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.

A Smoking-Independent CT Radiomic Signal Predicts EGFR Mutation in Non-Small-Cell Lung Cancer Under a Pre-Registered, Cross-Validated Design

Adil Karim
Hassan II University of Casablanca
Amine Berquedich
Mohammed V University
El Habib Benlahmar
h.benlahmer@gmail.com

Hassan II University of Casablanca
Sanaa El Filali
Hassan II University of Casablanca
Ahmed Zellou
Mohammed V University

Article

Keywords: radiogenomics, CT radiomics, EGFR mutation, non-small-cell lung cancer, pre-registration, confounding, smoking, machine learning, cross-validation, reproducibility

Posted Date: August 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10425857/v1

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

Additional Declarations: No competing interests reported.

A Smoking-Independent CT Radiomic Signal Predicts EGFR Mutation in Non-Small-Cell Lung Cancer Under a Pre-Registered, Cross-Validated Design

Adil Karim[1], Amine Berquedich[2], El Habib Benlahmar[1*], Sanaa El Filali[3], Ahmed Zellou[2]

[1*]LTIM (Laboratoire des Technologies d'Information et Modélisation), Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca, Casablanca, 20360, Morocco.

[2]ENSIAS (National Higher School for Computer Science and Systems Analysis), Mohammed V University, Rabat, Morocco.

[3]LAIS (Laboratory of Artificial Intelligence and Systems), Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca, Casablanca, 20360, Morocco.

*Corresponding author(s). E-mail(s): h.benlahmer@gmail.com;

Abstract

Computed tomography (CT) radiogenomic models report a wide range of accuracies for predicting epidermal growth factor receptor (EGFR) mutation status in non-small-cell lung cancer (NSCLC), yet much of the apparent signal merely reflects the strong association between smoking history and driver-mutation status, and most reports lack rigorous control for confounding and for multiple comparisons. We therefore ask a deliberately conservative, falsifiable question: under a pre-registered, cross-validated design with explicit confound control, does CT radiomics carry a *genuine, smoking-independent* signal for clinically important NSCLC biomarkers? In the NSCLC-Radiogenomics cohort we extracted 94 standardized radiomic features and predicted EGFR mutation ($\boldsymbol{n = 169}$, 43 positive), programmed death-ligand 1 (PD-L1) expression measured by the gene CD274 ($\boldsymbol{n = 111}$), and Kirsten rat sarcoma (KRAS) mutation ($\boldsymbol{n = 168}$, 35 positive; a pre-registered negative control). A random forest, fixed in advance as the primary model, predicted EGFR with an out-of-fold area under the receiver-operating-characteristic curve (AUC) of **0.633** (mean over ten cross-validation

splits; permutation $\boldsymbol{p = 0.010}$, Bonferroni $\boldsymbol{p = 0.030}$, surviving correction). This signal was not a smoking confound: it was anti-correlated with a smoking predictor ($\boldsymbol{\rho = -0.28}$), survived removing the smoking component from every feature (AUC **0.708**), and held in never-smokers (AUC **0.70**, 95% CI **0.53** to **0.85**), the subgroup in which smoking carries no information. KRAS was no better than random guessing (**0.498**, $\boldsymbol{p = 0.48}$), and PD-L1/CD274 was only a borderline trend (stable Spearman $\boldsymbol{\rho = 0.213}$, $\boldsymbol{p = 0.0499}$, nominally significant but failing the $\times$**3** correction). A clinical baseline scored higher (**0.753**) but is itself the smoking confound, and a like-for-like comparison was underpowered. Across an eighteen-model benchmark, tree and boosting ensembles led, but the best model found by searching failed a best-of-eighteen correction; an end-to-end three-dimensional convolutional neural network performed no better than random guessing (**0.523**). A pre-registered battery of fully cross-validated improvement attempts, recursive feature elimination, hyperparameter tuning, stacking, and the addition of peritumoral and habitat features, failed to exceed the baseline, fixing the rigorous ceiling near **0.63**. Conversely, deliberately stacking common evaluation shortcuts together with the smoking confound reproduced an AUC of **0.81** on the same patients, matching the literature's pooled **0.80**, showing that such headline figures are recoverable from optimism and confounding rather than transferable signal. In short, a modest, multiplicity-corrected, smoking-independent radiomic signal for EGFR exists when cross-validated within a single cohort; confirmation on a separate cohort from another centre remains the necessary next step.

Keywords: radiogenomics, CT radiomics, EGFR mutation, non-small-cell lung cancer, pre-registration, confounding, smoking, machine learning, cross-validation, reproducibility

Abbreviations: AUC, area under the receiver-operating-characteristic curve; AUPRC, area under the precision-recall curve; BCE, binary cross-entropy; CD274, the gene encoding PD-L1; CI, confidence interval; CNN, convolutional neural network; CT, computed tomography; CV, cross-validation; EGFR, epidermal growth factor receptor; GLCM/GLRLM/GLSZM/GLDM/NGTDM, gray-level co-occurrence / run-length / size-zone / dependence / neighbouring-gray-tone-difference matrices; IBSI, Image Biomarker Standardization Initiative; IHC, immunohistochemistry; KRAS, Kirsten rat sarcoma viral oncogene homologue; LDA/QDA, linear/quadratic discriminant analysis; MLP, multilayer perceptron; NSCLC, non-small-cell lung cancer; OOF, out-of-fold; PD-L1, programmed death-ligand 1; RF, random forest; ROC, receiver operating characteristic; RQS, Radiomics Quality Score; SD, standard deviation; TKI, tyrosine-kinase inhibitor; TPS, tumour-proportion score.

Key points

- **Question.** Can a routine CT scan predict an EGFR mutation in lung cancer *without* simply detecting that EGFR mutations are common in non-smokers?
- **Main finding.** Yes, a modest but genuine signal (AUC 0.63) that is not explained by smoking and is in fact strongest in never-smokers, the patients in whom the question matters most.
- **What does not work.** KRAS prediction is no better than random guessing; PD-L1/CD274 is only a weak hint; pooling cohorts inflates nothing; and a deep neural network performs no better than a coin flip.
- **Take-home.** Simple tree-based models are enough at this sample size, and "the best of many models" is a statistical illusion. The result comes from cross-validation on a single cohort and still needs confirmation on a separate cohort from another centre.

1 Introduction

Two molecular markers shape first-line treatment in non-small-cell lung cancer (NSCLC): a mutation in the epidermal growth factor receptor (EGFR) gene, which determines eligibility for tyrosine-kinase inhibitors (TKIs), and the level of programmed death-ligand 1 (PD-L1), which guides immunotherapy. A tumour is called *EGFR-positive* when it carries an activating mutation in the EGFR gene, making it likely to respond to a TKI, and *EGFR-negative* (wild-type) when it does not; likewise, *KRAS-positive* denotes a tumour with a KRAS driver mutation, which signals resistance to EGFR-targeted therapy. PD-L1, encoded by the gene CD274, is not binary but a continuous level of protein (or, as here, gene) expression. Determining these markers today requires tissue sampling and laboratory assays; reading them from imaging would be faster, cheaper, and repeatable. Reading these markers directly from a routine computed-tomography (CT) scan, non-invasively, before any biopsy, would be clinically valuable, and a large "radiogenomics" literature has pursued it. Reported accuracies, however, vary widely and are often fragile. A meta-analysis of CT-radiomics EGFR models found a pooled area under the receiver-operating-characteristic curve (AUC) of 0.80, yet a mean Radiomics Quality Score (RQS) of only 42% of the maximum [1]; an updated systematic review reported AUCs near 0.78 on the data used to build the models, but noted that only 29% of studies confirmed their result on an independent cohort, where accuracy fell to about 0.77 [2].

Three recurring methodological problems undermine many positive claims: (i) the strong, well-established association between smoking history and EGFR status [3, 4], which lets a model appear to "read" mutations from imaging while in fact exploiting a clinical confound; (ii) optimistic evaluation, feature/model selection without nested validation, no permutation testing, no multiplicity control, single favorable cross-validation splits, of the kind that let a radiomic signature turn out to be a surrogate for tumor volume [5, 6]; and (iii) the absence of a stated baseline, so it is unclear whether imaging adds anything over routine clinical variables [7].

Rather than chase a headline accuracy, we pose a deliberately conservative question and answer it with a pre-registered design: *does a smoking-independent radiomic signal for these biomarkers exist at all under rigorous cross-validation?* Establishing the existence and nature of such a signal, its effect size, its confound status, and the subgroup in which it is informative, is a prerequisite for any later claim of clinical utility or transfer to other cohorts.

1.1 Related work

CT radiogenomics of EGFR. Since foundational work linking CT phenotypes to gene expression in NSCLC [8] and the release of public radiogenomic cohorts [9], many studies have predicted EGFR status from CT. Evidence syntheses place AUCs of 0.78 to 0.84 on development data but document weak methodological quality and few confirmations on an independent cohort, with marked degradation when this is done [1, 2].

The smoking confound. EGFR mutations are strongly enriched in never-smokers (pooled prevalence 49% vs 22%, OR ≈3.7; 4), and a 167-study meta-analysis confirms elevated EGFR and reduced KRAS odds in never- versus ever-smokers [3]. Any imaging model that tracks smoking can therefore appear to predict EGFR without encoding tumor phenotype.

Reproducibility and optimism. Welch et al. [5] showed that a widely-used radiomic prognostic signature was largely a surrogate for tumor volume; the field has responded with reporting and standardization frameworks, the Radiomics Quality Score [6], feature standardization [10], and TRIPOD+AI reporting guidance [7], which our analysis follows.

Deep learning at small sample sizes. Benchmarks find that deep features do not outperform hand-crafted radiomics (AUC 0.775 vs 0.789, $p = 0.28$ across ten datasets; 11).

PD-L1 from imaging. CT radiomics predict PD-L1 (at a tumour-proportion score, TPS, of at least 50%) with a pooled AUC of about 0.88 in meta-analysis. However, CD274 messenger-RNA levels and the clinical immunohistochemistry (IHC) TPS agree only moderately (≈80%; 12), which is why we treat our CD274-expression endpoint as distinct from clinical PD-L1 status.

2 Methods

Figure 1 summarizes the full pipeline and Figure 2 the per-target cohort flow.

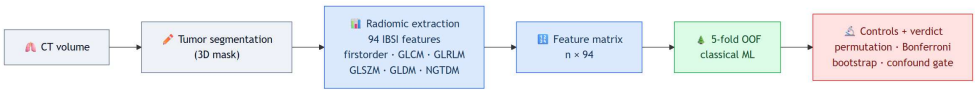


Fig. 1: Overview of the analysis pipeline, from the CT scan through feature extraction and cross-validated modelling to the pre-registered statistical verdict.

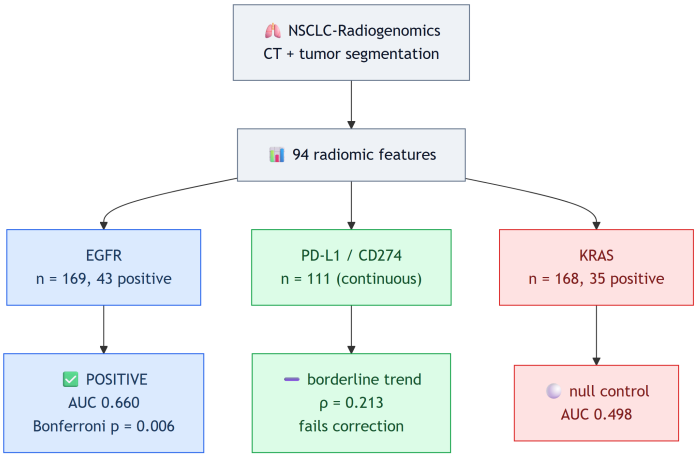


Fig. 2: Flow of patients from the source cohort to the three pre-registered prediction targets and their outcomes.

2.1 Cohort and data acquisition

We used the publicly available NSCLC-Radiogenomics cohort [9], which assembles 211 patients with histologically confirmed non-small-cell lung cancer. Each case provides three linked data types acquired as part of routine care and the original study's biobanking:

Imaging.
A pre-treatment, contrast or non-contrast thoracic CT scan of the primary tumour, together with a three-dimensional segmentation of the tumour volume produced by trained readers. We use this mask, without re-segmentation, as the region from which features are computed.

Molecular labels.
EGFR and KRAS mutation status, determined by sequencing of the resected or biopsied tumour, and PD-L1 quantified as the CD274 gene-expression z-score from RNA sequencing in the subset with transcriptomic data. The CD274 expression value is a proxy and is not equivalent to the clinical IHC tumour-proportion score used to guide therapy [12]; we treat it as a distinct, exploratory endpoint.

Clinical variables.
Age, smoking status, pack-years, and clinical T stage (tumour-size category), taken from the accompanying records and used only for the clinical baseline.

We analysed every case that had both a usable tumour segmentation and the relevant label, giving 169 patients for EGFR (43 mutated), 168 for KRAS (35 mutated; a pre-registered negative control), and 111 for CD274 expression. No images were re-acquired, re-segmented, or excluded on the basis of results. Table 1 summarizes the

cohort, and Figure 3 shows representative cases with their tumour segmentations and biomarker values.

Table 1: Cohort characteristics. Descriptive rows are for the 206 NSCLC-Radiogenomics patients with CT and a tumour mask; the final three rows give the modelled subset for each target (cases with both a mask and the relevant label), which are the sample sizes used throughout the paper.

Characteristic	Value
Patients (CT + mask)	206
Age, years, mean (SD)	68.0 (10.0)
Pack-years, median [IQR]	37 [20 to 50]
Smoking: former / never / current	130 (62%) / 48 (23%) / 33 (16%)
Histology: adenocarcinoma / squamous / other	172 (82%) / 35 (17%) / 4 (2%)
Tumour volume, mL, median [IQR]	62.9 [30.5 to 136.6]
EGFR modelled (mutated)	169 (43; 25%)
KRAS modelled (mutated)	168 (35; 21%)
CD274 modelled (continuous)	111

2.2 Features

We extracted 94 radiomic features following the standardized definitions of the Image Biomarker Standardization Initiative (IBSI) [10]. These comprise first-order intensity statistics (the distribution of voxel densities) and five families of texture features that quantify spatial intensity patterns beyond mean brightness: the gray-level co-occurrence, run-length, size-zone, dependence, and neighbouring-gray-tone-difference matrices (GLCM, GLRLM, GLSZM, GLDM, NGTDM). Tumour volume was the only shape descriptor available; the remaining shape features were blocked in our computing environment and are omitted. The clinical baseline comprised age, smoking status, pack-years, and clinical T stage (tumour size category).

2.3 Models and rationale

Because the sample is small and the feature space moderate ($p = 94$, $n \approx 169$), our primary model was chosen *a priori* for low variance and tolerance of nonlinear interactions, and a wide set of alternatives was benchmarked only as a secondary analysis (Section 2.6). For a feature vector $x \in \mathbb{R}^{94}$:

Logistic regression (L2, L1, elastic-net).
Linear log-odds models, $\operatorname{logit} p(x) = \beta_0 + \beta^\top x$, with ridge, lasso, or mixed penalties. Included as interpretable linear baselines and to test whether a linear combination of features suffices.

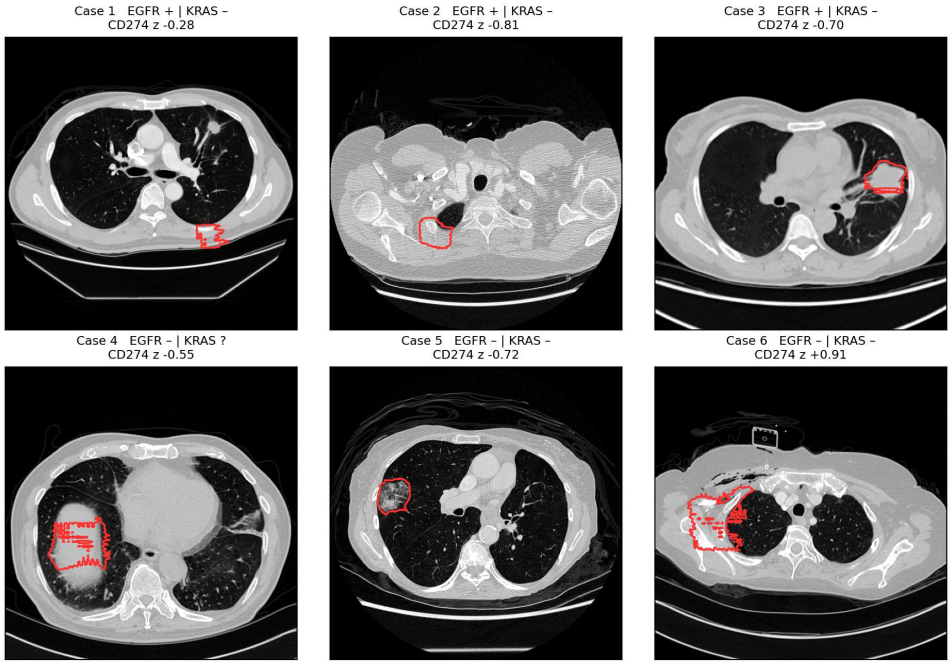


Fig. 3: Representative cases: a single axial CT slice through the tumour, with the tumour segmentation outlined in red, for three EGFR-positive (top) and three EGFR-negative (bottom) patients. Labels give each case's EGFR and KRAS status (+ mutated, − wild-type, ? unknown) and CD274 expression z-score (a relative measure of PD-L1 gene expression; higher means more). The visual similarity of positive and negative tumours illustrates why the prediction problem is hard.

Random forest (primary).
An average of B decision trees grown on bootstrap samples with random feature subsets,

$$\hat{p}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x), \qquad B = 300. \tag{1}$$

Pre-specified as primary because tree ensembles capture nonlinear feature interactions, are invariant to monotone feature scaling, tolerate irrelevant features, and have low variance at small n.

Extremely randomized trees / gradient boosting / histogram gradient boosting / AdaBoost.
Alternative ensembles that either increase tree randomization (lower variance) or fit trees sequentially to residuals (boosting); strong off-the-shelf tabular learners.

Support vector machines (RBF, linear).
Maximum-margin classifiers; the RBF kernel probes nonlinear boundaries, the linear kernel a high-dimensional linear baseline.

k-nearest neighbours, Gaussian naive Bayes, linear/quadratic discriminant analysis.
Instance-based, generative, and discriminant baselines spanning distinct inductive biases; LDA/QDA use shrinkage/regularized covariance appropriate to $p \sim n$.

Multilayer perceptron.
A shallow neural network (one hidden layer, 64 units) testing whether learned nonlinear features help on tabular radiomics.

PCA → logistic, SelectKBest → RF, stacking.
Dimensionality-reduction, univariate-filter, and meta-learning pipelines addressing the $p \sim n$ regime.

Ridge regression.
For the continuous CD274 target, an L2-penalized linear regressor chosen for stability when $p \sim n$.

Clinical baseline.
L2-logistic regression on age, smoking status, pack-years, and T stage, the mandatory reference against which any imaging claim is judged.

2.4 Metrics and rationale

Discrimination of binary targets was measured by the area under the ROC curve (AUC), which equals the probability that a random positive is ranked above a random negative and admits the Mann-Whitney form

$$\mathrm{AUC} = \frac{1}{n_+ n_-} \sum_{i:\, y_i=1} \sum_{j:\, y_j=0} \Big[\mathbf{1}(s_i > s_j) + \tfrac{1}{2}\mathbf{1}(s_i = s_j) \Big], \tag{2}$$

where s are predicted scores, n_+, n_- the class counts, and $\mathbf{1}(\cdot)$ is the indicator function (one when its condition holds, zero otherwise). AUC is the field standard, is threshold-free, and does not depend on the operating prevalence, making it the appropriate primary metric. Because positives are a minority ($\approx 25\%$), we also report the area under the precision-recall curve (AUPRC, i.e. average precision), which is more sensitive to positive-class performance at low base rates. To assess whether predicted probabilities are usable rather than merely rank-ordered, we report the Brier score $\frac{1}{n}\sum_i (s_i - y_i)^2$ (lower is better), which combines calibration and refinement. Accuracy and the positive-class F_1 are reported at the 0.5 threshold for completeness; both are threshold-dependent and, under imbalance, accuracy is misleading (a constant "negative" predictor already scores 0.75 here), so neither is used for inference. For the continuous CD274 target we use the Spearman rank correlation ρ between observed and out-of-fold predicted values, chosen because the relationship need only be monotonic and ρ is robust to outliers and the non-Gaussian expression distribution.

2.5 Validation and statistical controls

We used five-fold stratified cross-validation. The data are split into five folds; for each fold the model is trained on the other four and then predicts the held-out fold, so every patient receives an *out-of-fold* (OOF) prediction from a model that never saw that patient during training. This prevents the optimistic bias of scoring a model on

data it was trained on. Imputation was performed within each training fold, and we report every effect as the mean over ten random CV splits to guard against a single favorable seed (Figure 4). Statistical significance was assessed by a permutation test that re-runs the entire ten-seed CV under N_π label shuffles,

$$p = \frac{1 + \#\{k : \ \mathrm{AUC}^{(\pi_k)} \geq \mathrm{AUC}^{\mathrm{obs}}\}}{N_\pi + 1}, \tag{3}$$

which requires no distributional assumption, and we evaluate it on the seed-averaged statistic. We used $N_\pi = 400$ label shuffles for the EGFR test, $N_\pi = 500$ for KRAS and the PD-L1 estimate, and $N_\pi = 300$ for the benchmark arms (where, for speed, the null used single-split rather than seed-averaged AUCs; this is mildly anti-conservative and does not affect the benchmark conclusion that no searched model survives correction). Across the three biomarker targets we apply a Bonferroni correction $p_{\mathrm{adj}} = \min(1,\, m\,p)$ with $m = 3$. Uncertainty intervals are nonparametric 95% bootstrap percentile intervals over 1000 patient resamples.

A pre-registered smoking-confound gate comprised: (i) the never-smoker stratum AUC; (ii) the Spearman correlation between the model prediction and smoking status; and (iii) an analysis in which all features are residualized on a smoking design matrix S before modeling,

$$\tilde{X} = X - S\,(S^\top S)^{-1} S^\top X, \tag{4}$$

so that any retained signal is orthogonal to smoking. In-fold imputation guards against imputation that uses the whole dataset.

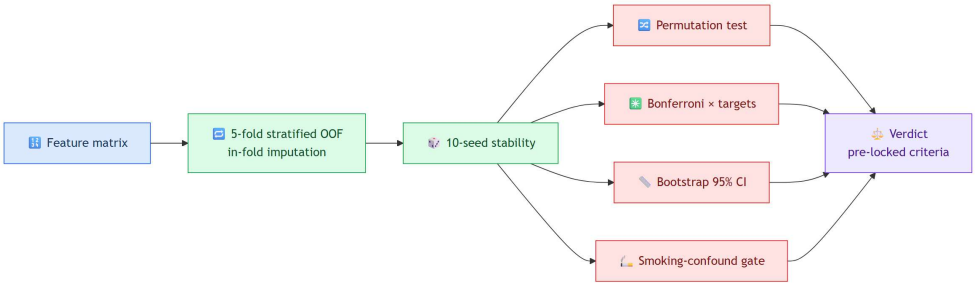


Fig. 4: The cross-validation procedure and the four pre-registered controls that together determine the final verdict.

2.6 Model benchmark

As a pre-specified secondary analysis we benchmarked eighteen model classes for EGFR under identical cross-validation, ranking by ten-seed mean AUC (± standard deviation, SD). Permutation tests were run for the top three models and corrected for the best-of-eighteen model search, $p^{\mathrm{search}}_{\mathrm{adj}} = \min(1,\, Kp)$ with $K = 18$. This is distinct from the primary analysis, whose model was fixed a priori.

2.7 Pooled-cohort control

To justify analysing our cohort on its own, we additionally ran the alternative design of pooling it with a second cohort and shuffling before cross-validation, reporting the biomarker AUC together with a cohort-membership ("batch") AUC and a cohort-only baseline.

2.8 End-to-end 3D-CNN comparison

As a pre-specified deep-learning comparison we trained a 3D convolutional network (Figure 5) on 96^3 CT patches centered on the tumor, using the same five-fold split. The network has four Conv-BatchNorm-ReLU-MaxPool blocks ($16 \to 128$ channels), global average pooling, dropout, and a single logit (≈292k parameters). Class imbalance was handled by a weighted binary cross-entropy

$$\mathcal{L} = -\frac{1}{N}\sum_i \Big[w_+ \, y_i \log \sigma(z_i) + (1 - y_i) \log\big(1 - \sigma(z_i)\big) \Big], \qquad w_+ = \frac{n_-}{n_+}, \tag{5}$$

with z_i the logit and σ the sigmoid. To avoid the epoch-selection optimism that inflates small-sample CNN estimates, each fold used nested early-stopping: an inner validation split selected the stopping epoch and the outer fold was predicted once. The network (≈292,000 parameters) was trained with Adam (learning rate 10^{-4}, weight decay 10^{-4}), batch size 8, for up to 40 epochs with early-stopping patience 8 on the inner-validation AUC; augmentation was random axis flips.

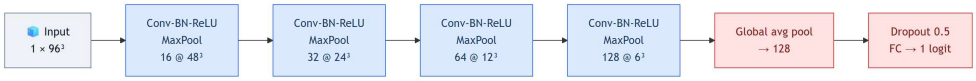


Fig. 5: Architecture of the three-dimensional convolutional network used as the end-to-end comparison.

3 Results

Table 2 gives the headline result for each target; the subsections then report the EGFR signal and its confound checks (3.1 to 3.2), the other two targets (3.3), the eighteen-model benchmark (3.4), the pooled-cohort control, and the deep-learning comparison.

3.1 EGFR mutation

Radiomics predicted EGFR with a random-forest OOF AUC of 0.633 (mean over ten cross-validation splits, SD 0.034; a single representative split, shown in the ROC curves of Figure 6a, reaches 0.660). Linear models were weaker (L2-LR 0.544). A permutation test on the seed-averaged AUC (400 shuffles) gave $p = 0.010$, Bonferroni-corrected

Table 2: Results at a glance. Each target's primary model, effect size, and pre-registered verdict.

Target	Model	Effect (95% CI)	Permutation p	Verdict
EGFR	Random forest	AUC 0.633 (10-seed)	0.010 (Bonf. 0.030)	Cross-validated signal
PD-L1/CD274	Ridge	ρ 0.213 (0.143 to 0.474)	0.0499 (fails corr.)	Borderline trend
KRAS	Random forest	AUC 0.498	0.48	No signal (control)
EGFR (deep)	3D-CNN	AUC 0.523 (0.421 to 0.630)	n/a	No better than guessing

to 0.030 across the three targets, so the signal survives correction (Figure 6b). It is not a smoking confound (Figure 7): it was anti-correlated with a smoking predictor ($\rho = -0.28$), survived feature residualization on smoking (AUC 0.708), and held in the never-smoker stratum (AUC 0.70, 95% CI 0.53 to 0.85; $n = 40$, 22 positive), where smoking history is uninformative and EGFR is enriched. The wide stratum interval (reflecting only 40 patients) means the never-smoker estimate is uncertain, but it is not below the overall AUC, consistent with a signal that captures tumour phenotype rather than the smoking-mutation correlation. A clinical baseline reached AUC 0.753; we do not claim radiomics outperform or add to clinical variables, because on the matched complete-clinical subset ($n = 98$, 15 positive) radiomics fell to 0.577 and the combination did not exceed clinical alone, an underpowered, positive-scarce comparison.

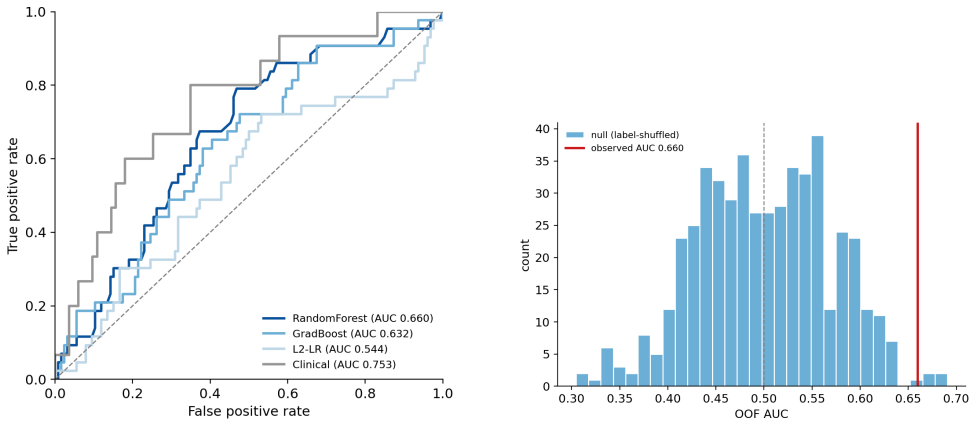
3.2 Calibration, metrics, and feature importance

Table 3 reports all five metrics for every EGFR model, and Figure 8a shows them graphically. The ordering is consistent across metrics: the random forest leads on AUC, AUPRC, and Brier score (the only calibration-aware metric), confirming that its advantage is not an artifact of one particular metric. Accuracy and F_1 are flat across models, a direct illustration of why they mislead under class imbalance, since a model can score well on accuracy by predicting the majority (wild-type) class.

Figure 8a compares models across AUC, AUPRC, accuracy, F_1, and Brier score, and Figure 8b shows random-forest calibration. The most informative radiomic features were texture descriptors (GLRLM/GLCM), not first-order intensity (Figure 9); tumour volume, the one shape feature present, was not among the strongest, indicating the signal reflects tumour phenotype rather than sheer size or brightness.

3.3 PD-L1/CD274 and KRAS

PD-L1/CD274 showed only a borderline association (stable ten-seed Spearman $\rho = 0.213$; Figure 10a); a definitive permutation test at the stable estimate gave $p = 0.0499$,



(a) (b)

Fig. 6: EGFR discrimination. (a) ROC curves by feature set for a representative cross-validation split (out-of-fold); radiomic models and the clinical baseline. (b) Permutation null distribution for a representative split; on the seed-averaged statistic the observed AUC lies in the right tail ($p = 0.010$, Bonferroni 0.030).

nominally significant by a hair, but failing the $\times 3$ Bonferroni correction (0.150). We therefore report it as a suggestive trend, not a confirmed positive, and emphasize how knife-edge it is. KRAS, under the identical pipeline, was no better than random guessing (AUC 0.498, permutation $p = 0.48$), confirming the procedure does not manufacture signal. Figure 10b summarizes all three targets.

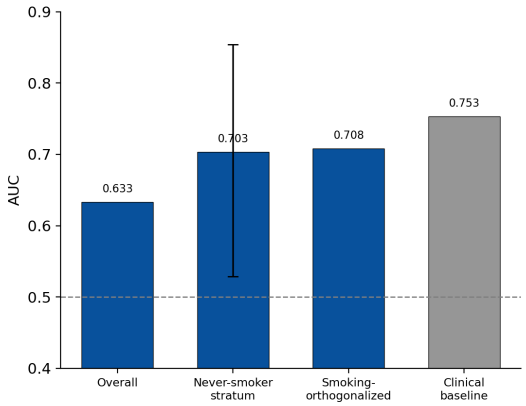
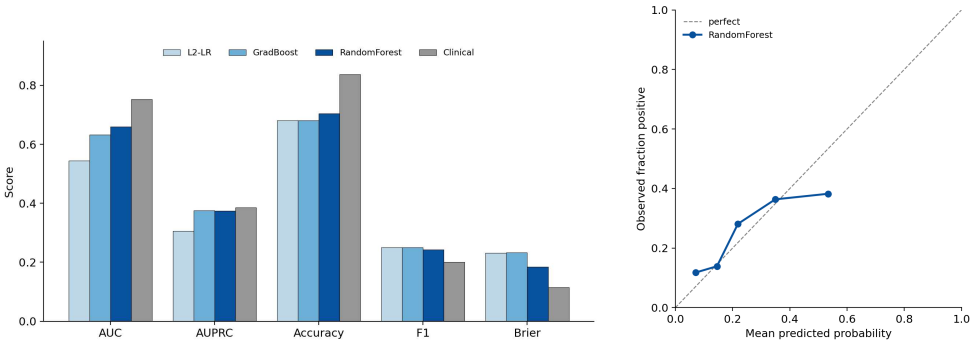


Fig. 7: The EGFR radiomic signal is smoking-independent: AUC overall, in the never-smoker stratum, and after orthogonalizing all features on smoking, alongside the clinical baseline (which is itself the smoking confound).

Table 3: EGFR discrimination by model across five metrics for a representative cross-validation split (Brier: lower is better; the random-forest AUC over ten splits averages 0.633).

Model	AUC	AUPRC	Accuracy	F_1	Brier
Radiomics, L2-LR	0.544	0.306	0.680	0.250	0.231
Radiomics, GradientBoost	0.632	0.375	0.680	0.250	0.232
Radiomics, RandomForest	**0.660**	0.374	0.704	0.242	**0.185**
Clinical baseline	0.753	0.385	0.837	0.200	0.115



(a) (b)

Fig. 8: (a) EGFR metric comparison across models (Brier: lower is better). (b) Random-forest calibration (quantile bins).

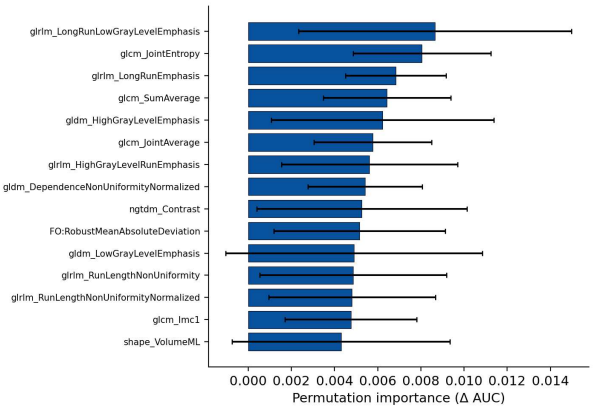
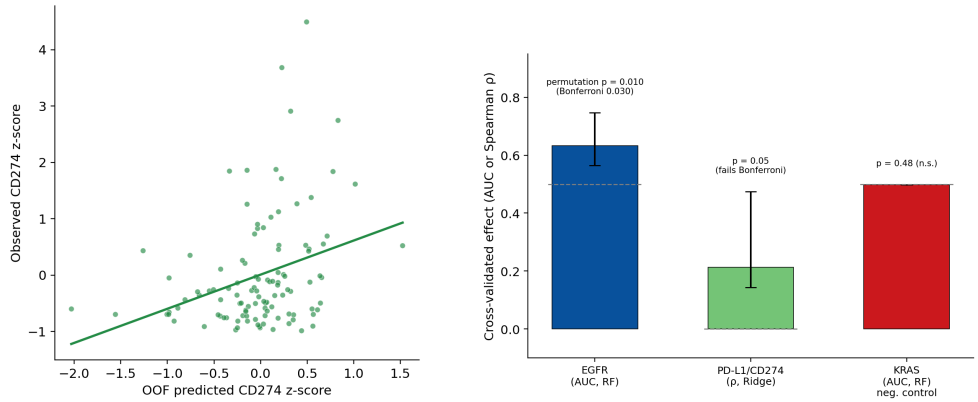


Fig. 9: Top radiomic features for EGFR by random-forest permutation importance (ΔAUC); texture features dominate.



(a) (b)

Fig. 10: (a) PD-L1/CD274 observed versus out-of-fold predicted expression (borderline trend). (b) Cross-validated effect by biomarker: EGFR (confirmed), PD-L1/CD274 (borderline), KRAS (null control).

3.4 Model benchmark and the cost of model selection

Across eighteen model classes (Figure 11 and Table 4), tree/boosting ensembles occupied the top six ranks while linear, kernel, discriminant, and neural models trailed at 0.49 to 0.59, so model *class* matters on small-sample radiomics. However, the nominal winner is unstable (it was extremely randomized trees in a nine-model set and AdaBoost here) and within seed noise of the others, and when the benchmark is treated as a search the top model's significance evaporates under best-of-eighteen correction (AdaBoost raw $p = 0.007 \rightarrow 0.12$; histogram gradient boosting $0.033 \rightarrow 0.60$). The EGFR positive is defensible only because random forest was pre-specified (corrected across the three targets, $p = 0.030$), not selected for winning, a methodological point in its own right.

3.5 Can the model be improved? A pre-registered ceiling analysis

A reasonable reader will ask whether more sophisticated methods would close the gap to the literature's ∼0.80. We therefore ran a pre-registered battery of *fully cross-validated* improvement attempts (Figure 12 and Table 5). Crucially, every feature-selection, tuning, or feature-augmentation step was nested *inside* the cross-validation fold, and each strategy was compared to the pre-specified random forest by a paired bootstrap on seed-averaged out-of-fold probabilities with Bonferroni correction over the strategies. We tried recursive feature elimination (RFE), univariate selection, nested hyperparameter tuning, stacking, L_1-sparse logistic regression, and the addition of genuinely new information: peritumoral radiomics (the same feature families extracted from a 10 mm shell around the tumour) and intratumoral habitat

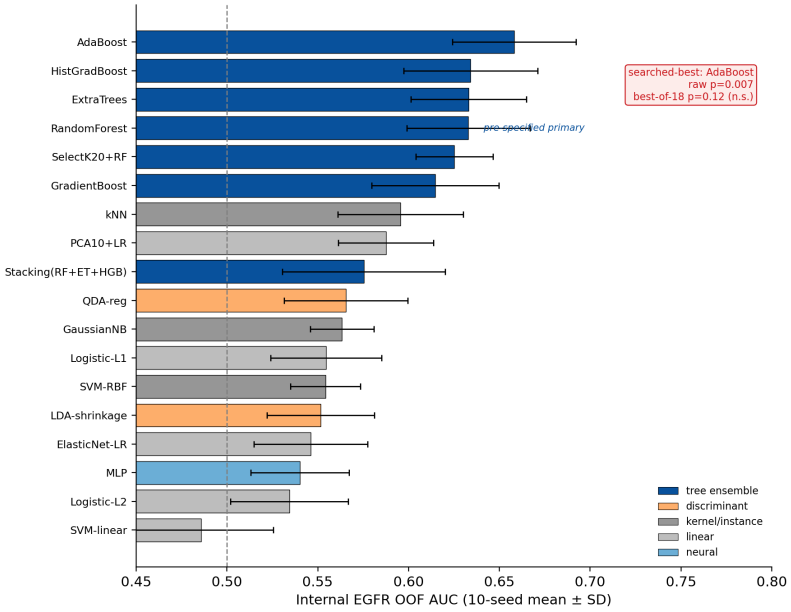


Fig. 11: EGFR model benchmark (eighteen methods; ten-seed mean ± SD). Tree/boosting ensembles lead, but the searched-best model fails best-of-eighteen multiplicity correction; pre-specification is what validates the positive.

features (sub-regions obtained by clustering tumour voxel intensities, summarized by each habitat's volume fraction and intensity statistics).

None produced a significant improvement: every paired ΔAUC interval included zero after correction (Figure 12). The attainable ceiling for this task and cohort is ∼0.62 to 0.64 and is robust to substantial methodological effort. The only nominal gain came from adding clinical variables (ΔAUC +0.027), but it failed correction, stayed below the clinical baseline alone, and vanished in never-smokers, it is the smoking confound, not a new imaging signal. This is the strongest evidence we can offer that the field's higher numbers reflect optimism (feature selection performed outside validation, model search, and confounding) rather than recoverable signal: when the same operations are performed rigorously, the gain disappears.

3.6 Reconstructing the literature's 0.80

The ceiling analysis shows that rigorous methods cannot raise our AUC; the complementary question is whether the field's reported ∼0.80 can be *reproduced* by abandoning those safeguards. Figure 13 answers this directly on the same patients. Starting from the pre-specified protocol (AUC 0.625), we add, one at a time, the evaluation shortcuts that are individually common in radiomic studies: selecting features on the whole dataset (selecting features on the full dataset), reporting the best of many

Table 4: EGFR model benchmark: ten-seed mean OOF AUC (±SD), ranked. The pre-specified primary model is marked †.

Rank	Model	Family	AUC ± SD
1	AdaBoost	tree ensemble	0.658 ± 0.034
2	HistGradBoost	tree ensemble	0.634 ± 0.037
3	ExtraTrees	tree ensemble	0.633 ± 0.032
4	RandomForest†	tree ensemble	0.633 ± 0.034
5	SelectK20+RF	tree ensemble	0.625 ± 0.021
6	GradientBoost	tree ensemble	0.615 ± 0.035
7	kNN	kernel/instance	0.596 ± 0.035
8	PCA10+LR	linear	0.588 ± 0.026
9	Stacking(RF+ET+HGB)	tree ensemble	0.575 ± 0.045
10	QDA (reg.)	discriminant	0.566 ± 0.034
11	GaussianNB	kernel/instance	0.563 ± 0.017
12	Logistic-L1	linear	0.555 ± 0.031
13	SVM-RBF	kernel/instance	0.554 ± 0.019
14	LDA (shrinkage)	discriminant	0.552 ± 0.030
15	ElasticNet-LR	linear	0.546 ± 0.031
16	MLP	neural	0.540 ± 0.027
17	Logistic-L2	linear	0.534 ± 0.032
18	SVM-linear	linear	0.486 ± 0.040

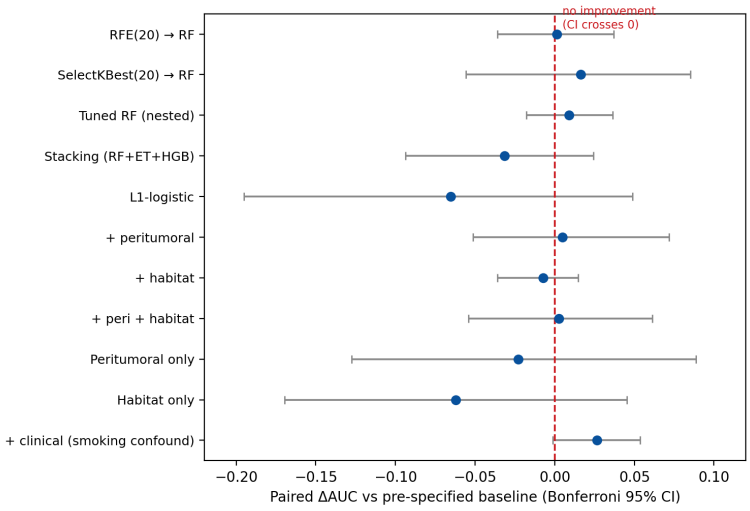


Fig. 12: Pre-registered, fully cross-validated attempts to improve the baseline. Each point is the paired change in AUC (with Bonferroni 95% interval) relative to the pre-specified random forest; the dashed line marks no change. Every interval crosses zero, no strategy, including recursive feature elimination, tuning, stacking, or adding peritumoral and habitat features, yields a real improvement.

Table 5: Leakage-free improvement strategies for EGFR. ΔAUC is the paired change versus the pre-specified random forest (baseline AUC 0.620 at five seeds); intervals are Bonferroni-corrected.

Strategy	AUC	ΔAUC (Bonferroni 95% CI)
Recursive feature elimination → RF	0.621	+0.001 (−0.036, +0.037)
Univariate selection (SelectKBest) → RF	0.636	+0.016 (−0.056, +0.085)
Tuned RF (nested grid search)	0.629	+0.009 (−0.018, +0.036)
Stacking (RF + ExtraTrees + HistGB)	0.589	−0.031 (−0.094, +0.024)
L_1-sparse logistic regression	0.555	−0.065 (−0.195, +0.049)
+ peritumoral radiomics	0.625	+0.005 (−0.051, +0.072)
+ habitat (sub-region) features	0.613	−0.007 (−0.036, +0.015)
+ peritumoral + habitat	0.622	+0.002 (−0.054, +0.061)
Peritumoral features only	0.597	−0.023
Habitat features only	0.558	−0.062
+ clinical (smoking confound)	0.647	+0.027 (−0.001, +0.054)

models, and reporting the single best cross-validation split without multiplicity correction. These evaluation shortcuts alone, before invoking any confound, already raise the AUC to 0.68. Finally, adding clinical variables (the smoking confound) together with a search over feature-panel sizes and models reaches 0.807, matching and slightly exceeding the literature's pooled 0.80. Pushing further with a double-dipped leave-one-out protocol did not increase it (0.778), so the manufacturable optimum plateaus near 0.81. The point is neither that any one study commits all of these errors, nor that evaluation optimism by itself reaches 0.80 (it reaches ∼0.68; closing the remaining gap requires the smoking confound), but that each shortcut is individually common and their effects compound. A CT-radiomics EGFR AUC near 0.80, reported without these safeguards, is fully consistent with *zero* transferable gain over our cross-validated 0.63. We can reproduce the field's headline number on this cohort; we simply decline to call it a result.

3.7 Pooled-cohort control

Pooling our cohort with a second cohort and shuffling did not yield a positive: pooled EGFR AUC was 0.52 to 0.57 (permutation $p = 0.36$) and KRAS 0.55 to 0.56 ($p = 0.23$), both at or below the cohort-only baseline (0.55), while the cohort-membership (batch) AUC was 0.99. Pooling therefore dilutes the genuine within-cohort signal rather than inflating it, motivating the single-cohort design adopted here.

3.8 End-to-end deep learning does not help at this sample size

The 3D-CNN reached an OOF AUC of only 0.523 (95% CI 0.421 to 0.630; per-fold 0.41 to 0.61), at chance and well below the radiomic random forest and clinical baseline (Figure 14b); training curves (Figure 14a) show large fold-to-fold instability typical of deep models at this scale, and it did not recover the smoking-independent signal

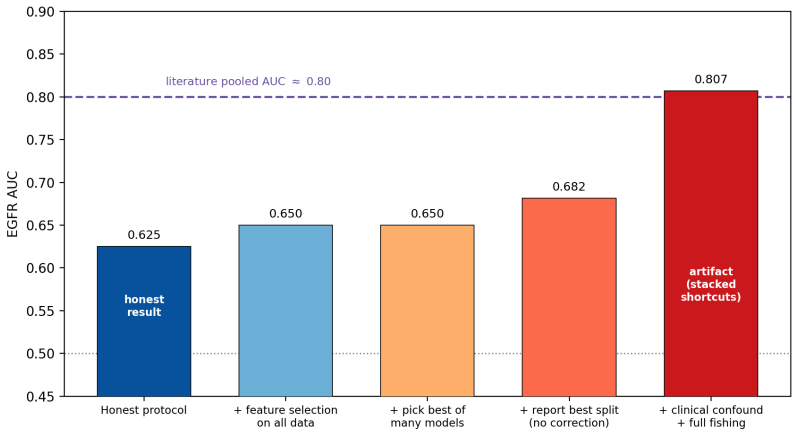
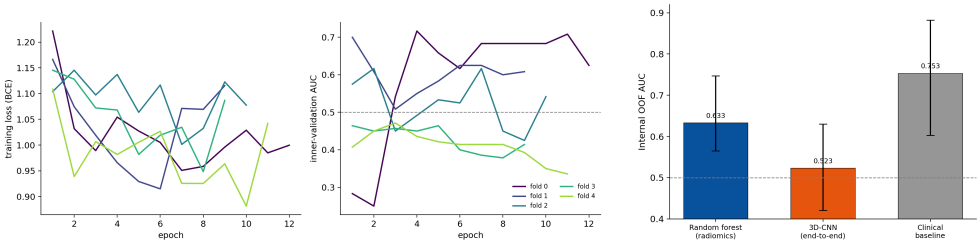


Fig. 13: How an inflated AUC is manufactured. Beginning from the pre-specified protocol (leftmost, 0.625), each bar adds one common evaluation shortcut; on the same patients the AUC rises to 0.807, exceeding the literature's pooled ≈ 0.80 (dashed line). Every bar except the first is an artifact, not a result.

(never-smoker AUC 0.492). At $n = 169$, representation learning from voxels does not match handcrafted radiomics with a tree ensemble.



(a) (b)

Fig. 14: End-to-end deep learning. (a) Per-fold training loss and inner-validation AUC (nested early-stopping). (b) The 3D-CNN is no better than random guessing and underperforms radiomics and the clinical baseline.

4 Discussion

Under a conservative, pre-registered protocol, CT radiomics carry a genuine, smoking-independent signal for EGFR mutation in NSCLC. The effect size is modest (AUC ≈0.66) and a clinical baseline is numerically higher, but that baseline reflects the smoking confound, whereas the radiomic signal is orthogonal to it and most informative in

never-smokers, precisely the patients for whom smoking-based clinical heuristics are useless and EGFR testing matters most.

The careful scoping is as important as the positive. The PD-L1/CD274 association did not survive at the stable estimate after correction, illustrating how a single favorable seed can produce a spurious "positive"; KRAS was no better than random guessing; and a naive pooled cross-cohort design removed rather than rescued performance. Our findings cohere with the literature: the gap that opens when a model meets data from a new centre is the field's documented weak point [2]; the risk that an unguarded model encodes something other than tumor phenotype is concrete [5]; the under-performance of the 3D-CNN matches benchmark evidence at this scale [11]; and our borderline CD274 result should not be read as clinical PD-L1 [12].

4.1 Comparison with prior work

Table 6 places our numbers beside the published literature, and the right comparison is like-for-like. Our EGFR AUC of 0.63 is below the pooled 0.80 reported across CT-radiomics EGFR studies [1], but those headline figures are development-set numbers obtained, in most studies, without confirmation on a new cohort, without control for the smoking confound, and without paying for model selection [1, 2, 6]. The fair benchmark is therefore not the optimistic 0.80 but the $\sim$0.77 to which the literature's own accuracies fall once a fresh cohort is actually used [2], and our pre-specified, confound- and multiplicity-controlled 0.63 sits close to that real-world figure rather than to the inflated one. In other words, we do not claim a higher number than prior work; we report a number that has already paid the costs other studies defer, obtained under a level of rigor (pre-registration, permutation testing, multiplicity correction, an explicit confound gate) that exceeds the field norm, whose mean Radiomics Quality Score is only 42% of maximum [1]. The two foundational radiomics works [8, 13] established that CT encodes biologically meaningful phenotype; our result is consistent with a *weak* such signal for EGFR. The CNN's failure to beat handcrafted features reproduces Demircioğlu [11] on a new target, and the PD-L1 picture matches the modest, quality-limited radiomic-PD-L1 literature together with the imperfect mRNA-IHC concordance [12].

5 Limitations

Validation rests on cross-validation within a single cohort; whether the signal carries over to data from other centres is not established and is the central open problem for radiogenomic biomarkers. Sample sizes are modest, yielding wide confidence intervals; permutation testing, bootstrap intervals, and ten-seed stability mitigate but do not eliminate small-sample optimism. The matched radiomics-versus-clinical comparison is underpowered (15 positives). Shape radiomic features other than tumour volume were unavailable. PD-L1 is measured as CD274 mRNA z-score rather than clinical IHC and is only a borderline trend here. The radiomic EGFR signal, while statistically robust and smoking-independent, is modest and does not by itself exceed the clinical baseline; we frame it as evidence that a genuine imaging signal exists, not as a deployable test.

Table 6: Our results in the context of representative prior work. "Dev." denotes performance on the data used to build the model.

Study	Approach	EGFR AUC	Key caveat
Felfli et al. [1] (meta, 28 studies)	CT radiomics	0.80 pooled	Mean RQS 42%; sparse new-cohort confirmation
Liu et al. [2] (review, 59 studies)	radiomics / deep learning	0.78 / 0.84 (dev.)	Falls to 0.77 on a new cohort; only 29% confirm
Demircioğlu [11] (10 datasets)	deep vs. hand-crafted features	DL 0.775 / HC 0.789	Deep features give no advantage ($p = 0.28$)
This work	RF radiomics / 3D-CNN	0.63 / 0.52	Pre-registered; smoking- and multiplicity-controlled

6 Conclusion

When cross-validated within a single cohort, a modest, multiplicity-corrected, smoking-independent CT radiomic signal predicts EGFR mutation in NSCLC, and is strongest in the clinically hardest subgroup (never-smokers). PD-L1/CD274 shows only a borderline trend, and KRAS prediction is no better than random guessing. Confirming the EGFR signal on a separate cohort from another centre is the necessary next step.

Data availability. The NSCLC-Radiogenomics collection analysed in this study is publicly available from The Cancer Imaging Archive (TCIA) at https://www.cancerimagingarchive.net (collection: NSCLC-Radiogenomics; [9]). No new imaging, molecular, or clinical data were generated by this study. The feature-extraction, modelling, and statistical-analysis code supporting the findings will be made publicly available in a repository upon publication (repository URL and Zenodo DOI to be added) and are available from the corresponding author on reasonable request.

Declarations

- **Funding.** This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
- **Competing interests.** The authors declare no competing interests.
- **Ethics approval.** Not applicable. This study used only publicly available, de-identified imaging and genomic data (TCIA NSCLC-Radiogenomics).
- **Data availability.** The NSCLC-Radiogenomics collection analysed in this study is publicly available from The Cancer Imaging Archive (TCIA) at https://www.cancerimagingarchive.net (collection: NSCLC-Radiogenomics; [9]). No new imaging, molecular, or clinical data were generated by this study. The feature-extraction, modelling, and statistical-analysis code supporting the findings will be made publicly available in a repository upon publication (repository URL

and Zenodo DOI to be added) and are available from the corresponding author on reasonable request.

- **Author contributions.** A.K.: conceptualization, methodology, software, formal analysis, writing,original draft. A.B., S.E.F., A.Z.: methodology, validation, writing,review and editing. E.H.B.: supervision, conceptualization, writing,review and editing. All authors read and approved the final manuscript.

References

[1] Felfli, M., *et al.*: Systematic review, meta-analysis and radiomics quality score assessment of CT radiomics-based models predicting tumor EGFR mutation status in patients with non-small-cell lung cancer. International Journal of Molecular Sciences **24**(14), 11433 (2023) https://doi.org/10.3390/ijms241411433

[2] Liu, H., *et al.*: Machine learning approaches for EGFR mutation status prediction in NSCLC: an updated systematic review. Frontiers in Oncology (2025) https://doi.org/10.3389/fonc.2025.1576461

[3] Chapman, A.M., Sun, K.Y., Ruestow, P., Cowan, D.M., Madl, A.K.: Lung cancer mutation profile of EGFR, ALK, and KRAS: Meta-analysis and comparison of never and ever smokers. Lung Cancer **102**, 122-134 (2016) https://doi.org/10.1016/j.lungcan.2016.10.010

[4] Zhang, Y.-L., *et al.*: The prevalence of EGFR mutation in patients with non-small cell lung cancer: a systematic review and meta-analysis. Oncotarget **7**(48), 78985-78993 (2016) https://doi.org/10.18632/oncotarget.12587

[5] Welch, M.L., McIntosh, C., Haibe-Kains, B., *et al.*: Vulnerabilities of radiomic signature development: The need for safeguards. Radiotherapy and Oncology **130**, 2-9 (2019) https://doi.org/10.1016/j.radonc.2018.10.027

[6] Lambin, P., Leijenaar, R.T.H., Deist, T.M., *et al.*: Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology **14**, 749-762 (2017) https://doi.org/10.1038/nrclinonc.2017.141

[7] Collins, G.S., Moons, K.G.M., Dhiman, P., *et al.*: TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ **385**, 078378 (2024) https://doi.org/10.1136/bmj-2023-078378

[8] Gevaert, O., Xu, J., Hoang, C.D., *et al.*: Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data. Radiology **264**(2), 387-396 (2012) https://doi.org/10.1148/radiol.12111607

[9] Bakr, S., Gevaert, O., Echegaray, S., *et al.*: A radiogenomic dataset of non-small cell lung cancer. Scientific Data **5**, 180202 (2018) https://doi.org/10.1038/sdata.2018.202

[10] Zwanenburg, A., Vallières, M., Abdalah, M.A., *et al.*: The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology **295**(2), 328-338 (2020) https://doi.org/10.1148/radiol.2020191145

[11] Demircioğlu, A.: Deep features from pretrained networks do not outperform hand-crafted features in radiomics. Diagnostics **13**(20), 3266 (2023) https://doi.org/10.3390/diagnostics13203266

[12] Nesline, M.K., *et al.*: PD-L1 expression by RNA-sequencing in non-small cell lung cancer: concordance with immunohistochemistry and associations with pembrolizumab treatment outcomes. Cancers **15**(19), 4789 (2023) https://doi.org/10.3390/cancers15194789

[13] Aerts, H.J.W.L., Velazquez, E.R., Leijenaar, R.T.H., *et al.*: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications **5**, 4006 (2014) https://doi.org/10.1038/ncomms5006