis-integrated radiomics model for identifying *KRAS* mutation status and supporting individualized risk stratification in patients with non-small cell lung cancer (NSCLC).

Methods This retrospective multicenter study enrolled 359 patients with pathologically confirmed NSCLC from two independent medical centers. Three predictive models were built: a clinical-feature-based model, a conventional whole-tumor radiomics model, and a habitat analysis-integrated radiomics model. All CT images were processed using a standardized preprocessing pipeline. Three machine-learning algorithms, including support-vector machines, random forests, and

ExtraTrees classifiers, were applied for model training and optimization. Model discriminative performance was assessed using the area under the receiver operating characteristic curve (AUC). Decision curve analysis (DCA) was performed to quantify the clinical net benefit. An individualized risk-prediction nomogram was constructed based on the optimal multi-feature fusion model.

Results Among all models, the habitat analysis-integrated radiomics model demonstrated the best predictive capability.Habitat-derived features and fused multidimensional features yielded significantly higher AUC values than clinical features alone or conventional whole-tumor radiomic features, indicating superior capacity to capture intratumoral heterogeneity. DCA demonstrated that the fusion model maintained favorable clinical net benefit over a wide range of probability thresholds. The visualized nomogram converted composite indicators of individual patients into cumulative risk scores, which could be further translated into quantitative probabilities of KRAS mutation to enable personalized risk stratification.

Conclusion We developed a habitat analysis-integrated radiomics fusion model allows accurate, non-invasive preoperative detection of KRAS mutation profiles in NSCLC. The corresponding nomogram exhibits promising potential for clinical translation.

Key points

1st Key point: A radiomics model integrated with habitat analysis was developed to preoperatively predict KRAS mutation status in NSCLC using multicenter CT data.

2nd Key point: The habitat-based model outperformed clinical and conventional whole-tumor radiomics models and achieved favorable net clinical benefit by DCA evaluation.

3rd Key point: This non-invasive nomogram enables personalized risk stratification, supporting preoperative molecular phenotyping and clinical decision-making for NSCLC patients.

Introduction

Lung cancer remains the leading cause of cancer-related death worldwide. Non-small cell lung cancer (NSCLC) accounts for more than 85% of all pulmonary malignant tumors, among which lung adenocarcinoma is the most important histological subtype. Kirsten rat sarcoma viral oncogene homolog (KRAS) mutation is one of the most common driver gene alterations in advanced NSCLC, with an incidence rate of about 30%, and there is a significant inter-population difference in the incidence rate between Western populations and Asian populations. For decades, KRAS has been regarded as an "undruggable" oncogenic driver, until researchers found a unique binding pocket in the G12C switch-II domain. This breakthrough promoted the research and development of sotorasib and adagrasib, which fundamentally changed the treatment strategy of KRAS-mutant NSCLC.

Precision oncology highly relies on molecular typing to guide targeted therapy and immunotherapy interventions. However, tissue biopsy, as the gold standard for genetic testing, has obvious clinical limitations: it is an invasive operation, the

ubiquitous intratumoral heterogeneity will bring inevitable sampling bias, and repeated molecular testing will bring huge time and cost burdens. Traditional whole-tumor radiomics extracts global imaging features, ignoring intratumoral subregional differences, which limits its ability to reflect heterogeneous tumor phenotypes. In contrast, radiomics integrated with habitat analysis can divide tumors into multiple subregions, extract region-specific radiomics features, and support refined quantitative assessment of intratumoral heterogeneity.

Most of the existing habitat-driven radiomics studies currently focus on predicting immunotherapy response or EGFR mutation status, and tools specifically used for preoperative KRAS typing are still relatively scarce. To fill this research gap, we conducted a retrospective multicenter study to construct a multi-dimensional prediction model integrating clinical variables, traditional radiomics features and habitat-derived features, aiming to establish a fast, reproducible, low-cost non-invasive auxiliary tool for preoperative KRAS risk stratification, so as to provide reference for individualized clinical decision-making of NSCLC patients.The overall study workflow is presented in Figure 1.

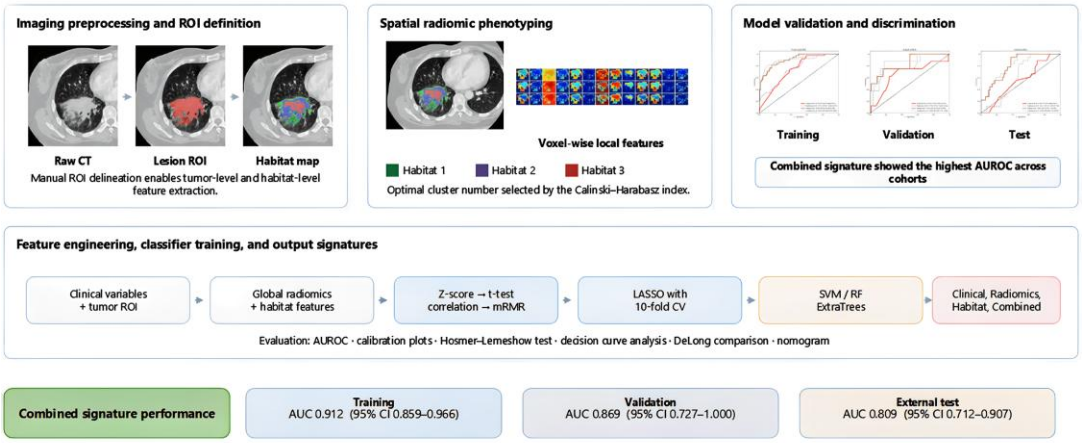

Figure 1. Flowchart of the study

Method

Study Population and Cohort Allocation

Patients who were pathologically diagnosed with NSCLC and received surgery or tissue biopsy in the Fourth Affiliated Hospital of Harbin Medical University from June 2020 to November 2023 were collected as the main cohort of this study (n = 245; training subset: n=172, internal validation subset: n=73). In addition, 114 NSCLC patients who were treated in Harbin Medical University Cancer Hospital from November 2017 to October 2023 were collected as an independent external test set, and a total of 359 patients were finally included. The inclusion criteria are as follows: (a) Thin-slice CT images within 14 days before surgery or biopsy can be obtained from the PACS system; (b) Genetic sequencing is performed on the tumor tissue obtained from surgery or biopsy, and the KRAS mutation status can be obtained; (c) Corresponding clinical data can be found in the electronic medical record system. The exclusion criteria are as follows: (a) No preoperative or pre-biopsy CT images in the PACS system; (b) Poor CT image quality or artifacts; (c) Affected

by factors such as atelectasis, ROI segmentation is difficult; (d) Missing or incomplete clinical data.

Since this study has a multicenter design, a random allocation strategy was adopted to reduce selection bias and ensure the balanced distribution of baseline characteristics of each subset. The total cohort was randomly divided into a training set and an internal validation set according to the ratio of 7:3. This randomization process was repeated iteratively until there was no statistically significant difference in key clinical variables between groups.

All stages of model development, including feature selection, hyperparameter tuning, and classifier training, were carried out only within the training set to avoid data leakage. The internal validation set was used as an independent held-out dataset for unbiased efficacy evaluation. To further evaluate the generalization of the model, this study also included an additional external dataset from different institutions.

Image Acquisition and Preprocessing

Segmentation of Lesions. Regions of interest (ROIs) encompassing the lesions were manually outlined on Computed tomography (CT)images using ITK-SNAP (version 3.8). Two radiologists specializing in chest imaging independently performed the segmentations. Any discrepancies were reconciled through consensus discussion, with a final adjudication provided by a senior radiologist possessing over 20 years of clinical experience. This multi-observer approach was adopted to enhance segmentation consistency and minimize inter-reader variability.

Standardization of Images. All images underwent resampling to achieve a uniform isotropic voxel size of 1 × 1 × 1 mm^3, thereby eliminating geometric discrepancies attributable to different scanners and acquisition settings. Intensity normalization was subsequently performed by adjusting the window width and level to fixed values of 1300 and -300 Hounsfield units, respectively, to standardize soft-tissue contrast and reduce inter-sequence variability. These preprocessing steps were implemented to augment the reproducibility of extracted features and bolster the robustness of subsequent quantitative analyses across diverse clinical centers.

Habitat-Based Radiomic Modeling

Identification of Habitat Subregions. To capture intratumoral heterogeneity, a voxel-wise clustering strategy was employed to delineate regions with homogeneous radiomic profiles. For each voxel, a set of local radiomic features—including first-order statistics and textural metrics derived from gray-level co-occurrence (GLCM), run-length (GLRLM), size zone (GLSZM), and neighborhood gray-tone difference (NGTDM) matrices—were computed within a predefined cubic neighborhood of 5 × 5 × 5 pixels.

Unsupervised K-means clustering was subsequently applied to partition each tumor into candidate habitats, with the cluster count varying between 3 and 10. The optimal number of clusters was ascertained by maximizing the Calinski–Harabasz (CH) index, which strikes a balance between internal cohesion and inter-cluster separation. The resultant habitat maps offer a spatial depiction of distinct radiomic phenotypes

(Supplementary 1).Representative intratumoral habitat patterns are visualized in Figure 2.

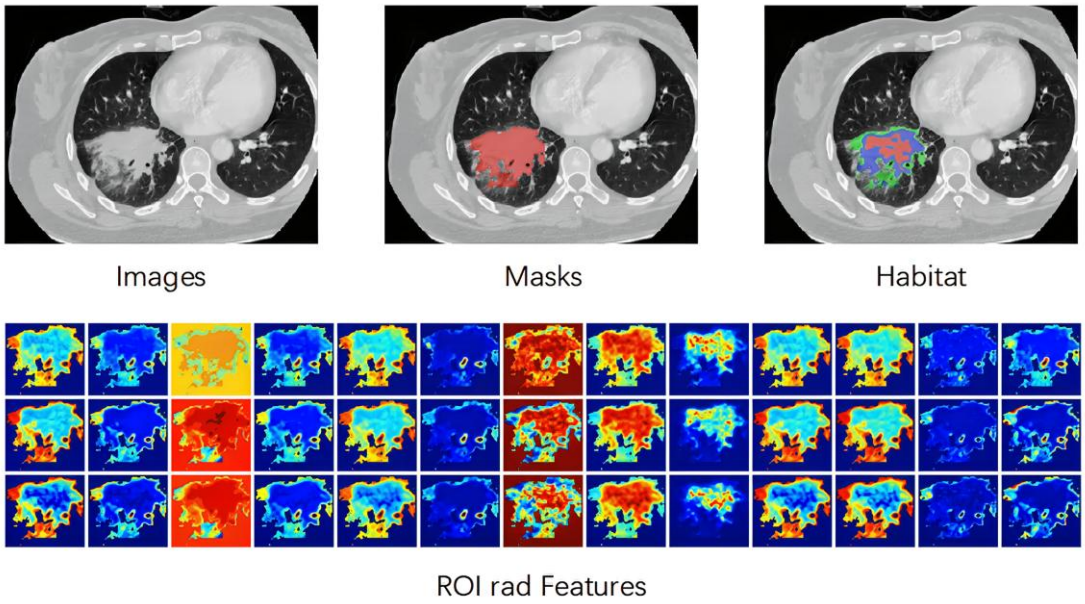

Figure 2. Representative habitat maps derived from voxel-wise clustering, illustrating spatially distinct intratumoral subregions.

Feature Extraction and Engineering

Radiomic attributes were categorized into three principal groups: morphological descriptors (characterizing lesion geometry), intensity-based statistics (summarizing voxel-value distributions), and texture features (quantifying spatial interrelationships via GLCM, GLRLM, GLSZM, and NGTDM).

Features were computed at two spatial scales: globally across the entire lesion, and locally within each habitat-defined subregion. To enhance spatial coherence, cluster assignments were refined using a K-nearest neighbor relabeling procedure. All feature extraction adhered to Image Biomarker Standardisation Initiative (IBSI) guidelines and was executed using PyRadiomics (version 3.0.1).

Habitat-specific signatures were derived by aggregating subregional radiomic features, while an early fusion approach was employed to concatenate features from various spatial compartments into a unified feature vector for subsequent modeling.

Feature Selection Pipeline

All features were standardized via Z-score normalization prior to selection. Univariate t-tests were initially applied to filter features associated with the outcome ($p < 0.05$). To mitigate redundancy, a Pearson correlation analysis was performed; for any pair exhibiting a correlation coefficient exceeding 0.9, the feature with the lower variance was excluded. The minimum redundancy maximum relevance (mRMR) algorithm was then utilized to rank the remaining features, retaining the top 32. Finally, least absolute shrinkage and selection operator (LASSO) regression, with its regularization parameter tuned via 10-fold cross-validation, was applied for further dimensionality reduction. This multi-step procedure yielded a compact, non-redundant set of predictive features.

Model Construction and Validation

Multiple signatures were developed to appraise the predictive contribution of different spatial information. All models conformed to the same preprocessing workflow and were trained using several machine learning algorithms, including support vector machine (SVM), RandomForest, and ExtraTrees.

- **Radiomics Signature:** Constructed using global tumor features. Features demonstrating intraclass correlation coefficients (ICC) > 0.75 were retained.

SVM served as the primary classifier, with Random Forest and ExtraTrees for comparative assessment.

- **Habitat Signature:** Formulated from features extracted from computationally defined habitat subregions; ICC filtering was omitted due to the algorithmic nature of region definition.

- **Clinical Signature:** All available clinical variables were directly employed as input features for model construction. The complete set of candidate variables was used to train the classifier, and the resultant model coefficients defined the clinical signature.

- **Combined Signature:** Integrates clinically significant variables ($p < 0.05$) with optimal habitat signature and radiomics signature. A nomogram was derived from this integrated model to facilitate individualized risk estimation.

Model performance was quantified primarily by the area under the receiver operating characteristic curve (AUC). Calibration was evaluated via calibration plots and the Hosmer–Lemeshow test, while decision curve analysis (DCA) was performed to gauge clinical net benefit.

Statistical Considerations

Statistical computations were performed using Python (version 3.7.12) on the OnekeyAI platform (version 3.1.8). Categorical variables were compared using the chi-square test, and continuous variables using independent-sample t-tests. Radiomic

feature extraction was carried out with PyRadiomics (version 3.0.1). Statistical modeling employed statsmodels (version 0.13.2), while machine learning classifiers were implemented using scikit-learn (version 1.0.2), LightGBM (version 3.3.2), and XGBoost (version 2.1.4). A two-sided p-value < 0.05 was considered statistically significant.

Results

Clinical Feature Analysis

Univariable and Multivariable Analyses. Initial univariable analysis assessed the individual association of each candidate clinical factor with the endpoint, providing odds ratios (ORs) and p-values. Variables with $p < 0.05$ were subsequently incorporated into a multivariable logistic regression model to identify features independently associated with the outcome after adjusting for potential confounders. This sequential strategy ensured that only robust, clinically relevant predictors were retained. Notably, age emerged as an independent predictor in the multivariable model (OR = 0.976, $p = 0.035$). Full outputs from univariable and multivariable logistic regression analyses are summarized in Table 1.

Table 1. Univariable and multivariable logistic regression analyses of clinical variables for outcome prediction.

feature_name	OR_UNI	95% CI_UNI	p_value_UNI	OR_MULTI	95% CI_MULTI	p_value_MULTI
Histologicaltype	0.164	0.0580-0.4630	0.004	0.353	0.1270-0.9840	0.095
Lobulation	0.308	0.2210-0.4290	0.000	0.551	0.1990-1.5250	0.336

feature_name	OR_UNI	95% CI_UNI	p_value_UNI	OR_MULTI	95% CI_MULTI	p_value_MULTI
Sex	0.352	0.2270-0.5460	0.000	1.563	0.7830-3.1240	0.288
Pleuralindentation	0.354	0.2560-0.4890	0.000	2.854	0.8350-9.7470	0.160
Lesiontype	0.363	0.2640-0.4970	0.000	0.925	0.5360-1.5950	0.813
Spiculation	0.370	0.2610-0.5260	0.000	1.979	0.7900-4.9580	0.222
Vacuole_Cavity_Airbronchogram	0.380	0.2440-0.5920	0.000	1.175	0.5830-2.3660	0.705
Smoking_Familytumorhistory	0.409	0.2800-0.5960	0.000	0.908	0.5240-1.5730	0.772
Stage	0.573	0.4720-0.6960	0.000	0.876	0.6210-1.2370	0.529
Lesionlocation	0.730	0.6640-0.8020	0.000	0.912	0.7660-1.0880	0.390
Tumorsize	0.731	0.6650-0.8030	0.000	0.950	0.7630-1.1820	0.699
Age	0.981	0.9760-0.9860	0.000	0.976	0.9580-0.9950	0.035

Habitat Signature Performance

Optimal Cluster Number Determination. The stability of the habitat-based partitioning was evaluated by varying the cluster count. The CH index, a metric of cluster validity, attained its maximum value at k=3, signifying that this configuration yielded the most distinct and well-separated subregions (Figure 3). Consequently, three habitats per lesion were used for all subsequent analyses.

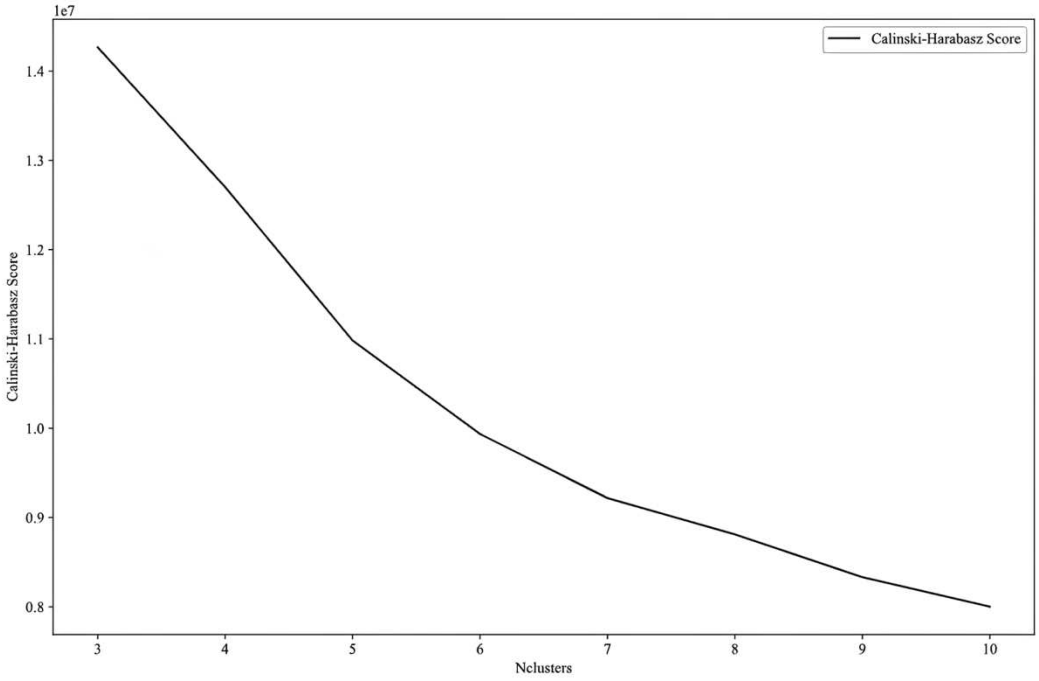

Figure 3. Identification of the optimal habitat cluster number. The Calinski–Harabasz index plotted against cluster counts from 3 to 10; the peak at three clusters indicates the preferred partitioning for subregional modeling.

Extracted Feature Inventory. From each lesion ROI, a total of 1,561 handcrafted radiomic features were extracted, encompassing shape, first-order intensity, and various texture matrix descriptors. Following the delineation of three intratumoral habitats via unsupervised clustering, feature extraction was performed independently for each subregion, yielding 4683 features (1561 × 3) per patient. This approach preserves spatially resolved radiomic information(Figure 4).

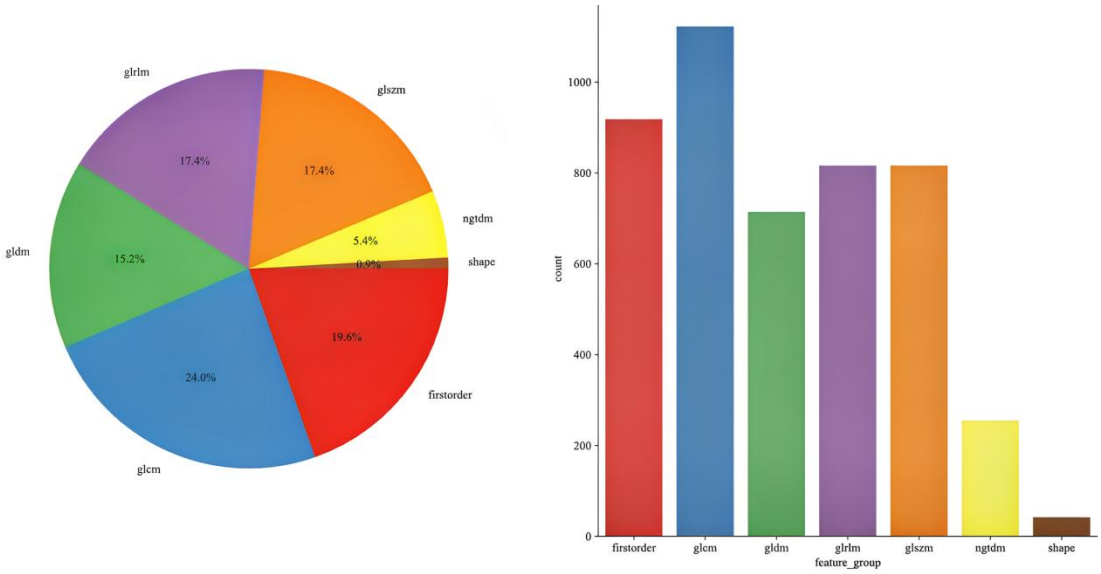

Figure 4. Composition and distribution of the radiomic feature space across different imaging domains and spatial subregions.

LASSO Dimensionality Reduction. To refine the feature set and curtail overfitting, LASSO logistic regression with 10-fold cross-validation was employed. The optimal penalty coefficient (λ) was selected based on the minimum cross-validated mean squared error. Figure 5 illustrates the coefficient trajectories of the candidate features and the MSE curve used for λ selection, alongside a histogram of the final selected features and their prevalence.

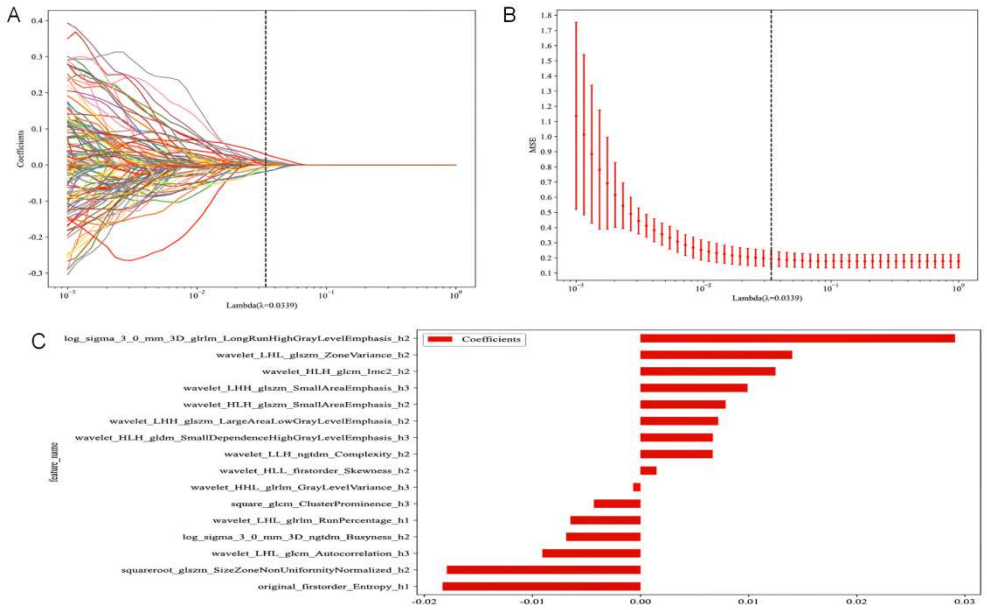

Figure 5. LASSO feature selection process. (A) Coefficient paths for radiomic features; (B) Cross-validation MSE curve for optimal λ determination; (C) Distribution of features retained after LASSO selection.

Among the three machine learning algorithms evaluated, ExtraTrees consistently achieved the highest discriminative performance across all cohorts. In the training cohort, ExtraTrees yielded an AUC of 0.902 (95% CI: 0.850–0.954), surpassing SVM (AUC = 0.773) and Random Forest (AUC = 0.854), with corresponding accuracy of 0.836, sensitivity of 0.816, and specificity of 0.843. This superior performance was maintained in the validation cohort, where ExtraTrees achieved an AUC of 0.863 (95% CI: 0.738–0.988), compared with 0.681 for SVM and 0.688 for Random Forest, along with a sensitivity of 1.000 and specificity of 0.619. In the external test cohort, ExtraTrees again outperformed the other classifiers, attaining an AUC of 0.807 (95% CI: 0.711–0.902), versus 0.670 for SVM and 0.538 for Random Forest, with a sensitivity of 0.944 and specificity of 0.623. Notably, SVM and Random Forest exhibited marked performance degradation from training to test cohorts, with AUC

drops of 0.103 and 0.316, respectively, whereas ExtraTrees demonstrated substantially better generalization with a more modest decline of 0.095.

These findings indicate that the habitat signature, when coupled with the ExtraTrees algorithm, provides robust and generalizable predictive capability across independent datasets. The superior performance and maintained stability of ExtraTrees relative to SVM and Random Forest suggest that ensemble tree-based methods with high randomization may be particularly well-suited for capturing the complex, non-linear relationships inherent in high-dimensional habitat radiomic data. Given its consistent discriminative performance across training, validation, and external test cohorts, the ExtraTrees-based habitat model was selected as the optimal classifier for subsequent comparative analyses. Detailed performance metrics for the three machine-learning classifiers derived from the habitat signature in different cohorts are summarised in Table 2. Corresponding ROC curves of these machine-learning models are displayed in Figure 6.

Table 2. Performance metrics of various machine learning classifiers utilizing the habitat signature across training, validation, and test cohorts.

model_name	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
SVM	0.794	0.773	0.677 - 0.869	0.632	0.843	0.545	0.884	train
SVM	0.451	0.681	0.495 - 0.866	1.000	0.381	0.170	1.000	val
SVM	0.506	0.670	0.530 - 0.811	0.889	0.406	0.281	0.933	test
RandomForest	0.830	0.854	0.787 - 0.922	0.737	0.858	0.609	0.916	train

RandomForest	0.775	0.688	0.502 - 0.875	0.625	0.794	0.278	0.943	val
RandomForest	0.529	0.538	0.395 - 0.681	0.556	0.522	0.233	0.818	test
ExtraTrees	0.836	0.902	0.850 - 0.954	0.816	0.843	0.608	0.939	train
ExtraTrees	0.662	0.863	0.738 - 0.988	1.000	0.619	0.250	1.000	val
ExtraTrees	0.690	0.807	0.711 - 0.902	0.944	0.623	0.395	0.977	test

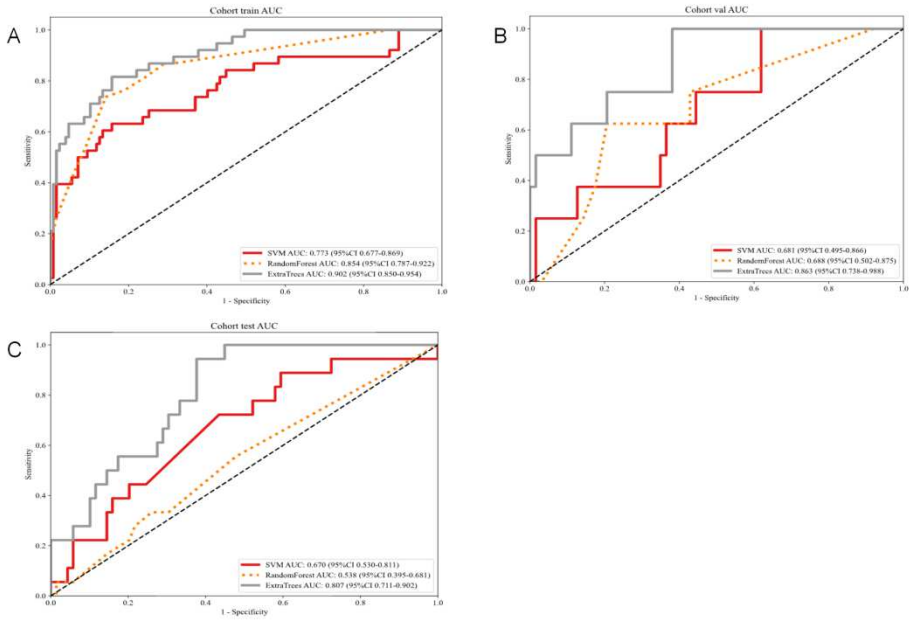

Figure 6. AUROC comparison across the four signatures for the training, validation, and test cohorts.(A) Train cohort; (B) Validation cohort; (C) Test cohort. ROC, receiver-operating characteristic.

Further comparative analyses among the four predictive signatures were then performed. Quantitative performance results of the clinical, radiomics, habitat, and combined signatures are provided in Table 3, while cross-cohort AUROC comparisons are visualized in Figure 7.

Table 3. Comparative predictive performance of the Clinical, Radiomics, Habitat, and Combined signatures.

Signature	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
Clinical	0.570	0.730	0.6489 - 0.8119	0.921	0.465	0.340	0.952	train
Radiomics	0.673	0.742	0.6564 - 0.8275	0.711	0.661	0.386	0.884	train
Habitat	0.836	0.902	0.8505 - 0.9543	0.816	0.843	0.608	0.939	train
Combined	0.891	0.912	0.8587 - 0.9660	0.789	0.921	0.750	0.936	train
Clinical	0.651	0.659	0.4207 - 0.8966	0.750	0.641	0.176	0.962	val
Radiomics	0.826	0.788	0.6095 - 0.9659	0.750	0.833	0.316	0.970	val
Habitat	0.826	0.841	0.6849 - 0.9978	0.750	0.833	0.316	0.970	val
Combined	0.826	0.869	0.7270 - 1.0000	0.875	0.821	0.333	0.985	val
Clinical	0.414	0.544	0.4056 - 0.6830	0.889	0.290	0.246	0.909	test
Radiomics	0.506	0.621	0.5036 - 0.7379	1.000	0.377	0.295	1.000	test
Habitat	0.690	0.807	0.7114 - 0.9022	0.944	0.623	0.395	0.977	test
Combined	0.632	0.809	0.7117 - 0.9066	1.000	0.536	0.360	1.000	test

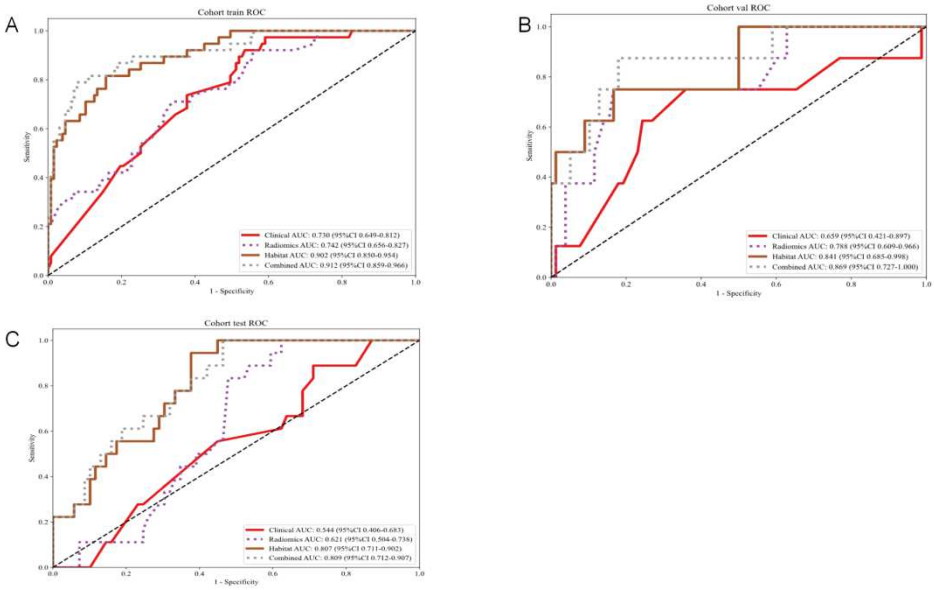

Figure 7. ROC curves for different machine learning models based on the habitat signature across all cohorts. (A) Training cohort; (B) validation cohort; (C) test cohort. ROC, receiver-operating characteristic.

Clinical Utility and Translation

Decision Curve Analysis. The clinical net benefit of each model was assessed via DCA. As shown in Figure 8, the Combined model consistently provided a higher net benefit over a wide range of threshold probabilities in both training and test cohorts, suggesting its superior potential for guiding clinical decisions.

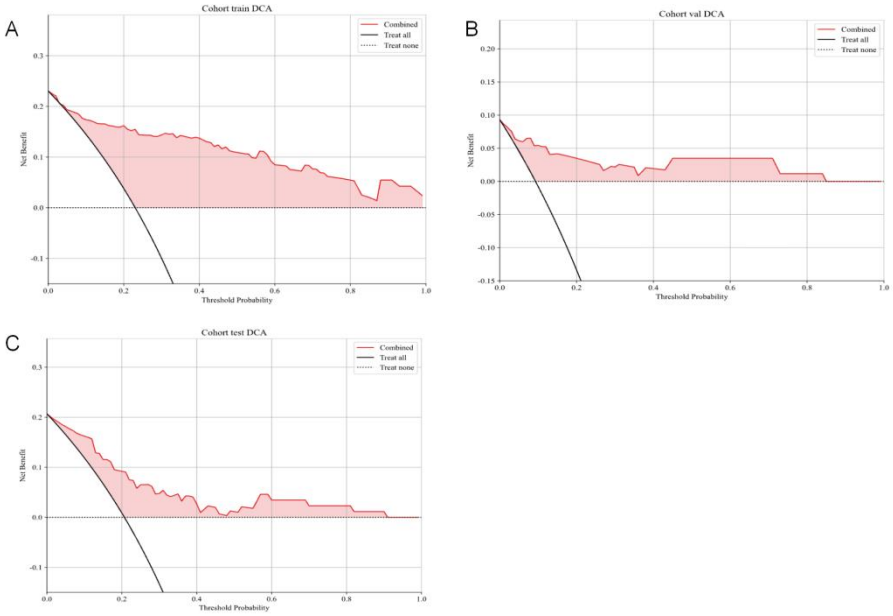

Figure 8. Decision curve analysis for all signatures. The Combined model demonstrates the greatest net benefit across a range of threshold probabilities in both training and test cohorts.(A) Training cohort; (B) validation cohort; (C) test cohort.

Nomogram for Individualized Prediction. To translate the Combined model into a clinically actionable tool, a nomogram was constructed. This visual instrument allows for individualized risk estimation by mapping patient-specific feature values to a total point score and corresponding probability of the outcome, as depicted in Figure 9.

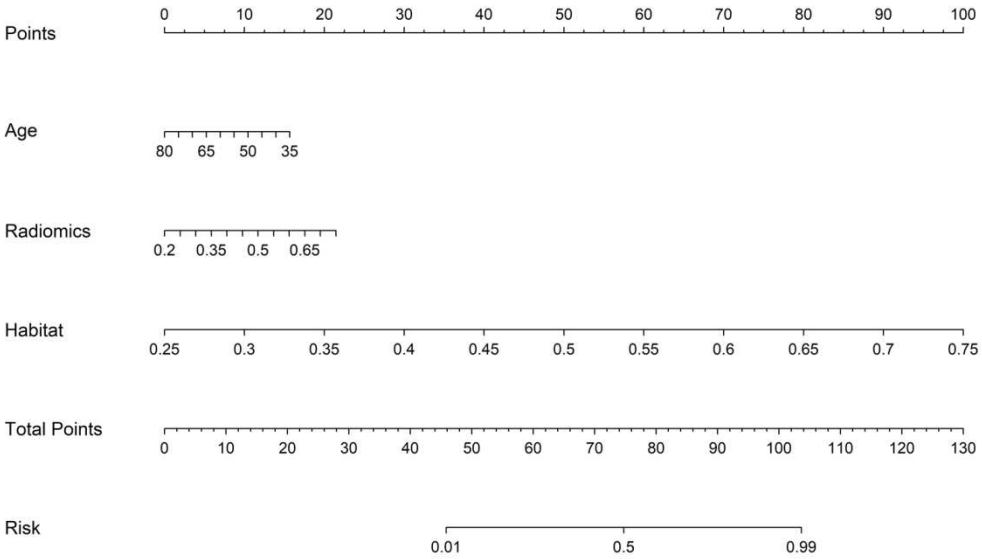

Figure 9. Nomogram based on the Combined signature for individualized risk estimation.

Calibration Assessment. The calibration of the Combined signature was evaluated using the Hosmer–Lemeshow test, yielding non-significant p-values of 0.197, 0.768, and 0.281 for the training, validation, and test cohorts, respectively, indicating excellent agreement between predicted probabilities and observed outcomes. The corresponding calibration curves are displayed in Figure S1.

Pairwise AUC Comparisons (DeLong Test). The DeLong test was employed for pairwise comparisons of AUCs to ascertain statistically significant differences in model performance. In both the training and test cohorts, the Habitat and Combined signatures exhibited significantly higher AUCs relative to the other signatures ($p < 0.05$), underscoring their superior discriminative ability.Pairwise statistical comparisons derived from the DeLong test are visualized as heatmaps in Figure S2.

Additional supporting performance analyses are provided in Figures S3-4 and Table S1–2.

Discussion

Lung cancer remains the leading cause of malignant tumor-related death worldwide. Driven by the development of precision oncology, targeted therapy has become the core treatment strategy for NSCLC patients carrying actionable driver mutations. Existing clinical evidence shows that KRAS mutation is one of the most common genetic abnormalities in advanced NSCLC, which is associated with poor overall prognosis and increased risk of targeted drug resistance [13]. Targeted therapy can significantly improve the survival time of patients with matched genotypes. In the era of single-agent chemotherapy, the median overall survival (OS) of patients with advanced NSCLC was only 7~9 months. In contrast, the novel KRAS G12C targeted inhibitor increased the median OS of treated patients to 17.5 months [14]. Advances in molecular typing, targeted therapy and immunotherapy have reshaped the diagnosis and treatment path of NSCLC, and individualized patient management has become a core clinical priority [15]. Therefore, clarifying the KRAS mutation status before surgery is crucial for formulating a reasonable follow-up treatment plan. However, routine clinical practice faces multiple practical obstacles: widespread intratumoral heterogeneity, sampling bias related to percutaneous biopsy, and high economic cost of repeated genetic testing. Therefore, fast, reproducible, low-cost non-invasive auxiliary biomarkers are urgently needed in clinical practice for preoperative KRAS genotype assessment.

The traditional whole-tumor radiomics workflow only extracts global imaging features, which cannot fully capture intratumoral spatial heterogeneity. A number of previous studies have combined machine learning algorithms with global radiomics features to construct prediction models for EGFR and KRAS mutations in NSCLC [16,17]. However, researchers still question the robustness of such single-region radiomics classifiers, emphasizing that the complex biological behavior driven by KRAS mutations requires subtype stratification analysis [18]. From the perspective of tumor ecology theory, tumors are heterogeneous ecosystems, and the biological phenotypes of different intratumoral subregions are different [19]. Radiomics integrated with habitat analysis can realize tumor subregion segmentation and subregional feature extraction, which is helpful for more accurate quantitative characterization of intratumoral heterogeneity. In this retrospective multicenter study, we constructed three types of prediction models. The results showed that compared with the models constructed only based on clinical variables or traditional radiomics, the model integrated with habitat analysis had higher accuracy and stability in distinguishing KRAS-mutant from wild-type NSCLC. The prediction efficacy of the multi-dimensional fusion model combining habitat features, radiomics features and clinical predictors was further improved. The above results indicate that habitat-derived features can effectively capture subregional heterogeneity information, have unique advantages in non-invasive KRAS typing, and can provide imaging basis for individualized treatment decision-making of NSCLC.

As reported by Ye et al. [20], habitat-based quantitative analysis can quantify intratumoral heterogeneity and predict the pathological complete response rate of neoadjuvant immunochemotherapy. Dong et al. [21] conducted a multicenter retrospective study including 630 pathologically confirmed patients with ground-glass nodules. They identified tumor habitats through K-means clustering, and extracted radiomics features from tumor parenchyma, 1mm peritumoral region, 2mm peritumoral region and habitat regions. This method can reliably quantify intratumoral heterogeneity and accurately distinguish in situ/minimally invasive adenocarcinoma from invasive lesions. Another multicenter study further verified that habitat radiomics can be used for invasive risk stratification of subsolid lung adenocarcinoma [22]. In summary, habitat analysis, as a powerful tool for quantifying tumor heterogeneity and non-invasively predicting tumor biological characteristics, has been increasingly validated.

A large number of preliminary studies have confirmed the potential of radiomics in non-invasively detecting driver gene alterations in NSCLC. Multiple research teams have extracted thousands of CT-derived radiomics features, confirming that radiomics features can predict EGFR and KRAS mutation status, especially suitable for patients who cannot tolerate invasive biopsy [23,24]. With the increasingly wide application of habitat analysis workflow in the field of thoracic oncology, Wu et al. [25] constructed a prediction model based on 248 patients with advanced NSCLC to evaluate the efficacy of immune checkpoint inhibitor therapy, among which habitat radiomics had the best prediction efficacy. Cai et al. [26] further combined deep

learning technology with habitat analysis workflow to predict immunotherapy response and conduct patient risk stratification. In addition, studies have constructed habitat-based models for EGFR mutation prediction and lung adenocarcinoma invasiveness assessment, confirming the feasibility of this technology [27–29]. It is worth noting that most of the published habitat radiomics studies focus on immunotherapy efficacy and EGFR typing, and few studies specifically use radiomics integrated with habitat analysis to predict KRAS mutation status in NSCLC. This multicenter study fills this gap in the existing literature.

On the basis of previous studies, this multicenter study focuses on KRAS driver mutations and uses habitat analysis to mine intratumoral spatial heterogeneity information. To promote clinical transformation, we generated a visualized nomogram based on the optimal multi-feature fusion model. Clinicians can calculate the total risk score through this nomogram, convert the score into the corresponding KRAS mutation probability, so as to carry out individualized risk stratification. In general, the radiomics fusion model integrated with habitat analysis can realize non-invasive, efficient and accurate prediction of KRAS mutation spectrum, provides an alternative evaluation strategy for patients who are not suitable for tissue biopsy, and has good clinical transformation value. Published cohort data show that there is significant subtype-level heterogeneity in KRAS-mutant NSCLC: the G12C subtype accounts for 41% of all KRAS mutations, which is associated with an increased risk of bone metastasis, followed by G12V and G12D subtypes; there is no significant difference in overall survival and progression-free survival among each subtype

406 [30,31]. A number of clinical studies have confirmed that smoking history and
407 occupational exposure are independent risk factors for KRAS mutation. Compared
408 with non-G12C patients, G12C carriers are mostly male, mostly have a smoking
409 history, and have a higher risk of tumor recurrence [32–35]. Consistent with our
410 multivariate regression results (OR=0.976, P=0.035), age is an independent predictor
411 of KRAS mutation. Previous studies have also shown that patients with G12C and
412 G12D mutations have different clinical manifestations and immunotherapy outcomes,
413 and tumors carrying G12D mutations often have poorer response to immunotherapy
414 [36]. In the past, chemotherapy combined with immunotherapy was the main first-line
415 treatment regimen for metastatic NSCLC. At present, targeted monotherapy can bring
416 significant survival benefits to patients with KRAS-mutant metastatic disease [37].
417 Simple binary classification of KRAS mutation status can no longer meet the demand
418 for refined stratified treatment. Based on the fusion model constructed in this study,
419 future research will develop prediction tools for specific subtypes to provide support
420 for stratified precision treatment and prognosis evaluation of patients with
421 KRAS-mutant NSCLC.Several limitations of this study should be acknowledged.
422 First, despite the multicenter design, limited sample size and potential
423 enrollment-related selection bias may restrict the generalizability of our findings.
424 Second, only CT-derived habitat and radiomic signatures were analyzed, without
425 multi-omics or pathological-immunohistochemical datasets; thus, the intrinsic
426 molecular mechanisms linking habitat imaging phenotypes to KRAS mutations

remain unclear. Future work will expand prospective multicenter cohorts and perform prospective trials for systematic model validation.

Conclusions

By constructing a habitat analysis-integrated radiomics fusion mode, it is possible to non-invasively assess KRAS gene mutations in non-small cell lung cancer and identify their risk stratification, thereby guiding clinicians to more comprehensively evaluate the patient's condition and select more appropriate treatment options.

Abbreviations

AUC: Area under the receiver-operating-characteristic curve

CH: Calinski–Harabasz

DCA: Decision-curve analysis

EGFR: Epidermal growth factor receptor

GLCM: Gray-level co-occurrence matrix

GLRLM: Gray-level run-length matrix

GLSZM: Gray-level size zone matrix

IBSI: Image Biomarker Standardisation Initiative

ICC: Intraclass correlation coefficients

KRAS: Kirsten rat sarcoma viral oncogene homolog

LASSO: Least absolute shrinkage and selection operator

mRMR: Minimum redundancy maximum relevance

NGTDM: Neighborhood gray-tone difference matrix

NSCLC: Non-small cell lung cancer

OR: Odds ratio

ROI: Regions of interest

SVM: Support vector machine

Acknowledgements Not applicable.

Authors' contributions

LH, JY, MW, ZT analyzed and interpreted the patient data regarding the Lung cancer. LH,JY,MW performed selection of radiomics features, and LH was a major contributor in writing the manuscript. All authors read and approved the final.

Funding This work was supported by the National Natural Science Foundation of China (Grant No.62376078) .

Data availability The datasets generated and analyzed during the current study are not publicly available due restrictions apply to the availability of these data but are available from the corresponding author on reasonable request.

Declarations This retrospective study was approved by the Research Ethics Committee of the fourth affiliated hospital of Harbin Medical University (Approval No:2024-LLSC-21). Owing to the retrospective nature of the study, informed consent was waived by the Research Ethics Committee of the fourth affiliated hospital of Harbin Medical University, and all clinical data were anonymized during analysis to maintain patient privacy. All methods were performed in accordance with the relevant guidelines and regulations of the Ethics Committee of the fourth affiliated hospital of Harbin Medical University. This study follows the latest revision of the Helsinki Declaration.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. Smolarz B, Łukasiewicz H, Piekarska E, Kołaciński R, Romanowicz H. Lung cancer—epidemiology, pathogenesis, treatment and molecular aspect (review of literature). *Int J Mol Sci*. 2025;26(5):2049. https://doi.org/10.3390/ijms26052049.
2. Siegel RL, Miller KD, Fuchs H, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin*. 2021;71:7–33. https://doi.org/10.3322/caac.21654.
3. Chen Z, Fillmore C, Hammerman P, et al. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat Rev Cancer*. 2014;14:535–546. https://doi.org/10.1038/nrc3775.
4. Guo C, Chénard-Poirier M, Roda D, et al. Intermittent schedules of the oral RAF–MEK inhibitor CH5126766/VS-6766 in patients with RAS/RAF-mutant solid tumours and multiple myeloma: a single-centre, open-label, phase 1 dose-escalation and basket dose-expansion study. *Lancet Oncol*. 2020;21:1478–1488. https://doi.org/10.1016/S1470-2045(20)30446-7.

5. Malhotra J, Nguyen D, Tan T, Semeniuk GB. Management of KRAS-mutated non-small cell lung cancer. *Clin Adv Hematol Oncol*. 2024;22(2):67–75.
6. Keohavong P, Di YP. Pulmonary inflammation and KRAS mutation in lung cancer. *Adv Exp Med Biol*. 2021;1303:71–87. https://doi.org/10.1007/978-3-030-63046-1_5.
7. Piva S, Ganzinelli M, Garassino MC, Caiola E, Farina G, Broggini M, Marabese M. Across the universe of K-RAS mutations in non-small-cell-lung cancer. *Curr Pharm Des*. 2014;20:3933–3943. https://doi.org/10.2174/13816128206661403141645l2.
8. Karimi N, Moghaddam SJ. KRAS-mutant lung cancer: targeting molecular and immunologic pathways, therapeutic advantages and restrictions. *Cells*. 2023;12(5):749. https://doi.org/10.3390/cells12050749.
9. Lv X, Li Y, Wang B, Wang Y, Xu Z, Hou D. Multisequence MRI-based radiomics signature as potential biomarkers for differentiating KRAS mutations in non-small cell lung cancer with brain metastases. *Eur J Radiol Open*. 2024;12:100548. https://doi.org/10.1016/j.ejro.2024.100548.
10. Torres-Jiménez J, Espinar JB, de Cabo HB, et al. Targeting KRASG12C in non-small-cell lung cancer: current standards and developments. *Drugs*. 2024;84:527–548. https://doi.org/10.1007/s40265-024-02030-7.
11. Colombino M, Palomba G, Casula M, Pisano M, Tore S, Serra R, Putzu C, Fadda GM, Maestrale GB, et al. KRAS mutations in non-small cell lung cancer: translational aspects, current therapies and challenges for future research. *Crit Rev Oncol Hematol*. 2025;215:104925. https://doi.org/10.1016/j.critrevonc.2025.104925.
12. Ceddia S, Landi L, Cappuzzo F. KRAS-mutant non-small-cell lung cancer: from past efforts to future challenges. *Int J Mol Sci*. 2022;23(16):9391. https://doi.org/10.3390/ijms23169391.
13. Giopanou I, Pintzas A. RAS and BRAF in the foreground for non-small cell lung cancer and colorectal cancer: similarities and main differences for prognosis and therapies. *Crit Rev Oncol Hematol*. 2020;146:102859. https://doi.org/10.1016/j.critrevonc.2019.102859.
14. Lu S, Wang Y, He J, et al. Glecirasib in KRAS G12C-mutated non-small-cell lung cancer: a phase 2b trial. *Nat Med*. 2025;31(11):1876–1883. https://doi.org/10.1038/s41591-025-03522-7.
15. Jeon H, Wang S, Song J, Gill H, Cheng H. Update 2025: management of non-small-cell lung cancer. *Lung*. 2025;203(1):53. https://doi.org/10.1007/s00408-025-00801-x.

16. Zhang T, Xu Z, Liu G, Jiang B, de Bock GH, Groen HJM, Vliegenthart R, Xie X. Simultaneous identification of EGFR, KRAS, ERBB2, and TP53 mutations by machine learning-derived three-dimensional radiomics. *Cancers (Basel)*. 2021;13(8):1814. https://doi.org/10.3390/cancers13081814.
17. Li J, Xing Y, Gao X, Ye Z, Wang M, Song F. A CT imaging-based deep learning model for predicting EGFR and KRAS mutations in non-small cell lung cancer. *Transl Lung Cancer Res*. 2026;15(1):4. https://doi.org/10.21037/tlcr-2025-1043.
18. Schöneck M, Rehbach N, Lotter-Becker L, Persigehl T, Lennartz S, Caldeira LL. Machine learning-based radiomics analysis for identifying KRAS mutations in non-small-cell lung cancer from CT images. *Life (Basel)*. 2025;15(1):83. https://doi.org/10.3390/life15010083.
19. Gatenby RA, Gillies RJ. A quantitative model of carcinogenesis: the ecological perspective. *Cancer Res*. 2008;68(19):7611–7616. https://doi.org/10.1158/0008-5472.CAN-08-1458.
20. Ye G, Wu G, Zhang C, Wang M, Liu H, Song E, Zhuang Y, Li K, Qi Y, Liao Y. CT-based quantification of intratumoral heterogeneity for predicting pathologic complete response to neoadjuvant immunochemotherapy. *Front Immunol*. 2024;15:1414954. https://doi.org/10.3389/fimmu.2024.1414954.
21. Dong N, Yan Y, Li Y, Li G, Wang P, Li L, Zhang H, Sheng H, Sun X. CT-based habitat radiomics for preoperative differentiation of adenocarcinoma in situ/minimally invasive adenocarcinoma from ground-glass invasive adenocarcinoma. *Front Oncol*. 2025;15:1660071. https://doi.org/10.3389/fonc.2025.1660071.
22. Li Y, Long Y, Zheng Y, Liang J, Lin W, Qing H, Zhou P, Liu J. Ternary-classification habitat model for invasiveness grading of subsolid lung adenocarcinoma on low-dose CT: a multicenter study. *AJR Am J Roentgenol*. 2025;225(5):e2533352. https://doi.org/10.2214/AJR.25.33352.
23. Chen Z, Gao S, Ding C, Luo T, Xu J, Xu S, Li S. CT-based non-invasive identification of common driver gene mutations in NSCLC. *Med Phys*. 2024;51(3):1872–1882. https://doi.org/10.1002/mp.16744.
24. Le NQK, Kha QH, Nguyen VH, Chen YC, Cheng SJ, Chen CY. Machine learning-based radiomics signatures for EGFR and KRAS mutation prediction. *Int J Mol Sci*. 2021;22(17):9254. https://doi.org/10.3390/ijms22179254.
25. Wu Y, Zhang W, Liang X, Zhang P, Zhang M, Jiang Y, Cui Y, Chen Y, Zhou W, Liang Q,et al. Habitat radiomics for predicting immunotherapy efficacy and immune adverse events in advanced NSCLC. *J Transl Med*. 2025;23(1):393. https://doi.org/10.1186/s12967-024-06057-y.

26. Caii W, Wu X, Guo K, Chen Y, Shi Y, Chen J. Deep learning integrated habitat analytical workflows for immunotherapy response prediction in NSCLC. *Cancer Immunol Immunother*. 2024;73(8):153. https://doi.org/10.1007/s00262-024-03724-3.
27. Wu J, Meng H, Zhou L, Wang M, Jin S, Ji H, Liu B, Jin P, Du C. Habitat radiomics fusion nomogram for stage I NSCLC EGFR mutation prediction: a multicenter study. *Sci Rep*. 2024;14(1):15877. https://doi.org/10.1038/s41598-024-66751-1.
28. Shang Y, Zeng Y, Luo S, Wang Y, Yao J, Li M, Li X, Kui X, Wu H, Fan K, et al. Tumor and peritumoral habitat radiomics for preoperative lung adenocarcinoma invasiveness prediction. *AJR Am J Roentgenol*. 2024;223(4):e2431675. https://doi.org/10.2214/AJR.24.31675.
29. Lin J, Yan Z, He L, Zhang H, Xie M. Radiogenomics prediction of *KRAS* and EGFR mutations in patients with non-small cell lung cancer. *Zhong Nan Da Xue Xue Bao Yi Xue Ban*. 2025;50(5):805–814. https://doi.org/10.11817/j.issn.1672-7347.2025.250026.
30. Urtecho SB, Provenzano L, Spagnoletti A, Bottiglieri A, Pircher C, Massa G, Sposetti C, Proto C, Brambilla M, Occhipinti M, et al. Real-world clinicogenomic correlates of KRAS-mutated NSCLC receiving immunotherapy. *Lung Cancer*. 2025;199:108051. https://doi.org/10.1016/j.lungcan.2024.108051.
31. Oya Y, Imaizumi K, Mitsudomi T. Next-generation KRAS inhibitors beyond sotorasib and adagrasib. *Lung Cancer*. 2024;194:107886. https://doi.org/10.1016/j.lungcan.2024.107886.
32. Ricciuti B, Alessi JV, Elkrief A, Wang X, Cortellini A, Li YY, Vaz VR, Gupta H, Pecci F, Barrichello A, et al. Clinicogenomic and immunophenotypic differences across KRAS mutant subtypes in NSCLC. *Ann Oncol*. 2022;33(10):1029–1040. https://doi.org/10.1016/j.annonc.2022.07.005.
33. Serna-Blasco R, Medina K, Sala MÁ, Aguiar D, Díaz-Serrano A, Antoñanzas M, González Larriba JL, Ocaña J, Mielgo X, Fernández I, et al. Occupational exposure correlates with KRAS mutational status in NSCLC. *Chest*. 2025;168(6):1528–1543. https://doi.org/10.1016/j.chest.2025.06.043.
34. Tamiya Y, Matsumoto S, Zenke Y, Yoh K, Ikeda T, Shibata Y, Kato T, Nishino K, Nakamura A, Furuya N, et al. Large-scale clinico-genomic profiling of KRAS G12C-mutated NSCLC in the LC-SCRUM-Asia cohort. *Lung Cancer*. 2023;176:103–111. https://doi.org/10.1016/j.lungcan.2022.12.019.
35. Cao H, Ma Z, Li Y, Zhang Y, Chen H. Radiological and prognostic characteristics of KRAS G12C-mutated lung adenocarcinoma stratified by

disease stage. *J Thorac Cardiovasc Surg*. 2023;166(6):e479–e499. https://doi.org/10.1016/j.jtcvs.2023.04.037.

36. Shahnam A, Davis A, Brown LJ, Sullivan I, Lin K, Ng C, Yeo N, Khoo T, Warburton L, Da Silva IP, et al. Real-world survival comparison between KRAS G12C and G12D mutant NSCLC patients. *Lung Cancer*. 2025;201:108421. https://doi.org/10.1016/j.lungcan.2025.108421.
37. Harris E, Thawani R. Contemporary management landscape of KRAS-mutated NSCLC. *Curr Probl Cancer*. 2024;51:101106. https://doi.org/10.1016/j.currproblcancer.2024.101106.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile.docx