

A multimodal CT-based machine-learning model for differentiating pulmonary sarcomatoid carcinoma from conventional non-small cell lung cancer: an internally validated study

Zhihong Chen

Ganzhou People's Hospital

Wenjian Tang

Ganzhou People's Hospital

Jiawei Feng

Ganzhou People's Hospital

Yuan Kang

Ganzhou People's Hospital

Yujin Yin

Ganzhou People's Hospital

Qing Chen

Ganzhou People's Hospital

Zhongjian Liao

Ganzhou People's Hospital

Xingting Qiu

Department of Medical Imaging, First Affiliated Hospital of Gannan Medical University, Ganzhou, 341000, China.

Ming Jia

Department of Radiology, Southern Medical University Nanfang Hospital, Guangzhou, 510515, China.

Jianping Zhong

Ganzhou People's Hospital

Shuhua Luo

luoshuhua@mail.gzsrmhy.com

Ganzhou People's Hospital

Research Article

Keywords:

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11005575/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background: Pulmonary sarcomatoid carcinoma (PSC) is difficult to distinguish from conventional non-small cell lung cancer (c-NSCLC) before pathological diagnosis. We evaluated whether clinical–CT semantic features, radiomics and deep-learning features can support this distinction.

Methods: After image quality control, 397 patients (83 PSC, 314 c-NSCLC) from three institutions were analysed. Six models (clinical-only, radiomics-only, deep-only, clinical+radiomics, clinical+deep and full fusion) were developed and evaluated with repeated nested cross-validation (five outer folds repeated 20 times, five inner folds), with preprocessing, feature selection, hyperparameter tuning, classifier selection and threshold selection confined to training folds. Discrimination, calibration and operating-point performance were estimated from patient-level out-of-fold predictions. The influence of CT acquisition heterogeneity was examined in sensitivity analyses.

Results: The prespecified full fusion model achieved an AUROC of 0.890 (95% CI 0.848–0.930) and an AUPRC of 0.739 (0.643–0.823; PSC prevalence 0.209). At a training-derived high-sensitivity threshold, sensitivity and specificity were 0.855 and 0.755. The clinical+radiomics model performed similarly (AUROC 0.895, 0.851–0.934), the clinical-only model reached 0.843 (0.796–0.887), and deep features alone were not discriminative (AUROC 0.533). Intratumoral low-attenuation area (LAA) presence and LAA ratio were independently associated with PSC. Acquisition parameters alone also discriminated the groups (AUROC 0.901); after adjustment for acquisition variables, radiomics retained incremental value over clinical features (Δ AUROC 0.075, $P < 0.001$), but residual confounding by source or protocol cannot be excluded.

Conclusions: Intratumoral LAA features and CT radiomics may help flag PSC before pathology in an internally validated setting, whereas Kinetics-pretrained deep features added no value. Because acquisition heterogeneity may contribute to imaging-based performance and a negative prediction cannot exclude PSC, the model should be regarded as exploratory and requires external validation with standardized acquisition before clinical use.

1. Background

Pulmonary sarcomatoid carcinoma (PSC) is a rare form of non-small cell lung cancer with epithelial and mesenchymal differentiation. [1,2] Clinical series describe aggressive behaviour and adverse outcomes, while case reports also document disseminated presentation. [3,4,5,6] These features make recognition before treatment clinically relevant.

PSC has substantial biological and therapeutic heterogeneity. Genomic and clinical studies have examined molecular alterations, immune characteristics and treatment outcomes. [7,8] Studies of advanced disease also report variable responses to immune checkpoint inhibition. [9,10]

MET exon 14 alterations have been investigated in PSC and broader NSCLC cohorts, providing a rationale for molecularly informed management. [11,12] Occasional EGFR-mutated spindle-cell carcinomas have also been reported. [13] These developments highlight the need to distinguish pathological subtypes rather than assume a uniform treatment strategy. [14]

CT studies of pulmonary pleomorphic carcinoma have examined central low attenuation, cavitation and other clinoradiological features in relation to prognosis. [15,16] PSC has also been characterized with fluorodeoxyglucose PET/CT. [17] However, imaging findings overlap with conventional NSCLC (c-NSCLC), and pathological confirmation remains necessary. [18]

Intratumoral low-attenuation area (LAA) and CT enhancement are candidate descriptors for distinguishing PSC from c-NSCLC. [18] This tumour-level construct differs from lung parenchymal low attenuation used to quantify emphysema in surgical NSCLC cohorts. [19]

Quantitative imaging has been studied across several lung-cancer tasks. CT radiomics has been applied to mediastinal lymph-node characterization, mutation prediction and immunotherapy-related outcomes. [20,21,22,23] Deep learning has also been applied to histopathology for immunotherapy prognosis. [24] These applications support further investigation but do not directly validate CT-based PSC diagnosis.

PSC-specific studies have combined CT texture and semantic features to predict MET exon 14 skipping. [25] Machine-learning and population-based models have also addressed distant metastasis and cancer-specific survival in PSC. [26,27] Their endpoints differ from the present diagnostic task.

We compared clinical–CT semantic, radiomic and deep-feature models for distinguishing PSC from c-NSCLC. We evaluated discrimination, calibration and operating thresholds using repeated nested cross-validation. We also examined LAA associations, feature robustness and the influence of CT acquisition heterogeneity.

2. Methods

2.1 Patients

This retrospective study included patients from January 2013 to June 2023. PSC eligibility required immunohistochemical confirmation of a sarcomatoid component exceeding 10% of the tumour, with mesenchymal and epithelial marker expression, and CT before treatment. Patients with treatment before CT, poor image quality or missing images were excluded. The comparator group comprised contemporaneous, untreated patients with histopathologically confirmed lung adenocarcinoma or squamous cell carcinoma following surgery or biopsy and available contrast-enhanced chest CT. PSC was coded 1 and c-NSCLC 0.

Patients were recruited from Southern Hospital of Southern Medical University, Ganzhou People's Hospital and The First Affiliated Hospital of Gannan Medical University. Patient-level centre identifiers

were not retained in the de-identified research dataset; therefore, centre-specific counts could not be reported and centre-held-out validation was not possible. [28,29]

2.2 Image quality control and cohort

Of 504 clinical records, 502 had paired images and masks. We excluded 100 windowed exports without original Hounsfield units (HU), one empty mask and four duplicate images. The final cohort comprised 397 patients: 83 PSC, 117 adenocarcinoma and 197 squamous cell carcinoma cases (Figure 1). Missing clinical values did not exclude otherwise eligible patients.

Each image and mask was checked for readability, nonempty tumour labels, physical-grid agreement and original CT intensity. Masks requiring geometric repair were resampled to the image grid with nearest-neighbour interpolation and reviewed on overlays. Duplicate detection used image pixel hashes and geometry, retaining one record per identical image group. The low padding value outside the field of view was distinguished from loss of HU throughout a windowed export. Representative repaired overlays are shown in Figure S1.

2.3 CT acquisition

The reported acquisition protocol used a 256-slice Revolution CT scanner (GE Healthcare), 100–120 kV, automatic tube current modulation and a 512×512 matrix. Reconstruction thickness was 1.25 mm. Non-contrast, arterial-phase and venous-phase acquisitions were described, with arterial and venous scans at 25 s and 60 s after contrast injection. In the available image files, slice spacing ranged from 0.6 to 10 mm, reflecting nonuniform spatial sampling across stored examinations (Table S1, Additional file 1). Image export format (background padding value) was used as a proxy for acquisition source. We call these categories export-format strata.

2.4 Clinical–CT semantic features

The binary outcome was PSC versus c-NSCLC, with PSC as the positive class for probability estimation. The clinical predictor set contained 12 prediagnostic clinical–CT variables: sex, smoking, central or peripheral tumor location, calcification, cavity or parenchymal vacuole, LAA presence, age, maximum tumor diameter, maximum plain CT attenuation, maximum arterial-phase CT attenuation, maximum venous-phase CT attenuation, and LAA ratio. Survival, treatment, postoperative, stage, invasion, metastasis, and symptom variables were not included in this model specification.

Sex was coded 1 for male and 0 for female. Presence/absence variables were coded 1 for present or yes and 0 for absent or no. Maximum tumour diameter was measured in cm.

LAA denotes low attenuation within the tumour, rather than lung parenchymal emphysema. Tang et al. described intratumoral low-attenuation, nonenhancing regions and an area ratio using the largest LAA and tumour sections, which need not be the same section. [18] We cite this framework to explain the recorded LAA presence and ratio variables. The specific source-cohort measurement protocol was not

retained, so a same-section ratio and the precise tumour-ROI procedure for the recorded maximum CT values could not be verified.

2.5 Segmentation and radiomics

Tumour volumes of interest were manually delineated. We checked each available mask against its CT grid; geometric repair and empty-mask exclusions followed the quality-control rules above.

Radiomic features were extracted with PyRadiomics 3.0.1 from the original CT and tumor mask. The pipeline included original-image features, Laplacian-of-Gaussian filtering at sigma values of 1, 2, and 3 mm, and wavelet-derived features. Images were resampled with B-spline interpolation to 1-mm isotropic spacing, with a fixed gray-level bin width of 25; intensity normalization was disabled. The radiomic table contained 1,130 feature columns. Extraction definitions and the parameter file are documented in the reproducible analysis code. [30] All available shape, first-order and texture classes were enabled; mask correction and label 1 were used. Diagnostic-prefixed outputs were stored separately and never entered the predictor matrix.

Feature robustness was assessed with a simplified mask-perturbation procedure. A spherical structuring element of radius one native voxel was used for three-dimensional erosion and dilation, followed by extraction with unchanged settings. Eroded masks that became empty or had fewer than 10 voxels were unavailable for that perturbation only; original feature values were retained. Perturbation reliability was summarized with a two-way random-effects, absolute-agreement, single-measure ICC (ICC[2,1]) across original, eroded, and dilated conditions, using feature-specific complete cases without imputation. [31]

This perturbation analysis addresses sensitivity to small mask changes; it is not interobserver agreement or repeat annotation. No second independent contour set was available. Features with ICC of at least 0.75 formed the radiomic eligibility pool; 344 features met that threshold for downstream fold-specific modeling. Clinical variables were not subjected to the radiomic ICC screen. [31]

2.6 Deep features

We extracted deep features using torchvision R3D-18 with official Kinetics-400 pretrained weights. The network remained frozen in evaluation mode, without fine-tuning. Global average pooling after layer four produced a 512-dimensional vector for downstream classification. [32]

We cropped each image to the mask bounding box with a 10-mm margin. Values below -1,000 HU were clamped before B-spline resampling to 1-mm spacing. Intensities were clipped to [-1,000, 400] HU and scaled to [0,1]. We replicated the volume into three channels, applied Kinetics channel normalization and resized it to $16 \times 112 \times 112$ using trilinear interpolation. Original image axes were preserved. Channel replication retained all pretrained first-layer convolution weights; we neither replaced nor trained the input convolution.

2.7 Model development and evaluation

We evaluated six models: Clinical-only, Radiomics-only, Deep-only, Clinical+Radiomics, Clinical+Deep and Full fusion. Candidate classifiers were L2-regularized logistic regression, radial-basis-function support-vector classification, random forest and histogram-based gradient boosting. The search sampled 30 configurations, with 8, 8, 7 and 7 allocated to these families, respectively. [28]

We used stratified five-fold outer cross-validation repeated 20 times. Each outer training set used stratified five-fold inner cross-validation to select the classifier and hyperparameters by AUROC. All six models shared the same outer splits. We fitted imputation, filtering, feature selection, scaling and threshold selection within training folds. [28]

The label-blind perturbation ICC screen defined eligible radiomic features before cross-validation. All remaining processing was fitted within the training folds. For radiomic and deep features, we removed zero-variance features and applied two-sided Mann–Whitney U screening at $P < 0.05$. We pruned features with absolute Spearman correlation above 0.90, retaining the feature with higher training AUROC. Balanced L1 logistic regression selected features, with C chosen in the inner search. We standardized image features using training-fold means and standard deviations. Deep features followed this process without ICC screening. Clinical categorical and continuous variables used training-fold mode and median imputation, respectively, without missingness indicators. We concatenated selected image features for early fusion and retained all clinical variables in models containing them. Table S2 provides the search space.

Primary classifiers used balanced class weights; histogram gradient boosting replaced unavailable LightGBM. Inner out-of-fold predictions within each outer training set defined two thresholds. We selected the maximum Youden index and the largest finite threshold achieving training sensitivity of at least 0.80. We then applied these thresholds to the outer test predictions. For patient-level classification, we compared each patient's average probability across repeats with the median of their fold-specific thresholds. No outer test outcome informed preprocessing, selection or threshold choice.

2.8 Statistics

We calculated primary probability metrics from patient-level out-of-fold predictions, averaged across repeats. Two-sided 95% percentile intervals used 2,000 patient-bootstrap samples with seed 2026. Each sample retained the patient's outcome and probability together. We checked that each point estimate fell within its interval. Comparisons used paired DeLong tests and paired patient-bootstrap intervals where specified. Repeat-level means and standard deviations describe variation across splits and are reported separately. [33]

Continuous variables are summarized as median (IQR) and compared with Mann–Whitney U tests. Categorical variables are reported as n (%) and compared with χ^2 or Fisher tests. Table 1 includes group-specific missingness.

We examined associations with PSC using unpenalized logistic regression. Variables with pooled univariable $P < 0.05$ entered the multivariable model. We used 20 imputations and pooled coefficients and variances using Rubin's rules with small-sample degrees of freedom. CT attenuation was scaled per 10 HU, LAA ratio per 0.1, age per year and maximum tumour diameter per cm. These association analyses were separate from cross-validated prediction.

We tested the added association of LAA ratio beyond LAA presence using a likelihood-ratio comparison in the same complete cases. We assessed incremental discrimination separately using out-of-fold predictions from the nested LAA models. We also estimated a descriptive full-cohort Youden cut point for arterial-phase attenuation. [33]

We calculated Brier score, calibration-in-the-large, calibration slope and decision curves using the same original probabilities as the discrimination analysis. Approximate training-fold prior correction was assessed only as a sensitivity analysis (Table S7 and Figure S4). [34,35,36,37]

We examined acquisition heterogeneity by adding acquisition variables and evaluating performance across export-format strata. The five-repeat sensitivity models reused the first five outer splits of the primary analysis. The acquisition-only and radiomics-plus-acquisition comparison used matched search capacity.

We reported AUROC, AUPRC with the prevalence baseline, sensitivity, specificity, PPV, NPV, F1, accuracy and confusion-matrix counts. Calibration measures were Brier score, calibration-in-the-large and calibration slope. Full-cohort apparent fits quantified optimism relative to nested estimates. Stratified AUC compared PSC–control pairs only within the same export-format stratum, weighting strata by their pair counts. Patient bootstrap resampling provided percentile intervals. Other sensitivity analyses examined correlation pruning, complete records, disputed LAA records and class weighting.

Analyses used Python 3.10.20, NumPy 1.26.4, scikit-learn 1.5.2, statsmodels 0.14.4 and SHAP 0.46.0. Extraction used PyRadiomics 3.0.1; the fixed seed was 2026. Code and environment records accompany the reproducible analysis. Supplementary Tables S1–S10 and Figures S1–S4 are provided in Additional file 1.

Additional analyses reused the first five outer splits of the full cohort. Before outcome-guided selection, we regressed each radiomic feature on export-format category, pixel type, voxel z/x/y spacing and matrix x/y/z dimensions. We used the residuals to develop Radiomics-only and Clinical+Radiomics models. Imputation and regression coefficients were fitted within each inner training fold and then refitted on the outer training set. We applied the fitted transformations unchanged to held-out patients. Clinical variables remained in Clinical+Radiomics. Separately, we trained Clinical-only and unadjusted Clinical+Radiomics models within the -1024 HU export-format stratum. Both used shared stratified five-fold outer splits repeated five times and the same inner search. We calculated AUROC, patient-bootstrap intervals and paired DeLong tests from out-of-fold predictions averaged across repeats.

3. Results

3.1 Cohort and clinical characteristics

The cohort comprised 397 patients: 83 with PSC and 314 with c-NSCLC, including 117 adenocarcinoma and 197 squamous cell carcinoma cases. Figure 1 summarizes exclusions from the 504 received records. Table 1 reports clinical characteristics and missingness; Table 2 and Table S9 report regression results. [29]

In multivariable analysis, LAA presence was associated with PSC (odds ratio 4.585; 95% CI 1.783–11.789; P=0.002). LAA ratio was also associated per 0.1 increase (odds ratio 1.245; 95% CI 1.026–1.510; P=0.026). Arterial-phase attenuation was not significant after adjustment (odds ratio per 10 HU 0.767; 95% CI 0.540–1.088; P=0.135).

Adding LAA ratio to LAA presence improved model fit among 371 complete records, comprising 63 PSC and 308 c-NSCLC cases. The likelihood-ratio statistic was 12.173 with 1 degree of freedom (P<0.001). Table S10 reports the incremental analysis and descriptive arterial-phase cut point. [18]

Table 1. Clinical–CT semantic characteristics by pathological diagnosis

Variable / source code	PSC	c-NSCLC	LUAD	LUSC	P
Age	62.000 (56.000, 68.000); missing 0	62.000 (55.000, 68.000); missing 0	60.000 (54.000, 67.000); missing 0	62.000 (55.000, 68.000); missing 0	0.550
Maximum tumor diameter (cm)	5.500 (4.100, 7.250); missing 0	4.450 (3.300, 6.175); missing 0	3.800 (2.800, 5.300); missing 0	4.700 (3.700, 6.500); missing 0	<0.001
Maximum plain CT value	35.000 (30.500, 39.500); missing 20	34.000 (30.000, 40.000); missing 6	33.000 (29.000, 39.000); missing 6	35.000 (30.000, 40.000); missing 0	0.591
Maximum arterial CT value	45.000 (37.500, 52.500); missing 20	57.500 (47.000, 69.000); missing 6	58.000 (48.500, 70.000); missing 6	57.000 (46.000, 69.000); missing 0	<0.001
Maximum venous CT value	51.000 (41.000, 58.500); missing 20	64.000 (53.000, 76.000); missing 7	69.000 (55.250, 79.750); missing 7	61.000 (52.000, 72.000); missing 0	<0.001
LAA ratio	0.268 (0.123, 0.541); missing 20	0.000 (0.000, 0.105); missing 6	0.000 (0.000, 0.000); missing 6	0.000 (0.000, 0.215); missing 0	<0.001
Sex – female	15 (18.1%); missing 0	66 (21.0%); missing 0	45 (38.5%); missing 0	21 (10.7%); missing 0	0.554
Sex – male	68 (81.9%); missing 0	248 (79.0%); missing 0	72 (61.5%); missing 0	176 (89.3%); missing 0	0.554
Smoking – 0.0	23 (27.7%); missing 0	93 (29.6%); missing 0	58 (49.6%); missing 0	35 (17.8%); missing 0	0.734
Smoking – 1.0	60 (72.3%); missing 0	221 (70.4%); missing 0	59 (50.4%); missing 0	162 (82.2%); missing 0	0.734
Central/peripheral location – 0.0	54 (65.1%); missing 0	234 (74.5%); missing 0	112 (95.7%); missing 0	122 (61.9%); missing 0	0.086
Central/peripheral location – 1.0	29 (34.9%); missing 0	80 (25.5%); missing 0	5 (4.3%); missing 0	75 (38.1%); missing 0	0.086
Calcification – 0.0	72 (86.7%); missing 0	261 (83.1%); missing 0	99 (84.6%); missing 0	162 (82.2%); missing 0	0.424
Calcification – 1.0	11 (13.3%); missing 0	53 (16.9%); missing 0	18 (15.4%); missing 0	35 (17.8%); missing 0	0.424
Cavity – 0.0	66 (79.5%); missing 0	255 (81.2%); missing 0	93 (79.5%); missing 0	162 (82.2%); missing 0	0.727

Variable / source code	PSC	c-NSCLC	LUAD	LUSC	P
Cavity – 1.0	17 (20.5%); missing 0	59 (18.8%); missing 0	24 (20.5%); missing 0	35 (17.8%); missing 0	0.727
LAA presence – 0.0	12 (19.0%); missing 20	216 (70.1%); missing 6	95 (85.6%); missing 6	121 (61.4%); missing 0	<0.001
LAA presence – 1.0	51 (81.0%); missing 20	92 (29.9%); missing 6	16 (14.4%); missing 6	76 (38.6%); missing 0	<0.001

Continuous variables: median (IQR); categorical variables: n (%) among observed records. Missing values are reported separately. PSC versus c-NSCLC tests are Mann–Whitney U or χ^2 /Fisher tests as appropriate. Sex: 1 male, 0 female. Presence/absence variables: 1 present/yes, 0 absent/no. Tumour diameter is in cm.

Table 2. Multivariable associations with PSC

Variable	Contrast / increment	Adjusted OR (95% CI)	P
Maximum tumour diameter	Per cm	0.968 (0.849–1.105)	0.634
Maximum arterial CT value	per 10 HU	0.767 (0.540–1.088)	0.135
Maximum venous CT value	per 10 HU	0.989 (0.704–1.388)	0.947
LAA ratio	per 0.1	1.245 (1.026–1.510)	0.026
LAA presence	1 vs 0	4.585 (1.783–11.789)	0.002

PSC is the positive outcome. Estimates are pooled across multiple imputations using Rubin’s rules. Variables meeting the univariable screening criterion entered the multivariable model. LAA ratio: per 0.1 increase; CT attenuation: per 10 HU. Full univariable and multivariable estimates are in Table S9. Binary presence contrasts are 1 versus 0; tumour diameter is per cm.

3.2 Model performance and ablation

Full fusion was the prespecified primary model and achieved AUROC 0.890 (95% CI 0.848–0.930). Clinical+Radiomics performed similarly, with AUROC 0.895 (0.851–0.934). AUROCs were 0.845 (0.799–0.888) for Radiomics-only, 0.843 (0.796–0.887) for Clinical-only, 0.842 (0.795–0.885) for Clinical+Deep and 0.533 (0.455–0.606) for Deep-only.

Clinical+Radiomics exceeded Clinical-only by 0.051 AUROC (95% DeLong CI 0.017–0.086; P=0.003). Full fusion minus Clinical+Radiomics was –0.004 (95% CI –0.010 to 0.001; P=0.148). Clinical+Deep did not improve on Clinical-only. Figure 4 presents the paired comparisons. [33]

Full fusion exceeded Clinical-only by 0.047 AUROC (95% DeLong CI 0.014–0.080; $P=0.005$). Its AUPRC was 0.739 (0.643–0.823), against a prevalence baseline of 0.209. Table 3 and Figure 2 present all six models.

At the Youden threshold, Full fusion sensitivity was 0.771 (0.679–0.859) and specificity was 0.815 (0.772–0.855). At the high-sensitivity threshold, sensitivity was 0.855 (0.775–0.928), specificity was 0.755 (0.708–0.801) and NPV was 0.952 (0.924–0.976). Table 4 reports both thresholds, predictive values, F1 and confusion-matrix counts for every model.

Table 3. Discrimination and original-probability calibration

Model	AUROC (95% CI)	AUPRC (95% CI)	Brier (95% CI)	CITL (95% CI)	Slope (95% CI)
Clinical-only	0.843 (0.796– 0.887)	0.622 (0.524– 0.719)	0.127 (0.110– 0.143)	-0.578 (-0.866–-0.333)	1.152 (0.905– 1.477)
Radiomics-only	0.845 (0.799– 0.888)	0.650 (0.551– 0.744)	0.117 (0.098– 0.136)	-0.274 (-0.567–-0.021)	1.123 (0.902– 1.409)
Deep-only	0.533 (0.455– 0.606)	0.281 (0.210– 0.376)	0.171 (0.151– 0.191)	-0.299 (-0.570–-0.063)	0.295 (-0.241– 0.799)
Clinical+Radiomics	0.895 (0.851– 0.934)	0.752 (0.657– 0.833)	0.096 (0.079– 0.113)	-0.252 (-0.545– 0.010)	1.218 (0.977– 1.556)
Clinical+Deep	0.842 (0.795– 0.885)	0.602 (0.499– 0.704)	0.129 (0.113– 0.145)	-0.604 (-0.889–-0.367)	1.213 (0.958– 1.562)
Full fusion	0.890 (0.848– 0.930)	0.739 (0.643– 0.823)	0.099 (0.081– 0.116)	-0.326 (-0.612–-0.065)	1.247 (1.008– 1.580)

Estimates use original patient-level out-of-fold PSC probabilities, averaged across repeats. Patient bootstrap resampling provided percentile intervals. CITL: calibration-in-the-large; ideal CITL and slope are zero and one, respectively. AUPRC prevalence baseline: 0.209.

Table 4. Classification at two training-derived operating points

Model / point	Sensitivity	Specificity	PPV	NPV	F1	TP / FP / TN / FN
Clinical-only A: Youden	0.855 (0.778– 0.927)	0.697 (0.648– 0.747)	0.428 (0.351– 0.500)	0.948 (0.918– 0.974)	0.570 (0.491– 0.641)	71 / 95 / 219 / 12
Radiomics-only A: Youden	0.663 (0.557– 0.767)	0.774 (0.728– 0.822)	0.437 (0.352– 0.520)	0.897 (0.859– 0.931)	0.526 (0.443– 0.604)	55 / 71 / 243 / 28
Deep-only A: Youden	0.494 (0.386– 0.600)	0.576 (0.522– 0.630)	0.236 (0.173– 0.297)	0.812 (0.760– 0.861)	0.319 (0.245– 0.391)	41 / 133 / 181 / 42
Clinical+Radiomics A: Youden	0.783 (0.697– 0.870)	0.799 (0.754– 0.843)	0.508 (0.419– 0.594)	0.933 (0.903– 0.962)	0.616 (0.534– 0.686)	65 / 63 / 251 / 18
Clinical+Deep A: Youden	0.843 (0.764– 0.919)	0.685 (0.634– 0.734)	0.414 (0.338– 0.486)	0.943 (0.912– 0.971)	0.556 (0.477– 0.624)	70 / 99 / 215 / 13
Full fusion A: Youden	0.771 (0.679– 0.859)	0.815 (0.772– 0.855)	0.525 (0.437– 0.613)	0.931 (0.900– 0.959)	0.624 (0.546– 0.697)	64 / 58 / 256 / 19
Clinical-only B: High sensitivity	0.807 (0.718– 0.890)	0.729 (0.682– 0.777)	0.441 (0.360– 0.517)	0.935 (0.903– 0.963)	0.570 (0.491– 0.641)	67 / 85 / 229 / 16
Radiomics-only B: High sensitivity	0.819 (0.731– 0.896)	0.653 (0.602– 0.704)	0.384 (0.314– 0.455)	0.932 (0.898– 0.963)	0.523 (0.449– 0.592)	68 / 109 / 205 / 15
Deep-only B: High sensitivity	0.843 (0.759– 0.916)	0.057 (0.033– 0.085)	0.191 (0.151– 0.232)	0.581 (0.400– 0.743)	0.312 (0.254– 0.367)	70 / 296 / 18 / 13
Clinical+Radiomics B: High sensitivity	0.831 (0.750– 0.908)	0.764 (0.717– 0.809)	0.483 (0.401– 0.566)	0.945 (0.916– 0.971)	0.611 (0.533– 0.684)	69 / 74 / 240 / 14

Model / point	Sensitivity	Specificity	PPV	NPV	F1	TP / FP / TN / FN
Clinical+Deep B: High sensitivity	0.843 (0.764– 0.919)	0.691 (0.640– 0.739)	0.419 (0.343– 0.491)	0.943 (0.913– 0.971)	0.560 (0.483– 0.629)	70 / 97 / 217 / 13
Full fusion B: High sensitivity	0.855 (0.775– 0.928)	0.755 (0.708– 0.801)	0.480 (0.400– 0.560)	0.952 (0.924– 0.976)	0.615 (0.537– 0.686)	71 / 77 / 237 / 12

A: maximum Youden index; B: largest training-fold threshold with sensitivity at least 0.80. Values are estimates (95% CI), except confusion-matrix counts. Each patient's average out-of-fold probability is compared with their median training-fold threshold.

3.3 Calibration and decision curves

Using the original out-of-fold probabilities averaged across repeats, Full fusion had a Brier score of 0.099 (0.081–0.116), calibration-in-the-large -0.326 (-0.612–-0.065) and slope 1.247 (1.008–1.580). Clinical-only had a Brier score of 0.127 (0.110–0.143), calibration-in-the-large -0.578 (-0.866–-0.333) and slope 1.152 (0.905–1.477). Table 3 reports original-probability calibration for all six models. Negative calibration-in-the-large indicates risk overestimation, while positive values indicate underestimation. Approximate prior correction, evaluated only as a sensitivity analysis, caused systematic risk underestimation (calibration-in-the-large 0.66–0.90) and increased every Brier score, supporting recalibration in the target population before clinical use (Table S7 and Figure S4). [36]

Figure 3 shows calibration and decision curves using the original out-of-fold probabilities averaged across repeats, with treat-all and treat-none strategies. [37]

3.4 Robustness and sensitivity analyses

Of 1,130 radiomic features, 344 met the perturbation ICC threshold of at least 0.75. Median ICC was 0.622, with first and third quartiles of 0.378 and 0.791 (Table S4). [31]

Table S3 lists features retained in at least half the outer training folds and their radiomic ICC values. Table S5 reports classifier-selection frequencies, and Figure S2 shows the leading feature-selection frequencies. Repeat-level means and standard deviations are in Table S6.

Full fusion apparent AUROC was 1.000, compared with nested AUROC 0.890, giving optimism of 0.110. Table S8 reports optimism for all models.

Full fusion/Radiomics-only AUROCs were 0.882/0.823, 0.878/0.836 and 0.889/0.832 at the lower, primary and higher correlation-pruning thresholds, respectively (Figure S3). The primary-threshold sensitivity analysis reused the first five repeats.

In complete-case analyses, Clinical-only AUROC was 0.830 and Full fusion AUROC was 0.860 (n=370). After excluding disputed LAA records, Clinical-only AUROC was 0.848 (n=392). Table S8 provides uncertainty intervals.

Unweighted Full fusion had AUROC 0.896 and calibration-in-the-large 0.195. At the fixed probability threshold, sensitivity/specificity were 0.506/0.971 without weighting and 0.554/0.946 with weighting (Table S8).

The initial acquisition-only model had AUROC 0.808 (0.750–0.862). Stratified AUROCs were Clinical-only 0.831 (0.780–0.879), Radiomics-only 0.843 (0.788–0.891), Clinical+Radiomics 0.893 (0.848–0.932) and Full fusion 0.889 (0.844–0.927). Clinical-only point estimates were similar across export-format strata. Table S8 gives stratum-specific counts and intervals.

In the matched five-repeat analysis, acquisition-only AUROC was 0.901 (0.861–0.937). Adding acquisition variables yielded AUROCs of 0.919 (95% CI 0.880–0.953) for Full fusion, 0.926 (0.889–0.958) for Clinical+Radiomics and 0.917 (0.883–0.948) for Radiomics-only.

Radiomics plus acquisition exceeded acquisition-only by 0.016 AUROC (95% CI –0.004 to 0.037; paired DeLong P=0.129). Table S8 presents this matched-search comparison.

Adding acquisition variables increased Full fusion AUROC by 0.042 (95% CI 0.019–0.065; P<0.001). Clinical+Radiomics increased by 0.048 (95% CI 0.024–0.074; P<0.001), relative to its matched five-repeat baseline.

After acquisition adjustment by linear regression, Radiomics-only AUROC was 0.894 (95% CI 0.847–0.935) and Clinical+Radiomics AUROC was 0.917 (95% CI 0.881–0.949). Their respective gains over the matched original models were 0.058 (95% CI 0.010–0.104; P=0.015) and 0.039 (95% CI 0.010–0.067; P=0.007). Adjusted Clinical+Radiomics exceeded matched Clinical-only by 0.075 AUROC (95% CI 0.031–0.117; P<0.001). The –1024 HU export-format stratum included 185 patients: 58 PSC and 127 controls. Models trained within this stratum achieved AUROCs of 0.831 (95% CI 0.763–0.893) for Clinical-only and 0.940 (95% CI 0.903–0.969) for Clinical+Radiomics. The paired gain was 0.110 (95% CI 0.049–0.174; P<0.001). Table S8 reports these five-repeat sensitivity analyses.

3.5 Exploratory model explanations

We used TreeExplainer to explain PSC probability from the final Full fusion histogram gradient boosting model fitted to all 397 patients. The background comprised 100 patients sampled without replacement using seed 2026. Figure 5 shows the summary and dependence plots in original clinical units. Each clinical dependence panel includes 371 observed values and omits 26 missing values. LAA-presence SHAP contributions were zero in this fitted model, despite its separate adjusted regression association.

4. Discussion

Clinical–CT semantic features and radiomics supported internal discrimination of PSC from c-NSCLC. LAA presence and ratio remained independently associated with PSC. Clinical+Radiomics performed similarly to Full fusion, while Kinetics-pretrained deep features provided no observed gain. These findings extend the investigation of CT semantic and texture measurements in PSC. [18,25]

Repeated nested cross-validation evaluated each patient without using that patient's outcome for model development. Averaging predictions across repeats reduced dependence on a single split. This approach follows reporting and evaluation principles for prediction models. [28,36] Previous PSC models addressed metastasis or survival rather than this diagnostic distinction, so their performance cannot be directly compared with ours. [26,27]

LAA presence and ratio retained adjusted associations with PSC, and adding ratio improved model fit beyond presence alone. These observations complement prior work examining LAA and enhancement in PSC. [18] However, the regression association does not imply that LAA presence drove the final fusion model; its SHAP contributions were zero.

Central low attenuation has been interpreted as necrosis in CT studies of pulmonary pleomorphic carcinoma. [16] Our data lacked lesion-level pathological correlation, so we could not establish its substrate here. A reproducible protocol should specify tumour boundaries, attenuation criteria, the ratio denominator and partial-volume handling. [18] Lung parenchymal emphysema represents a different anatomical construct and should not define intratumoral LAA. [19]

Acquisition parameters alone discriminated the groups well (AUROC 0.901), raising concern about source or protocol effects. Scanner-dependent radiomic variation has been documented in phantom and patient CT data. [38] We therefore evaluated export-format strata, removed linear acquisition associations from radiomic features within training folds and rebuilt models within one export-format stratum. Radiomics retained incremental value over clinical features after adjustment (Δ AUROC 0.075, $P < 0.001$) and within-stratum training (Δ AUROC 0.110, $P < 0.001$). These findings support an additional radiomic contribution, although unmeasured source or protocol differences may still influence it.

ComBat has been evaluated for harmonizing CT radiomic features across imaging protocols in phantom and lung-cancer datasets. [39] We did not apply ComBat; our sensitivity analysis removed measured linear acquisition associations within training folds. Future studies could evaluate harmonization alongside standardized acquisition and verified source labels.

The fixed extraction settings and training-fold feature selection make the radiomic pipeline reproducible. Mask perturbation provided a practical robustness screen, retaining features less sensitive to small contour changes. This screen complements the predictive analysis without measuring agreement between independent readers. [30,31]

The deep branch used a frozen network pretrained on Kinetics human-action videos. [32] Its fixed 512-dimensional representation showed no independent discrimination or additional benefit in fusion.

Models Genesis investigated self-supervised pretraining on three-dimensional medical images and transfer to medical-imaging tasks. [40] Such pretraining merits separate evaluation, but our results do not establish whether it would improve PSC classification.

These models may support diagnostic triage by flagging patients for further pathological assessment. At the high-sensitivity threshold, the observed NPV reflects the study prevalence. A negative prediction cannot exclude PSC. Independent evaluation of calibration and clinical net benefit is needed before choosing an operating threshold. [36,37]

4.1 Limitations

Acquisition heterogeneity is the principal limitation. Linear adjustment and export-format analyses cannot remove all scanner, reconstruction, protocol, referral or patient-mix effects; adjustment may also remove correlated biological information. Patient-level centre identifiers were not retained, preventing centre-held-out evaluation. This retrospective study therefore provides internal validation only. Excluding 100 windowed exports without original HU may introduce selection bias because exclusions differed across histological groups. The image phase of the stored volumes used for segmentation was not recorded in the available dataset. Only one mask per patient was available; contour perturbation cannot establish interobserver reproducibility. The specific LAA and maximum-CT-value measurement protocols were not retained, limiting reproducibility of these semantic variables. Kinetics pretraining differs substantially from CT, and medical-image pretraining was not evaluated. The 83 PSC cases limit subgroup precision; class imbalance and missing measurements further constrain interpretation. The high-sensitivity threshold targets training sensitivity rather than guaranteeing sensitivity in new patients. Weighting comparisons also involved feature and classifier reselection, so they do not isolate the effect of class weights. Calibration and decision curves depend on the study case mix; probabilities require external recalibration before clinical use. We did not evaluate prospective clinical impact.

5. Conclusions

Intratumoral LAA features and CT radiomics may help flag PSC before pathological diagnosis. Kinetics-pretrained deep features added no observed value. The models remain exploratory and require external validation with standardized acquisition and recalibration before clinical use; a negative prediction cannot exclude PSC.

Declarations

Ethics approval and consent to participate

This retrospective study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Ganzhou People's Hospital (Approval No. PJB2025-403-01); the requirement for informed consent was waived owing to the retrospective design.

Consent for publication

Not applicable.

Data availability

The datasets analyzed during the current study are not publicly available due to restrictions related to patient privacy and ethical requirements but are available from the corresponding author on reasonable request and with permission from the relevant ethics committee. De-identified patient-level out-of-fold predictions are provided as Additional file 3.

Funding

This research was funded by the Ganzhou Science and Technology Planning Project (No. GZ2024YLJ026),the Science and Technology Program of the Health Commission of Jiangxi Province (No. 2025110045), the Ganzhou Science and Technology Planning Project (No. 2025YLCCE0069, GZ2024ZSF064).

Authors' contributions

CZ and TW conceived and designed the study, performed the data analysis, and drafted the manuscript. FJ, KY and YY contributed to data collection and image analysis. ZJ and LS supervised the study, contributed to the interpretation of the results, and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors have no acknowledgements to declare.

Use of artificial intelligence

An AI-assisted language model was used for English-language editing, organizational support, and assistance with code-based statistical checking. All source-data verification, analytical specifications, statistical outputs, interpretations, and final manuscript text were independently reviewed and approved by the authors. No synthetic data or AI-generated clinical images were used.

Competing interests

The authors declare no competing interests.

Code availability

The analysis code is provided as Additional file 2.

References

1. Kwon HJ, Lee S, Han YB, Lee J, Kwon S, Kim H, et al. Genomic Landscape of Pulmonary Sarcomatoid Carcinoma. *Cancer Res Treat*. 2024;56(2):442-454. doi:10.4143/crt.2023.764.
2. Kharouaa B, Hayoune A, Gartini S, Rhazari M, Thouil A, Kouismi H. Pulmonary Sarcomatoid Carcinoma: What Makes This Rare Lung Cancer So Challenging? *Cureus*. 2024;16(6):e63045. doi:10.7759/cureus.63045.
3. Hou J, Xing L, Yuan Y. A clinical analysis of 114 cases of sarcomatoid carcinoma of the lung. *Clin Exp Med*. 2018;18(4):555-562. doi:10.1007/s10238-018-0517-2.
4. Ung M, Rouquette I, Filleron T, Taillandy K, Bouchet L, Bennouna J, et al. Characteristics and Clinical Outcomes of Sarcomatoid Carcinoma of the Lung. *Clin Lung Cancer*. 2016;17(5):391-397. doi:10.1016/j.clcc.2016.03.001.
5. Park JS, Lee Y, Han J, Kim HK, Choi YS, Kim J, et al. Clinicopathologic outcomes of curative resection for sarcomatoid carcinoma of the lung. *Oncology*. 2011;81(3-4):206-213. doi:10.1159/000333095.
6. Carter RE, Denning KL, Goebel LJ. A Rare Case of Metastatic Pulmonary Sarcomatoid Carcinoma. *Cureus*. 2025;17(7):e89143. doi:10.7759/cureus.89143.
7. Li X, Wang D, Zhao Q, Ren D, Ren F, Chen G, et al. Clinical Significance and Next-Generation Sequencing of Chinese Pulmonary Sarcomatoid Carcinoma. *Sci Rep*. 2017;7(1):3947. doi:10.1038/s41598-017-04296-2.
8. Wei Y, Wang L, Jin Z, Jia Q, Brcic L, Akaba T, et al. Biological characteristics and clinical treatment of pulmonary sarcomatoid carcinoma: a narrative review. *Transl Lung Cancer Res*. 2024;13(3):635-653. doi:10.21037/tlcr-24-127.
9. Xie M, Chu T, Dong X, Wang H, Chu Q, Cai X, et al. Heterogeneity in advanced pulmonary sarcomatoid carcinoma and its efficacy to immune checkpoint inhibitors. *Eur J Cancer*. 2024;209:114260. doi:10.1016/j.ejca.2024.114260.
10. Birsén G, Gounant V, Girard N, Cadranel J, Monnet I, Raynaud-Donzel C, et al. Efficacy of first-line immune checkpoint inhibitors in pulmonary sarcomatoid carcinoma. *Transl Lung Cancer Res*. 2024;13(9):2212-2221. doi:10.21037/tlcr-24-263.
11. Vuong HG, Ho ATN, Altibi AMA, Nakazawa T, Katoh R, Kondo T. Clinicopathological implications of MET exon 14 mutations in non-small cell lung cancer - A systematic review and meta-analysis. *Lung Cancer*. 2018;123:76-82. doi:10.1016/j.lungcan.2018.07.006.
12. Li Y, Gao L, Ma D, Qiu T, Li W, Li W, et al. Identification of MET exon14 skipping by targeted DNA- and RNA-based next-generation sequencing in pulmonary sarcomatoid carcinomas. *Lung Cancer*. 2018;122:113-119. doi:10.1016/j.lungcan.2018.06.001.
13. Ikushima H, Sakatani T, Masuda Y, Morikawa T, Usui K. Lung spindle cell carcinoma harbouring a constitutively active epidermal growth factor receptor mutation. *Respirol Case Rep*. 2019;7(2):e00395. doi:10.1002/rcr2.395.
14. Boland JM, Mansfield AS, Roden AC. Pulmonary sarcomatoid carcinoma-a new hope. *Ann Oncol*. 2017;28(7):1417-1418. doi:10.1093/annonc/mdx276.

15. Nishida A, Abiru H, Hayashi H, Uetani M, Matsumoto K, Tsuchiya T, et al. Clinicoradiological outcomes of 33 cases of surgically resected pulmonary pleomorphic carcinoma: correlation with prognostic indicators. *Eur Radiol.* 2016;26(1):25-31. doi:10.1007/s00330-015-3811-3.
16. Fujisaki A, Aoki T, Kasai T, Kinoshita S, Tomoda Y, Tanaka F, et al. Pleomorphic Carcinoma of the Lung: Relationship Between CT Findings and Prognosis. *AJR Am J Roentgenol.* 2016;207(2):289-294. doi:10.2214/AJR.15.15542.
17. Luo ZH, Luo XY, Qi WL, Liu Q. 18F-FDG PET/CT imaging in pulmonary sarcomatoid carcinoma. *Front Oncol.* 2024;14:1334156. doi:10.3389/fonc.2024.1334156.
18. Tang W, Wen C, Pei Y, Wu Z, Zhong J, Peng J, et al. Preoperative CT findings and prognosis of pulmonary sarcomatoid carcinoma: comparison with conventional NSCLC of similar tumor size. *BMC Med Imaging.* 2023;23(1):105. doi:10.1186/s12880-023-01065-8.
19. Colombi D, Adebajo GAR, Delfanti R, Chiesa S, Morelli N, Capelli P, et al. Association between Mortality and Lung Low Attenuation Areas in NSCLC Treated by Surgery. *Life (Basel).* 2023;13(6):1377. doi:10.3390/life13061377.
20. Sha X, Gong G, Qiu Q, Duan J, Li D, Yin Y. Discrimination of mediastinal metastatic lymph nodes in NSCLC based on radiomic features in different phases of CT imaging. *BMC Med Imaging.* 2020;20(1):12. doi:10.1186/s12880-020-0416-3.
21. Le NQK, Kha QH, Nguyen VH, Chen YC, Cheng SJ, Chen CY. Machine Learning-Based Radiomics Signatures for EGFR and KRAS Mutations Prediction in Non-Small-Cell Lung Cancer. *Int J Mol Sci.* 2021;22(17):9254. doi:10.3390/ijms22179254.
22. Agazzi GM, Ravanelli M, Roca E, Medicina D, Balzarini P, Pessina C, et al. CT texture analysis for prediction of EGFR mutational status and ALK rearrangement in patients with non-small cell lung cancer. *Radiol Med.* 2021;126(6):786-794. doi:10.1007/s11547-020-01323-7.
23. Wu Y, Zhang W, Liang X, Zhang P, Zhang M, Jiang Y, et al. Habitat radiomics analysis for progression free survival and immune-related adverse reaction prediction in non-small cell lung cancer treated by immunotherapy. *J Transl Med.* 2025;23(1):393. doi:10.1186/s12967-024-06057-y.
24. Wang Y, Ju X, Hua R, Chen J, Dai X, Liu L, et al. Deep learning analysis of histopathological images predicts immunotherapy prognosis and reveals tumour microenvironment features in non-small cell lung cancer. *Br J Cancer.* 2024;131(11):1833-1845. doi:10.1038/s41416-024-02856-8.
25. Miao L, Qiu T, Li Y, Li J, Jiang X, Liu M, et al. Predicting MET exon 14 skipping mutation in pulmonary sarcomatoid carcinoma by whole-tumour texture analysis combined with clinical and conventional contrast-enhanced computed tomography features. *Transl Lung Cancer Res.* 2024;13(6):1232-1246. doi:10.21037/tlcr-24-56.
26. Yi X, Xu W, Tang G, Zhang L, Wang K, Luo H, et al. Individual risk and prognostic value prediction by machine learning for distant metastasis in pulmonary sarcomatoid carcinoma: a large cohort study based on the SEER database and the Chinese population. *Front Oncol.* 2023;13:1105224. doi:10.3389/fonc.2023.1105224.

27. Yi X, Zhu B, Zhang J, Tang G, Luo H, Zhou X. Novel model for cancer-specific survival of patients with pulmonary sarcomatoid carcinoma: A population-based analysis and external validation. *Asian J Surg.* 2024;47(1):184-194. doi:10.1016/j.asjsur.2023.07.002.
28. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med.* 2015;162(1):55-63. doi:10.7326/M14-0697.
29. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ.* 2015;351:h5527. doi:10.1136/bmj.h5527.
30. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339.
31. Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, et al. Assessing robustness of radiomic features by image perturbation. *Sci Rep.* 2019;9(1):614. doi:10.1038/s41598-018-36938-4.
32. Carreira J, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2017. p. 6299-6308. doi:10.1109/CVPR.2017.502.
33. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44(3):837-845. doi:10.2307/2531595.
34. King G, Zeng L. Logistic regression in rare events data. *Political Analysis.* 2001;9(2):137-163. doi:10.1093/oxfordjournals.pan.a004868.
35. Elkan C. The foundations of cost-sensitive learning. In: *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01)*, volume 2; 2001; Seattle, WA, USA. p. 973-978. doi:10.5555/1642194.1642224.
36. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2.
37. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26(6):565-574. doi:10.1177/0272989X06295361.
38. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol.* 2015;50(11):757-765. doi:10.1097/RLI.0000000000000180.
39. Mahon RN, Ghita M, Hugo GD, Weiss E. ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. *Phys Med Biol.* 2020;65(1):015010. doi:10.1088/1361-6560/ab6177.
40. Zhou Z, Sodha V, Pang J, Gotway MB, Liang J. Models Genesis. *Med Image Anal.* 2021;67:101840. doi:10.1016/j.media.2020.101840.

Figures

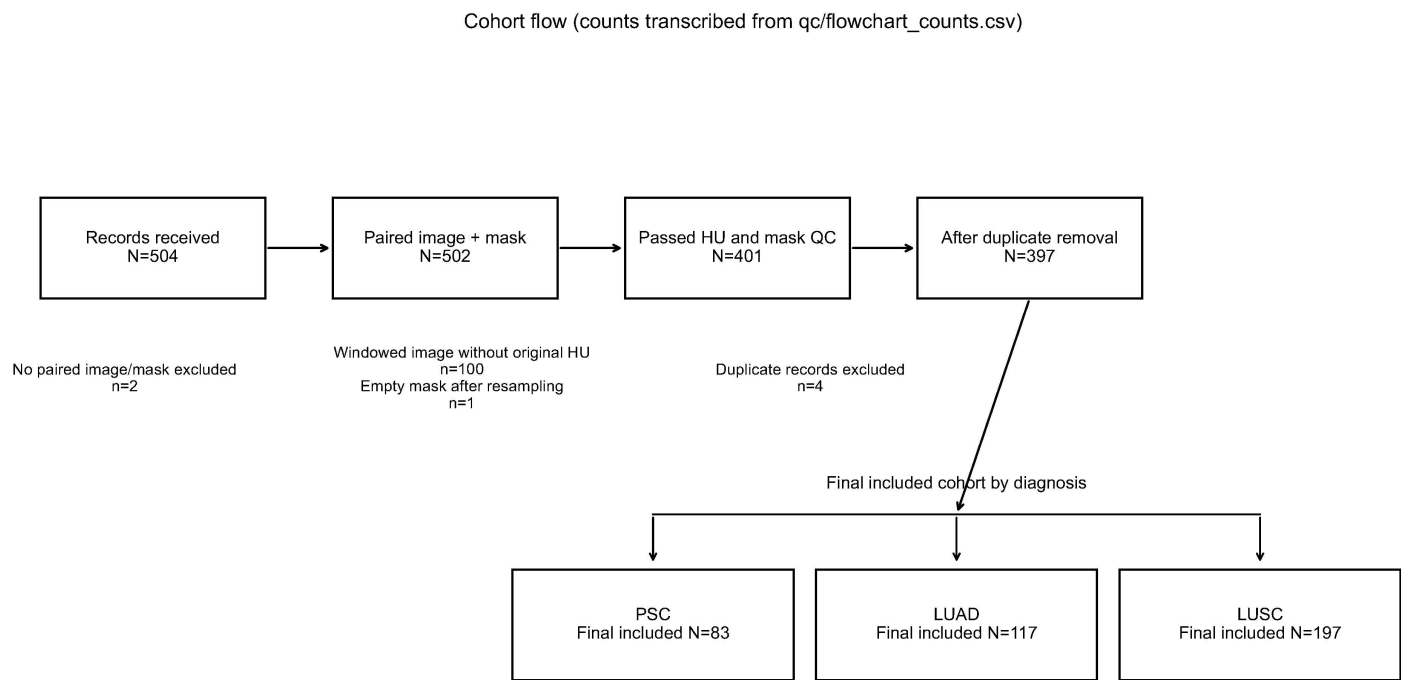


Figure 1

Patient flow through image–mask matching and quality control. PSC is the positive class; adenocarcinoma and squamous cell carcinoma form the c-NSCLC comparison group.

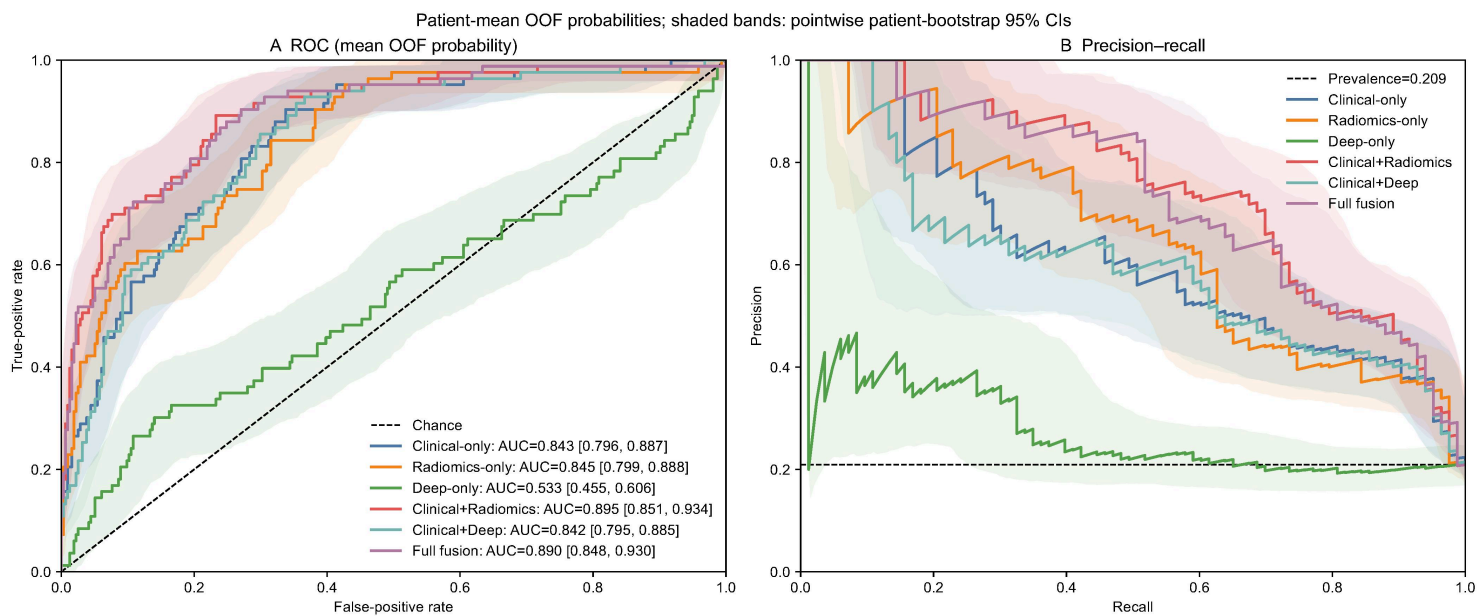


Figure 2

Discrimination of the six models. A receiver operating characteristic curves with patient-bootstrap confidence bands; B precision–recall curves with the PSC prevalence baseline. Curves use patient-level out-of-fold predictions, averaged across repeats.

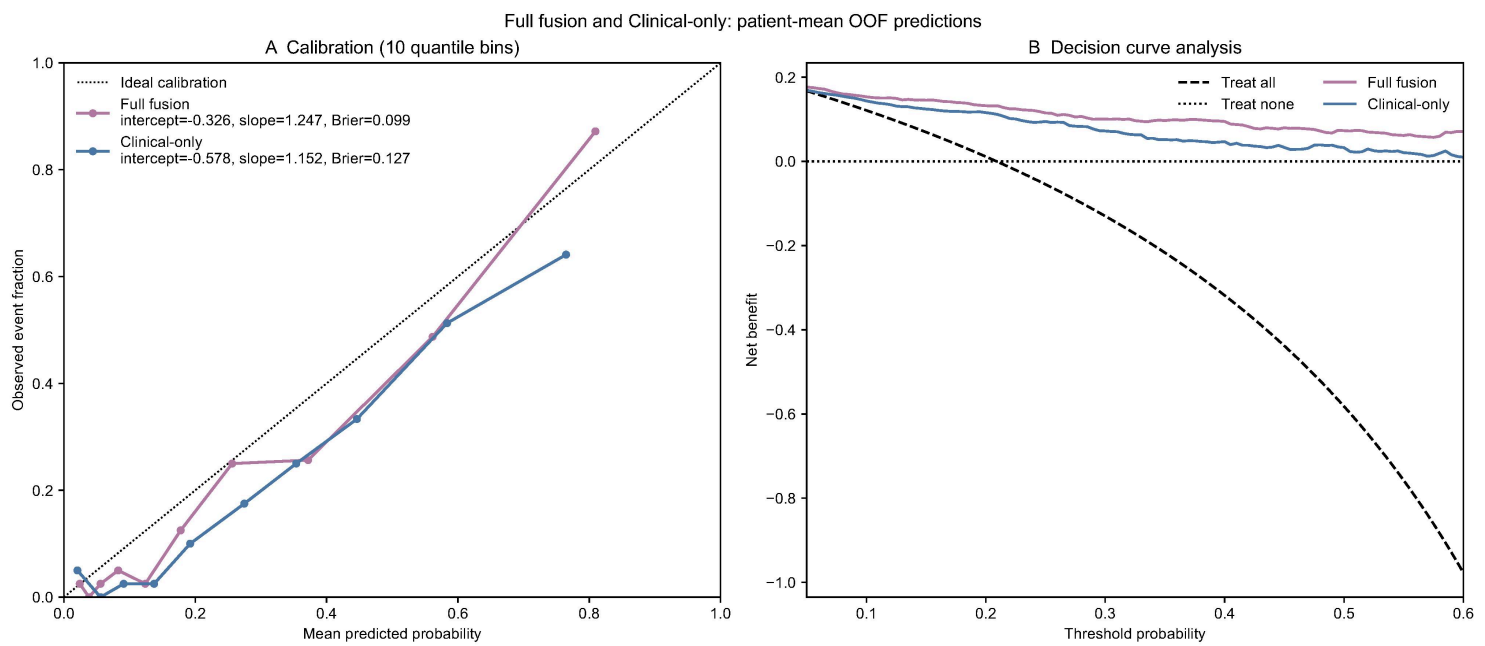


Figure 3

Calibration and decision curves using original probabilities. A calibration curves for Full fusion and Clinical-only, with calibration-in-the-large, slope and Brier score. B decision curves, including treat-all and treat-none strategies. Both panels use patient-level out-of-fold predictions, averaged across repeats, without recalibration.

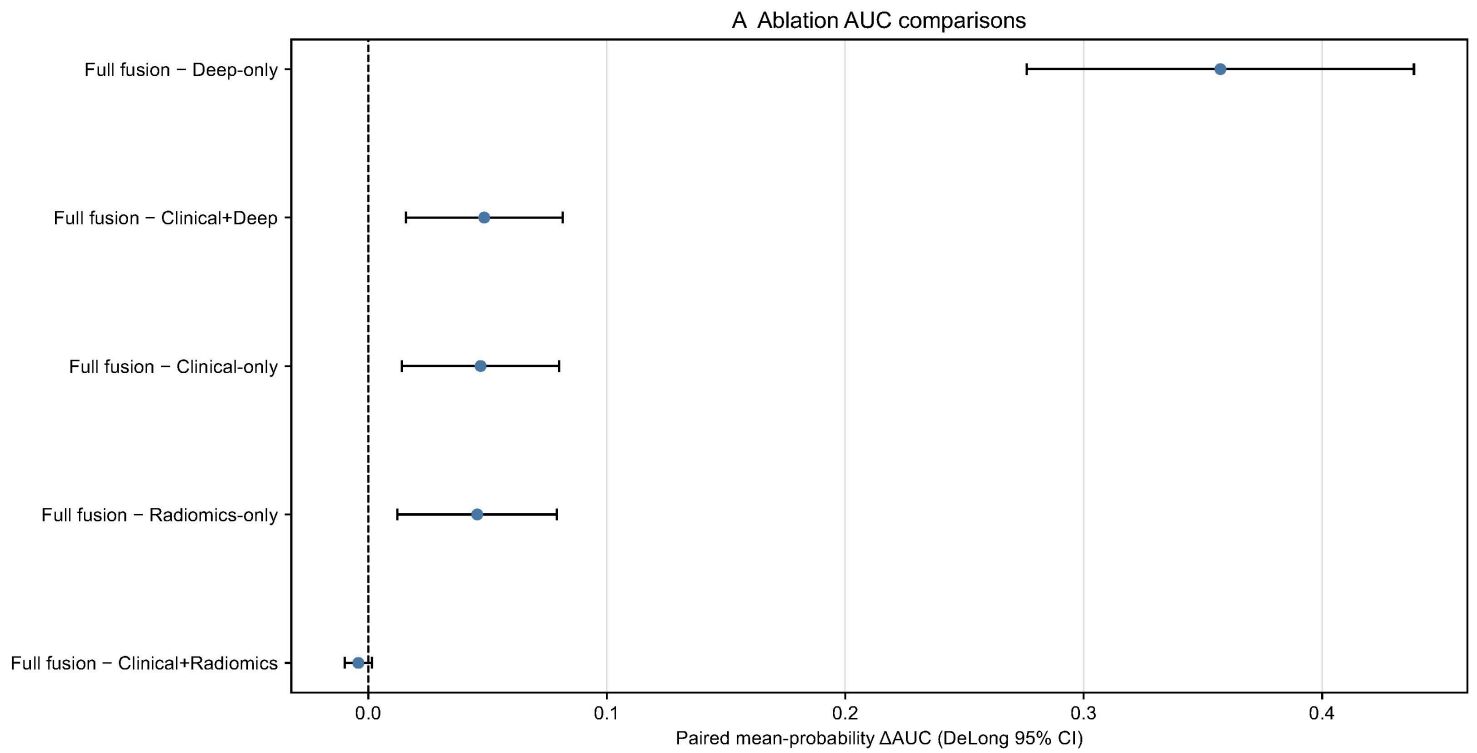
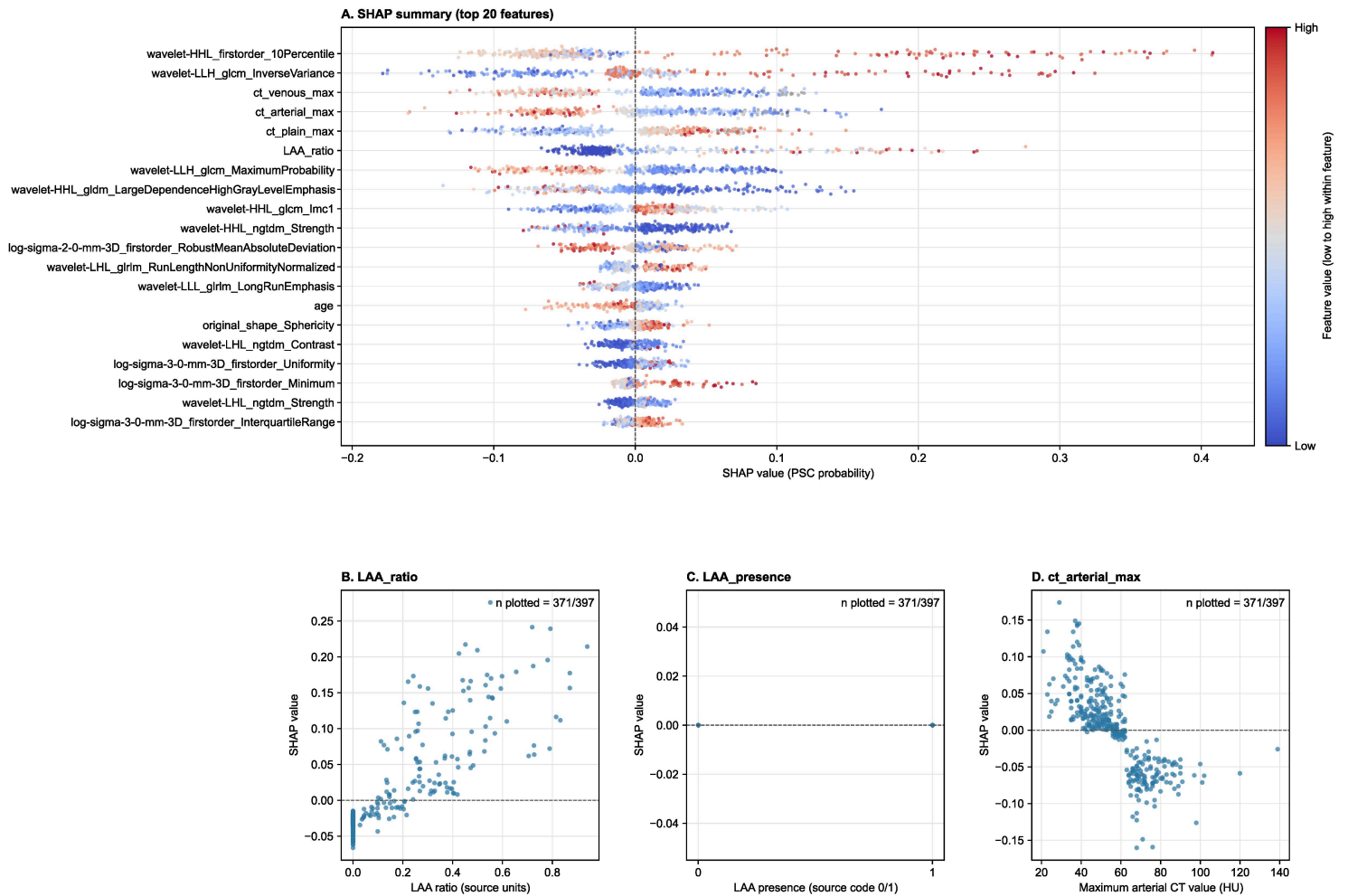


Figure 4

Paired AUROC differences between Full fusion and the other models. Estimates use patient-level out-of-fold predictions, averaged across repeats. Confidence intervals and paired DeLong tests are reported in the corresponding result table.



Panel A ranks the complete transformed feature set by mean absolute SHAP value and displays the top 20; per-case SHAP values for every transformed feature are saved. Panels B-D use raw clinical values on the x-axis. Cases with a missing source value are excluded only from that dependence panel. Deep feature IDs retain their source names; deep features do not have direct biological interpretability.

Figure 5

SHAP explanations of PSC probability from the final Full fusion model. A summary; B LAA ratio; C LAA presence; D maximum arterial-phase attenuation. The model was histogram gradient boosting fitted to the full cohort and explained using TreeExplainer. Clinical axes show original units. LAA-presence contributions were zero in this fitted model; deep-feature identifiers have no direct biological interpretation. Background: 100 patients sampled without replacement, seed 2026; explanation cohort: 397 patients. Each clinical dependence panel uses 371 observed values and omits 26 missing raw values.

Supplementary Files

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

- [Additionalfile1SupplementaryMaterial.docx](#)

- [Additionalfile2analysiscode.zip](#)
- [Additionalfile300Fpredictions.csv](#)