

# Predictive Modeling of Lung Cancer Subtype with Integration of Clinical and Thoracic Imaging Features to Guide Early Brain Metastasis Management

Jino Park

[jinop@hs.uci.edu](mailto:jinop@hs.uci.edu)

University of California, Irvine

Paul Buclez

California Northstate University College of Medicine

Jessica Lubisich

University of California, Irvine

Krishna Hanubal

University of California, Irvine

Andrea Inda

University of California, Irvine

Caressa Hui

University of California, Irvine

Jeremy Harris

University of California, Irvine

Aaron Simon

University of California, Irvine

---

## Research Article

**Keywords:** Brain metastasis, EGFR, non-small cell lung cancer, small cell lung cancer, predictive modeling, machine learning

**Posted Date:** August 15th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10581040/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

**Version of Record:** A version of this preprint was published at Journal of Neuro-Oncology on October 5th, 2026. See the published version at <https://doi.org/10.1007/s11060-026-05812-z>.

# Abstract

**Purpose:** Optimal treatment for synchronous brain metastases at lung cancer diagnosis depends on subtype, but histologic/molecular profiling is often delayed. We developed machine-learning models using clinical and imaging data to classify lung cancer into three subtypes with distinct brain metastasis treatment implications: small cell lung cancer (SCLC), EGFR-mutated non-small cell lung cancer (NSCLC), and EGFR–wild-type NSCLC.

**Methods:** We retrospectively identified patients with metastatic lung cancer diagnosed from 2016–2025 at a single institution. Patients without brain metastases at diagnosis formed the training cohort; those with synchronous brain metastases formed the validation cohort. Clinical variables and thoracic imaging features from diagnostic chest CT reports were used to train LASSO logistic regression and random forest models, evaluated via AUC and standard diagnostic metrics. Feature contributions were assessed using logistic regression coefficients and random forest permutation importance.

**Results:** Of 305 patients, 182 were used for training and 123 for validation. Logistic regression achieved AUCs of 0.900, 0.797, and 0.799 for EGFR-mutated NSCLC, SCLC, and EGFR–wild-type NSCLC, respectively; random forest achieved 0.914, 0.766, and 0.808. Adding imaging features to clinical variables improved prediction for EGFR-mutated NSCLC and SCLC. Contributory imaging features included fibrosis, emphysema, mediastinal lymphadenopathy, pleural attachment, and vessel encasement, among others. Three-way prediction showed highest positive predictive value for EGFR-mutated NSCLC and high negative predictive value for SCLC.

**Conclusion:** Combined clinical and chest imaging features can help predict lung cancer subtype, supporting early multidisciplinary discussions for patients with synchronous brain metastases.

## Introduction

Brain metastases represent a pressing oncologic development, requiring urgent management to prevent highly morbid and potentially irreversible sequelae. Decision making can be particularly challenging when patients present with synchronous brain metastases at the time of a new lung cancer diagnosis, as the optimal treatment strategy may depend on tumor histology and molecular subtype. For example, except in true emergencies, patients with small cell lung cancer (SCLC) may derive limited benefit from surgical resection owing to SCLC's exquisite radiosensitivity.<sup>1–3</sup> Conversely, patients with large tumors from non-small cell lung cancer (NSCLC) may benefit from surgical resection prior to radiotherapy to reduce mass effect and improve local control.<sup>4–6</sup> Meanwhile, select patients with subsets of NSCLC harboring driver mutations may be able to defer upfront radiotherapy to trial a tyrosine kinase inhibitor (TKI) with intracranial activity.<sup>7–10</sup>

Prompt identification of lung cancer subtype is critical to optimizing brain metastasis management. However, time required to obtain biopsies, process pathological specimens, and assess for actionable biomarkers can delay diagnosis by days to weeks.<sup>11</sup> In this context, a predictive tool capable of

estimating lung cancer subtype using readily available clinical and radiographic features could provide meaningful guidance for early treatment planning and multidisciplinary discussion. In the present study, we developed predictive models integrating clinical and thoracic computed tomography (CT) imaging features to classify lung cancer subtype among three clinically relevant groups: SCLC, EGFR-mutated NSCLC, and EGFR-wild-type NSCLC.

## Methods

### Study Cohort

The study was approved by the institutional review board. The records of patients diagnosed with metastatic lung cancer between 2016-2025 were retrospectively obtained from a single institutional database. Those with initially localized disease or who received prior cancer treatment were excluded. Eligible patients were required to have available radiology reports of the initial diagnostic chest CT with or without contrast and brain MRI with and without contrast performed within two months of diagnosis, as well as documented histologic and molecular classification of lung cancer subtype. Patients were categorized into three groups according to final histologic and molecular diagnoses: small cell lung cancer (SCLC), EGFR-mutated non-small cell lung cancer (EGFR+), and EGFR-wild-type non-small cell lung cancer (EGFR-). Of all possible driver mutations, only EGFR was selected for independent categorization because (1) with exception of KRAS, it is significantly more prevalent than other driver mutations and (2) unlike KRAS, multiple CNS penetrating targeted therapies are approved for EGFR+ patients.<sup>12,13</sup>

### Data Partitioning

To maximize the availability of training data, the study was designed to include all patients with synchronous presentation of metastatic lung cancer, not only those with brain metastases. To maintain the relevance of the findings to the synchronous brain metastasis population, the dataset was partitioned into two cohorts for the primary analysis. The first cohort included patients without brain metastases at diagnosis, while the second cohort included patients with brain metastases at diagnosis. The first cohort was used to train the model while the second cohort was used for testing. A sensitivity analysis was also run to assess whether presence vs. absence of brain metastasis patients in the training cohort had an impact on model performance.

### Clinical and Imaging Feature Selection

Variables that may help distinguish lung cancer subtypes were extracted from the electronic medical record. These included baseline clinical information and radiographic features reported in literature to be associated with SCLC or EGFR+. <sup>1,14-22</sup> Clinical variables included sex, age at diagnosis, race, smoking history, pack-years, and prior history of lung disease such as emphysema, chronic obstructive pulmonary disease or interstitial lung disease. Thoracic imaging features were obtained from the initial diagnostic chest CT reports. To standardize interpretation and minimize bias, an institutional, firewalled large

language model (Anthropic Claude Sonnet 4.6) blinded to patient identifiers and tumor subtype was used to determine the presence or absence of the following predefined imaging features: location (central vs peripheral), pleural attachment, cavitation, lymphangitic spread, endobronchial component, fibrosis, miliary pattern, emphysema, T3 or T4 classification, multifocality, mediastinal lymphadenopathy, bulky nodal involvement, vessel encasement, superior vena cava (SVC) involvement, and osseous metastases. A standardized prompt was engineered to extract imaging features and is shown in supplemental document 1.

## **Model Development and validation**

Two machine learning approaches were evaluated to predict tumor subtype. The first approach used logistic regression with least absolute shrinkage and selection operator (LASSO) regularization, which allows for feature selection and reduces overfitting by penalizing less informative variables. The second approach employed a random forest classifier, an ensemble learning method that constructs multiple decision trees and aggregates predictions to improve classification performance. To determine the incremental predictive value of incorporating imaging features, the models were developed and tested using two different feature sets. The first included clinical variables alone, while the second incorporated both clinical and imaging features.

A one-versus-rest classification strategy was used to evaluate model performance for each lung cancer subtype. Model performance was assessed using the area under the receiver operating characteristic curve (AUC) as the primary metric of discrimination. For each one-versus-rest subtype comparison, predictors with nonzero LASSO coefficients were entered into an unpenalized multivariable logistic regression model to estimate adjusted odds ratios and 95% confidence intervals. Random forest feature importance was assessed using out-of-bag permuted predictor delta error, defined as the increase in out-of-bag prediction error after randomly permuting each predictor, with larger values indicating greater contribution to model performance. DeLong's test was used to compare performances between models.

Positive predictive value (PPV) and negative predictive value (NPV) were calculated to evaluate clinical utility. Because AUC analyses were performed using one-versus-rest subtype comparisons, an additional 3-way prediction analysis was performed for interpretability. For this analysis, each patient in the testing cohort was assigned to the subtype with the highest predicted probability among EGFR+, SCLC, and EGFR-, and 3-way confusion matrices were generated for the LASSO logistic regression and random forest models.

Machine learning models were developed and evaluated using MATLAB R2025b (The MathWorks Inc., Natick, MA, USA) with the Statistics and Machine Learning Toolbox, utilizing the Classification Learner framework for model training, cross-validation, and performance evaluation. Additional statistical analyses were performed using R version 4.5.2 (R Foundation for Statistical Computing, Vienna, Austria). A simplified study design is shown in figure 1.

## Data Partitioning Sensitivity Analyses

Following the completion of the primary analysis, additional sensitivity analyses were performed to determine the appropriateness of the primary data partitioning approach. To assess whether model training on patients without brain metastases substantially affected performance when applied to patients with brain metastases, we performed a supplemental size-matched sensitivity analysis. Sixty patients with brain metastases were randomly selected and held out as a common testing cohort. Models were then trained using either 60 additional patients with brain metastases or 60 randomly selected patients without brain metastases. Testing AUCs for EGFR+, SCLC, and EGFR- classification were compared between the two training strategies using paired DeLong tests.

## Results

A total of 324 patients with metastatic lung cancer at initial presentation were identified from 2016 to 2025. After exclusion of 19 patients without available chest CT or brain MRI reports, 305 patients were included in the final analysis. Patients without brain metastases at diagnosis were used for model training and included 182 patients, consisting of 57 patients with SCLC, 66 with EGFR+, and 59 with EGFR-. Patients with brain metastases at diagnosis were used as the testing cohort and included 123 patients, consisting of 19 patients with SCLC, 53 with EGFR+, and 51 with EGFR-.

Baseline characteristics of the training and testing cohorts are shown in Table 1. The two cohorts were generally well balanced across most clinical variables. The training and testing cohorts had similar sex distribution, with women comprising 46% and 44% of patients, respectively. The mean age was 65.6 years in the training cohort and 63.1 years in the testing cohort. Smoking history was present in 64% of the training cohort and 58% of the testing cohort, and mean pack-year history was 22.5 and 18.6, respectively. Prior lung disease was more common in the training cohort than the testing cohort, occurring in 51% versus 39% of patients. Most imaging features were also similar between cohorts. Cavitation was more common in the testing cohort than the training cohort, occurring in 17% versus 7% of patients, while fibrosis was less common in the testing cohort, occurring in 15% versus 25%.

In the primary analysis, LASSO logistic regression and random forest models were trained using the cohort without brain metastases and tested in the independent cohort with brain metastases. Models were evaluated using both clinical variables alone and combined clinical plus imaging variables. Across both modeling approaches, incorporation of imaging features generally improved discrimination compared with clinical variables alone, as shown in Figure 2. For the LASSO logistic regression model, the combined clinical and imaging model achieved testing AUCs of 0.900 for EGFR+, 0.797 for SCLC, and 0.799 for EGFR-. In comparison, the clinical-only LASSO model achieved testing AUCs of 0.863 for EGFR+, 0.684 for SCLC, and 0.761 for EGFR-. The largest relative improvement with addition of imaging features was seen for SCLC classification, where AUC improved from 0.684 to 0.797 (DeLong's test,  $p=0.022$ ). The EGFR+ model showed strong discrimination with clinical variables alone, but performance further improved with addition of imaging features (DeLong's test,  $p=0.018$ ). For the EGFR- group, the

difference in AUC did not reach statistical significance (DeLong's test,  $p=0.132$ ). For the random forest model, the combined clinical and imaging model achieved testing AUCs of 0.914 for EGFR+, 0.766 for SCLC, and 0.808 for EGFR-. The clinical-only random forest model achieved testing AUCs of 0.888 for EGFR+, 0.690 for SCLC, and 0.783 for EGFR-. Similar to the LASSO model, the addition of imaging features provided the greatest gain for SCLC classification, and the DeLong's test showed statistical improvement for EGFR+ ( $p=0.031$ ) and SCLC ( $p=0.008$ ) but not EGFR- ( $p=0.199$ ). Overall, the random forest and LASSO models demonstrated similar performance (DeLong's test, EGFR+  $p=0.372$ ; SCLC  $p=0.563$ ; EGFR-  $p=0.615$ ), while the highest discrimination was observed for the EGFR+ subtype.

To assess overall prediction performance in the testing cohort, 3-way confusion matrices were generated by assigning each patient to the subtype with the highest predicted one-vs-rest probability among EGFR+, SCLC, and EGFR- disease, as shown in Figure 3. Using combined clinical and imaging features, the LASSO model demonstrated PPVs of 82.7%, 26.9%, and 64.4% for EGFR+, SCLC, and EGFR-, respectively, with corresponding NPVs of 85.9%, 87.6%, and 71.8%. The random forest model demonstrated PPVs of 83.3%, 29.6%, and 69.0% for EGFR+, SCLC, and EGFR-, respectively, with corresponding NPVs of 88.4%, 88.5%, and 72.8%. Overall, both models showed the strongest PPV for EGFR+ classification, while SCLC classification had lower PPV but preserved relatively high NPV in the 3-way prediction framework. Of the six patients who were incorrectly classified as EGFR+, three harbored other driver mutations associated with non-smoking related lung cancer (Supplementary Table 1).

In the sensitivity analysis, model performance was not statistically different when training was performed using size-matched cohorts of patients with brain metastasis versus patients without (Supplemental Table 2), suggesting that training cohort composition did not affect performance on the test set for either the LASSO logistic regression or Random Forest model.

To determine which features contributed to subtype classification, LASSO-selected predictors were refit in multivariable logistic regression models to estimate odds ratios and 95% confidence intervals, demonstrated in figure 4. For EGFR+, Asian race, pleural attachment, and miliary pattern were associated with higher odds of EGFR-mutated disease, while smoking history, greater pack-year history, and fibrosis were associated with lower odds, though only Asian race, smoking history, and fibrosis remained statistically significant on multivariable analysis. For SCLC, LASSO-selected predictors included smoking history, pack-year history, emphysema, SVC involvement, and pleural attachment. Greater pack-year history and SVC involvement were associated with higher odds of SCLC, while pleural attachment was associated with lower odds on multivariable analysis. Finally, smoking history, cavitation, and pleural attachment were associated with higher odds of EGFR-. In contrast, older age, lymphangitic spread, and SVC involvement were associated with lower odds.

In figure 5, random forest feature importance was evaluated using out-of-bag permuted predictor delta error, which estimates how much model error increases when each predictor is randomly shuffled and provides a measure of which variables contributed most to model performance. Importantly, it does not indicate whether a feature increased or decreased the likelihood of a given subtype. For EGFR+, the most

important predictors were pack-years, smoking history, prior lung disease, Asian race, and emphysema. Additional contributors included fibrosis, White race, sex, pleural attachment, and miliary pattern. For SCLC, the most important random forest predictors were pack-years, pleural attachment, SVC involvement, smoking history, and prior lung disease. Emphysema, White race, vessel encasement, cavitation, and mediastinal lymphadenopathy also contributed to model performance to a lesser extent. For EGFR-, the highest-ranking random forest predictors were pack-years, smoking history, cavitation, pleural attachment, and SVC involvement. Other important predictors included White race, lymphangitic spread, vessel encasement, prior lung disease, and Asian race.

## Discussion

Contemporary management of brain metastases for patients with advanced lung cancer depends significantly on histological and molecular tumor features. For patients who present with synchronous brain metastatic disease, comprehensive classification introduces delays that may compromise outcomes or lead to non-optimal treatment strategies. Methods of predicting tumor subtype with immediately available information may help guide decision making in this critical period.

In this study, we developed and tested predictive models dependent only upon clinical information and reported findings from chest CT imaging, data that is readily obtainable for hospitalized patients in this clinical scenario. The goal of the models was to distinguish amongst EGFR-mutated NSCLC, SCLC, and EGFR-wild-type NSCLC in patients presenting with new lung cancer diagnoses and synchronous brain metastases. Both LASSO logistic regression and random forest models were tested, and both models were trained using either clinical data alone, or clinical data supplemented with CT imaging data. To maximize available data for training and testing, patients with synchronous metastatic disease but without brain metastases were used for training, while patients with brain metastases were reserved for testing. The validity of this data partitioning decision was assessed with a sensitivity analysis.

Both logistic regression and random forest models demonstrated good discriminatory performance, particularly for identifying EGFR-mutated disease, with neither modeling approach demonstrating superiority. Clinical variables alone were sufficient to maximize discriminatory power for the EGFR-cohort, while the addition of CT imaging features improved discrimination of EGFR+ and SCLC. The 3-way prediction analysis further supports the potential clinical utility of these models. Both LASSO logistic regression and random forest models demonstrated the highest PPV for EGFR-mutated NSCLC, with PPVs exceeding 80% and NPVs approaching 90%, suggesting that patients predicted to have EGFR-mutated NSCLC may represent a particularly informative group for early multidisciplinary discussion. In contrast, SCLC classification was associated with lower PPV but relatively preserved NPV, indicating that the models may be more useful for ruling out SCLC rather than for definitively assigning small cell histology. The specific feature analyses provide additional context for these performance patterns. Features such as pleural attachment and miliary pulmonary metastases were associated with EGFR-mutated NSCLC, consistent with previously described imaging phenotypes of this molecular subtype.<sup>18,20,21</sup> Similarly, features such as superior vena cava involvement and central thoracic disease

distribution were associated with small cell histology.<sup>1,22</sup> These findings highlight the potential value of thoracic imaging features as biomarkers of tumor histology, and taken together, suggest that readily available chest CT features provide complementary information beyond clinical variables alone and may improve early subtype estimation.

Several limitations and strengths of the study should be acknowledged. First, this single institutional retrospective study is subject to selection bias based on the geographic location, which may have a higher relative incidence of EGFR+ lung cancer compared with other regions of the United States and globe. This may inflate the positive predictive value of the model compared with regions with lower relative incidence of EGFR+ lung cancer and high incidence of smoking related lung cancer. In addition, imaging features were derived from clinical radiology reports. While this may increase the feasibility of implementing this approach in the clinical setting, it also means that the reports were not structured in a standardized way, potentially introducing noise into the model via mis-identification of presence or absence of features. However, the effect of this noise is unlikely to bias the predictions in a systematic way and would more likely lead to underestimation than overestimation of model performance. Future work will include investigation of effects of standardization of image interpretation as well as of more quantitative but less intuitive approaches such as radiomics analysis. Finally, patients were stratified into cohorts based on the presence or absence of brain metastases at presentation. This approach may introduce confounders, as patients presenting with brain metastases may represent a population with more aggressive tumor biology and more advanced systemic disease, which could influence the thoracic imaging characteristics observed. To address this limitation, we performed a supplemental size-matched sensitivity analysis which showed no significant difference in performance w the different training cohorts. However, this analysis was limited by small sample size and class imbalance, especially for SCLC, and therefore should be cautiously interpreted as a supportive evidence that cohort partitioning did not influence model performance.

## Conclusion

Predictive models integrating clinical and chest CT imaging features can discriminate among EGFR-mutated NSCLC, small cell lung cancer, and EGFR-wild-type NSCLC with good accuracy. Model performance remained robust when applied to patients presenting with synchronous brain metastases, a clinical scenario in which early subtype estimation may influence treatment decisions.

These findings suggest that clinical and radiographic information available at initial presentation may help guide early multidisciplinary discussions regarding intracranial management in metastatic lung cancer. Further validation and prospective implementation studies are warranted to determine whether such predictive tools can meaningfully improve clinical decision-making and patient outcomes.

## Declarations

**Data Availability:** The data analyzed for this study are not publicly available because they contain information derived from patient medical records. Deidentified data may be available from the corresponding author upon reasonable request.

**Ethics Approval:** This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Internal Review Board of University of California Irvine.

**Consent to Participate:** Informed consent was obtained from all individual participants included in the study.

### **Declaration of generative AI and AI-assisted technologies in the manuscript preparation process:**

During the preparation of this work, the author used Anthropic Claude Sonnet 4.6 for grammar proofreading. The author(s) reviewed and edited the output as needed and take full responsibility for the content of the published article.

**Funding:** None

## **Author Contribution**

J.P. wrote the main manuscript text, prepared figures, and contributed to methodology, data curation, and formal analysis. P.B. and J.L. contributed to the original manuscript draft, data curation and analysis. K.H. and A.I. contributed to data curation and manuscript review & editing. C.H. and J.H. contributed to methodology and manuscript review & editing. A.S. provided project conceptualization and supervision and contributed to manuscript review & editing.

## **References**

1. Rudin CM, Brambilla E, Faivre-Finn C, Sage J (2021) Small-cell lung cancer. *Nat Rev Dis Primers* 7(1):3. 10.1038/s41572-020-00235-0
2. Slotman B, Faivre-Finn C, Kramer G et al (2007) Prophylactic cranial irradiation in extensive small-cell lung cancer. *N Engl J Med* 357(7):664–672. 10.1056/NEJMoa071780
3. Takahashi T, Yamanaka T, Seto T et al (2017) Prophylactic cranial irradiation versus observation in patients with extensive-disease small-cell lung cancer: a multicentre, randomised, open-label, phase 3 trial. *Lancet Oncol* 18(5):663–671. 10.1016/S1470-2045(17)30230-9
4. Brown PD, Ballman KV, Cerhan JH et al (2017) Postoperative stereotactic radiosurgery compared with whole brain radiotherapy for resected metastatic brain disease (NCCTG N107C/CEC-3): a multicentre, randomised, controlled, phase 3 trial. *Lancet Oncol* 18(8):1049–1060. 10.1016/S1470-2045(17)30441-2
5. Mahajan A, Ahmed S, McAleer MF et al (2017) Post-operative stereotactic radiosurgery versus observation for completely resected brain metastases: a single-centre, randomised, controlled,

- phase 3 trial. *Lancet Oncol* 18(8):1040–1048. 10.1016/S1470-2045(17)30414-X
6. Kayama T, Sato S, Sakurada K et al (2018) Effects of Surgery With Salvage Stereotactic Radiosurgery Versus Surgery With Whole-Brain Radiation Therapy in Patients With One to Four Brain Metastases (JCOG0504): A Phase III, Noninferiority, Randomized Controlled Trial. *J Clin Oncol* Published online June 20:JCO2018786186. 10.1200/JCO.2018.78.6186
  7. Lee CK, Lefresne S, Soon YY et al (2026) Osimertinib and Stereotactic Radiosurgery for Brain Metastases in EGFR-Mutated Lung Cancer: The STARLET Joint Analysis of OUTRUN and LUOSICNS Randomized Trials. *J Thorac Oncol* Published online January 12:103549. 10.1016/j.jtho.2026.01.001
  8. Miyawaki E, Kenmotsu H, Mori K et al (2019) Optimal Sequence of Local and EGFR-TKI Therapy for EGFR-Mutant Non-Small Cell Lung Cancer With Brain Metastases Stratified by Number of Brain Metastases. *Int J Radiation Oncology\*Biology\*Physics* 104(3):604–613. 10.1016/j.ijrobp.2019.02.051
  9. Magnuson WJ, Lester-Coll NH, Wu AJ et al (2017) Management of Brain Metastases in Tyrosine Kinase Inhibitor–Naïve Epidermal Growth Factor Receptor–Mutant Non–Small-Cell Lung Cancer: A Retrospective Multi-Institutional Analysis. *JCO* 35(10):1070–1077. 10.1200/JCO.2016.69.7144
  10. Yamaguchi H, Wakuda K, Fukuda M et al (2021) A Phase II Study of Osimertinib for Radiotherapy-Naïve Central Nervous System Metastasis From NSCLC: Results for the T790M Cohort of the OCEAN Study (LOGIK1603/WJOG9116L). *J Thorac Oncol* 16(12):2121–2132. 10.1016/j.jtho.2021.07.026
  11. Robert NJ, Nwokeji ED, Espirito JL et al (2021) Biomarker tissue journey among patients (pts) with untreated metastatic non-small cell lung cancer (mNSCLC) in the U.S. Oncology Network community practices. *JCO* 39(15suppl):9004–9004. 10.1200/JCO.2021.39.15\_suppl.9004
  12. Chevallier M, Borgeaud M, Addeo A, Friedlaender A (2021) Oncogenic driver mutations in non-small cell lung cancer: Past, present and future. *World J Clin Oncol* 12(4):217–237. 10.5306/wjco.v12.i4.217
  13. Reungwetwattana T, Nakagawa K, Cho BC et al (2018) CNS Response to Osimertinib Versus Standard Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors in Patients With Untreated *EGFR* -Mutated Advanced Non–Small-Cell Lung Cancer. *JCO* 36(33):3290–3297. 10.1200/JCO.2018.78.3118
  14. Shi Y, Au JSK, Thongprasert S et al (2014) A prospective, molecular epidemiology study of EGFR mutations in Asian patients with advanced non-small-cell lung cancer of adenocarcinoma histology (PIONEER). *J Thorac Oncol* 9(2):154–162. 10.1097/JTO.0000000000000033
  15. Zhang H, Cai W, Wang Y, Liao M, Tian S (2019) CT and clinical characteristics that predict risk of EGFR mutation in non-small cell lung cancer: a systematic review and meta-analysis. *Int J Clin Oncol* 24(6):649–659. 10.1007/s10147-019-01403-3
  16. Lee D, Rho JY, Kang S, Yoo KJ, Choi HJ (2016) CT findings of small cell lung carcinoma: Can recognizable features be found? *Med (Baltim)* 95(47):e5426. 10.1097/MD.00000000000005426

17. Lee HJ, Kim YT, Kang CH et al (2013) Epidermal growth factor receptor mutation in lung adenocarcinomas: relationship with CT characteristics and histologic subtypes. *Radiology* 268(1):254–264. 10.1148/radiol.13112553
18. Togashi Y, Masago K, Kubo T et al (2011) Association of diffuse, random pulmonary metastases, including miliary metastases, with epidermal growth factor receptor mutations in lung adenocarcinoma. *Cancer* 117(4):819–825. 10.1002/cncr.25618
19. Liu Y, Kim J, Qu F et al (2016) CT Features Associated with Epidermal Growth Factor Receptor Mutation Status in Patients with Lung Adenocarcinoma. *Radiology* 280(1):271–280. 10.1148/radiol.2016151455
20. Wu SG, Hu FC, Chang YL et al (2013) Frequent *EGFR* mutations in nonsmall cell lung cancer presenting with miliary intrapulmonary carcinomatosis. *Eur Respir J* 41(2):417–424. 10.1183/09031936.00006912
21. Sekine A, Katano T, Oda T et al (2018) Miliary lung metastases from non-small cell lung cancer with Exon 20 insertion: A dismal prognostic entity: A case report. *Mol Clin Oncol* 9(6):673–676. 10.3892/mco.2018.1730
22. Miyauchi E, Motoi N, Ono H et al (2015) Distinct Characteristics of Small Cell Lung Cancer Correlate With Central or Peripheral Origin: Subtyping Based on Location and Expression of Transcription Factor TTF-1. *Medicine* 94(51):e2324. 10.1097/MD.0000000000002324

## Tables

Table 1

Baseline clinical and imaging characteristics of the training and testing cohorts.

| Variable            | Training (n = 182) | Testing (n = 123) | p-value |
|---------------------|--------------------|-------------------|---------|
| Lung Cancer Subtype |                    |                   |         |
| SCLC                | 57                 | 19                |         |
| EGFR+               | 66                 | 53                |         |
| EGFR-               | 59                 | 51                |         |
| Clinical            |                    |                   |         |
| Sex                 |                    |                   | 0.769   |
| Female (%)          | 83 (46)            | 54 (44)           |         |
| Male (%)            | 99 (54)            | 69 (56)           |         |
| Age, y              |                    |                   | 0.089   |
| Mean (SD)           | 65.6 (9.3)         | 63.1 (11.1)       |         |
| Median (Range)      | 66 (42–99)         | 64 (18–90)        |         |
| Race                |                    |                   | 0.055   |
| White (%)           | 82 (45)            | 40 (33)           |         |
| Asian (%)           | 72 (40)            | 65 (53)           |         |
| Other (%)           | 28 (15)            | 18 (15)           |         |
| Smoking history     |                    |                   | 0.248   |
| Yes (%)             | 117 (64)           | 71 (58)           |         |
| No (%)              | 65 (36)            | 52 (42)           |         |
| Pack-years          |                    |                   | 0.201   |
| Mean (SD)           | 22.5 (25)          | 18.6 (22)         |         |
| Prior lung disease  |                    |                   | 0.038   |
| Yes (%)             | 93 (51)            | 48 (39)           |         |
| No (%)              | 89 (49)            | 75 (61)           |         |
| Imaging             |                    |                   |         |
| Location            |                    |                   | 0.456   |
| Central (%)         | 78 (43)            | 44 (36)           |         |

| <b>Variable</b>         | <b>Training (n = 182)</b> | <b>Testing (n = 123)</b> | <b>p-value</b> |
|-------------------------|---------------------------|--------------------------|----------------|
| Peripheral (%)          | 42 (23)                   | 33 (27)                  |                |
| Unspecified (%)         | 62 (34)                   | 46 (37)                  |                |
| Pleural attachment      |                           |                          | 0.744          |
| Yes (%)                 | 75 (41)                   | 53 (43)                  |                |
| No (%)                  | 107 (59)                  | 70 (57)                  |                |
| Cavitation              |                           |                          | 0.004          |
| Yes (%)                 | 12 (7)                    | 21 (17)                  |                |
| No (%)                  | 170 (93)                  | 102 (83)                 |                |
| Lymphangitic spread     |                           |                          | 0.439          |
| Yes (%)                 | 47 (26)                   | 27 (22)                  |                |
| No (%)                  | 135 (74)                  | 96 (78)                  |                |
| Endobronchial component |                           |                          | 0.115          |
| Yes (%)                 | 31 (17)                   | 30 (24)                  |                |
| No (%)                  | 151 (83)                  | 93 (76)                  |                |
| Fibrosis                |                           |                          | 0.033          |
| Yes (%)                 | 45 (25)                   | 18 (15)                  |                |
| No (%)                  | 137 (75)                  | 105 (85)                 |                |
| Miliary                 |                           |                          | 0.166          |
| Yes (%)                 | 9 (5)                     | 11 (9)                   |                |
| No (%)                  | 173 (95)                  | 112 (91)                 |                |
| Emphysema               |                           |                          | 0.182          |
| Yes (%)                 | 54 (30)                   | 28 (23)                  |                |
| No (%)                  | 128 (70)                  | 95 (77)                  |                |
| T3 or T4                |                           |                          | 0.468          |
| Yes (%)                 | 128 (70)                  | 81 (66)                  |                |
| No (%)                  | 54 (30)                   | 42 (34)                  |                |
| Multifocal              |                           |                          | 0.304          |
| Yes (%)                 | 120 (66)                  | 74 (60)                  |                |

| Variable                       | Training (n = 182) | Testing (n = 123) | p-value |
|--------------------------------|--------------------|-------------------|---------|
| No (%)                         | 62 (34)            | 49 (40)           |         |
| Mediastinal lymphadenopathy    |                    |                   | 0.809   |
| Yes (%)                        | 112 (61)           | 74 (60)           |         |
| No (%)                         | 70 (39)            | 49 (40)           |         |
| Bulky nodal involvement        |                    |                   | 0.432   |
| Yes (%)                        | 63 (35)            | 48 (39)           |         |
| No (%)                         | 119 (65)           | 75 (61)           |         |
| Vessel encasement              |                    |                   | 0.743   |
| Yes (%)                        | 40 (22)            | 29 (24)           |         |
| No (%)                         | 142 (78)           | 94 (76)           |         |
| superior vena cava involvement |                    |                   | 0.533   |
| Yes (%)                        | 12 (7)             | 6 (5)             |         |
| No (%)                         | 170 (93)           | 117 (95)          |         |
| Osseous Metastases             |                    |                   | 0.321   |
| Yes (%)                        | 44 (24)            | 36 (29)           |         |
| No (%)                         | 138 (76)           | 87 (71)           |         |

## Figures

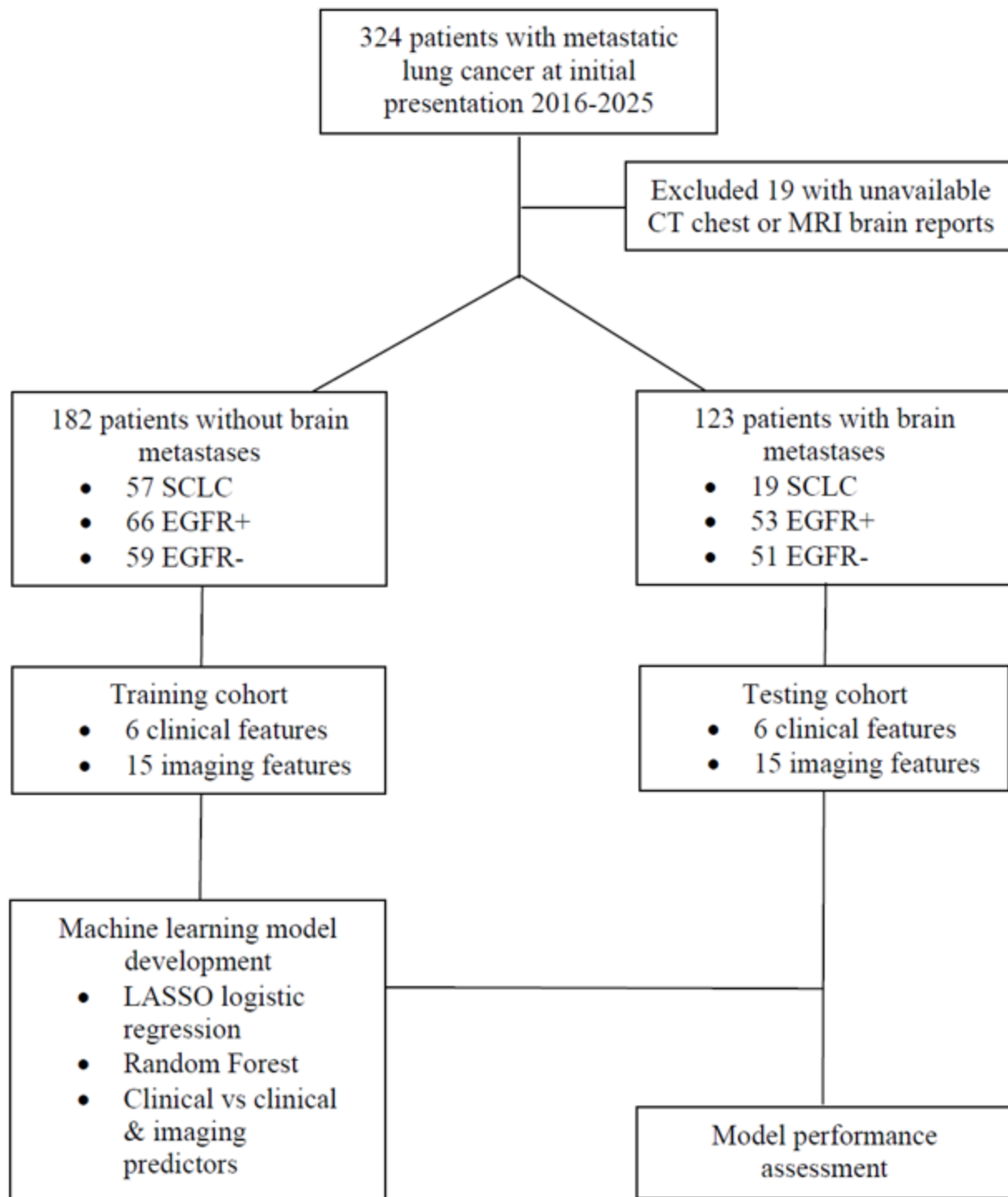


Figure 1

Patient selection and study design.

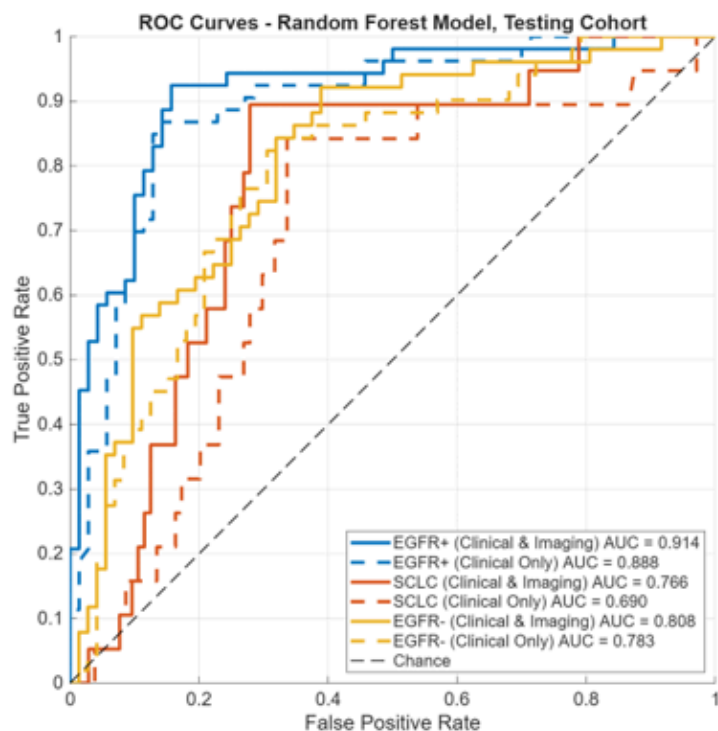
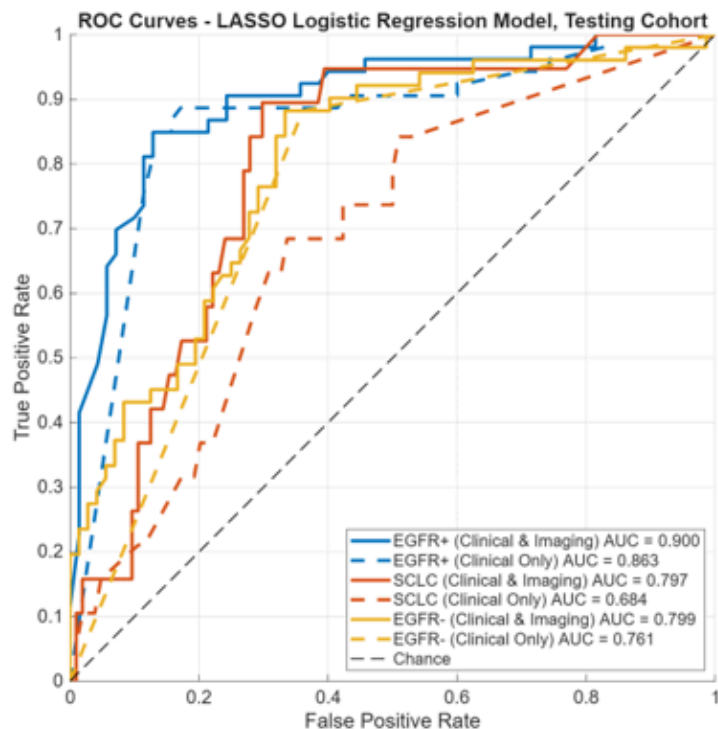
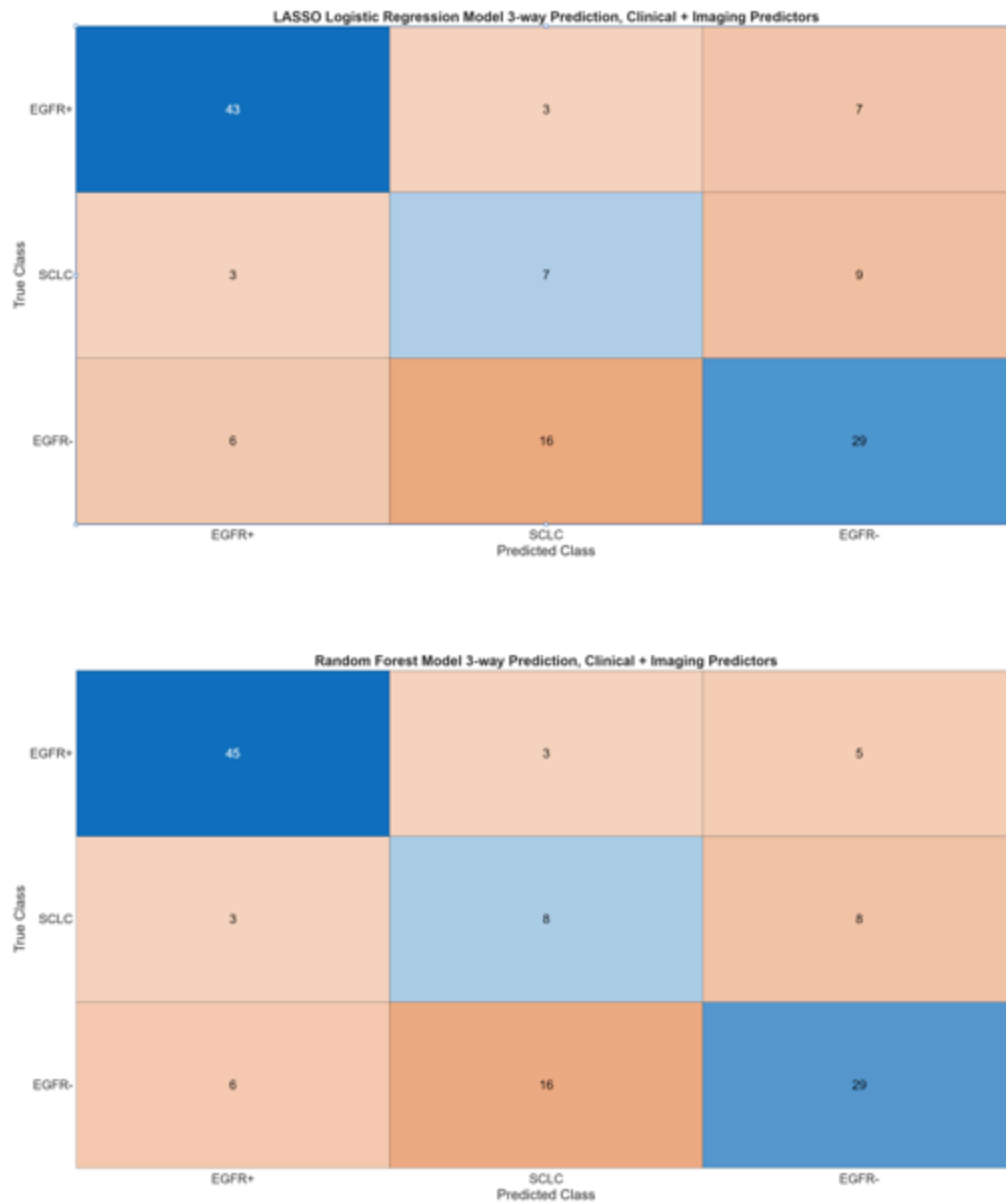


**Figure 2**

Testing cohort receiver operating characteristic curves for (a) LASSO logistic regression and (b) random forest models. Dashed curves represent model performances utilizing clinical variables only, and the corresponding solid curves illustrate the improvement in lung cancer subtype predictability when imaging features are incorporated.



|             | EGFR+ LR | EGFR+ RF | SCLC LR | SCLC RF | EGFR- LR | EGFR- RF |
|-------------|----------|----------|---------|---------|----------|----------|
| PPV         | 82.7%    | 83.3%    | 26.9%   | 29.6%   | 64.4%    | 69.0%    |
| NPV         | 85.9%    | 88.4%    | 87.6%   | 88.5%   | 71.8%    | 72.8%    |
| Sensitivity | 81.1%    | 84.9%    | 36.8%   | 42.1%   | 56.9%    | 56.9%    |
| Specificity | 87.1%    | 87.1%    | 81.7%   | 81.7%   | 77.8%    | 81.9%    |

**Figure 3**

Three-way prediction confusion matrices for the testing cohort using combined clinical and imaging features. Confusion matrices are shown for (a) LASSO logistic regression and (b) random forest models, with (c) corresponding performance metric values shown for each subtype.

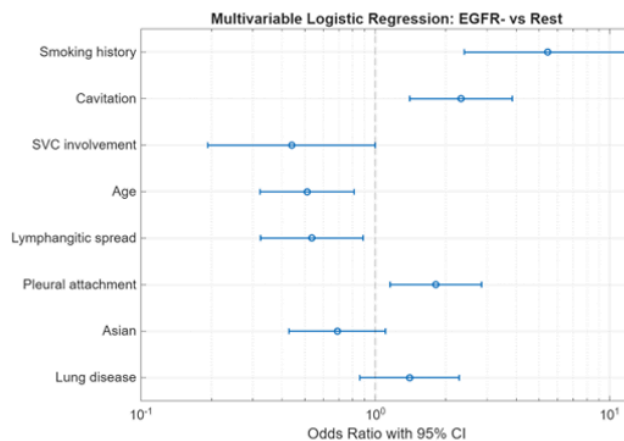
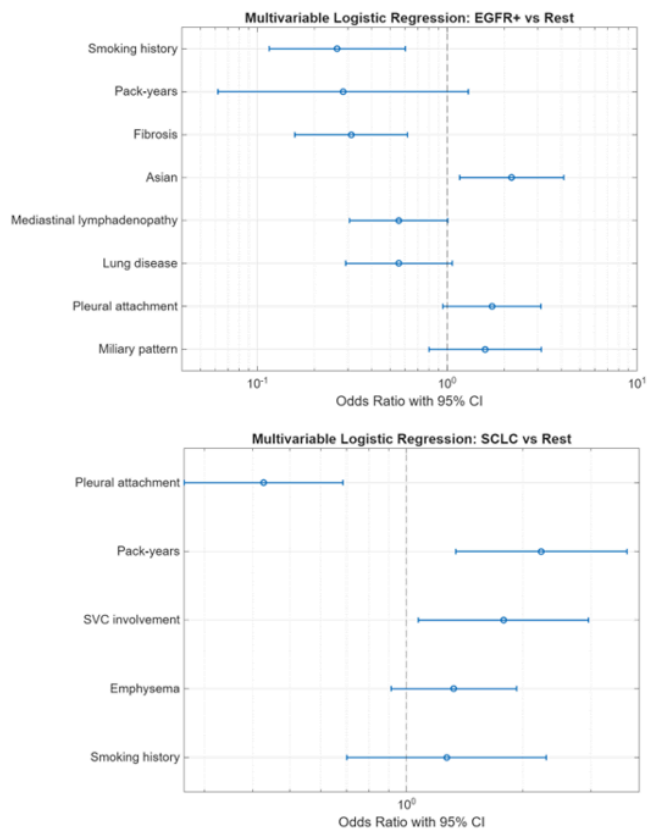


**Figure 4**

Forest plot showing odds ratios of multivariable logistic regression for clinical and imaging predictors of lung cancer subtype (a) EGFR+, (b) SCLC, and (c)EGFR-. Included variables are selected by LASSO penalization.

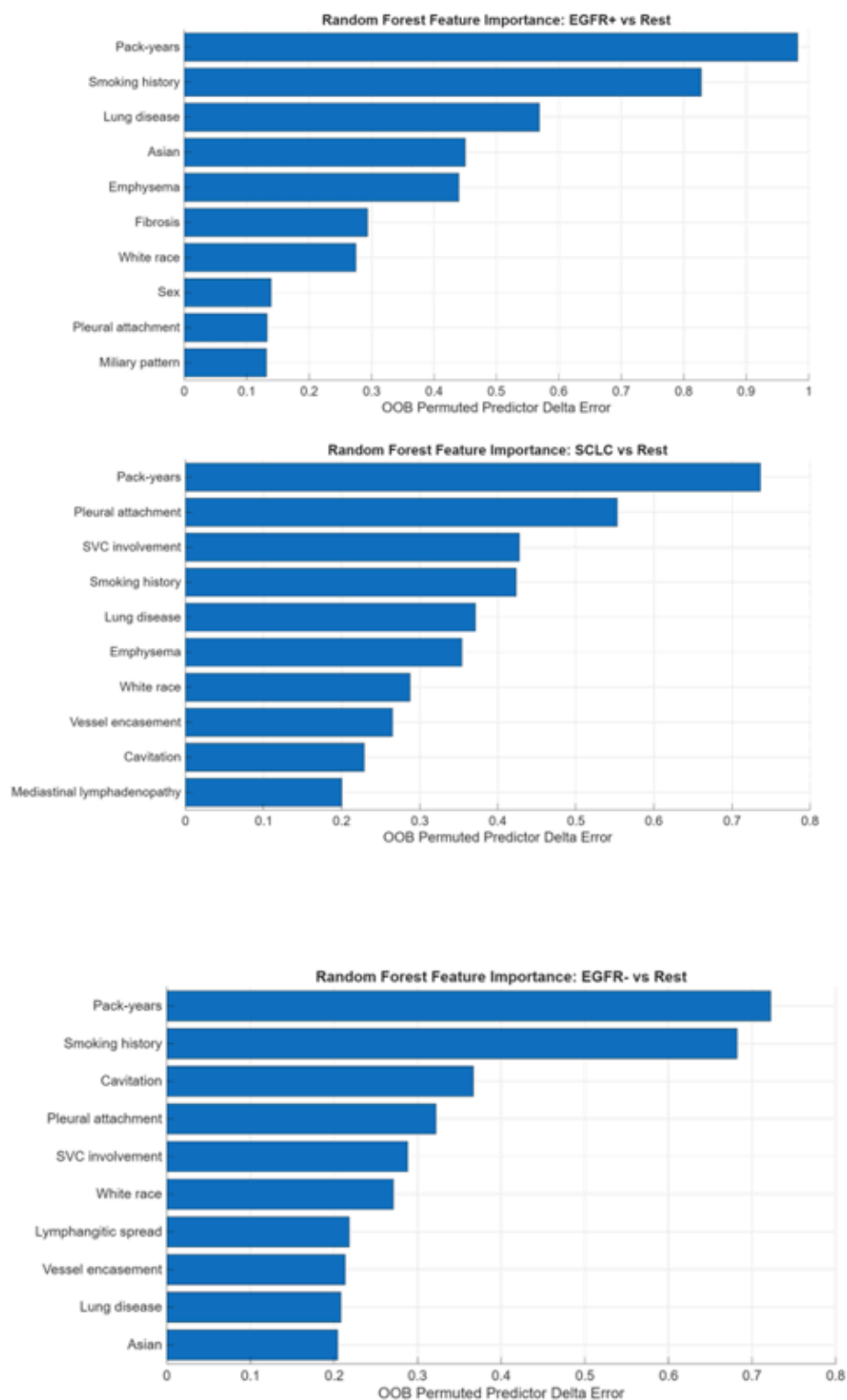


**Figure 5**

Top 10 important features of random forest model (a) EGFR+, (b) SCLC, and (c)EGFR-. Feature importance was calculated using out-of-bag permuted predictor delta error, where higher values indicate variables that contributed more strongly to model classification performance. Bars represent the relative importance of each clinical or imaging feature in the subtype-specific random forest models.

## Supplementary Files

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

- [Sup1Alprompt.docx](#)
- [Supplementaltable1.docx](#)
- [Supplementaltable2.docx](#)