Respiratory Research

RESEARCH

Open Access

# TRANS: a prediction model for EGFR mutation status in NSCLC based on radiomics and clinical features

Zhigang Chen[1†], Huiying Lu[1†], Ao Liu[2†], Jia Weng[1], Lei Gan[1], Lina Zhou[3], Xiao Ding[4,5] and Shicheng Li[1*]

## Abstract

**Background** Early detection of epidermal growth factor receptor (EGFR) is critical for guiding therapeutic decisions in non-small-cell lung cancer (NSCLC). The study aims to develop a predictive model for EGFR mutations with multicohort data.

**Methods** The study enrolled 254 NSCLC patients of four cohorts: the Affiliated Hospital of Qingdao University (AHQU, $n=54$), the Second Affiliated Hospital of Soochow University (SAHSU, $n=78$), TCGA-NSCLC ($n=91$), and CPTAC-NSCLC ($n=31$). Radiomic features were extracted using the LIFEx software. The least absolute shrinkage and selection operator (LASSO) algorithm was utilized to select predictive features of CT radiomics, clinical data, and RNA sequencing, which were evaluated using receiver operating characteristic (ROC) curves. A nomogram was developed by integrating predictive features. Biological functions were analyzed utilizing RNA sequencing data.

**Results** Eight radiomic features, four clinical features, and seven genomic features were selected to construct distinct signatures. Through internal 5-fold cross-validation, the first two signatures demonstrated notable discrimination capabilities for distinguishing between mutated and wild-type EGFR, resulting in area under the curve (AUC) values of 0.79 (±0.08) and 0.74 (±0.06), respectively. The combination of clinical variables and radiomics signature resulted in an increased AUC of 0.84 (±0.01). This combined model was named TRANS, representing TTF-1, radiomic signature, AE1/AE3, NapsinA, and stage, which uses radiomics and routine immunohistochemistry markers as inputs. High-risk TRANS was observed to be associated with poor overall survival, and showed relationships with high T cell infiltration and response to PD-1 immunotherapy.

**Conclusions** The TRANS model demonstrated favorable ability in predicting EGFR mutation status in NSCLC, providing a valuable approach for optimizing therapeutic strategies in clinical practice.

**Keywords** EGFR mutation, NSCLC, Radiomics, Clinical features, Nomogram

[†]Zhigang Chen, Huiying Lu and Ao Liu contributed equally to this work.

*Correspondence:
Shicheng Li
lishcheng@126.com
[1]Department of Oncology, The Second Affiliated Hospital of Soochow University, Suzhou, China
[2]Department of Thoracic Surgery, The Affiliated Hospital of Qingdao University, Shandong, China
[3]Health Management Center, The Second Affiliated Hospital of Soochow University, Suzhou, China
[4]State Key Laboratory of Common Mechanism Research for Major Diseases, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Suzhou, Jiangsu, China
[5]Key Laboratory of Pathogen Infection Prevention and Control (Peking Union Medical College), Ministry of Education, Beijing, China

## Background

In recent years, the management of non-small cell lung cancer (NSCLC) has underscored the importance of identifying specific mutation status. In NSCLC, the epidermal growth factor receptor (EGFR) status needs to be timely assessed before treatment, because EGFR mutation status is a critical biomarker that determines eligibility for targeted therapy with EGFR-TKIs (tyrosine kinase inhibitors), which offer superior response rates, progression-free survival, and better quality of life compared to standard chemotherapy for patients with activating mutations [1]. Recent advances in molecular profiling have improved treatment response prediction [2, 3], underscoring the importance of EGFR mutation detection [4]. However, these approaches are expensive, time-consuming, and lack the spatial resolution to describe intra-tumor heterogeneity and complexity of the tumor microenvironment [5].

Recent studies have demonstrated that features identified from clinical imaging, such as computed tomography (CT) scans, can function as biomarkers, revealing diagnostic and prognostic associations through correlations with pathological or molecular references, treatment response, and survival outcomes [6]. Currently, radiomic-based models utilizing CT scans present a promising non-invasive strategy for detecting EGFR mutation status in NSCLC patients [7]. Radiomic-based approaches can capture abundant information and multi-dimensionality of the entire tumor and its surrounding tissue.

Although there is evidence supporting the effectiveness of radiomics in predicting EGFR mutation status in NSCLC, numerous related studies have faced limitations, particularly in model interpretability [8]. In domains like healthcare, explaining the input features or how the prediction model works can be as critical as the model's performance itself. An interpretable model can provide insights into biological pathways and molecular mechanisms, potentially leading to new scientific discoveries [9]. The objective of the study was to develop an interpretable predictive model for EGFR mutation, providing a quantitative and visual nomogram for clinical application.

## Methods

### Study design

In the retrospective analysis, radiomics analyses were conducted in four distinct cohorts: the Affiliated Hospital of Qingdao University cohort (AHQU) ($n = 54$), the Second Affiliated Hospital of Soochow University cohort (SAHSU) ($n = 78$), and the Cancer Imaging Archive dataset (TCIA) with TCGA-NSCLC cohort ($n = 91$) and CPTAC-NSCLC cohort ($n = 31$). Figure 1 displays the details of four enrolled cohorts in our study. Table S1 (Additional file 1) shows the endpoints of the four cohorts.

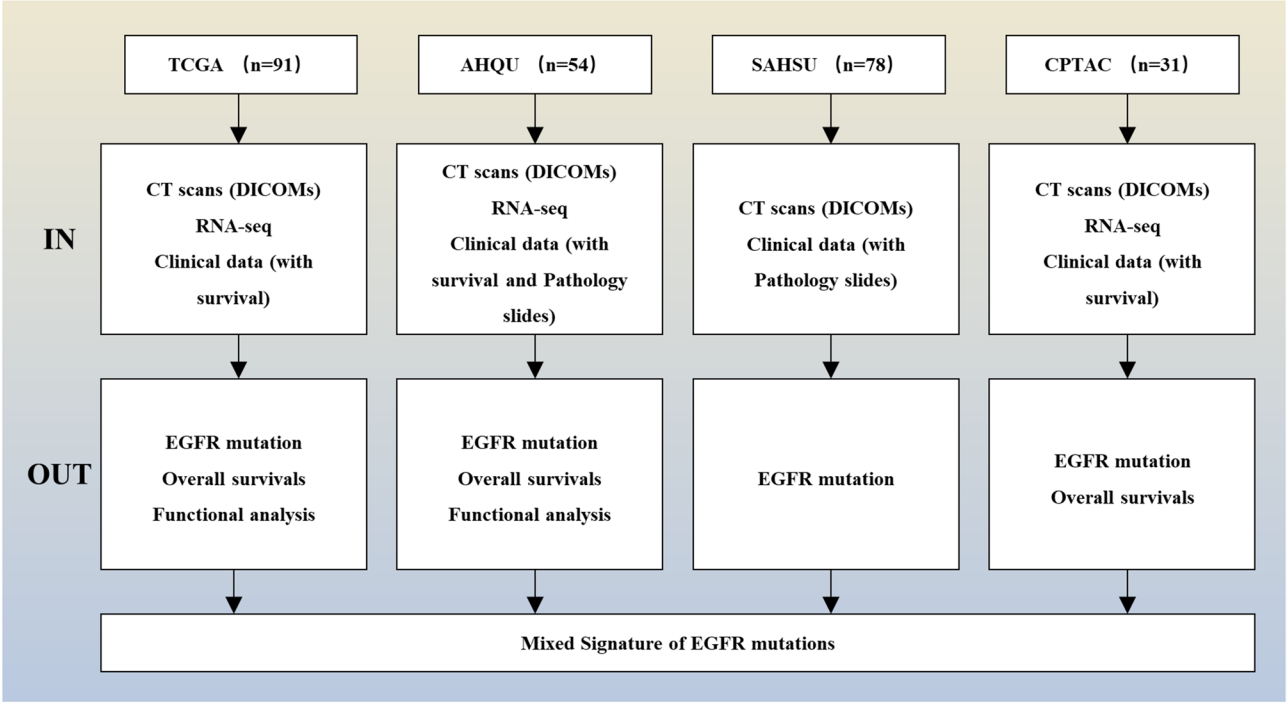


**Fig. 1** Design of the study. AHQU, the Affiliated Hospital of Qingdao University; SAHSU, the Second Affiliated Hospital of Soochow University; TCGA, The Cancer Genome Atlas; CPTAC, Clinical Proteomic Tumor Analysis Consortium; RNA-seq, RNA sequencing; DICOM, Digital Imaging and Communications in Medicine

Patients in AHQU and SAHSU cohorts were enrolled if they followed the criteria: [1] diagnosed with lung squamous cell carcinoma or lung adenocarcinoma; [2] underwent surgery or biopsy within 4 weeks after the CT scan; [3] the lesion had a maximum diameter of more than 1 cm to minimize the partial volume effect. Patients were excluded if they met the following criteria: [1] history of other malignancies, $n = 5$; [2] absence of EGFR testing and immunohistochemistry, $n = 8$; [3] incomplete case records, $n = 3$; [4] poor imaging quality, $n = 1$; [5] target lesions difficult to delineate on CT scan, such as overlap with adjacent lesions or surrounding tissues, or diffuse lesions, $n = 9$; [6] age under 18 years. In total, 132 patients from the two hospitals were enrolled (54 and 78, respectively). Included clinical features are age, gender, lesion location, clinical staging, and histopathological markers.

### CT scan protocol

Patients in the AHQU and SAHSU cohort were scanned using one of three multidetector CT machines: GE Revolution CT, GE CT 750 HD, and Philips Brilliance iCT. All scans were conducted in the supine position at full inspiration. The scanning parameters included 120 kVp, 100–450 mA, detector collimation of either 64 * 0.625 mm or 128 * 0.625 mm, and a field of view of 350 * 350 mm.

### CT segmentation and structuring

Figure 2 illustrates the radiomic pipeline. For all cohorts, CT data in the DICOM format was imported into LIFEx software [10] (Fig. S1 in Additional file 1). Four experienced radiologists, unaware of patient information, outlined the lesions of the primary tumors on the lung window as a volume of interest (VOI). The VOIs were resampled initially to $1 \times 1 \times 1$ mm³ voxels, and then rescaled from −1,000 to 3,000 Hounsfield units (HU) to a fixed bin size of 10 HU. After resampling and rescaling, a total of 201 radiomics features, including shape features, first-order statistics, and second-order statistics (four gray-level matrices), were extracted. To address potential variability introduced by different CT scanners, we implemented ComBat harmonization, a non-parametric Bayesian method that adjusts for site-specific variations while preserving biological differences. The reproducibility of image manually labeling was assessed with the interclass correlation coefficient (ICC) to evaluate

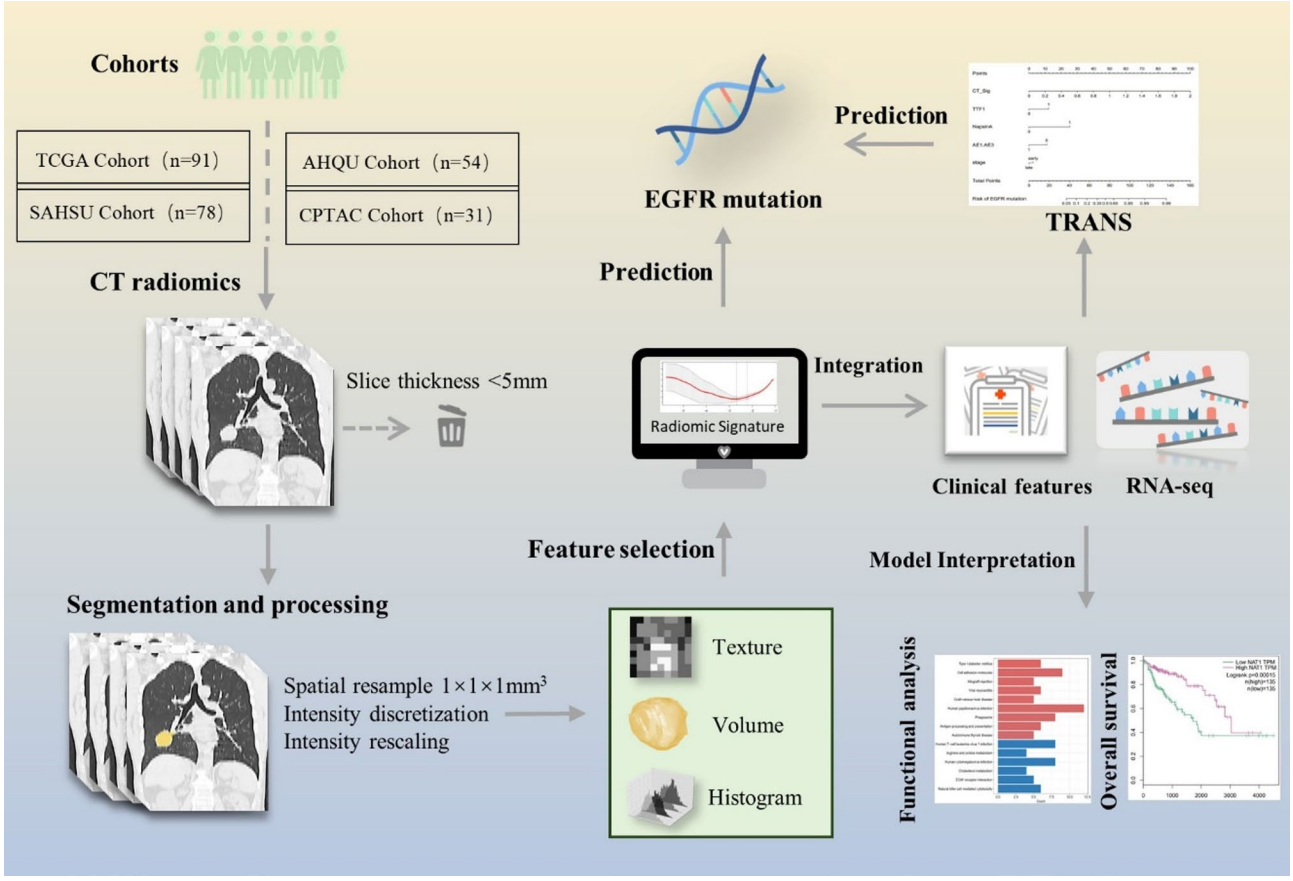


**Fig. 2** The radiomics workflow of the study. AHQU, the Affiliated Hospital of Qingdao University; SAHSU, the Second Affiliated Hospital of Soochow University; TCGA, The Cancer Genome Atlas; CPTAC, Clinical Proteomic Tumor Analysis Consortium; RNA-seq, RNA sequencing

inter-observer agreement. The radiomic features with ICC value < 0.75 were considered to have bad reproducibility and were excluded (see Fig. S2 in Additional file 1). A total of 143 radiomic features were included in the following study (details in Table S2).

#### EGFR mutation detection

Patients in AHQU and SAHSU underwent EGFR detection on histological specimens through either surgical resection or biopsy. The real-time polymerase chain reaction (PCR) method was employed to identify mutations in exons 18–21 of the EGFR gene using the Roche Cobas DNA sample preparation and EGFR mutation detection kit. PCR analysis was conducted using the Roche Cobas Z480 system. If a mutation was detected in EGFR exons 18–21, the tumor was classified as mutant EGFR; otherwise, it was classified as wild-type EGFR. Mutation data for patients in the TCGA and CPTAC cohorts were downloaded and analyzed using maftools version 2.18 (presented in Fig. S3 and S4).

#### RNA sequencing

RNA-seq data included 91 samples from the TCGA cohort and 31 samples from the CPTAC cohort. Tissues collected during surgery in the AHQU cohort were used for RNA sequencing. We extracted RNA with the RNeasy Mini Kit (Qiagen, Germany), and quantified RNA concentration with a Quantus fluorometer (Promega, Madison, WI). mRNAs were enriched and reverse-transcribed into cDNAs to construct sequencing libraries with the NEBNext Ultra RNA library kit (New England Biolabs). Quality-checked libraries underwent ligation with sequencing adapters and were sequenced on the Illumina Novaseq 6000 platform. In SAHSU cohort, RNA sequencing was not available. When constructing the RNA-based model, we categorized samples in the TCGA cohort into a wild-type EGFR group and a mut-type EGFR group. Cut-off was set at log2 (fold change) > 2 to identify differentially expressed genes (DEGs). Adjusted P-value < 0.05 was considered statistically significant.

#### Feature selection and model construction

After data preprocessing, 143 radiomic features, 14 clinical features, and 166 genes (details shown in Table S3) were used for feature selection. Well-performed signatures (in cross-validation, the mean AUC > 0.70) would be integrated to improve the predictive ability. All cohorts were enrolled in the radiomics model, while only the AHQU and SAHSU cohorts with available pathology results were used for the clinical model. The genomic-based model used all cohorts except for the SAHSU cohort. The LASSO algorithm is good at dealing with high-dimensional data of collinearity issues and would be utilized to extract predictable features following data splitting. A Python package was developed to fit and predict using the LASSO regression model (alpha = 1.0, max_iter = 1000, tol = $1^{e-4}$). The code was provided in Code S1.

For model optimization, the radiomics score and clinical features were integrated to form a multivariate logistic regression model (then named TRANS) and visually represented using a predictive nomogram. Diagnostic ability of the model was evaluated through 5-fold cross-validation and ROC curves with scikit-learn 1.4.0 in Python. In this procedure, the dataset was partitioned into 5 equal subsets, with 4 folds used for training and 1 fold for validation in each iteration. Cohorts were partitioned using stratified random sampling to maintain consistent EGFR mutation prevalence between training and validation sets, with stratification applied during 5-fold cross-validation to preserve class balance. Feature preprocessing and selection were performed independently within each training fold to prevent data leakage and ensure unbiased performance assessment. The TRANS model was developed using the AHQU cohort as the training set and independently validated using the SAHSU cohort as the external test set.

#### Biological function analysis for TRANS

To explore the biological implications of TRANS, we stratified the TCGA cohort into high-risk and low-risk subgroups based on the risk scores calculated using the model's coefficients. DEGs in the two subgroups were identified using the aforementioned methods. Biological functional analysis was then conducted on the identified DEG. Pathway analysis was done using KEGG (Kyoto Encyclopedia of Genes and Genomes) database and visualized by clusterProfiler (version 4.0) [11]. The protein-protein interaction (PPI) network from STRING was analyzed and visualized through Cytoscape (version 3.7.1), and the top nodes were calculated using cytoHubba [12].

#### Assessment of immune infiltration

The infiltration levels of immune cells (including 22 classes) were quantified by the CIBERSORT algorithm [13] in high- and low-risk TRANS group. During this phase, the data comprising normalized gene expression values were analyzed combined with the LM22 signature and 1,000 permutations. A Lollipop graph illustrated the correlation between immune cells and the TRANS model. The immune scores referenced in the study include three established metrics: Immune Score, ESTIMATE Score, and Stromal Score. These scores were calculated using the ESTIMATE algorithm [14], which is a widely used computational method that uses gene expression signatures to infer the fraction of stromal and

immune cells in tumor samples. Box plots were generated using immune scores from both TRANS groups.

### Prediction of immunotherapy response

The Cancer Immunome Atlas (TCIA) database [15] offers the function of comprehensive immunogenomic analyses. Cancer immunogenicity was assigned a quantitative score ranging from 0 to 10, referred to as the immunophenoscore (IPS). IPS was used to calculate the response to ICIs (immune checkpoint inhibitors) in the high- and low-risk TRANS.

### Statistics

Statistical analyses and visualizations were performed using SPSS software (version 25.0, IBM Corp., Armonk, NY), GraphPad Prism (version 9.0), Python (version 3.7), and R software (version 3.5.3). Survival analysis and plotting were conducted using survival (version 3.5) [16] and survminer (version 0.4.9) [17] in the R software. $P < 0.05$ was considered significant. The prognostic capabilities of the models were demonstrated by Kaplan-Meier curves and log-rank test. Patients from TCGA, CPTAC, and AHQU cohort were included in the survival analysis. Accuracy (ACC) is a fundamental metric for measuring classification model performance, calculated as ACC = (TP + TN) / (TP + TN + FP + FN), where TP represents true positives, TN true negatives, FP false positives, and FN false negatives. Precision measures the proportion of correct positive predictions and is calculated as Precision = TP / (TP + FP). Recall (also known as sensitivity or true positive rate) measures the proportion of actual positives correctly identified, calculated as Recall = TP / (TP + FN). The F1 score provides a balance between precision and recall, particularly useful for imbalanced datasets, and is calculated as F1 = 2 × (Precision × Recall) / (Precision + Recall).

## Results

### Baseline information

Characteristics of enrolled patients are summarized in Table 1. Specifically, there was a large proportion of females (81.5%), lung adenocarcinoma (50/54), and early-stage patients in the AHQU cohort. EGFR mutations were less common in the TCGA (9/91) and CPTAC (3/31) cohort compared to the other two cohorts. The median overall survivals (OS) were 27 months in TCGA, 59 months in AHQU, and 58 months in CPTAC.

### Eight selected radiomic features

The LASSO regression model was employed for feature selection of radiomic, clinical, and genomic data. Eight radiomic features were extracted to construct the radiomic signature (Fig. 3), including Zone% of gray-level size zone matrix, IntensityHistogramMode of HISTOGRAM, LowGrayLevelZoneEmphasis of gray-level size zone matrix, Busyness of neighboring gray tone

**Table 1** Characteristics of four enrolled cohorts

| Features | TCGA(*N* = 91) | AHQU(*N* = 54) | SAHSU(*N* = 78) | CPTAC(*N* = 31) |
|---|---|---|---|---|
| **Age (year)** | | | | |
| Median (25%,75%) | 70 (61–74) | 63 (54–70) | 64 (57–74) | 69 (64–73) |
| **Female** | 38.04% | 81.5% | 42.3% | 35.48% |
| **Histology** | | | | |
| Adenocarcinoma | 54 | 50 | 60 | 10 |
| Squamous carcinoma | 37 | 4 | 18 | 21 |
| **Stage** | | | | |
| I-II | 45 | 54 | 35 | 28 |
| III-IV | 17 | 0 | 43 | 3 |
| Not Reported | 29 | 0 | 0 | 0 |
| **EGFR** | | | | |
| Mut | 9 | 34 | 41 | 3 |
| Wild | 82 | 20 | 37 | 28 |
| **Laterality** | | | | |
| Left | - | 14 | 20 | 16 |
| Right | - | 40 | 58 | 15 |
| **Vital status** | | | | |
| Alive | 36 | 45 | - | 23 |
| Dead | 29 | 9 | - | 5 |
| Not Reported | 26 | 0 | - | 3 |
| **OS (month)** | | | | |
| Median (25%,75%) | 27(8–43) | 59(52–66) | - | 58(18–61) |

AHQU, the Affiliated Hospital of Qingdao University; SAHSU, the Second Affiliated Hospital of Soochow University; TCGA, The Cancer Genome Atlas; CPTAC, Clinical Proteomic Tumor Analysis Consortium; OS, Overall survival

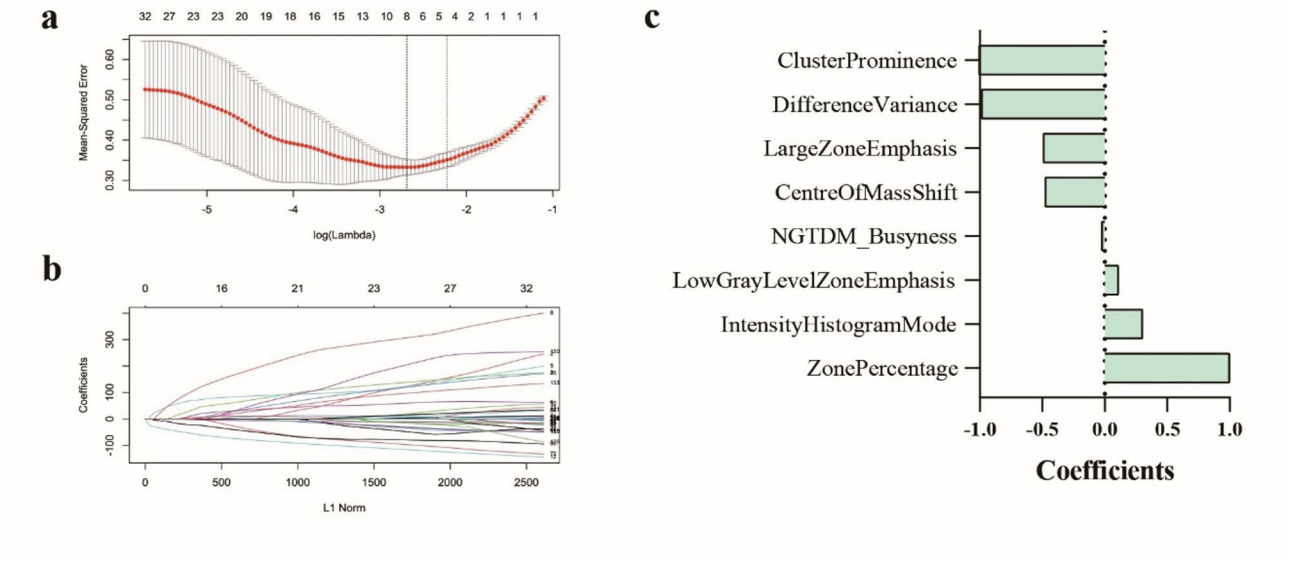


**Fig. 3** Feature selection with the LASSO model. (**a**, **b**) LASSO method for screening of radiomics features; (**c**) Coefficients of eight selected features

**Table 2** Radiomic features of different EGFR groups

| Feature | EGFR mutant | EGFR wild | *p* | BH_p |
|---|---|---|---|---|
| Radiomic score | 1.109(0.877–1.298) | 0.887(0.520–0.964) | < 0.01 | < 0.05 |
| GLSZM_Zone% | 0.0479(0.0199–0.132) | 0.022(0.0129–0.083) | < 0.001 | < 0.005 |
| HISTOGRAM_IntensityHistogramMode | 24(13–26) | 23(11–24) | < 0.05 | < 0.05 |
| GLSZM_LowGrayLevelZoneEmphasis | 0.0117(0.005–0.012) | 0.0068(0.004–0.008) | < 0.05 | < 0.05 |
| NGTDM_Busyness | 0.989(0.531–4.726) | 1.031(0.661–5.126) | < 0.05 | < 0.05 |
| MORPHOLOGICAL_CentreOfMassShift | 0.192(0.121–0.321) | 0.299(0.145–0.521) | < 0.001 | < 0.005 |
| GLSZM_LargeZoneEmphasis | 18,677(713–179678) | 29,625(1217–211138) | < 0.001 | < 0.005 |
| GLCM_DifferenceVariance | 1.19(0.685–2.729) | 1.22(0.751–2.827) | < 0.05 | < 0.05 |
| GLCM_ClusterProminence | 1597(560–3229) | 2459(962–4619) | < 0.001 | < 0.005 |

GLSZM, gray-level size zone matrix; NGTDM, neighboring gray tone difference matrix; GLCM, gray level co-occurrence matrix; BH, Benjamini-Hochberg

difference matrix, CentreOfMassShift of MORPHOLOGICAL, LargeZoneEmphasis of gray-level size zone matrix, DifferenceVariance of gray level co-occurrence matrix, and ClusterProminence of gray level co-occurrence matrix. The four selected clinical features comprised AE1/AE3, NapsinA, TTF-1, and clinical staging (details in Fig.S5). Genomic features were MFSD4A, STEAP4, SOX21, SERPINB4, SPRR1A, MAGEA4, and KRT14 (details in Fig.S6). The coefficient of each feature was provided in the Table S4 and S5.

Table 2 displays the median and interquartile range of radiomic score and eight CT features, which showed significant differences between the EGFR-mutant and the EGFR wild-type groups ($p < 0.05$). NSCLC with EGFR mutations exhibited a higher radiomic score compared to those with wild-type EGFR (radiomic score = 1.109 (0.877–1.298) vs. 0.887 (0.520–0.964)).

### Performance of the selected features

We assessed the predictive ability of radiomics signature, clinical signature, and genomic signature in determining EGFR mutation status. The radiomics model performed well, achieving an AUC of 0.79 (± 0.08) (Fig. 4a) in 5-fold validation. The genomic and clinical signature achieved AUCs of 0.68 (± 0.08) (Fig. 4b) and 0.74 (± 0.06) (Fig. 4c) in 5-fold validation, respectively. We discarded the underperforming gene variables and integrated radiomics signature with clinical features. The combined model named TRANS exhibited the best predictive performance of the AUC of 0.84 (± 0.01) in differentiating EGFR mutation status in 5-fold validation (Fig. 4d). The predictive abilities are presented in Table 3 and Table S7. The mean accuracy of the TRANS model reached 0.901.

### Nomogram construction and validation

A nomogram of TRANS was created for representing individualized predictions. The AHQU and SAHSU cohorts were included in the section. Patients in both cohorts were randomly divided into a training set and a validation set. Using the training set, we visualized output and weight of each variable (Fig. 5a). Calibration curves of the nomogram indicated good agreement between predicted and actual results in both training and validation sets (Fig. 5b and c). The Hosmer-Lemeshow test was employed specifically to assess the goodness-of-fit

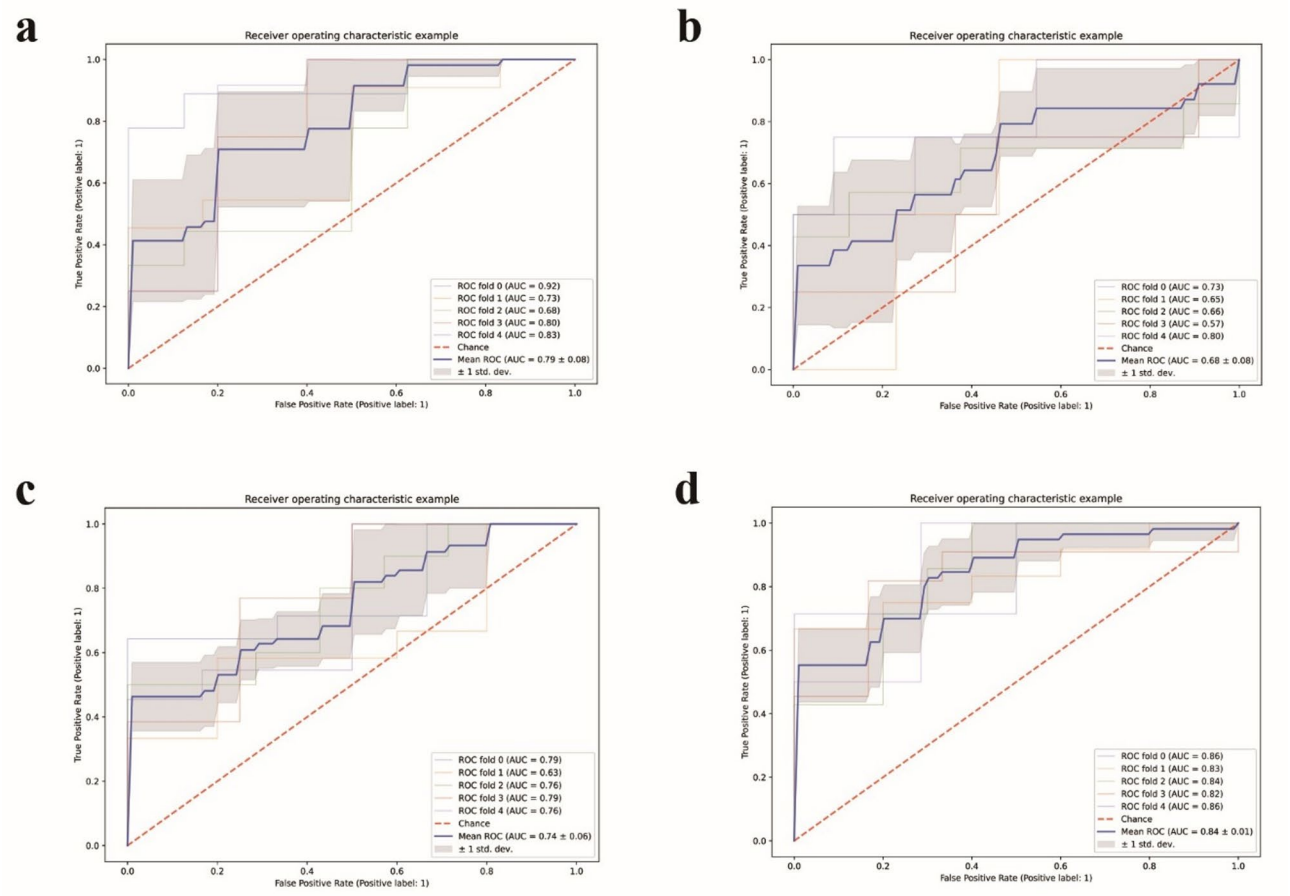


**Fig. 4** Performance of signatures. 5-fold ROC curves of radiomics signature (**a**), genomic signature (**b**), clinical signature (**c**) and TRANS model (**d**). TRANS: TTF-1, radiomic signature, AE1/AE3, NapsinA, and stage. ROC, Receiver Operating Characteristic; AUC, Area Under the Curve; std. dev, Standard Deviation

**Table 3** The predictive abilities of the signatures

| Models | AUC (SD, 95%CI) | Accuracy* |
|---|---|---|
| Genomic signature | 0.68(+-0.08, 0.662–0.698) | 0.744 |
| Radiomics signature | 0.79 (+-0.08, 0.771–0.809) | 0.851 |
| Clinical signature | 0.74(+-0.06, 0.727–0.753) | 0.873 |
| TRANS | 0.84(+-0.01, 0.838–0.842) | 0.901 |

AUC, Area Under Curve; SD, Standard Deviation; CI, confidence interval; TRANS: TTF-1, Radiomic Signature, AE1/AE3, NapsinA, and Stage. *Mean accuracy of 5 folds

of our predictive model. This statistical test evaluates whether the observed event rates match expected event rates in population subgroups. The non-significant p-values in both training ($p = 0.717$) and validation sets ($p = 0.201$) indicate that there is no evidence of poor fit, supporting the reliability of our model's probability estimates across different risk levels. The model was named TRANS, including TTF-1, Radiomic Signature, AE1/AE3, NapsinA, and Stage (the first letters of its components). Kaplan-Meier analysis suggested that the low-risk TRANS group had longer overall survival compared to the high-risk TRANS group ($p = 0.011$) (Fig. S7).

### Biological function of TRANS

To make TRANS explainable, we calculated risk scores of TCGA samples and grouped them into high- and low-risk groups, from which 134 DEGs were identified (84 up-regulated and 50 down-regulated, Fig. 6a). Nine hub genes, including SERPIND1, APOH, HABP2, ORM1, AZGP1, FGA, FGG, FGB, and FGL1, were identified using cytoHubba and were mainly enriched in the function of extracellular matrix (ECM) (Fig. 6b and c). In NSCLC, the interplay between ECM and EGFR involves the regulation of ECM stiffness, angiogenesis, cell survival, adhesion, migration, and metastasis, indicating that TRANS is potentially related with EGFR mutation through ECM regulation. Although these findings provide biological plausibility for TRANS, correlation between radiomics features and genomic alterations alone is insufficient for clinical decision support. The translation of these radiogenomic relationships into actionable clinical information requires further validation to demonstrate their direct impact on treatment selection.

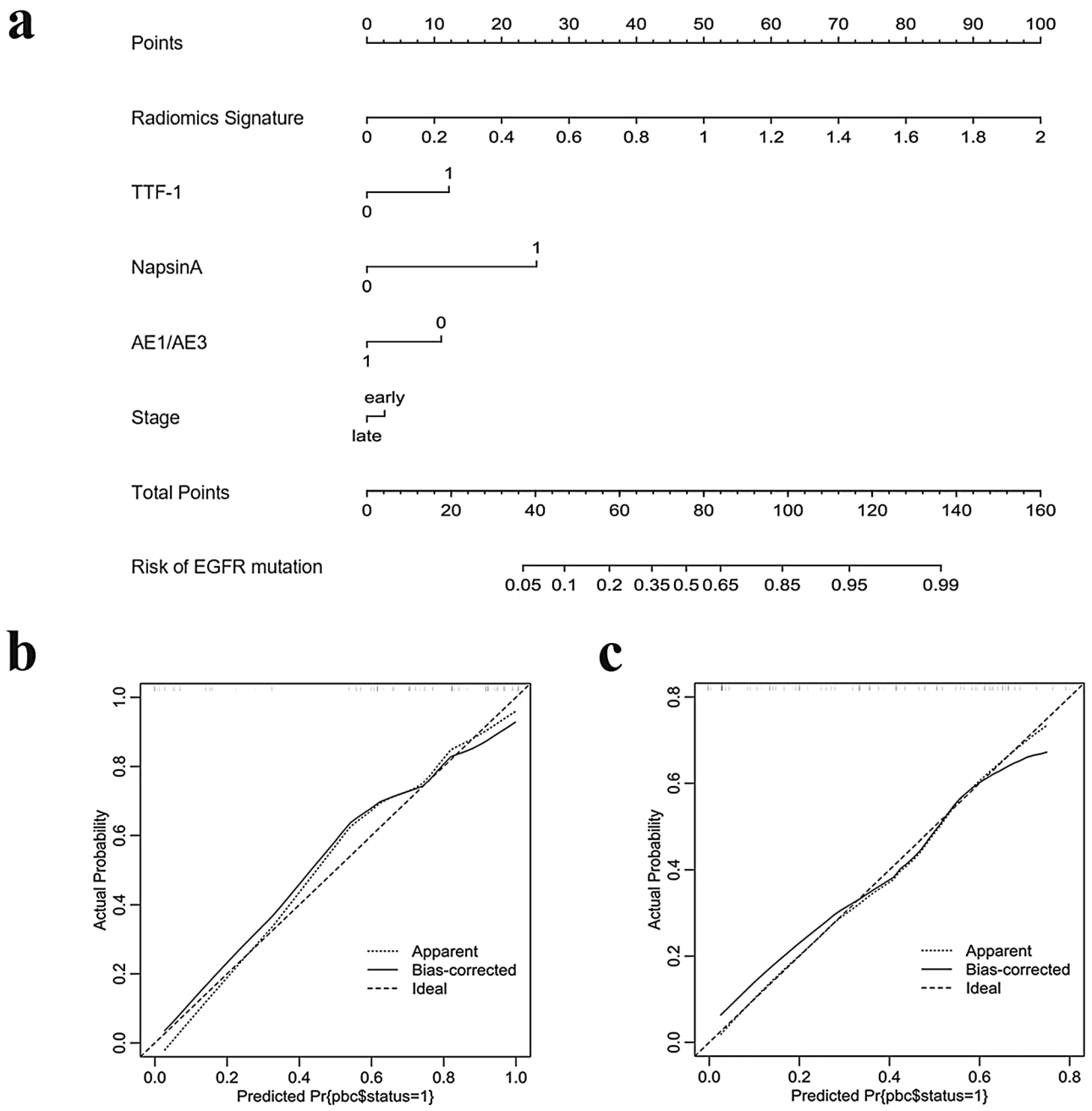


**Fig. 5** Nomogram of TRANS. (**a**) TRANS, including TTF-1, Radiomic Signature, AE1/AE3, NapsinA, and Stage; (**b**) Calibration curve of the nomogram in the training cohort. (c) Calibration curve of the nomogram in the validation cohort

#### Immune heterogeneity of TRANS

The results from the CIBERSORT algorithm showed that CD8 + T cells and activated CD4 + memory T cells were positively related with TRANS. Conversely, M2 macrophages and Neutrophils were negatively related with TRANS (Fig. 7a). The ESTIMATE score, immune score, and stromal score increased as the TRANS risk score increased (Fig. 7b and d). Additionally, we utilized TCIA to calculate patients' susceptibility to immunotherapy and found that the high-risk TRANS exhibited a higher IPS when PD-1 was positive, suggesting that the high-risk TRANS may be more sensitive to PD-1 targeted ICIs (Fig. 7e and h). While these immune correlations suggest potential biological mechanisms and treatment implications, the establishing associations alone is insufficient for clinical implementation. Further prospective validation is needed to confirm whether TRANS can reliably predict immunotherapy response in clinical practice.

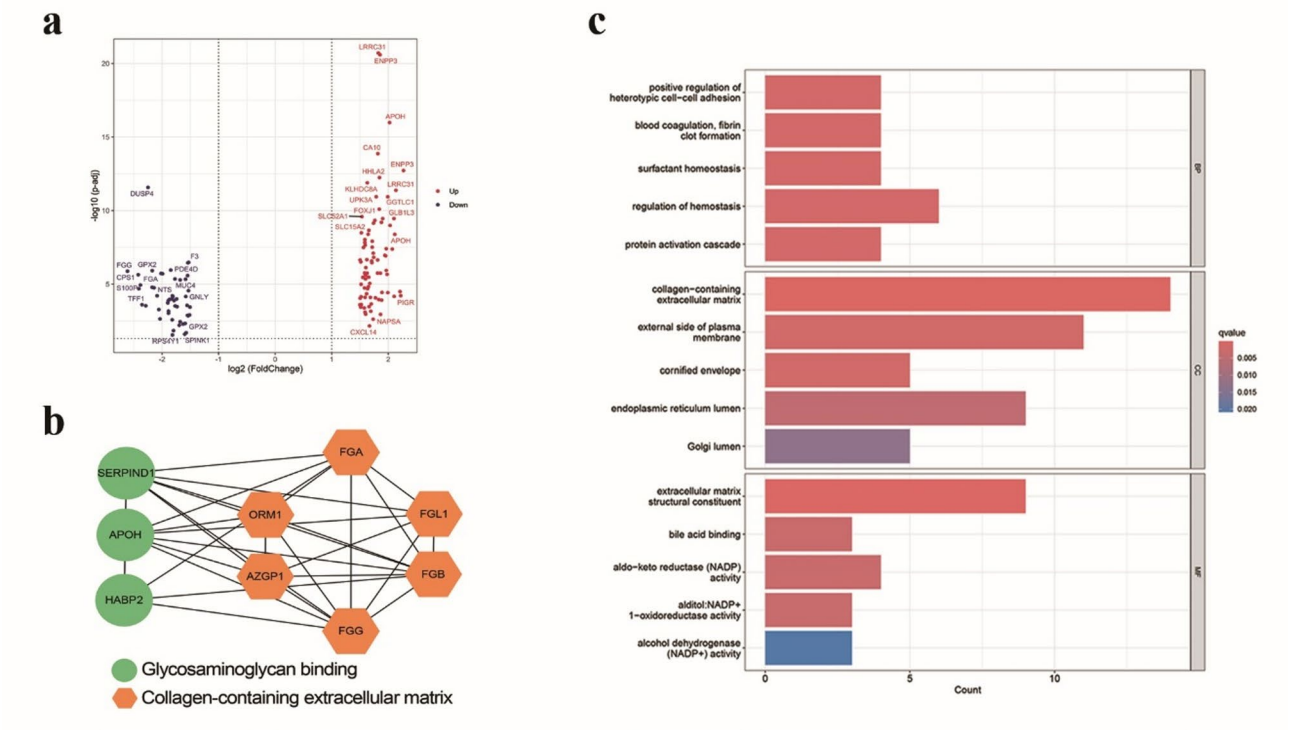


**Fig. 6** Biological function of TRANS based on RNA-seq data. (**a**) Volcano map of differential expressed genes (DEGs); (**b**) Hub genes identified by cytoHubba and their biological functions; (**c**) KEGG enrichment of the DEGs

## Discussion

Rapid detection of EGFR mutation status is crucial for selecting suitable therapy in patients with NSCLC. Conventional methods of detecting EGFR mutations are invasive and have limitations, including challenges such as sampling difficulties and tissue availability, as well as tumor heterogeneity in spatial dimensions [18]. Non-invasive radiomics features have demonstrated the ability to predict EGFR mutations. However, relying solely on medical imaging data limits the utilization of other available data types' potential benefits. In the study, we developed a nomogram representing TRANS for noninvasively predicting EGFR mutations in NSCLC patients by integrating the radiomics signature with four independent clinicopathological predictors, which yielded the highest performance with an AUC of 0.84.

While gene sequencing remains the gold standard for mutation detection, it faces challenges due to tumor heterogeneity and invasive biopsy procedures [19]. Moreover, gene sequencing alone cannot predict patients' prognosis or sensitivity to immunotherapy. Non-invasive approaches have gained attention in recent years. For example, liquid biopsy, particularly circulating tumor DNA (ctDNA), offers rapid and comprehensive EGFR detection [20]. However, ctDNA, as a representative way of liquid biopsy, demonstrates suboptimal sensitivity and specificity [21]. Here, we presented the nomogram visualizing a computed tomography (CT)-based model named TRANS, intended as a supportive clinical tool to improve decision-making. TRANS could complement biopsy sequencing, as the EGFR status predicted by TRANS significantly correlates with patient prognosis and response to PD-1 immunotherapy. Our finding that the TRANS-high group showed better response to PD-1 immunotherapy despite predicting EGFR mutations highlights the complexity of tumor biology. This suggests that certain EGFR-mutant tumors possess immunogenic properties that make them susceptible to checkpoint inhibition, challenging the conventional understanding of EGFR mutations as universal predictors of immunotherapy resistance. While biopsies remain the gold standard, our model leverages readily available radiomic features and routine immunohistochemical markers that are already part of standard diagnostic workups.

CT examinations provide insights into EGFR gene status while elucidating tumor development and progression through imaging [22]. Recent studies have shown the relationship between CT radiomics and EGFR mutations in patients with NSCLC. As reported by Tang et al. [23], peripheral location and pleural indentation signs on CT scans were associated with EGFR-T790M mutation. Conversely, Liu et al. found close associations between EGFR mutations and specific CT radiomic features related to texture heterogeneity, gray-level size zone patterns, and intensity distributions [24]. Wang et al. also found that

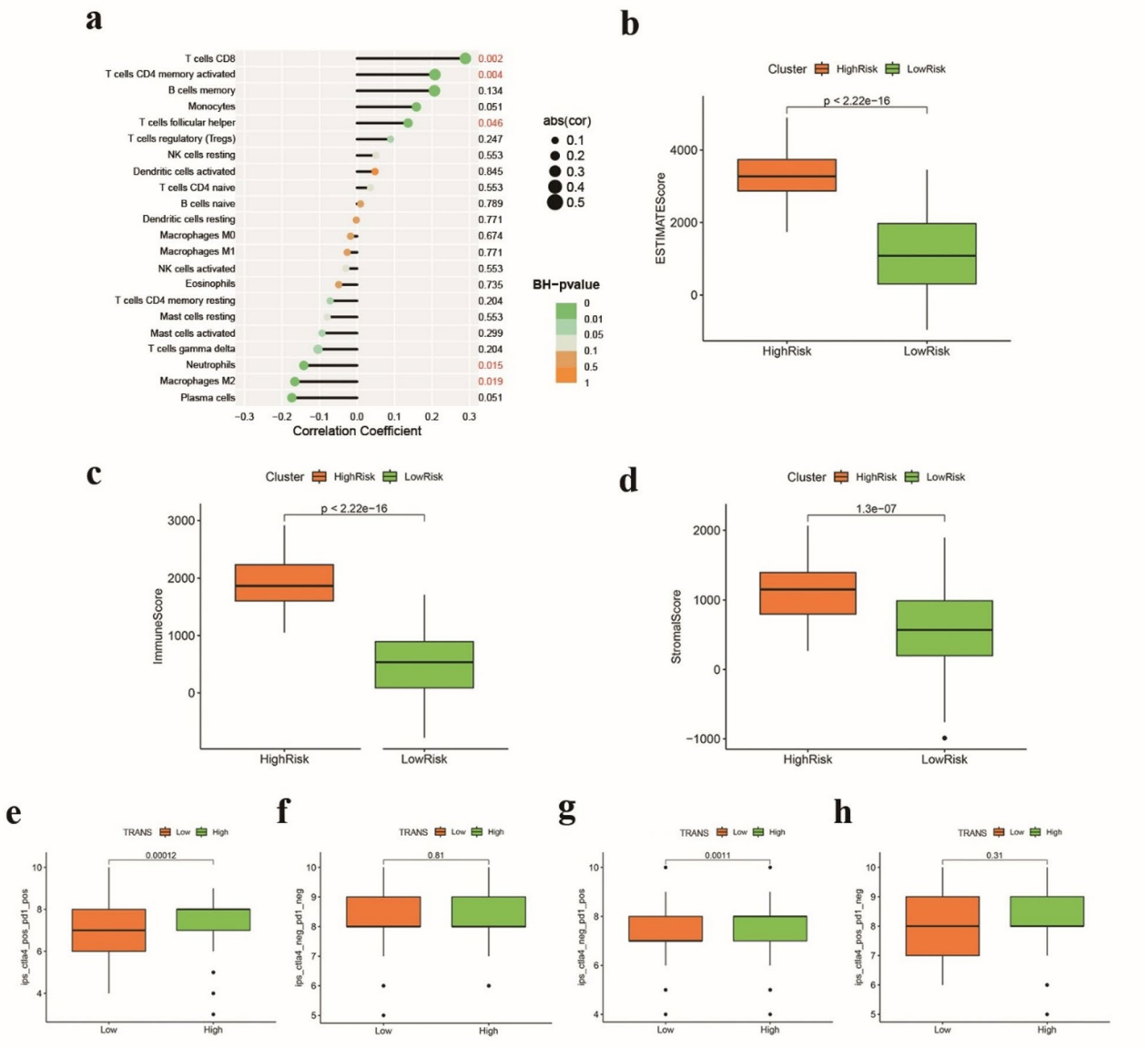


**Fig. 7** Immune heterogeneity between TRANS-high and TRANS-low groups. (**a**) The correlation between infiltration levels of 22 immune cells and TRANS analyzed by CIBERSORT algorithm; (**b**-**d**) The ESTIMATE score, immune score and stromal score in TRANS-high and TRANS -low groups, respectively; (**e**-**h**) The relative probabilities of responding to anti-CTLA-4 antibody and anti-PD-1 antibody in the TRANS-high and TRANS-low groups

preoperative CT images not only correlated with EGFR mutation status but also provided insights into patients' PD-L1 expression [25]. We summarized a total of 18 recent studies on predicting EGFR mutation (details in Table S6), which displayed the AUC, sample size, and study center in each study. The highest AUC was 0.972 in a single-center study conducted by Wu et al. [26], but the sample size was limited to 67. Among multicenter studies with over 200 samples, the highest AUC was 0.800 with a high standard error of 0.383 [27]. Our study demonstrated relatively high performance with an AUC of 0.84 with multicohort data, proving that clinicopathological factors are important indicators of EGFR mutation and may enhance the predictive ability of the model.

Although connecting genomic dysregulations and radiomics is valuable, it has been underexplored [28]. Consequently, the relationship between radiomics and molecular mechanism or immunotherapy in lung cancer remains unexplored. To our knowledge, this study is one of the first to integrate clinicopathological factors and radiomics signatures in predicting EGFR mutations in NSCLC and explaining the model's biological function by genomic data. The TRANS model achieves non-invasive EGFR prediction by integrating readily available clinical and radiological data, eliminating the need for additional

tissue sampling beyond what is already collected in standard care. This approach maintains predictive performance while avoiding the invasiveness, cost, and time constraints associated with genomic testing. In the study, we identified genes differentially expressed between the TRANS-low and TRANS-high groups. The TRANS-high group exhibited better response to PD-1 immunotherapy with significant infiltration of CD8 + T cells and activated CD4 memory cells, which were previously unreported. EGFR mutations may have distinct impacts on immunotherapy response, necessitating further investigation. Furthermore, genomic network analysis and KEGG pathway enrichment indicated that TRANS is significantly enriched in the function of extracellular matrix (ECM). Whilst the progression of NSCLC hinges upon driver mutations, it is equally influenced by the intricate interplay within the ECM, especially involving the EGFR molecule [29]. In the context of NSCLC, the dynamic interplay between the ECM and EGFR orchestrates pivotal processes such as ECM stiffness modulation, angiogenesis, and metastasis. Additionally, certain tumor-promoting ECM constituents, exemplified by glycoproteins and proteoglycans, intensify EGFR activation and PTEN loss [30]. The findings yield a novel perspective into the manner in which EGFR mutations influence CT modifications in NSCLC patients.

The study has several limitations. Firstly, differences between devices could impact the results, in spite of limiting the acquisition of images to three CT systems and preprocessing them before segmentation. Secondly, tumor segmentation was semi-automatic that was time-consuming for the radiologists. Thirdly, the limited sample size highlights the need for more prospective studies to enhance the generalizability and robustness of the constructed model. Furthermore, the developed prediction model must be evaluated on external data to avoid overfitting.

## Conclusions

In summary, we provided a novel strategy for assessing EGFR mutation status by integrating radiomics and clinicopathological data, which can assist clinicians in optimizing relevant treatment strategies for patients with NSCLC.

### Abbreviations

| | |
|---|---|
| NSCLC | Non-Small Cell Lung Cancer |
| EGFR | Epidermal Growth Factor Receptor |
| CT | Computed Tomography |
| OS | Overall Survival |
| TCGA | The Cancer Genome Atlas |
| TCIA | The Cancer Immunome Atlas |
| CPTAC | Clinical Proteomic Tumor Analysis Consortium |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| ROC | Receiver Operating Characteristic |
| AUC | Area Under the Curve |
| DEGs | Differentially Expressed Genes |
| SD | Standard Deviation |
| TRANS | TTF-1, Radiomic Signature, AE1/AE3, NapsinA, and Stage |
| AHQU | The Affiliated Hospital of Qingdao University |
| SAHSU | The Second Affiliated Hospital of Soochow University |
| TCGA | The Cancer Genome Atlas |
| CPTAC | Clinical Proteomic Tumor Analysis Consortium |
| RNA-seq | RNA sequencing |
| DICOM | Digital Imaging and Communications in Medicine |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12931-025-03287-6.

Supplementary Material 1: Additional file 1: Table S1. Endpoints of the four cohorts

Supplementary Material 2: Additional file 2: Table S2. Specific categories and interclass correlation coefficient (ICC) of enrolled radiomics features.

Supplementary Material 3: Additional file 3: Table S3. DEGs in feature selection.

Supplementary Material 4: Additional file 4: Table S4. Coefficients in the radiomics signature.

Supplementary Material 5: Additional file 5: Table S5. Coefficients in the genomic and clinical signature.

Supplementary Material 6: Additional file 6: Table S6. Recent studies on EGFR prediction with radiomics.

Supplementary Material 7: Additional file 7: Table S7. Full evaluation of different signatures.

Supplementary Material 8: Additional file 8: Fig S1. The snapshot of the LIFEx software.

Supplementary Material 9: Additional file 9: Fig S2. Interclass correlation coefficient of the radiomics feature.

Supplementary Material 10: Additional file 10: Fig S3. Mutation files of the CPTAC-NSCLC cohort.

Supplementary Material 11: Additional file 11: Fig S4. Mutation files of the TCGA-NSCLC cohort.

Supplementary Material 12: Additional file 12: Fig S5. LASSO regression for the clinical features.

Supplementary Material 13: Additional file 13: Fig S6. LASSO regression for the genomic features.

Supplementary Material 14: Additional file 14: Fig S7. Survival analysis for signatures.

Supplementary Material 15: Additional file 15: Additional Algorithms.

### Acknowledgements

Not applicable.

### Author contributions

SCL contributed to the conception and analysis data. HYL and AL contributed to the acquisition and interpretation of data. ZGC, HYL, LG and AL labelled VOIs in CT images. LNZ and SCL performed bioinformatics analysis. XD, JW and SCL contributed to the design of the work. SCL and ZGC wrote the draft of the manuscript. All authors contributed to the manuscript writing and revising, gave final approval of the manuscript as submitted and agreed to be accountable for all aspects of the work.

### Funding

This work was supported by the Research project of Jiangsu Provincial Health Commission (YI2023016), the Science and Technology Program of Suzhou City (SLJ2021012), Pre-Research Project of the National Natural Science Foundation of China (SDFEYGJ2014 and SDFEYBS2210).

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

## Declarations

**Ethical approval and consent to participate**
For data from public database (TCGA [31, 32] and CPTAC [33, 34] database), all ethics approvals and informed consents to participate were obtained by the original studies. All procedures performed in studies involving human participants were approved by the medical ethics committee of the Affiliated Hospital of Qingdao University (No. QYFY WZLL 28143) and the medical ethics committee of the Second Affiliated Hospital of Soochow University (No. JD-HG-2023-63). The study was conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from all subjects involved in the study. Clinical trial number: not applicable.

**Consent for publication**
Informed consents were obtained from all individual participants included in the study.

**Competing interests**
The authors declare no competing interests.

Received: 8 April 2024 / Accepted: 23 May 2025
Published online: 05 June 2025

**References**

1. Fu K, Xie F, Wang F, Fu L. Therapeutic strategies for EGFR-mutated non-small cell lung cancer patients with osimertinib resistance. J Hematol Oncol. 2022;15(1):173.
2. Ou SI, Hong JL, Christopoulos P, Lin HM, Vincent S, Churchill EN, et al. Distribution and detectability of EGFR exon 20 insertion variants in NSCLC. J Thorac Oncol. 2023;18(6):744–54.
3. Liu X, Xi X, Xu S, Chu H, Hu P, Li D, Zhang B, Liu H, Jiang T, Lu Z. Targeting T cell exhaustion: emerging strategies in non-small cell lung cancer. Front Immunol. 2024;15:1507501.
4. Xu S, Lu Z. Exploring FNDC4 as a biomarker for prognosis and immunotherapy response in lung adenocarcinoma. Asian J Surg 2024 Sep 13:S1015-9584(24)02098–0.
5. Aran V, Omerovic J. Current approaches in NSCLC targeting K-RAS and EGFR. Int J Mol Sci. 2019;20(22).
6. Li M, Xu G, Chen Q, Xue T, Peng H, Wang Y, et al. Computed Tomography-based radiomics nomogram for the preoperative prediction of tumor deposits and clinical outcomes in Colon cancer: a multicenter study. Acad Radiol. 2023;30(8):1572–83.
7. Mu W, Jiang L, Zhang J, Shi Y, Gray JE, Tunali I, et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. Nat Commun. 2020;11(1):5228.
8. Wang TW, Hsu MS, Lin YH, Chiu HY, Chao HS, Liao CY et al. Application of radiomics in prognosing lung Cancer treated with epidermal growth factor receptor tyrosine kinase inhibitors: A systematic review and Meta-Analysis. Cancers (Basel). 2023;15(14).
9. Luo J, Pan M, Mo K, Mao Y, Zou D. Emerging role of artificial intelligence in diagnosis, classification and clinical management of glioma. Semin Cancer Biol. 2023;91(Electronic):1096–3650.
10. Nioche C, Orlhac F, Boughdad S, Reuze S, Goya-Outi J, Robert C, et al. LIFEx: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Res. 2018;78(16):4786–9.
11. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov (Camb). 2021;2(3):100141.
12. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. CytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol. 2014;8(Suppl 4):S11.
13. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.
14. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, Mills GB, Verhaak RG. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
15. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, Hackl H, Trajanoski Z. Pan-cancer Immunogenomic analyses reveal Genotype-Immunophenotype relationships and predictors of response to checkpoint Blockade. Cell Rep. 2017;18(1):248–62.
16. Terry M, Therneau, Patricia M, Grambsch. (2000). Modeling Survival Data: Extending the Cox Model. Springer, New York. ISBN 0-387-98784-3.) and survminer (version 0.4.9).
17. Kassambara A, Kosinski M, Biecek P. (2024). survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.5.0.
18. Sun W, Yuan X, Tian Y, Wu H, Xu H, Hu G, et al. Non-invasive approaches to monitor EGFR-TKI treatment in non-small-cell lung cancer. J Hematol Oncol. 2015;8:1756–8722. (Electronic)):95.
19. Vu TN, Nguyen HN, Calza S, Kalari KR, Wang L, Pawitan Y. Cell-level somatic mutation detection from single-cell RNA sequencing. Bioinformatics. 2019;35(22):4679–87.
20. Yi X, Ma J, Guan Y, Chen R, Yang L, Xia X. The feasibility of using mutation detection in ctdna to assess tumor dynamics. Int J Cancer. 2017;140(12):2642–7.
21. Pessoa LS, Heringer M, Ferrer VP. CtDNA as a cancer biomarker: A broad overview. Crit Rev Oncol Hematol. 2020;155:103109. (1879– 0461 (Electronic)).
22. Wu G, Jochems A, Refaee T, Ibrahim A, Yan C, Sanduleanu S, et al. Structural and functional radiomics for lung cancer. Eur J Nucl Med Mol Imaging. 2021;48(12):3961–74.
23. Tang X, Li Y, Shen LT, Yan WF, Qian WL, Yang ZG. CT radiomics predict EGFR-T790M resistance mutation in advanced Non-Small cell lung Cancer patients after progression on First-line EGFR-TKI. Acad Radiol. 2023;30(11):2574–87.
24. Liu Q, Sun D, Li N, Kim J, Feng D, Huang G, et al. Predicting EGFR mutation subtypes in lung adenocarcinoma using (18)F-FDG PET/CT radiomic features. Transl Lung Cancer Res. 2020;9(3):549–62.
25. Wang C, Xu X, Shao J, Zhou K, Zhao K, He Y et al. Deep learning to predict EGFR mutation and PD-L1 expression status in Non-Small-Cell lung Cancer on computed tomography images. J Oncol. 2021;2021(1687–8450 (Print)):5499385.
26. Wu S, Shen G, Mao J, Gao B. CT radiomics in predicting EGFR mutation in Non-small cell lung cancer: A single institutional study. Front Oncol. 2020;10:2234. 943X (Print)).
27. Gong J, Fu F, Ma X, Wang T, Ma X, You C et al. Hybrid deep multi-task learning radiomics approach for predicting EGFR mutation status of non-small cell lung cancer in CT images. Phys Med Biol. 2023;68(24).
28. Wang S, Yu H, Gan Y, Wu Z, Li E, Li X, et al. Mining whole-lung information by artificial intelligence for predicting EGFR genotype and targeted therapy response in lung cancer: a multicohort study. Lancet Digit Health. 2022;4(5):e309–19.
29. Hassanein SS, Abdel-Mawgood AL, Ibrahim SA. EGFR-Dependent extracellular matrix protein interactions might light a candle in cell behavior of Non-Small cell lung Cancer. Front Oncol. 2021;11:2234. 943X (Print)).
30. Zuo YH, Gao WN, Xie YJ, Yang SY, Zhou JT, Liang HH, et al. Tumor PKCdelta instigates immune exclusion in EGFR-mutated non-small cell lung cancer. BMC Med. 2022;20(1):470.
31. Albertina B, Watson M, Holback C, Jarosz R, Kirk S. The Cancer Genome Atlas Lung Adenocarcinoma Collection(TCGA-LUAD)(Version 4) [Data set]. The Cancer Imaging Archive. Published online 2016.
32. Kirk S, Lee Y, Kumar P, Filippini J. The Cancer genome atlas lung squamous cell carcinoma collection (TCGA-LUSC) (Version 4) [Data set]. The Cancer Imaging Archive. Published online 2016.
33. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The Clinical Proteomic Tumor Analysis Consortium Lung Squamous Cell Carcinoma Collection (CPTAC-LSCC) (Version 15) [Data set]. The Cancer Imaging Archive. Published online 2018.
34. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The Clinical Proteomic Tumor Analysis Consortium Lung Adenocarcinoma Collection (CPTAC-LUAD) (Version 12) [Data set]. The Cancer Imaging Archive. Published online 2018.

## Publisher's note