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.

CT radiomics combined with 3D deep learning interpretable model to predict the efficacy of PD-1 inhibitor plus platinum-based chemotherapy in advanced NSCLC: a three-center retrospective cohort study

Weiming Zhao
The First Affiliated Hospital of Bengbu Medical University
Xinyu Lu
The First Affiliated Hospital of Bengbu Medical University
Xilong Zhou
The First Affiliated Hospital of Bengbu Medical University
Wenjing Fan
The First Affiliated Hospital of Bengbu Medical University
Yang Luo
The First Affiliated Hospital of Anhui Medical University
Qiuyang Zhu
North University of China
Yaqi Wang
The First Affiliated Hospital of Bengbu Medical University
Huanhuan Zhou
The First Affiliated Hospital of Bengbu Medical University
Yuxin Zheng
The First Affiliated Hospital of Bengbu Medical University
Shuhua Li
The First Affiliated Hospital of Bengbu Medical University
Wei Li
The First Affiliated Hospital of Bengbu Medical University
Fang Su
The First Affiliated Hospital of Bengbu Medical University
Zongyu Xie
Xiezongyu2@163.com

The First Affiliated Hospital of Bengbu Medical University

Research Article

Keywords: Non-small cell lung cancer, Immunochemotherapy resistance, Multimodal artificial intelligence, Radiomics, Deep learning, Treatment response prediction

Posted Date: November 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11255070/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Platinum-based chemotherapy combined with PD-1 inhibitors serves as standard systemic treatment for unresectable stage IIIB–IV non-small-cell lung cancer (NSCLC) in both first-line and later-line settings; however, primary resistance causes substantial treatment-response heterogeneity. Tissue-based biomarkers including PD-L1 expression and tumor mutational burden (TMB) are limited by invasive biopsy requirements and high costs, whereas routine thin-slice chest CT non-invasively captures intratumoral spatial heterogeneity. We constructed a multimodal interpretable AI model integrating clinical variables, three-dimensional (3D) radiomics and pre-trained 3D-DenseNet latent features to predict immunochemotherapy response and evaluate its cross-center generalizability.

Methods

We retrospectively enrolled 378 advanced NSCLC patients receiving PD-1 inhibitor plus platinum-based chemotherapy from three tertiary hospitals. Three-dimensional regions of interest (ROIs) of primary tumors were manually double-blinded segmented. Following three-step feature filtering, four models were built: clinical-only, radiomics-only, standalone 3D deep learning (DL), and multimodal fusion. Center 1 data were split into training (n = 66) and internal validation (n = 35) sets; Centers 2 and 3 formed two fully independent external test sets (n = 81 and n = 196). Model performance was evaluated via area under the receiver operating characteristic curve (AUC) and decision curve analysis (DCA). Ablation, multi-classifier and subgroup analyses assessed robustness. SHAP and 3D Grad-CAM enabled model interpretation; TCGA-TCIA datasets supported immunological mechanism exploration, and an interactive prediction platform was developed.

Results

Only pathological subtype and treatment line exerted clinical predictive effects. The fusion model achieved competitive AUCs: training 0.909, internal validation 0.946, external-test-1 0.848, external-test-2 0.893. Although standalone 3D DL attained marginally higher AUC in training cohort, fusion outperformed clinical-only and radiomics-only models across all cohorts (all $P < 0.05$). 3D DL features represented the major predictive contributor. The model maintained AUC > 0.87 across most subgroups; calibration and DCA verified favorable consistency and clinical net benefit. SHAP-Grad-CAM visualized decision logic, and transcriptomic data associated CT-defined high-risk phenotypes with immunosuppressive microenvironments and reduced $CD8^+$ T-cell infiltration.

Conclusions

This low-cost interpretable multimodal AI based on routine chest-CT non-invasively predicts immunochemotherapy resistance for advanced NSCLC across treatment lines. Validated by two external cohorts, it demonstrates complementary predictive value of radiomic and DL signatures and offers a standardized imaging tool for personalized immunotherapy stratification.

Introduction

Lung cancer remains the leading cause of cancer incidence and mortality worldwide. NSCLC accounts for ~ 85% of all lung cancer cases, and about 70% of patients present with unresectable stage IIIB-IV disease at diagnosis (1, 2). Combining immune checkpoint inhibitors (ICIs) with platinum-based chemotherapy is the standard systemic therapy for advanced NSCLC in first-line and selected later-line settings and significantly prolongs progression-free survival (PFS) and overall survival (OS) (3, 4). However, only 40%-55% of patients achieve durable disease control (DCR), while nearly half develop primary immune resistance and experience rapid tumor progression (5, 6). This resistance not only deprives patients of survival benefits but also causes irreversible immune-related adverse reactions and medical costs, highlighting the urgent need for non-invasive pre-treatment response stratification tools with multi-center reproducibility (7, 8).

Current predictive biomarkers have inherent limitations. PD-L1 expression, tumor mutational burden (TMB) and driver-gene alterations are assessed from invasive biopsy and often yield false-negative results due to intratumoral heterogeneity (9–11). Peripheral blood inflammatory markers are susceptible to confounding factors such as infection and hormone administration, with poor cross-cohort repeatability (12). Multicenter retrospective cohorts frequently lack standardized molecular testing data, as in our three centers, retaining only basic clinical records without gene or PD-L1 results (13, 14). Sequencing-based assays such as TMB testing are also prohibitively expensive for routine use in resource-limited settings (15).

Pre-treatment thin-slice chest plain CT is a standard examination for advanced NSCLC that captures intratumoral three-dimensional (3D) heterogeneity and serves as a carrier for predictive models (16, 17).

CT-based analysis falls into handcrafted radiomics and 2D/3D deep learning (DL). Although many single-center studies have investigated imaging predictors of immunotherapy response, most share fundamental limitations constraining their innovation and rigor (18, 19). Few models combine interpretable handcrafted texture features with non-linear 3D deep latent representations; handcrafted radiomics quantifies macro tumor fibrosis/necrosis, while pre-trained 3D-DenseNet mines micro volumetric immune microenvironment signals, yet their complementary effects are rarely validated (20, 21). Moreover, most studies randomly split single-center data without fully independent cross-regional external test sets, increasing overfitting risk and limiting generalizability (22, 23). A complete interpretability and robustness chain is also often lacking: few studies combine SHAP and Grad-CAM to address DL's black-box nature or perform ablation, multi-classifier and subgroup analyses (24–26). Most

imaging studies remain purely statistical without transcriptomic annotation, failing to link CT phenotypes with anti-tumor immune pathways.

To address these gaps, we developed a multimodal fusion model integrating clinical variables, 3D handcrafted radiomics, and latent features from a MedicalNet-pretrained 3D-DenseNet, based on a retrospective cohort of 378 patients from three tertiary hospitals. Using single-center data for training/internal validation and two independent external cohorts for generalizability assessment, we assessed discrimination, calibration, clinical utility, interpretability, robustness, and transcriptomic relevance. The dual-branch 3D fusion framework extracted handcrafted radiomic and DL features in parallel, with ablation experiments confirming their complementary predictive value.

An explainable AI (XAI) workflow incorporating SHAP and 3D Grad-CAM was implemented to dissect global and pixel-level predictive signals and alleviate DL black-box limitations. Closed-loop robustness assessments covering multiple classifiers, clinical subgroups and feature ablation confirmed that model performance was not tied to any specific algorithm or patient subset. We further used public TCGA-TCIA paired transcriptomic-imaging datasets to map CT-derived tumor heterogeneity onto immune-related pathways without requiring molecular specimens. Finally, we constructed a nomogram, performed decision curve analysis (DCA), and built a lightweight interactive web platform for clinical translation.

We hypothesized that the multimodal fusion model would yield better discrimination between primary immunochemotherapy resistance and disease control than clinical-only, radiomics-only, or standalone 3D DL models; preserve stable AUCs across two geographically separated external cohorts, with 3D deep latent features as the primary predictor and radiomics as complementary; and that our XAI pipeline could localize intratumoral CT heterogeneity driving model outputs. Finally, we hypothesized that elevated CT-detected tumor heterogeneity would correlate with an immunosuppressive tumor microenvironment.

Methods

2.1 Study Population and Ethical Approval

This study enrolled 378 patients with unresectable stage IIIB/IIIC/IV NSCLC who received PD-1 inhibitor plus platinum-based chemotherapy between January 2019 and December 2024 at three tertiary hospitals (Center 1, n = 101; Center 2, n = 81; Center 3, n = 196). All patients underwent 1-mm thin-slice plain chest CT within 7 days before treatment. The regimen was considered first-line if given as initial systemic anti-tumour therapy and later-line if after prior systemic regimens. ECOG performance status was recorded. Inclusion criteria: (1) histologically/cytologically confirmed NSCLC; (2) stage IIIB/IIIC/IV without radical surgical indication; (3) PD-1 inhibitor plus platinum doublet as first or subsequent regimen; (4) 1-mm thin-slice chest CT within 1 week before treatment; (5) complete baseline records, follow-up ≥ 12 weeks, and evaluable response per RECIST 1.1.

Exclusion criteria: (1) prior thoracic radiotherapy or systemic anticancer therapy; (2) severe CT artifacts precluding complete 3D ROI segmentation; (3) follow-up < 12 weeks or loss to follow-up without imaging response; (4) long-term high-dose glucocorticoid use or concurrent autoimmune disease. The protocol was approved by each center's institutional review board; informed consent was waived due to the retrospective design.

Center abbreviations: Center 1, The First Affiliated Hospital of Bengbu Medical University; Center 2, The First Affiliated Hospital of Anhui Medical University; Center 3, Shandong Provincial Cancer Hospital.

2.2 CT Acquisition, Preprocessing and Tumor Delineation

Non-contrast chest CT used 64/256-slice spiral scanners (120 kV, 100–150 mA, 1-mm slice thickness/interval, 350 mm FOV, 512×512 matrix). Mediastinal sequences were used for manual segmentation of the primary tumor ROI in ITK-SNAP, excluding mediastinal lymph nodes, pleural effusion, and peritumoral normal lung. Two radiologists (> 5 years'thoracic experience) independently performed double-blind 3D delineation; 20% of cases were re-segmented to assess inter-rater consistency (ICC > 0.75 acceptable), with disagreements resolved by consensus. Preprocessing included resampling to isotropic 1-mm^3 voxels, Z-score normalization, artifact removal, and retention of intra-ROI data.

2.3 Dual-branch Feature Extraction: Handcrafted 3D Radiomics+3D Deep Latent Features

2.3.1 Handcrafted 3D Radiomic Features

Using Python 3.7.12 and PyRadiomics on the OmniMedAI platform, 1834 3D radiomic features (first-order, morphology, GLCM/GLRLM/GLSZM/GLDM textures, wavelet high-order) were extracted.

2.3.2 3D Deep Learning Latent Features

The intra-ROI CT volume was cropped to the tumor bounding box, resized to 96×96×96 voxels, and voxel-wise Z-score normalized. A MedicalNet-pretrained 3D-DenseNet121 was used with the backbone frozen and only fully connected layers fine-tuned on the training cohort (Adam, learning rate = 1×10^{-4}, batch size = 8, epochs = 100, categorical cross-entropy loss). A 128-dimensional nonlinear latent feature vector was extracted from the penultimate fully connected layer as abstract volumetric signatures of the tumor microenvironment (27).

2.4 Three-Step Feature Screening Pipeline

To avoid data leakage, feature reduction was performed only on the training cohort. (1) Univariate logistic regression (LR) retained features with $P < 0.05$; (2) among features with pairwise Pearson correlation $|r|\geq0.8$, the one with the smaller P-value was kept; (3) LASSO with 10-fold cross-validation retained non-zero coefficients using λ_{min}. Selection was applied separately to clinical, radiomic, and DL feature pools, and filtered subsets were concatenated into the final multimodal set.

2.5 Dataset Partition and Class Imbalance Correction

Center 1 (n = 101) was stratified by treatment response and randomly split into training (n = 66, 70%) and internal validation (n = 35, 30%). Center 2 (n = 81) and Center 3 (n = 196) served as two independent external test cohorts without parameter tuning. The binary endpoint was DCR, comprising complete response (CR), partial response (PR), or stable disease (SD), versus progressive disease (PD), defined as primary immune resistance. Response was assessed at the 12-week follow-up per RECIST 1.1, the routine first imaging time-point, to identify primary rather than acquired resistance. PD was the positive class. Because the training set was imbalanced, SMOTE oversampling was applied to augment the minority DCR class (28, 29). Internal and external validation cohorts retained the raw real-world distribution without resampling.

2.6 Four Comparative Predictive Models

Four LR-based models with distinct feature inputs were constructed: (1) clinical-only (baseline variables with univariate significance: pathological subtype and treatment line); (2) radiomics-only (LASSO-selected Rad-score); (3) 3D DL (latent features from 3D-DenseNet121); (4) multimodal fusion (concatenated clinical, radiomic, and DL features), the primary model.

2.7 Comprehensive Model Evaluation System

Bootstrap resampling (1000 iterations) estimated AUC with 95% CIs; the DeLong test compared AUCs. Accuracy, sensitivity, specificity, PPV, NPV, F1-score, and MCC were computed.

2.7.1 Discriminative ability and incremental value

ROC curves assessed the fusion model's incremental value over single-modal models in each cohort.

2.7.2 Calibration assessment

Calibration curves and the Hosmer-Lemeshow test assessed predicted vs. observed probabilities ($P > 0.05$ indicating good calibration).

2.7.3 Clinical decision curve and nomogram construction

DCA evaluated net clinical benefit across thresholds versus "treat all" and "treat none". A nomogram was derived from the fusion model's regression coefficients for bedside risk scoring.

2.7.4 Model robustness validation

(1) The four feature sets were also fitted with six additional classifiers (K-nearest neighbors (KNN), support vector machine (SVM), random forest (RF), extra-trees, naive Bayes, XGBoost) to test classifier dependence.

(2) Subgroup analysis stratified by age, sex, smoking history, pathological type, TNM stage, and ECOG in the merged external cohort reported bootstrap AUC with 95% CI per subgroup.

(3) Each ablation removed one feature block (clinical, radiomic, or DL) from the final multimodal set, leaving other blocks and settings unchanged; no re-selection was performed.

2.7.5 Interpretable artificial intelligence analysis

A two-tier XAI pipeline was used. (1) SHAP global interpretation: mean bar and swarm plots were used to visualize global feature contributions and nonlinear relationships. (2) Local interpretation: SHAP waterfall plots for representative DCR and primary-resistant cases decomposed per-feature contributions. Grad-CAM generated heatmaps from 3D DL outputs to localize tumor regions driving model decisions.

2.8 TCGA-TCIA Transcriptomic Mechanism Analysis

Matched transcriptomic-imaging LUAD/LUSC data were downloaded from TCGA and TCIA. Fusion-model risk scoring stratified TCGA samples into high/low CT risk subgroups. DESeq2 was used for DEG screening; GO/KEGG enrichment and GSEA focused on T-cell exhaustion, IFN-γ response, and tumor fibrosis pathways. ssGSEA quantified immune cell infiltration to compare the immune microenvironment between risk subgroups.

2.9 Interactive Web Prediction Platform

A lightweight web interface based on the LR fusion model accepted clinical and imaging inputs and output predicted probabilities of all four models for one-click risk evaluation.

2.10 Statistical Software

Delineation used ITK-SNAP; feature extraction and the 3D DL network were implemented via the OmniMedAI platform (Based on Python 3.9, PyRadiomics and PyTorch, https://www.omnimai.com). Statistical analyses, plots, survival analysis, SHAP, and DCA were conducted using R version 4.2.1. All statistical tests were two-sided, with $P < 0.05$ defined as statistically significant.

Figure 1 illustrates the patient screening workflow and analytical pipeline. This study is reported in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement.

Results

3.1 Baseline Cohort Characteristics and Univariate Clinical Screening

3.1.1 Clinical Baseline Features

A total of 378 stage IIIB/IIIC/IV NSCLC patients receiving PD-1 combined platinum chemotherapy were enrolled, including 168 DCR patients and 210 primary-resistant patients; the 12-week progression rate was 55.6%. Baseline demographic and clinical information of training, internal validation and two external cohorts are summarized in Table 1. Inter-cohort comparison showed significant differences in

age, smoking history, ECOG score, treatment line and treatment response (all P<0.05), while sex and pathological/TNM distribution exhibited no statistical disparity (P>0.05). Univariate LR (Figure 2A) showed that only pathological subtype and treatment line were significantly associated with disease control. Of note, despite response-stratified random splitting, the internal validation cohort had only two progressive-disease patients because of the limited Center 1 sample size, resulting in 94.3% disease-control cases; this imbalance should be considered when interpreting metrics such as positive predictive value.

3.1.2 Radiomic Feature Screening

Three-step filtering of the 1834 raw 3D radiomic features retained eight independent predictive signatures: wavelet_LLH_ngtdm_Coarseness, wavelet_LLH_glcm_DifferenceVariance, lbp_3D_k_glrlm_ShortRunEmphasis, lbp_3D_k_glszm_HighGrayLevelZoneEmphasis, exponential_glszm_GrayLevelNonUniformityNormalized, lbp_3D_m1_glszm_GrayLevelVariance, log_sigma_2_0_mm_3D_glcm_Correlation, and exponential_gldm_DependenceNonUniformityNormalized. The category composition of the initial feature set and LASSO regularization-path diagrams are presented in Supplementary Figure 1.

3.1.3 Multicollinearity Control of Final Feature Pool

Pairwise correlation heatmap (Figure 2B) and VIF distribution plot (Figure 2C) showed all Pearson coefficients<0.6 and VIF<5 among filtered clinical, radiomic, and DL features, indicating no severe multicollinearity and supporting stable modeling.

3.2 Discriminative Performance of Four Modal Models Across All Cohorts

Table 2 summarizes the discriminative metrics of the four models across all cohorts, and Figure 3 displays the ROC curves and confusion matrices. In the training set, the standalone 3D DL model achieved the highest AUC (0.945), followed by the fusion model (0.909), whereas the clinical-only model performed worst (0.694). In internal validation, the fusion model reached its maximum AUC of 0.946 (95% CI: 0.856–0.987), outperforming the DL model (0.909). In external test set 1 (Center 2) and test set 2 (Center 3), it maintained robust AUCs of 0.848 (95% CI: 0.755-0.915) and 0.893 (95% CI: 0.837-0.936), while the clinical-only model declined to 0.562 (95% CI: 0.487-0.634) in test set 2, indicating poor cross-center generalizability.

The fusion model outperformed the clinical-only and radiomics-only models in all cohorts (DeLong, all P<0.05). Compared with the standalone DL model, it achieved higher AUCs in internal validation and test set 1 but was comparable in test set 2 (P>0.05), suggesting that fusion improved generalization without compromising the dominant DL contribution.

3.3 Calibration, DCA and Nomogram of Fusion Model

Calibration curves of the two external cohorts (Figure 4A-C) showed close overlap between predicted risk and actual event rate, with non-significant Hosmer-Lemeshow tests (P>0.05). DCA (Figure 4D-F) revealed superior net clinical benefit of the fusion model over “treat all/treat none” strategies across all probability thresholds. The nomogram (Figure 4G) translated regression coefficients into bedside scoring rules for rapid individualized prediction.

3.4 SHAP Global Contribution and Ablation Analysis

SHAP mean bar and swarm plots (Figure 5A,B) consistently indicated that 3D deep latent features carried the largest global predictive weight, followed by radiomic signatures, while clinical variables contributed marginally. Ablation heatmap (Figure 5C) showed that removing DL features caused the steepest drops in AUC, F1, and MCC; removing radiomic features caused moderate decay, whereas removing clinical variables barely altered discrimination.

3.5 Case-Specific Local Interpretable Visualization

Two representative patients (DCR, Figure 6A; primary resistant, Figure 6B) were selected for local XAI with paired 3D Grad-CAM heatmaps and SHAP waterfall plots. Responder tumors exhibited widespread high-activation Grad-CAM signals, with negative SHAP contributions from DL features pushing predictions toward DCR. Resistant tumors showed only scattered low-weight heatmap regions, with positive DL-feature SHAP contributions driving prediction toward PD.

3.6 TCGA-TCIA Transcriptomic Mechanism Validation

Utilizing the TCGA-LUAD and TCGA-LUSC public transcriptome datasets in conjunction with TCIA imaging data (after excluding cases with missing data or failed delineation; n=98; LUAD: 60; LUSC: 38), we replicated the fusion model’s risk stratification logic to define high- and low-imaging-risk subgroups. This yielded 62 DEGs ($P<0.05$, $|logFC|>1$; Figure 7A). Kaplan-Meier analysis (Figure 7B) showed significantly shorter overall survival in the high-risk subgroup (log-rank $P<0.0001$). GO bubble chart (Figure 7C) showed enrichment mainly in cilium assembly and microtubule movement. GSEA (Figure 7D,E) revealed overrepresentation of antigen receptor signaling, leukocyte-mediated immunity, and allograft rejection pathways. ssGSEA (Figure 7F) demonstrated decreased effector CD8+ and memory T cells with elevated regulatory T cells in the high-risk subgroup.

3.7 Cross-Classifier Robustness Analysis

To assess algorithm dependence, the clinical-only, radiomics-only, and multimodal fusion models were tested across seven classifiers; the 3D DL model, based on a single fixed DenseNet121 architecture, was not subject to this multi-classifier test.

As summarized in Supplementary Table 1 and visualized in the performance radar chart (Supplementary Figure 2), clinical-only models yielded poor AUCs (0.560-0.614) and occasionally degenerated to single-class predictions. Radiomics-only models showed high variability, peaking only with XGBoost

(AUC=0.895). The fusion model maintained stable AUCs (0.866-0.915) across all seven classifiers, with the primary LR model achieving an AUC of 0.893 and XGBoost reaching 0.915. This consistency confirms the algorithm-independent predictive advantage of tri-modal fusion.

3.8 Clinical Subgroup Stratification Performance

Subgroup analysis stratified by age, gender, smoking, histology, TNM, and ECOG was performed on the combined external cohorts, with subgroup AUCs listed in Table 3 and corresponding forest plot+probability violin plots in Supplementary Figure 3. All subgroups maintained AUC>0.87 except the ECOG 2 subgroup (n=13; AUC 0.722, 95% CI: 0.476–1.000). Stage IIIB/IIIC and ECOG 0 achieved optimal discrimination (AUC 0.925 and 0.936). Violin plots showed minimal overlap between DCR and PD predicted probability distributions, confirming stable performance across heterogeneous real-world populations.

3.9 Interactive Prediction Web Tool

Supplementary Figure 4 demonstrates the practical application of the web platform based on the LR fusion model. Entering routine clinical variables, radiomic features, and the DL score with a single click immediately provides predicted probabilities of all four models, enabling rapid pre-treatment identification of patients at high risk of primary resistance.

Discussion

4.1 Intrinsic Complementary Logic of Multimodal Feature Fusion

Most prior models relied on a single imaging modality or single-center internal validation, risking overfitting and poor external reproducibility. We integrated handcrafted 3D radiomics with pretrained 3D-DenseNet latent features and validated the model in two geographically independent external cohorts, reducing selection bias from scanning equipment and population heterogeneity.

Ablation and SHAP analyses showed that the two imaging branches were complementary: handcrafted radiomics captured macroscopic fibrotic and necrotic phenotypes, whereas deep networks extracted subtle nonlinear volumetric signals invisible to the naked eye; combining both produced consistent discrimination gains across all test cohorts (30–32). Cross-classifier experiments further showed that clinical-only and radiomics-only inputs were unstable across seven classifiers, whereas the fusion model maintained stable AUCs (0.866–0.915) regardless of classifier, indicating that complementary feature integration—rather than any single feature family—provided robust, standardized input for multicenter modeling (33, 34).

Although the fusion model had a slightly lower training AUC than the standalone 3D DL model, it achieved comparable or superior AUCs in internal and external cohorts, suggesting improved cross-

center generalizability and reduced overfitting.

The internal-validation cohort contained only two PD cases, explaining its low positive predictive value as a consequence of extreme class imbalance rather than model weakness.

4.2 Clinical Transformation Value of Complete Interpretable Pipeline

The black-box nature of DL impedes clinical adoption. Whereas most prior studies reported only global feature importance, our three-level interpretability framework combined SHAP global statistics, single-case waterfall plots, and 3D Grad-CAM heatmaps, enabling clinicians to trace each prediction and enhancing acceptability. Together with calibration curves, DCA, a nomogram, and an interactive web tool, this formed a complete clinical translation chain.

Because the pipeline requires only routine baseline records and thin-slice CT, without costly PD-L1 or sequencing assays, it may be suitable for primary hospitals and large retrospective multicenter cohorts.

4.3 Biological Insights from TCGA-TCIA Cohort Analysis

Because our retrospective cohort lacked paired tumor sequencing, we used public TCGA/TCIA datasets to explore the biological basis of the CT-defined risk phenotype. Transcriptomic analysis showed that the high-risk subgroup had enrichment of T-cell activation pathways but significantly reduced effector CD8 + T-cell infiltration, suggesting an immune microenvironment with enhanced activation signals yet impaired effector function—an exhausted state during immune editing. These findings provide indirect transcriptomic evidence for the chain “imaging heterogeneity-immune escape-treatment resistance” and may partly explain the poor response to immunochemotherapy in high-risk patients (35).

4.4 Clinical Guidance from Subgroup Results

Subgroup analysis was robust across age, sex, smoking history, histology, TNM stage, and ECOG 0–1 strata (AUC > 0.87); only the ECOG 2 subgroup had a lower AUC (0.722) due to its small size. Thus, the model is most reliable in fit patients (ECOG 0–1) and stage IIIB/IIIC disease. For patients with poor performance status, predictions should complement comprehensive clinical evaluation rather than serve as the sole basis for decisions.

4.5 Comparative Advantages Versus Published Literature

Three strengths distinguish this study from prior imaging-based studies: (1) two fully independent external test cohorts rather than single-center random splits, reducing overfitting risk and strengthening generalizability evidence; (2) a dual-branch 3D architecture combining handcrafted radiomics and DL features, with ablation quantifying their complementarity; and (3) an integrated analytical chain covering robustness, interpretability, clinical utility, and transcriptomic mechanisms. In addition, whereas many published models require molecular biomarkers, our pipeline relies only on routine clinical data and CT, lowering the barrier for real-world implementation.

4.6 Limitations and Future Perspectives

Several limitations should be acknowledged. The retrospective design may have introduced selection bias, and prospective multicenter validation is needed. Only the primary tumor ROI was segmented; peritumoral stroma and mediastinal lymph nodes were not included, potentially omitting systemic immune-related imaging signals. Standardized PD-L1 and TMB data were unavailable, preventing head-to-head comparison with established molecular biomarkers. The web tool is currently a local prototype not integrated with hospital PACS or electronic medical record systems.

Future work should include prospective multicenter validation; expansion of segmentation to peritumoral and nodal regions; centralized collection of PD-L1 and sequencing data for direct comparison with the imaging model; integration of biopsy specimens with CT imaging to validate imaging-immunology associations; and development of an automated, PACS/EMR-integrated clinical interface.

Conclusions

Using a three-center cohort of 378 advanced NSCLC patients, we developed an interpretable multimodal AI model integrating clinical baseline, 3D radiomics, and 3D-DenseNet deep features to non-invasively predict primary resistance to PD-1 plus platinum chemotherapy from standard thin-slice CT alone. Validated by two independent external cohorts and robustness tests, radiomic and DL features provided complementary gains beyond clinical variables. Combined with hierarchical XAI and transcriptomic evidence, this low-cost pipeline offers a reproducible framework with discrimination, generalizability, and interpretability for individualized immunotherapy stratification. Prospective validation is warranted before clinical deployment.

Declarations

1. Ethics approval and consent to participate

This retrospective multicenter study was led and ethically reviewed by the Institutional Review Board of The First Affiliated Hospital of Bengbu Medical University (Ethics approval number: 2023-440). The other two participating centers, The First Affiliated Hospital of Anhui Medical University and Shandong Provincial Cancer Hospital, provided paper-based institutional approval for participation in this multicenter project. Written informed consent was waived for this retrospective analysis using de-identified medical records. All study procedures were carried out in accordance with the principles of the 1964 Helsinki declaration and its subsequent amendments.

2. Consent for publication

Not applicable. No identifiable individual patient data, clinical images or case-specific personal information are presented in this manuscript.

3.Availability of data and materials

The datasets generated and analysed during the current study are not publicly available due to hospital patient-privacy restrictions. Reasonable requests for de-identified aggregated data may be addressed to the corresponding author upon justified request. The publicly available TCGA-TCIA dataset used in this work is accessible via their official repository.

4. Competing interests

The authors declare that they have no competing interests.

5.Funding

This study was supported by the Anhui Provincial Medical Imaging Discipline (Professional) Leader Cultivation Project (DTR2024028 to Zongyu Xie), the Longhu Talent Project of Bengbu Medical University (LH250302002 to Zongyu Xie).

This work was supported by the Anhui Provincial Health Research Key Project Supported by Provincial Fiscal Funds (AHWJ2023A10110 to Fang Su)

This study was supported by Excellent Research and Innovation Team of the Anhui Provincial Department of Education (2025AHGXZK10019 to Wei Li) and Anhui Transformation Program of Clinical Medicine Research (202304295107020081 to Wei Li)

6.Authors' contributions

W.Z. and X.L. contributed equally to this work. W.Z., X.L., W.L., F.S., and Z.X. conceived and designed the study. W.Z., X.L., X.Z., W.F., Y.L., Y.W., H.Z., and Y.Z. collected and curated the clinical and imaging data. X.Z., W.F., Y.L., Y.W., H.Z., and Y.Z. performed tumor segmentation and image preprocessing. Q.Z. developed the imaging data dimensionality reduction platform used for radiomic and deep learning feature processing. W.Z., X.L., and Z.X. developed the prediction models and performed statistical analyses. W.Z., X.L., and F.S. conducted the interpretability and transcriptomic analyses and drafted the manuscript. S.L. and W.L. contributed to clinical data interpretation and critical revision of the manuscript. F.S., W.L., and Z.X. supervised the project and acquired funding. All authors reviewed and approved the final manuscript.

7. Acknowledgements

Not applicable.

8. Authors' information

Not applicable.

References

1. Pasquier C, Khalifa J. Traitement standard du cancer bronchique non à petites cellules localement avancé non résécable. Bull Cancer. 2025;112(3):S352–5363. 10.1016/s0007-4551(25)00158-4.
2. Liang B, Tong C, Nong J, Zhang Y. Histological Subtype Classification of Non-Small Cell Lung Cancer with Radiomics and 3D Convolutional Neural Networks. J Imaging Inf Med. 2024;37(6):2895–909. 10.1007/s10278-024-01152-4.
3. Kang J, Zhang J, Tian Z, Xu Y, Li J, Li M. The efficacy and safety of immune-checkpoint inhibitors plus chemotherapy versus chemotherapy for non-small cell lung cancer: An updated systematic review and meta-analysis. PLoS ONE. 2024;19(2):e0276318. 10.1371/journal.pone.0276318.
4. Ferrara R, Imbimbo M, Malouf R, Paget-Bailly S, Calais F, Marchal C, et al. Single or combined immune checkpoint inhibitors compared to first-line platinum-based chemotherapy with or without bevacizumab for people with advanced non-small cell lung cancer. Cochrane Database Syst Reviews. 2021;2021(4). 10.1002/14651858.CD013257.pub3.
5. He K, Xu F, Xiang L, Luo Y. Albumin–globulin ratio is a predictive biomarker of antitumour effect of immune checkpoint inhibitors in cancer patients. Ann Med. 2025;57(1). 10.1080/07853890.2025.2591219.
6. Kaushik I, Ramachandran S, Zabel C, Gaikwad S, Srivastava SK. The evolutionary legacy of immune checkpoint inhibitors. Sem Cancer Biol. 2022;86:491–8. 10.1016/j.semcancer.2022.03.020.
7. Zhu Y, Duong L, Lu X, Lu X. Cancer-cell-intrinsic mechanisms shaping the immunosuppressive landscape of prostate cancer. Asian J Androl. 2023;25(2):171–8. 10.4103/aja202283.
8. Hong X, Hong T, Liu X, Yao P, Lin W, Liu G, et al. UBE2C-Mediated CD147-CTLA-4 Axis Promotes T Cell Exhaustion and Immunosuppressive Microenvironment in Bladder Cancer. Cancer Sci. 2026. 10.1111/cas.70430.
9. Scafetta R, Donato M, Curigliano G, Pantano F. Immunotherapy in rare histologies of breast cancer: challenges, opportunities, and future perspectives. Curr Opin Oncol. 2025;37(6):570–9. 10.1097/cco.0000000000001184.
10. Jiang Y, Wang N, Zeng Q, Liu J, Liu X, Cao G, et al. From Variant to Biomarker in NSCLC Immunotherapy Resistance: Multiomics Evidence Chains and Accountable AI Integration. Hum Mutat. 2026;2026(1). 10.1155/humu/6434376.
11. Lyu S, Reynders K, Laven IEWG, Faes C, Weynand B, Hillen LM, et al. Intratumor immune heterogeneity across simulated core biopsy sizes in non-small cell lung cancer: an exploratory study. Translational Lung Cancer Res. 2026;15(7):200–200. 10.21037/tlcr-2026-1-0017.
12. Lu W, Waters M, Larby J, Shibeeb S, Bozinovski S, Selemidis S, et al. Peripheral Blood Inflammatory Markers and Immune Cell Profiles in Smokers With and Without COPD. Pulmonary Therapy. 2026;12(2):637–52. 10.1007/s41030-026-00367-w.
13. Zhao P, Zhao S, Huang C, Li Y, Wang J, Xu J, et al. Efficacy and Safety of Polatuzumab Vedotin Plus Rituximab, Cyclophosphamide, Doxorubicin and Prednisone for Previously Untreated Diffuse Large

B-Cell Lymphoma: A Real-World, Multi-Center, Retrospective Cohort Study. Hematol Oncol. 2024;43(1). 10.1002/hon.70017.

14. Shen C-I, Chiang C-L, Shiao T-H, Luo Y-H, Chao H-S, Huang H-C, et al. Real-world evidence of the intrinsic limitations of PCR-based EGFR mutation assay in non-small cell lung cancer. Sci Rep. 2022;12(1). 10.1038/s41598-022-17394-7.
15. Fang J, Li Q, Xu N, Yang X, Zhang Q, Chen Y, et al. Systemic inflammation biomarkers can identify high tumor mutation burden in lung adenocarcinoma. BMC Cancer. 2025;25(1). 10.1186/s12885-025-14894-3.
16. Li J, Shi Q, Yang Y, Xie J, Xie Q, Ni M, et al. Prediction of EGFR mutations in non-small cell lung cancer: a nomogram based on 18F-FDG PET and thin-section CT radiomics with machine learning. Front Oncol. 2025;15. 10.3389/fonc.2025.1510386.
17. Haro A, Wakasu SHO, Kozuma Y, Tsukamoto S, Hamatake M. Investigation of the Biological Properties of Non-small Cell Lung Cancer Using Three-dimensional Computed Tomography Images. Cancer Diagnosis Prognosis. 2025;5(4):453–60. 10.21873/cdp.10458.
18. Wang C, Li Y, Ji Y, Yu K, Qin C, Liu L, et al. Predictions of Response in Non-small Cell Lung Cancer Patients Treated with Immune Checkpoint Inhibitors Using Clinical Data, Deep Learning, and Radiomics. Acad Radiol. 2026;33(1):236–54. 10.1016/j.acra.2025.09.037.
19. Vanguri RS, Luo J, Aukerman AT, Egger JV, Fong CJ, Horvat N, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. Nat Cancer. 2022;3(10):1151–64. 10.1038/s43018-022-00416-8.
20. Lee J, Kolluru C, Gharaibeh Y, Prabhu D, Zimin VN, Bezerra H, et al. Automatic A-line coronary plaque classification using combined deep learning and textural features in intravascular OCT images. Proc SPIE Int Soc Opt Eng. 2020;11315:1131513. 10.1117/12.2549066.
21. Shaukat A, Farhan S, Javed R, Idris S, Alamri FS, Ara A, et al. HyFusion-X: hybrid deep and traditional feature fusion with ensemble classifiers for breast cancer detection using mammogram and ultrasound images. Sci Rep. 2025;15(1). 10.1038/s41598-025-22262-1.
22. Elahi R, Karami P, Amjadzadeh M, Nazari M. Radiomics-based artificial intelligence (AI) models in colorectal cancer (CRC) diagnosis, metastasis detection, prognosis, and treatment response prediction. Abdom Radiol. 2025;51(4):2153–66. 10.1007/s00261-025-05201-6.
23. Sharma A, Sharma A, Arora P, Kumar A, Sharma AK. Triple-Negative Breast Cancer and Artificial Intelligence: Current Paradigms in Diagnosis, Therapy Prediction, Prognosis, and Challenges. Curr Pharm Biotechnol. 2026;27. 10.2174/0113892010413048251103143500.
24. Esser-Skala W, Fortelny N. Reliable interpretability of biology-inspired deep neural networks. npj Syst Biology Appl. 2023;9(1). 10.1038/s41540-023-00310-8.
25. Guo KH, Chaudhari NN, Jafar T, Chowdhury NF, Bogdan P, Irimia A. Alzheimer's Disease Neuroimaging Initiative. Anatomic Interpretability in Neuroimage Deep Learning: Saliency Approaches for Typical Aging and Traumatic Brain Injury. Neuroinformatics. 2024;22(4):591–606. 10.1007/s12021-024-09694-2.

26. Barić D, Fumić P, Horvatić D, Lipic T. Benchmarking Attention-Based Interpretability of Deep Learning in Multivariate Time Series Predictions. Entropy. 2021;23(2). 10.3390/e23020143.
27. Chen S, Ma K, Zheng Y. Med3D. Transfer Learning for 3D Medical Image Analysis. arXiv preprint arXiv:1904.00625; 2019. https://doi.org/10.48550/arXiv.1904.00625
28. Zhou Y, Wu J, Xu X, Shi G, Liu P, Jiang L. Investigating perioperative pressure injuries and factors influencing them with imbalanced samples using a Synthetic Minority Over-sampling Technique. Biosci Trends. 2025;19(2):173–88. 10.5582/bst.2025.01013.
29. Liu B, Cai Q, Zhao X, Su H, Lin Z, Wu J, et al. A nomogram based on radiomic features from peri-prostatic adipose tissue for predicting bone metastasis in first-time diagnosed prostate cancer patients. Adipocyte. 2025;14(1). 10.1080/21623945.2025.2517583.
30. Guo L, Fu Y, Tan S, Wang Q, Zhang Y, Huang X, et al. RadioGuide-DCN: A Radiomics-Guided Decorrelated Network for Medical Image Classification. Bioengineering. 2025;13(1). 10.3390/bioengineering13010046.
31. Ho WLJ, Fetisov N, Hall LO, Goldgof D, Schabath MB. Utilizing Clinicopathological and Radiomic Features for Risk Stratification of Lung Cancer Recurrence. Acad Radiol. 2025;32(9):5557–66. 10.1016/j.acra.2025.04.062.
32. Du L, Tang G, Che Y, Ling S, Chen X, Pan X. Fusing radiomics and deep learning features for automated classification of multi-type pulmonary nodule. Med Phys. 2025;52(7). 10.1002/mp.17901.
33. Komissarov L, Manevski N, Groebke Zbinden K, Sach-Peltason L. Explainable Graph Neural Networks in Chemistry: Combining Attribution and Uncertainty Quantification. J Chem Inf Model. 2025;65(14):7516–28. 10.1021/acs.jcim.5c01003.
34. Li Y, Liu J, Wen X, Niu X, Wei L, Yuan J, et al. Augmenting medical image specific foundation model with classical radiomic signatures for improved nasopharyngeal carcinoma stage classification. Front Med Technol. 2026;8. 10.3389/fmedt.2026.1863203.
35. Cai MC, Zhao X, Cao M, Ma P, Chen M, Wu J, et al. T-cell exhaustion interrelates with immune cytolytic activity to shape the inflamed tumor microenvironment. J Pathol. 2020;251(2):147–59. 10.1002/path.5435.

Tables

Tables 1 to 3 are available in the Supplementary Files section.

Figures

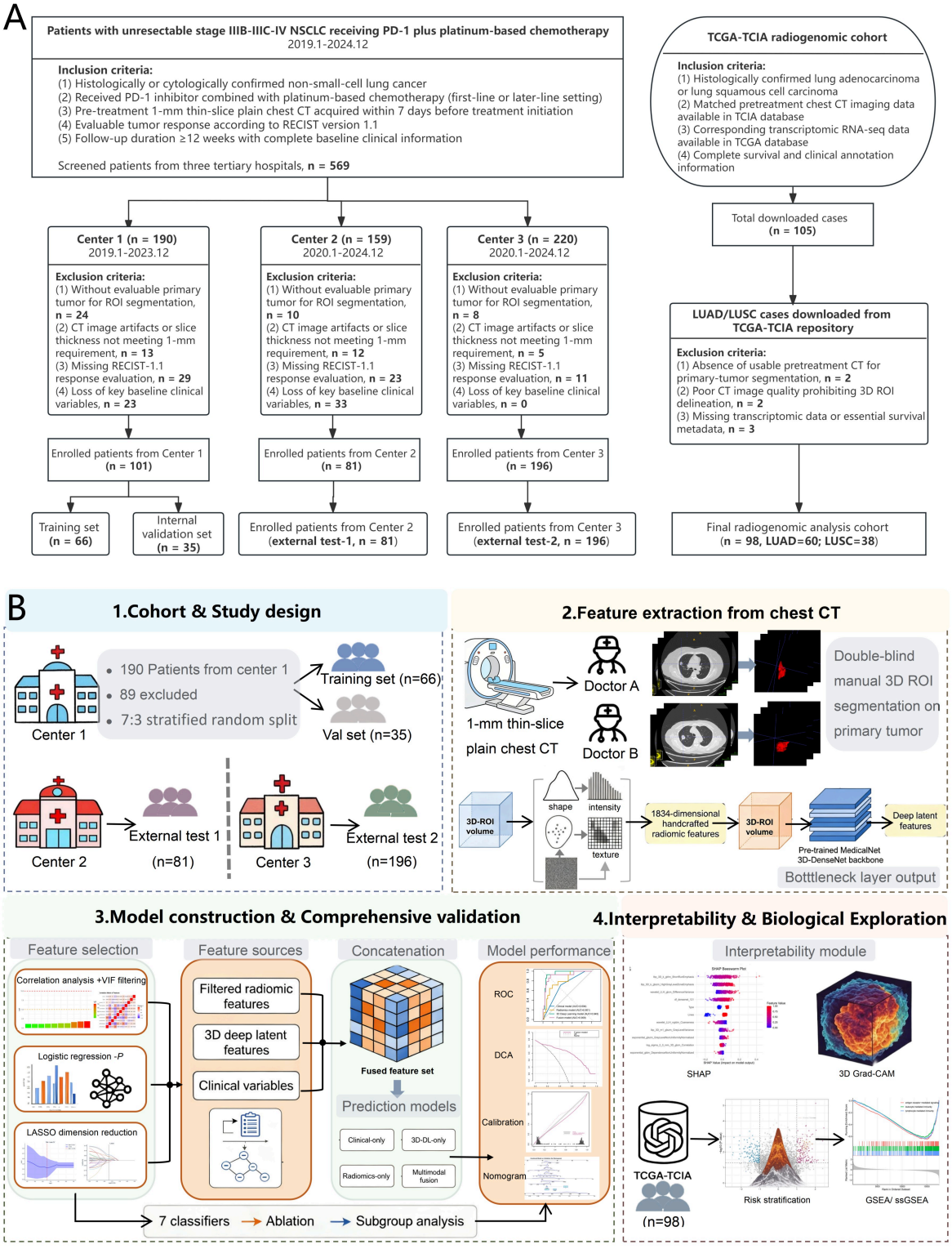


Figure 1

Study recruitment and analytical workflow. (A) Flowchart illustrating patient screening, inclusion-exclusion criteria, and cohort partitioning for the three-center retrospective NSCLC cohort, including sample sizes for training, internal-validation and two independent external test sets. (B) Schematic overview of the analytical pipeline: CT preprocessing, dual-branch feature extraction, multi-step feature filtering, multimodal feature concatenation, comparative model construction,

multi-dimensional performance assessment, XAI interpretation, TCGA-TCIA transcriptomic exploration and web-tool development.

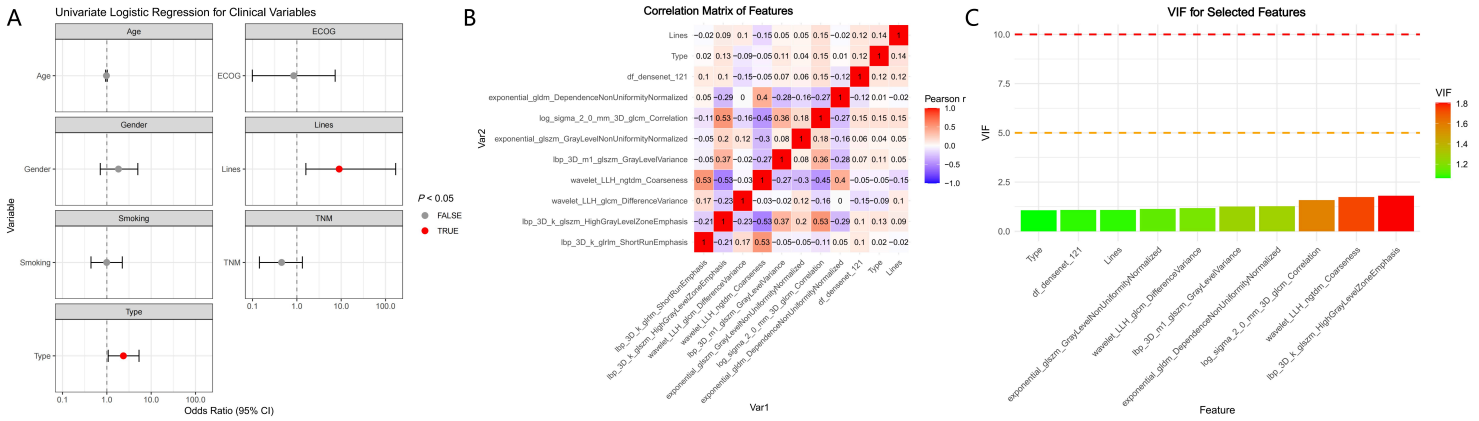


Figure 2

Univariate screening and feature-pool quality control. (A) Forest plot from univariate logistic-regression analysis for clinical predictors of immunochemotherapy response. (B) Pairwise Pearson correlation heatmap of final selected clinical, radiomic and deep-learning features. (C) Variance inflation factor (VIF) distribution plot assessing multicollinearity among retained predictors. All pairwise correlation coefficients <0.6 and VIF values< 5 indicate acceptable feature-set stability for downstream modelling.

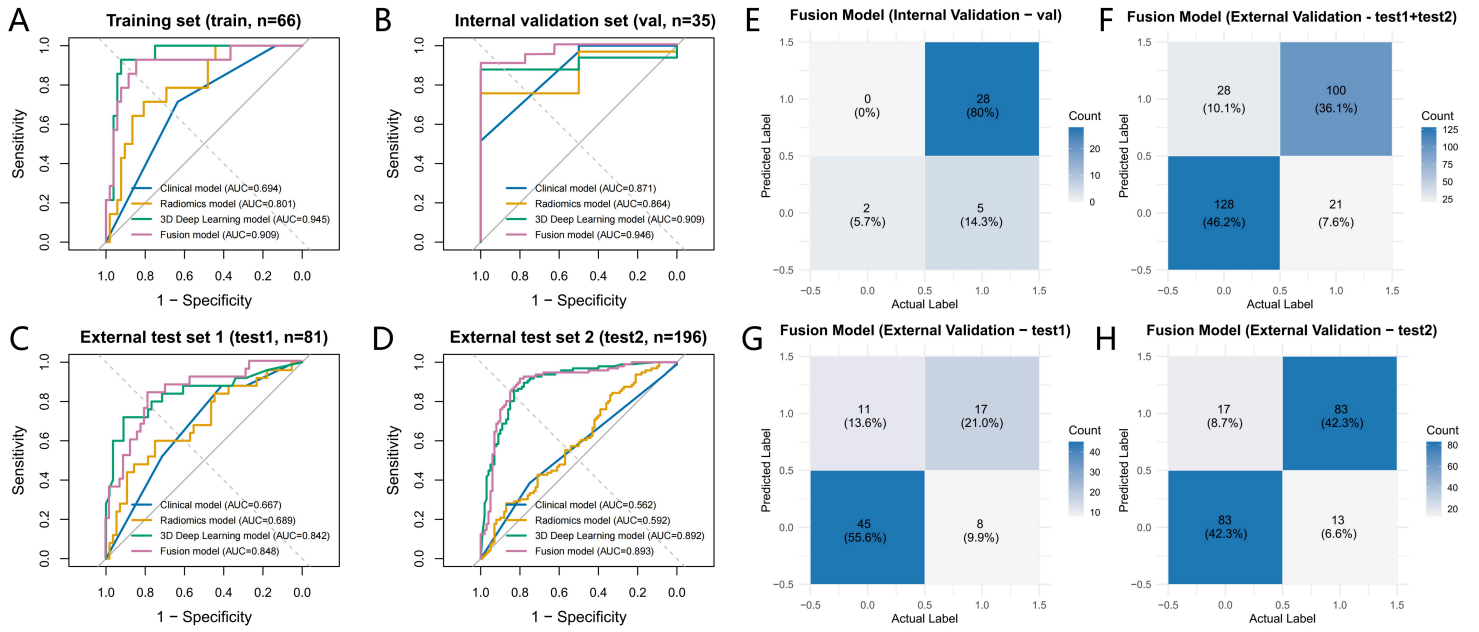


Figure 3

Model discrimination across multi-center cohorts. (A-D) Receiver-operating characteristic (ROC) curves of clinical-only, radiomics-only, 3D deep-learning and multimodal fusion models in training, internal validation, external test-1 and external test-2 cohorts, respectively. Corresponding AUC values are annotated for each model. (E-H) Confusion matrices of the multimodal fusion model in

internal-validation, combined external test cohorts, external test-1 and external test-2, displaying absolute case counts and proportional distribution of predicted versus true treatment-response status.

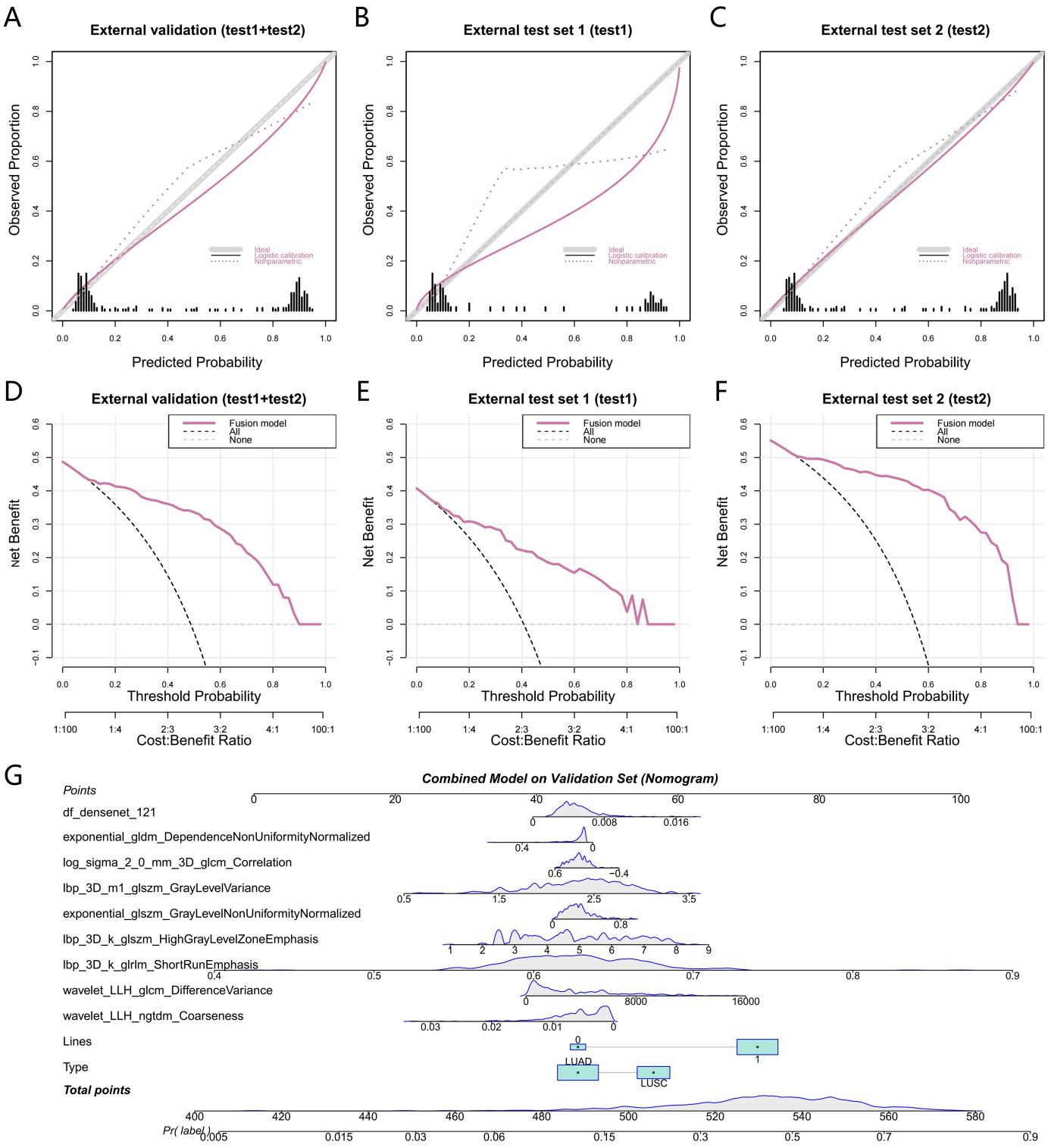


Figure 4

Calibration, clinical utility and nomogram of fusion model. (A-C) Calibration curves of the multimodal fusion model in combined external cohort (test1+test2), external test-1 and external test-2, illustrating

the agreement between predicted probability and observed response proportion. (D-F) Decision curve analysis (DCA) quantifying net clinical benefit across threshold probabilities for three external cohorts. (G) Visual nomogram of the multimodal fusion model built on validation dataset for individualized prediction of immunochemotherapy primary resistance.

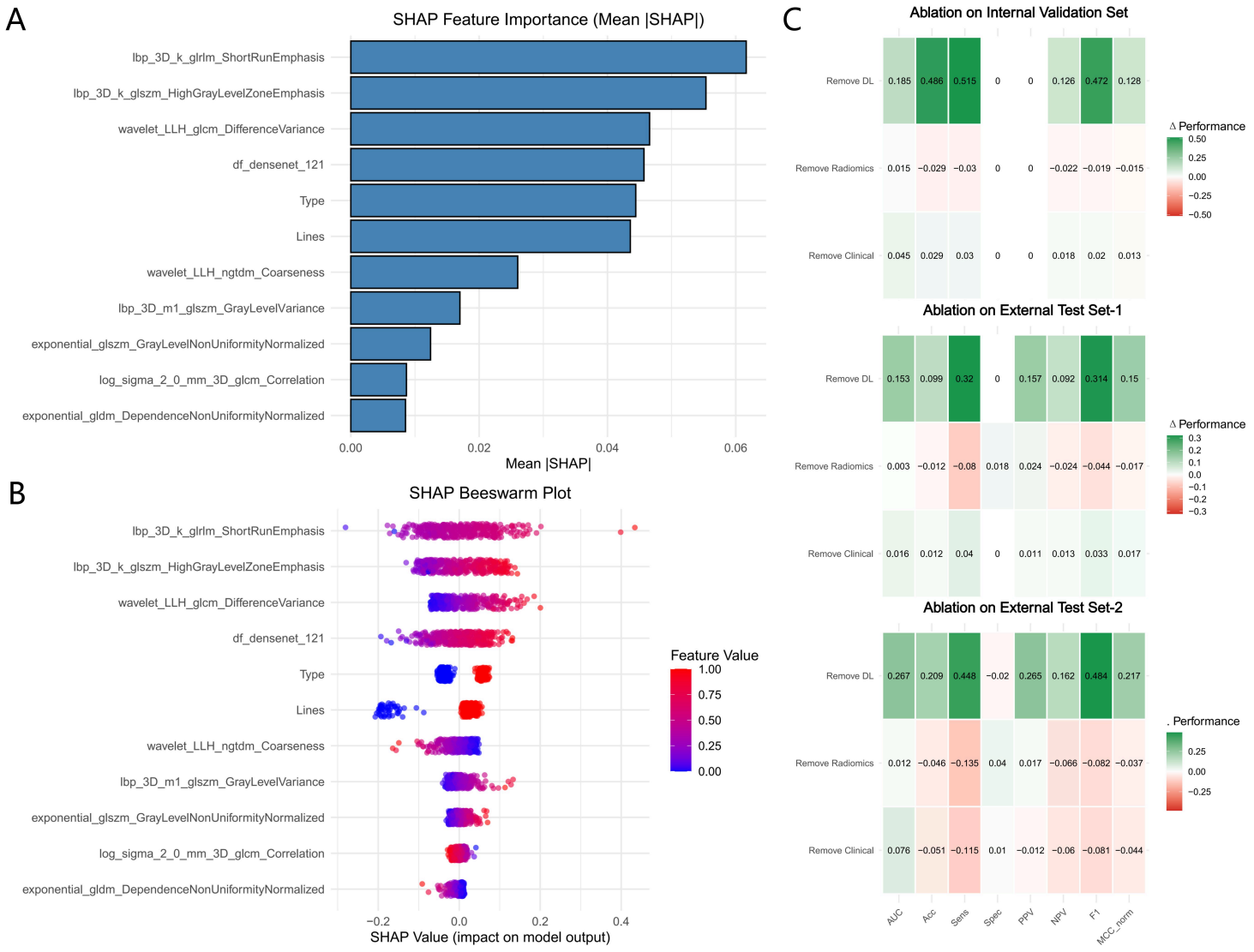


Figure 5

SHAP interpretability and ablation-experiment analysis. (A) SHAP-based mean absolute feature-importance bar plot showing the predictive contribution of each variable in the multimodal fusion model. (B) SHAP beeswarm plot visualizing feature-value distribution and corresponding SHAP values for individual patients. (C) Ablation-experiment heatmaps displaying performance changes (ΔPerformance) after removing deep-learning, radiomic or clinical components in internal-validation, external test-1 and external test-2 cohorts.

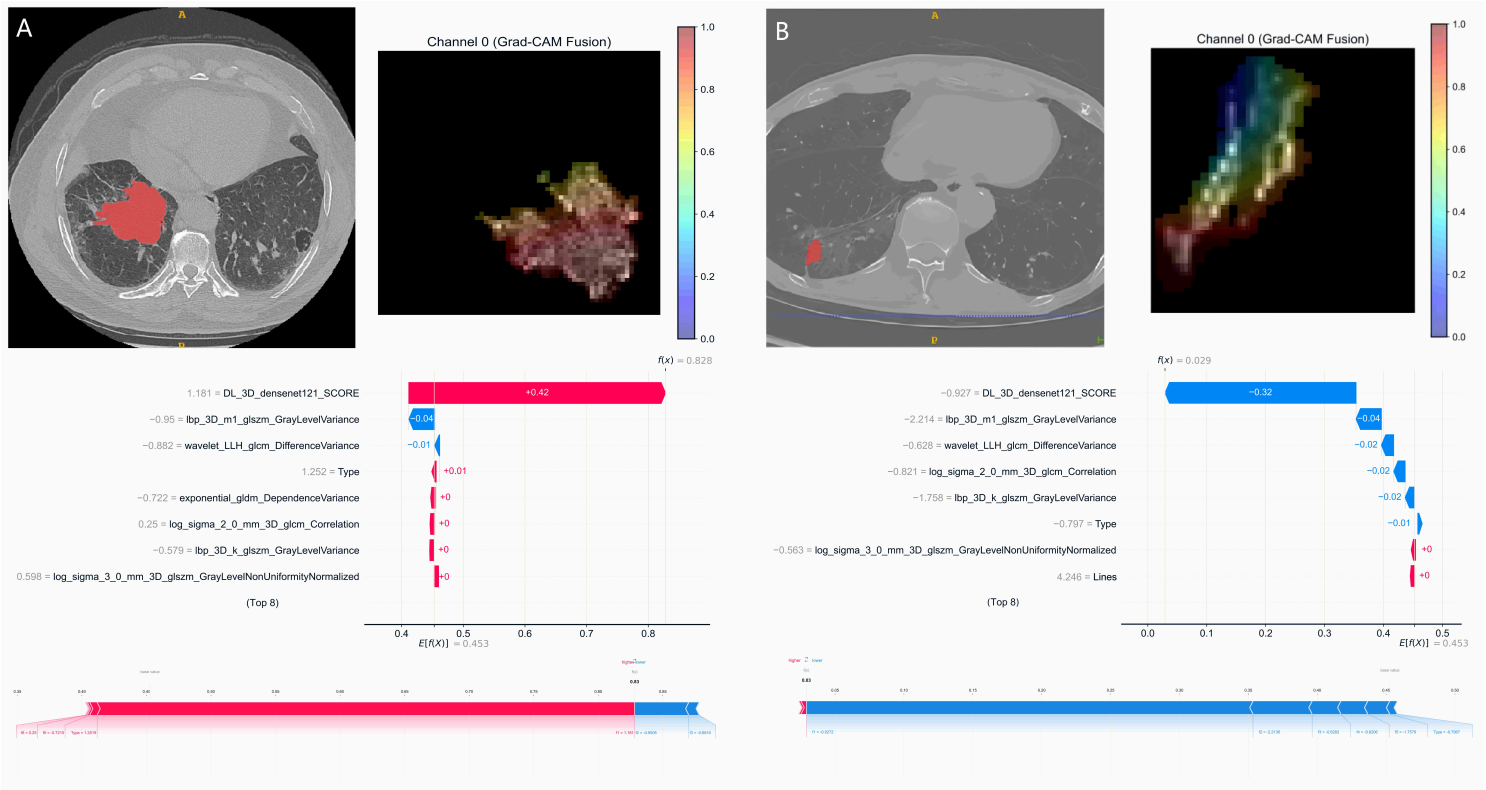


Figure 6

Case-level Grad-CAM and SHAP waterfall visualization. (A) A non-smoking male NSCLC patient with immunochemotherapy resistance. (B) A smoking male NSCLC patient achieving favorable immunochemotherapy response. Each panel includes original chest CT with tumor segmentation mask, Grad-CAM heatmap highlighting tumor regions contributing to model decision-making, patient-specific SHAP waterfall plot quantifying directional feature contributions, and feature-dependence summary plot. Positive SHAP values shift prediction toward immunochemotherapy resistance, while negative SHAP values favor treatment response.

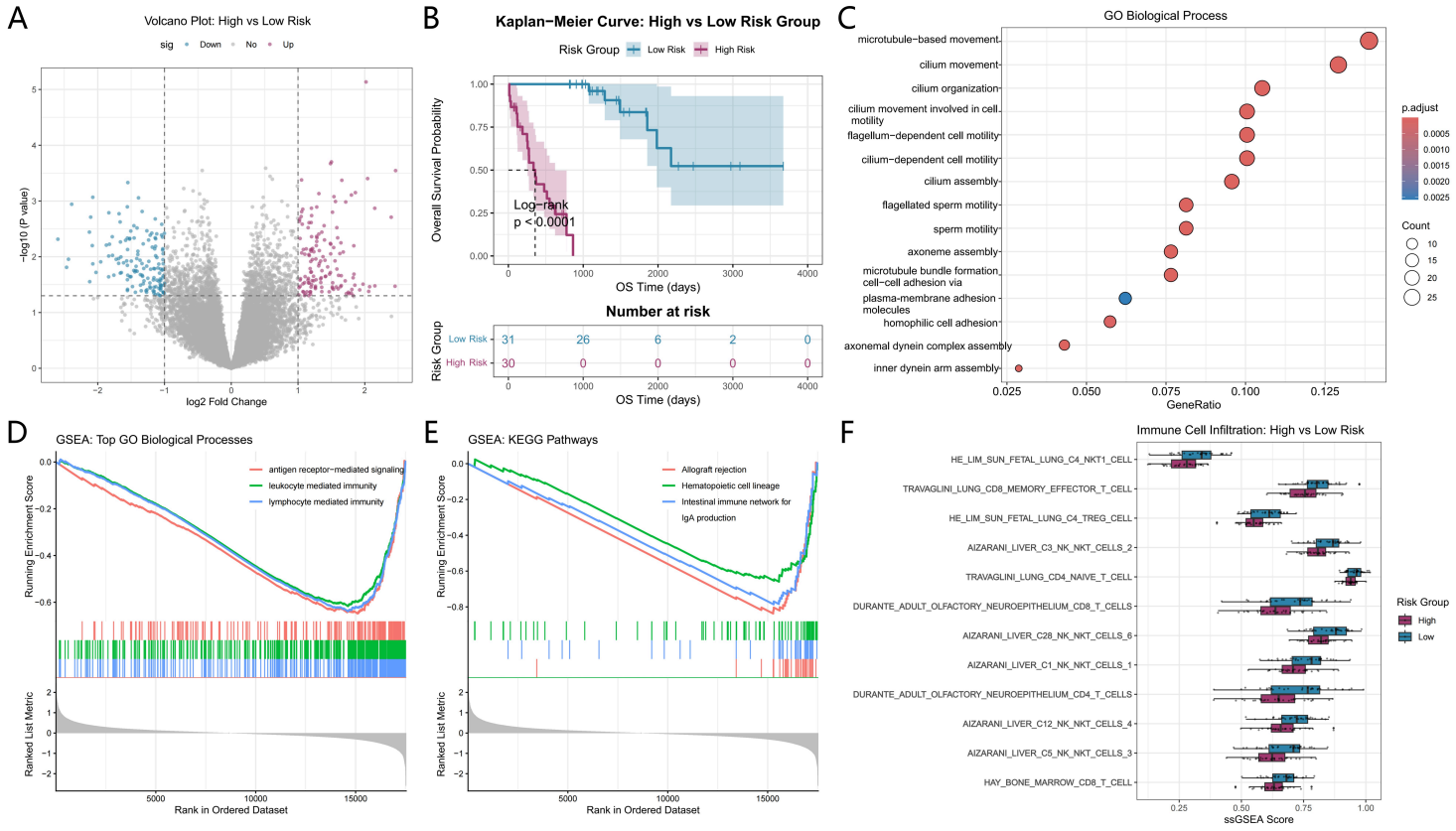


Figure 7

TCGA-TCIA transcriptomic validation of imaging risk phenotypes. Analysis was performed on TCGA-LUAD/LUSC cases with matched TCIA CT imaging data (n=98), stratified by our imaging-derived risk model. (A) Volcano plot showing differentially expressed genes between high-risk and low-risk subgroups. (B) Kaplan-Meier overall-survival curve demonstrating poorer survival outcomes in the high-risk subgroup. (C) Bubble plot for enriched GO biological-process terms. (D-E) GSEA enrichment plots for representative GO and KEGG immune-related pathways. (F) ssGSEA boxplots comparing immune-cell infiltration abundance between high-risk and low-risk subgroups. The high-risk imaging phenotype correlated with an immunosuppressive tumor microenvironment and reduced anti-tumor immune-cell infiltration.

Supplementary Files

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

- SFig1.LASSO.jpg
- SFig2.7Radar.jpg
- SFig3.Sub.jpg
- SFig4.Web.jpg
- SupplementaryTable.docx

- Tables.docx