frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 29 October 2025
DOI 10.3389/fonc.2025.1682633

OPEN ACCESS

EDITED BY
Yeon Wook Kim,
Seoul National University, Republic of Korea

REVIEWED BY
Konstantinos Gioutsos,
Inselspital University Hospital Bern, Switzerland
Xiangkui Li,
Harbin University of Science and Technology, China

*CORRESPONDENCE
Yu-Sheng Shu
18051061999@yzu.edu.cn
Xiao-Lin Wang
18051063909@yzu.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 09 August 2025
ACCEPTED 17 October 2025
PUBLISHED 29 October 2025

CITATION
Ren Q-L, Lin L, Chu K, Xu X-R, Wang H-J, Wu J, You J-Z, Hu J-X, Wang X-L and Shu Y-S (2025) Development and validation of machine learning models for predicting STAS in stage I lung adenocarcinoma with part-solid and solid nodules: a two-center study. *Front. Oncol.* 15:1682633. doi: 10.3389/fonc.2025.1682633

Development and validation of machine learning models for predicting STAS in stage I lung adenocarcinoma with part-solid and solid nodules: a two-center study

Qing-Lin Ren[1†], Liu Lin[2†], Kai Chu[3], Xin-Rong Xu[4], Hui-Jun Wang[1], Jun Wu[4], Jin-Zhi You[5], Jun-Xi Hu[4], Xiao-Lin Wang[4*] and Yu-Sheng Shu[4*]

[1]Department of Graduate School, Dalian Medical University, Dalian, China, [2]Department of Thoracic Surgery, Wuxi People's Hospital, Wuxi, China, [3]Department of Graduate School, Xuzhou Medical University, Xuzhou, China, [4]Department of Thoracic Surgery, Northern Jiangsu People's Hospital, Yangzhou, China, [5]Department of Thoracic Surgery, The Affiliated Suqian Hospital of Xuzhou Medical University, Suqian, China

Background: This study aimed to preoperatively predict spread through air spaces (STAS) in stage I lung adenocarcinoma presenting as part-solid and solid nodules by leveraging clinical features and machine learning models, thereby guiding surgical decision-making and enhancing patient counseling.
Methods: A total of 473 patients were retrospectively enrolled, including 353 from our center and 120 from an validation cohort. Predictive features were selected using maximum relevance minimum redundancy (mRMR) and least absolute shrinkage and selection operator (LASSO) algorithms. Seven machine learning models—logistic regression, random forest, support vector machine (SVM), extreme gradient boosting (XGBoost), adaptive boosting (AdaBoost), light gradient boosting machine (LightGBM), and category boosting (CatBoost)—were developed and evaluated using receiver operating characteristic curves, calibration plots, and decision curve analysis (DCA). Feature importance was assessed using Shapley Additive Explanations (SHAP). A web-based nomogram was constructed for clinical application.
Result: STAS was present in 44.76% of the training set and 50.83% of the validation cohort. Seven predictors were selected to construct the predictive models. The XGBoost model demonstrated superior performance with an AUC of 0.889 (95% CI, 0.852–0.926) in training and 0.856 (95% CI, 0.789–0.928) in validation. The calibration curves in training and validation set exhibited good agreement between the predictions and actual observations. The Decision Curve Analyses (DCA) provide significant clinical utility. SHAP analysis identified the most important predictors for STAS as CEA, vascular convergence, proGRP, age, AFP, smoking history, and CTR.

Conclusion: The XGBoost model provides robust preoperative prediction of STAS and may assist clinicians in optimizing surgical strategies for patients with stage I lung adenocarcinoma.

KEYWORDS

spread through air spaces, lung adenocarcinoma, machine learning, solid and part solid component, surgical strategy

Introduction

Lung cancer remains one of the most commonly diagnosed malignancies and the leading cause of cancer-related mortality globally. In China, it accounts for approximately 1.06 million new cases and 0.73 million deaths annually (1). Lung adenocarcinoma (LUAD) is the predominant histological subtype, comprising approximately 85% of non-small cell lung cancer (NSCLC) cases (2, 3). For stage I NSCLC patients undergoing curative R0 resection, 5-year recurrence-free survival (RFS) and overall survival (OS) range from 62.5%–64.7% and 78.7%–81.9%, respectively (4). Despite achieving negative resection margins, early-stage LUAD patients continue to face high locoregional recurrence rates.

Spread through air spaces (STAS) is a newly recognized form of invasion in lung cancer, first proposed in the 2015 WHO classification. It is defined as micropapillary clusters, solid nests, or single cancer cells infiltrating into air spaces beyond the main tumor edge (5, 6), STAS is an independent predictor of poor outcomes in stage I NSCLC and is associated with increased recurrence and reduced survival (7, 8). It also raises recurrence risk in LUAD patients treated with limited resection (9, 10).

With advances in imaging technology, lung cancer is being detected earlier, and smaller nodules are considered suitable for sublobar resection, providing surgeons with more opportunities to choose this approach. providing surgeons with more opportunities to opt for sublobar resection. However, several studies have shown that, compared to lobectomy, sublobar resection is associated with lower RFS and OS in patients with STAS-positive tumors. Patients who underwent wedge resection had significantly worse RFS and OS than those who underwent lobectomy (11). Thus, lobectomy is associated with better outcomes for STAS-positive T1 LUAD compared to sublobar resection (12, 13). Therefore, accurately selecting the surgical approach preoperatively is critical.

Given the importance of preoperative STAS identification for optimal surgical decision-making, enhancing the accuracy of these predictions is crucial for improving patient outcomes. Machine learning (ML) models have gained significant traction in disease prediction due to their ability to process high-dimensional data efficiently (14, 15). However, further optimization and validation of ML models for preoperative STAS prediction are needed to improve their clinical applicability.

We developed several ML models for preoperative STAS prediction using clinical and radiological data from our institution, followed by external validation in an independent cohort from another medical center. The primary goal of this study is to accurately identify STAS preoperatively, facilitating precise surgical decisions, improving patient prognosis, and providing valuable insights for clinical treatment strategies.

Abbreviations: AdaBoost, Adaptive boosting; AFP, Alpha-fetoprotein; AUC, Area under the receiver operating characteristics curve; CA125, Carbohydrate antigen 125; CatBoost, Categorical boosting; CEA, Carcinoembryonic antigen; CTR, Consolidation/Tumor Ratio; cyfra21-1, Cytokeratin 19 Fragment 21-1; DCA, Decision curve analysis; DLCO, Diffusing Capacity of the Lungs for Carbon Monoxide; FEV1, Forced Expiratory Volume in 1 second; GGO, Pure Ground-Glass Opacity; LASSO, Least absolute shrinkage and selection operator; LightGBM, Light Gradient Boosting Machine; LLL, Left Lower Lobe; LUAD, Lung adenocarcinoma; LUL, Left Upper Lobe; ML, Machine Learning; MLR, Monocyte to Lymphocyte Ratio; mRMR, Maximum Relevance Minimum Relevance; NLR, Neutrophil to Lymphocyte Ratio; NSCLC, Non-Small Cell Lung Cancer; NSE, Neuron-Specific Enolase; OS, Overall survival; PLR, Platelet to Lymphocyte Ratio; proGRP, Pro-Gastrin-Releasing Peptide; RFS, Recurrence-Free Survival; RLL, Right Lower Lobe; RML,Right Middle Lobe; RUL, Right Upper Lobe; SHAP, Shapley Additive Explanations; STAS, Spread Through Air Spaces; SVM, Support Vector Machine; WBC, White Blood Cell; XGBoost, Extreme Gradient Boosting.

Materials and methods

Patients

Clinical data were collected from 158 cases of stage I lung adenocarcinoma with STAS admitted to Northern Jiangsu People's Hospital between January 2021 and June 2025. These cases were compared with clinical data from 195 stage I lung adenocarcinoma patients without STAS (the flowchart of this study is shown in Figure 1). This study was approved by the Ethics Committee of Northern Jiangsu People's Hospital.

Inclusion criteria

1. Lung adenocarcinoma confirmed by pathology with or without STAS and single nodule.
2. Solid or part-solid nodule.

3. No preoperative radiotherapy, chemotherapy or targeted therapy.
4. No history of other malignant tumors.
5. No lymph node metastases or distant metastases.
6. Complete clinical data and CT images available.
7. Maximum tumor diameter ≤ 4 cm on CT.

Exclusion criteria

1. Rare histological variants of lung adenocarcinoma.
2. Pure ground-glass opacity (GGO) nodules.
3. Multiple nodules.
4. Preoperative radiotherapy, chemotherapy, or targeted therapy (neoadjuvant therapy).
5. History of other malignancies.
6. Incomplete clinical data or unavailable CT images.
7. Lymph node metastases or distant metastases.
8. Maximum tumor diameter on CT images >4cm on CT.

Based on the postoperative pathological results, patients were classified as either STAS-positive or STAS-negative.

Radiological and histological evaluation

Based on findings in the lung window, the solid component was defined as a patch that completely obscures the underlying lung parenchyma and the GGO component was defined as a hazy area of increased lung attenuation with preserved bronchial and vascular margins. The part-solid nodule was defined as a lesion containing both GGO and solid components, solid nodule was defined as a lesion consisting solely of solid components. In histological evaluation, we focused on stage I lung adenocarcinoma with nodules consisting of solid components, excluding pure GGO nodules.

CT features evaluated included pleural invasion, vascular invasion, spiculation, lobulation, vascular convergence, pleural traction, pleural indention, air bronchogram, vacuole, and consolidation/tumor ratio (CTR). CTR was quantified as the ratio of the tumor consolidation diameter to the total diameter. These CT features were defined and assessed according to previous reports (16–19). Two radiologists with over 10 years of chest imaging experience validated the reliability of the radiological assessments. Only features with good inter-observer agreement between the two radiologists were included in subsequent analyses. Both radiologists were blinded to the patients' STAS status.

Clinical data collection

Clinical information was obtained through the hospital's medical record system. The clinicopathologic features included the age at surgery, sex, smoking history, BMI, pathologic tumor (T) stage (seventh edition of the lung cancer staging system) (20), tumor location, operative methods, forced expiratory volume in 1 s (FEV1), and diffusing capacity of the carbon monoxide (DLCO). Laboratory findings on admission included serum levels of white blood cell (WBC), neutrophil, lymphocyte, monocyte, platelet, albumin, neutrophil to lymphocyte ratio (NLR), monocyte to lymphocyte ratio (MLR), platelet to lymphocyte ratio (PLR), carbohydrate antigen 125 (CA125), alpha-Fetoprotein (AFP), carcinoembryonic antigen (CEA), neuron-specific enolase (NSE), cytokeratin 19 Fragment 21-1(cyfra21-1) and pro-Gastrin-Releasing Peptide (proGRP) within 2 weeks prior to surgery. In this study, patients who underwent lobectomy, segmentectomy, or wedge resection were reviewed. Variables with >30% missingness were removed. Likewise, patients whose records exceeded this threshold across candidate predictors were excluded from model development. For the remaining data, missing values were imputed using the mode for categorical variables and the mean for continuous variables.

We aimed to develop a machine learning model to predict the presence of STAS in stage I lung adenocarcinoma patients using preoperative indicators. Thus, operative methods were excluded from the construction of the machine learning model.

External validation

A validation cohort from Wuxi People's Hospital, affiliated with Nanjing MedicalFI University, was included for external validation. This cohort included 120 patients, 59 of whom were pathologically confirmed as STAS-positive and 61 as STAS-negative, all meeting the inclusion and exclusion criteria.

Predictive model construction and evaluation

Maximum relevance minimum redundancy (mRMR) was applied to the initial feature set to reduce data dimensionality (21). Subsequently, the least absolute shrinkage and selection operator (LASSO) logistic regression was employed to identify key features and develop a predictive model for STAS (22).

To reduce dimensionality while retaining non-redundant information, we first applied maximum relevance minimum redundancy (mRMR) using the mRMRe framework (mutual-information–based relevance with redundancy penalization). We used ensemble selection with feature_count = 15 and solution_count = 1, targeting the binary outcome (STAS positive or negative). The retained candidates were then passed to LASSO logistic regression (glmnet) with $\alpha = 1$. The regularization parameter λ was tuned by 10-fold stratified cross-validation, minimizing binomial deviance. We used λ_min = 0.03469344 for subsequent selection, yielding seven non-zero predictors (CTR, age, smoking history, AFP, CEA, proGRP, vascular convergence).

Seven predictive models were constructed, including logistic regression, random forest, support vector machines (SVM), extreme gradient boosting (XGBoost), adaptive boosting (AdaBoost), light gradient boosting machine (LightGBM), and category boosting

(CatBoost) (23–26) and were trained on the selected clinical features using stratified cross-validation (primary metric AUC). Logistic regression (logit link) reported class probabilities without additional penalties. Random Forest used the Gini criterion with 500 trees and CV-tuned mtry (final mtry = 5). SVM (RBF kernel) tuned C and γ (selected: C = 1×10^{6}, γ = 1×10^{-6}). XGBoost (objective = binary: logistic) tuned learning rate (η), tree depth, min_child_weight, and subsample by 10-fold CV with early stopping, then trained for nrounds = 1000. AdaBoost used mfinal = 10 based on the training-error plateau. LightGBM (metric = AUC) used max_depth = 8 and nrounds = 1000 with leaf/sampling controls chosen by CV. CatBoost (loss = Logloss, eval_metric = AUC) trained with iterations = 1000, learning_rate = 0.1, depth = 8, while l2_leaf_reg and rsm were tuned by CV. After hyperparameter tuning within the training set by stratified 10-fold CV (primary metric AUC), each model was refitted on the full training set and evaluated once on an validation set; we report both cross-validation and external validation performance.

Model performance was evaluated using area under the receiver operating characteristics curve (AUC) and decision curve analysis (DCA) to assess both discriminatory power and clinical utility. Additionally, Shapley additive explanations (SHAP) analysis was employed to interpret model results and assess feature importance, with particular focus on the top-performing machine learning model.

Statistical methods

Statistical analyses were performed using R software (version 4.3.1). Variables with a missing data rate exceeding 20% and outliers were excluded, with remaining missing values addressed via multiple imputation. Continuous variables following a normal distribution were expressed as mean ± standard deviation (x ± s) and compared using the t-test. For non-normally distributed continuous variables, data were presented as median (25th percentile, 75th percentile) [M (P25, P75)] and analyzed using the Mann-Whitney U test. Categorical variables were summarized as counts and percentages [n (%)] and compared using the chi-square test or Fisher's exact test, as appropriate.

Results

Baseline characteristics of patients

A total of 473 stage I lung adenocarcinoma patients, with either STAS-positive or STAS-negative status, met the eligibility criteria for this study. These patients were divided into the training set (n = 353; data from Northern Jiangsu People's Hospital) and the independent validation cohort (n = 120; data from Wuxi People's Hospital, affiliated with Nanjing Medical University).

In the training set, 158 patients were STAS-positive, accounting for 44.76%. In the validation set, 61 patients were STAS-positive, representing 50.83%. Demographic data of all patients were thoroughly examined before modeling.

Table 1 summarizes the baseline characteristics and perioperative serum variables of the 473 patients. Variables such as age, CEA, CYFRA 21-1, proGRP, NLR, CTR, gender, smoking history, pleural invasion, and vascular convergence demonstrated statistically significant differences between groups in the training set ($p< 0.05$) but not in the validation set. CEA was the only variable that showed significant differences in both cohorts.

Selection of variables

We first employed mRMR for initial variable screening to maximize the correlation between features while minimizing inter-feature redundancy, followed by LASSO to identify key STAS-related variables (Figures 2A, B). The selected predictive variables were: CTR, age, smoking history, AFP, CEA, proGRP, and vascular convergence.

Model construction and validation

Using the 7 selected variables, seven machine learning algorithms—logistic regression, random forest, SVM, XGBoost, AdaBoost, LightGBM, and CatBoost—were employed to construct and validate predictive models.

the AdaBoost model demonstrated the highest discriminative ability, with an area under the curve (AUC) of 0.963 (95% CI: 0.947–0.980), along with a sensitivity of 0.886, specificity of 0.908, positive predictive value (PPV) of 0.886, and negative predictive value (NPV) of 0.907. However, the Hosmer-Lemeshow test indicated poor calibration ($P = 1.33 \times 10^{-15}$), suggesting a significant discrepancy between the predicted probabilities and the observed outcomes.

In contrast, the XGBoost model exhibited strong and more balanced performance across both the training and validation sets. In the training set, XGBoost achieved an AUC of 0.889 (95% CI: 0.852–0.926), with a sensitivity of 0.810, specificity of 0.805, PPV of 0.701, and NPV of 0.840. In the validation cohort, the model demonstrated an AUC of 0.856 (95% CI: 0.789–0.928), with sensitivity, specificity, PPV, and NPV of 0.738, 0.881, 0.865, and 0.765, respectively (see Figure 3). Its calibration curve demonstrated close agreement between observed and predicted risks (Figure 4), indicating good calibration. A comprehensive summary of all models is provided in Table 2. The calibration plot demonstrated close agreement between observed and predicted outcomes, indicating good predictive accuracy of the model. (see Figure 4). A comprehensive performance summary for each model is presented in Table 2.

The DCA revealed all seven models consistently provided a net benefit greater than 0 across a range of threshold probabilities (0.0 to 0.8), suggesting their potential clinical utility in guiding STAS prediction decisions (Figure 5). The net benefit rate for all models remained above 0 in both the training and validation sets. The XGBoost model maintained a favorable net benefit within the clinically relevant cost-benefit ratio range (1:4 to 4:1), highlighting

TABLE 1 Characteristics baseline of patients in train set and validation set.

	Train-total (n = 353)	STAS negative (n = 195)	STAS positive (n = 158)	*P* value	Validation -total (n = 120)	STAS negative (n = 59)	STAS positive (n = 61)	*P* value
Age, (years)				0.004				0.426
Mean ± SD	63.13 (10.01)	61.82 (20.30)	64.74 (10.15)		64.18 (9.44)	65.08 (9.35)	63.31 (9.52)	
Gender, No. (%)				0.012				0.102
Female	186 (52.69)	115 (58.97)	71 (44.94)		57 (47.50)	33 (55.93)	24 (39.34)	
male	167 (47.31)	80 (41.03)	87 (55.06)		63 (52.50)	26 (44.07)	37 (60.66)	
BMI (Kg/m^2)				0.773				0.257
Mean ± SD	25.12 (3.42)	25.21 (3.49)	25.01 (3.33)		23.85 (2.73)	23.62 (2.87)	24.08 (2.59)	
smoke, No. (%)				0.013				0.187
no	244 (69.12)	146 (74.87)	98 (62.03)		69 (57.50)	38 (64.41)	31 (50.82)	
yes	109 (30.88)	49 (25.13)	60 (37.97)		51 (42.50)	21 (35.59)	30 (49.18)	
FEV1, (L)				0.612				0.589
Mean ± SD	2.33 (0.89)	2.29 (0.63)	2.39 (1.13)		2.34 (0.71)	2.30 (0.64)	2.38 (1.08)	
DLCO, (mmol/min/kPa)				0.856				0.785
Mean ± SD	6.04 (1.71)	6.01 (1.74)	6.07 (1.68)		6.12 (1.51)	6.10 (1.68)	6.14 (1.81)	
CE125, (U/ml)				0.761				0.916
Mean ± SD	11.26 (10.47)	10.68 (4.99)	11.97 (14.63)		11.15 (11.46)	9.79 (8.90)	12.48 (13.42)	
AFP, (ng/ml)				0.636				0.364
Mean ± SD	3.01 (2.42)	3.15 (3.05)	2.83 (1.28)		2.90 (1.33)	2.73 (1.08)	3.06 (1.52)	
CEA, (ng/ml)				0.004				0.048
Mean ± SD	3.30 (3.73)	2.83 (2.39)	3.88 (4.86)		3.92 (7.23)	4.38 (7.59)	3.47 (6.88)	
NSE, (ng/ml)				0.266				0.209
Mean ± SD	13.73 (3.76)	13.92 (4.05)	13.49 (3.36)		11.64 (3.66)	11.35 (3.66)	11.93 (3.67)	
cyfra.21.1, (ng/ml)				0.038				0.238
Mean ± SD	2.29 (1.01)	2.19 (1.00)	2.41 (1.02)		2.94 (1.41)	2.90 (1.55)	2.99 (1.27)	
proGRP, (pg/ml)								
Mean ± SD	42.15 (23.98)	39.16 (21.27)	45.83 (26.56)	0.002	41.22 (18.87)	42.24 (23.15)	40.23 (13.65)	0.395
WBC, ($\times 10^9$/L)				0.689				0.283
Mean ± SD	5.97 (1.83)	5.96 (1.77)	5.98 (1.90)		5.84 (1.58)	5.61 (1.33)	6.06 (1.78)	
Neut, ($\times 10^9$/L)				0.347				0.389
Mean ± SD	3.92 (3.44)	4.01 (4.37)	3.81 (1.72)		3.57 (3.04)	3.19 (0.90)	3.94 (4.15)	
Lymphocyte, ($\times 10^9$/L)				0.093				0.313
Mean ± SD	1.69 (0.56)	1.73 (0.52)	1.66 (0.61)		1.84 (0.62)	1.78 (0.62)	1.89 (0.62)	
Monocyte, ($\times 10^9$/L)				0.571				0.472
Mean ± SD	0.39 (0.20)	0.39 (0.17)	0.40 (0.23)		0.49 (0.19)	0.46 (0.14)	0.51 (0.23)	
Platelet, ($\times 10^9$/L)				0.594				0.781

(Continued)

TABLE 1 Continued

	Train-total (n = 353)	STAS negative (n = 195)	STAS positive (n = 158)	*P* value	Validation -total (n = 120)	STAS negative (n = 59)	STAS positive (n = 61)	*P* value
Mean ± SD	190.08 (62.55)	193.10 (65.32)	186.35 (58.93)		213.53 (61.84)	212.73 (52.61)	214.31 (70.06)	
Albumin, (g/L)				0.317				0.063
Mean ± SD	45.56 (13.87)	44.68 (4.11)	46.65 (20.21)		39.06 (2.70)	39.53 (2.59)	38.61 (2.75)	
NLR				0.05				0.852
Mean ± SD	2.56 (2.26)	2.49 (2.53)	2.65 (1.89)		2.09 (1.44)	1.96 (0.80)	2.21 (1.86)	
MLR				0.232				0.783
Mean ± SD	0.25 (0.13)	0.24 (0.11)	0.26 (0.14)		0.29 (0.16)	0.28 (0.10)	0.30 (0.20)	
PLR				0.378				0.350
Mean ± SD	120.12 (47.88)	118.19 (46.27)	122.49 (49.84)		127.27 (54.44)	131.45 (55.08)	123.22 (53.96)	
CTR				0.012				0.219
Mean ± SD	0.88 (0.54)	0.84 (0.58)	0.91 (0.48)		0.87 (0.52)	0.84 (0.58)	0.90 (0.48)	
Maximum solid component diameter (cm)				0.214				0.001
	1.75 (1.40)	1.72 (1.32)	1.79 (1.48)		1.82 (1.72)	1.57 (1.64)	2.06 (1.66)	
T, No. (%)				0.333				0.367
T1A	43 (12.18)	28 (14.36)	15 (9.49)		6 (5.00)	5 (8.47)	1 (1.64)	
T1B	203 (57.51)	114 (58.46)	89 (56.33)		46 (38.33)	23 (38.98)	23 (37.70)	
T1C	71 (20.11)	34 (17.44)	37 (23.42)		46 (38.33)	21 (35.59)	25 (40.98)	
T2A	36 (10.20)	19 (9.74)	17 (10.76)		22 (18.33)	10 (16.95)	12 (19.67)	
Pleural invasion, No. (%)				0.028				0.322
no	307 (86.97)	177 (90.77)	130 (82.28)		75 (62.50)	40 (67.80)	35 (57.38)	
yes	46 (13.03)	18 (9.23)	28 (17.72)		45 (37.50)	19 (32.20)	26 (42.62)	
Vascular invasion, No. (%)				0.343				0.013
no	338 (95.75)	189 (96.92)	149 (94.30)		106 (88.33)	57 (96.61)	49 (80.33)	
yes	15 (4.25)	6 (3.08)	9 (5.70)		14 (11.67)	2 (3.39)	12 (19.67)	
Spiculation, No. (%)				0.237				0.727
no	136 (38.53)	81 (41.54)	55 (34.81)		54 (45.00)	28 (47.46)	26 (42.62)	
yes	217 (61.47)	114 (58.46)	103 (65.19)		66 (55.00)	31 (52.54)	35 (57.38)	
Lobulation, No. (%)				0.297				0.747
no	98 (27.76)	59 (30.26)	39 (24.68)		80 (66.67)	38 (64.41)	42 (68.85)	
yes	255 (72.24)	136 (69.74)	119 (75.32)		40 (33.33)	21 (35.59)	19 (31.15)	
Vascular convergence, No. (%)				<0.001				0.532
no	88 (24.93)	70 (35.90)	18 (11.39)		77 (64.17)	40 (67.80)	37 (60.66)	
yes	265 (75.07)	125 (64.10)	140 (88.61)		43 (35.83)	19 (32.20)	24 (39.34)	
Pleural traction, No. (%)				0.685				0.740
no	193 (54.67)	109 (55.90)	84 (53.16)		46 (38.33)	24 (40.68)	22 (36.07)	

(Continued)

TABLE 1 Continued

	Train-total (n = 353)	STAS negative (n = 195)	STAS positive (n = 158)	*P* value	Validation -total (n = 120)	STAS negative (n = 59)	STAS positive (n = 61)	*P* value
yes	160 (45.33)	86 (44.10)	74 (46.84)		74 (61.67)	35 (59.32)	39 (63.93)	
Pleural indentation, No. (%)				0.86				0.826
no	253 (71.67)	141 (72.31)	112 (70.89)		49 (40.83)	23 (38.98)	26 (42.62)	
yes	100 (28.33)	54 (27.69)	46 (29.11)		71 (59.17)	36 (61.02)	35 (57.38)	
Air bronchogram, No. (%)				0.142				0.881
no	271 (76.77)	156 (80.00)	115 (72.78)		98 (81.67)	49 (83.05)	49 (80.33)	
yes	82 (23.23)	39 (20.00)	43 (27.22)		22 (18.33)	10 (16.95)	12 (19.67)	
Vacuole, No. (%)				0.841				0.708
no	231 (65.44)	129 (66.15)	102 (64.56)		97 (80.83)	49 (83.05)	48 (78.69)	
yes	122 (34.56)	66 (33.85)	56 (35.44)		23 (19.17)	10 (16.95)	13 (21.31)	
Tumor location, No. (%)				0.128				0.756
LUL	94 (26.63)	53 (27.18)	41 (25.95)		36 (30.00)	15 (25.42)	21 (34.43)	
LLL	78 (22.10)	35 (17.95)	43 (27.22)		22 (18.33)	12 (20.34)	10 (16.39)	
RUL	103 (29.18)	64 (32.82)	39 (24.68)		26 (21.67)	15 (25.42)	11 (18.03)	
RML	19 (5.38)	8 (4.10)	11 (6.96)		4 (3.33)	2 (3.39)	2 (3.28)	
RLL	59 (16.71)	35 (17.95)	24 (15.19)		32 (26.67)	15 (25.42)	17 (27.87)	
Operative mode, No. (%)				<0.001				0.900
wedge resection	50 (14.16)	40 (20.51)	10 (6.33)		17 (14.17)	9 (15.25)	8 (13.11)	
sublobar resection	47 (13.31)	29 (14.87)	18 (11.39)		7 (5.83)	3 (5.08)	4 (6.56)	
lobectomy	256 (72.52)	126 (64.62)	130 (82.28)		96 (80.00)	47 (79.66)	49 (80.33)	

SD, Standard deviation; STAS, Spread through air space; WBC, white blood cell; RLL, Lower Lobe; FEV1, forced expiratory volume in 1 second; DLCO, diffusing capacity of the carbon monoxide; WBC, white blood cell; NLR, neutrophil to lymphocyte ratio; MLR, monocyte to lymphocyte ratio; PLR, platelet to lymphocyte ratio; CA125, carbohydrate antigen 125; AFP, alpha-Fetoprotein; CEA, carcinoembryonic antigen; NSE, Neuron-Specific Enolase; cyfra21-1, cytokeratin 19 Fragment 21-1; proGRP, pro-Gastrin-Releasing Peptide.

its robustness and practical value in clinical settings. To clarify threshold selection and clinical use, we focused on 0.20 – 0.80 as a prespecified, clinically plausible range based on the accepted trade-off between missing STAS and unnecessary escalation. Within this range—particularly around 0.30–0.50—XGBoost offers the most consistent net benefit, supporting model-guided escalation when the consequence of missing STAS is considered high, whereas higher thresholds (≥0.60) help limit overtreatment.

Taken together, while AdaBoost delivered the highest AUC, however, its calibration performance was markedly unsatisfactory, as indicated by the Hosmer–Lemeshow test ($P = 1.33\times10^{-15}$) and the corresponding calibration curves (see Supplementary Figure 1), which revealed a considerable discrepancy between predicted and observed outcomes. In contrast, XGBoost provided a more balanced performance: it maintained strong discrimination while demonstrating good calibration, with predicted risks well aligned with actual probabilities (see Figure 3). From a clinical perspective, accurate probability estimates are crucial for risk stratification and surgical decision-making, where over- or underestimation of risk may lead to inappropriate treatment choices. Therefore, despite its slightly lower AUC, XGBoost was chosen as the preferred model.

Model explanation

To enhance clinical interpretability, we used SHAP to quantify both the direction and magnitude of each feature's contribution to the final XGBoost model's predictions. Figure 6A summarizes global importance as the mean absolute SHAP value for each feature, with CEA contributing the most overall, followed by vascular convergence, proGRP, age, AFP, smoking history, and CTR. In the beeswarm plot (Figure 6B), features are ordered by mean absolute SHAP value (global importance). Horizontal position reflects the SHAP value for each case (positive values increase the predicted probability of STAS; negative values decrease it), while color encodes the feature value (yellow = higher value, purple = lower value). Consistent with the biological rationale, higher CEA values (yellow points) cluster toward positive SHAP values, indicating that elevated CEA is associated with an increased predicted risk of STAS. A similar positive directionality is observed for vascular convergence and CTR (higher values tend to push predictions toward higher STAS probability). proGRP and AFP exhibit modest positive contributions at higher values, and smoking history (ever vs.

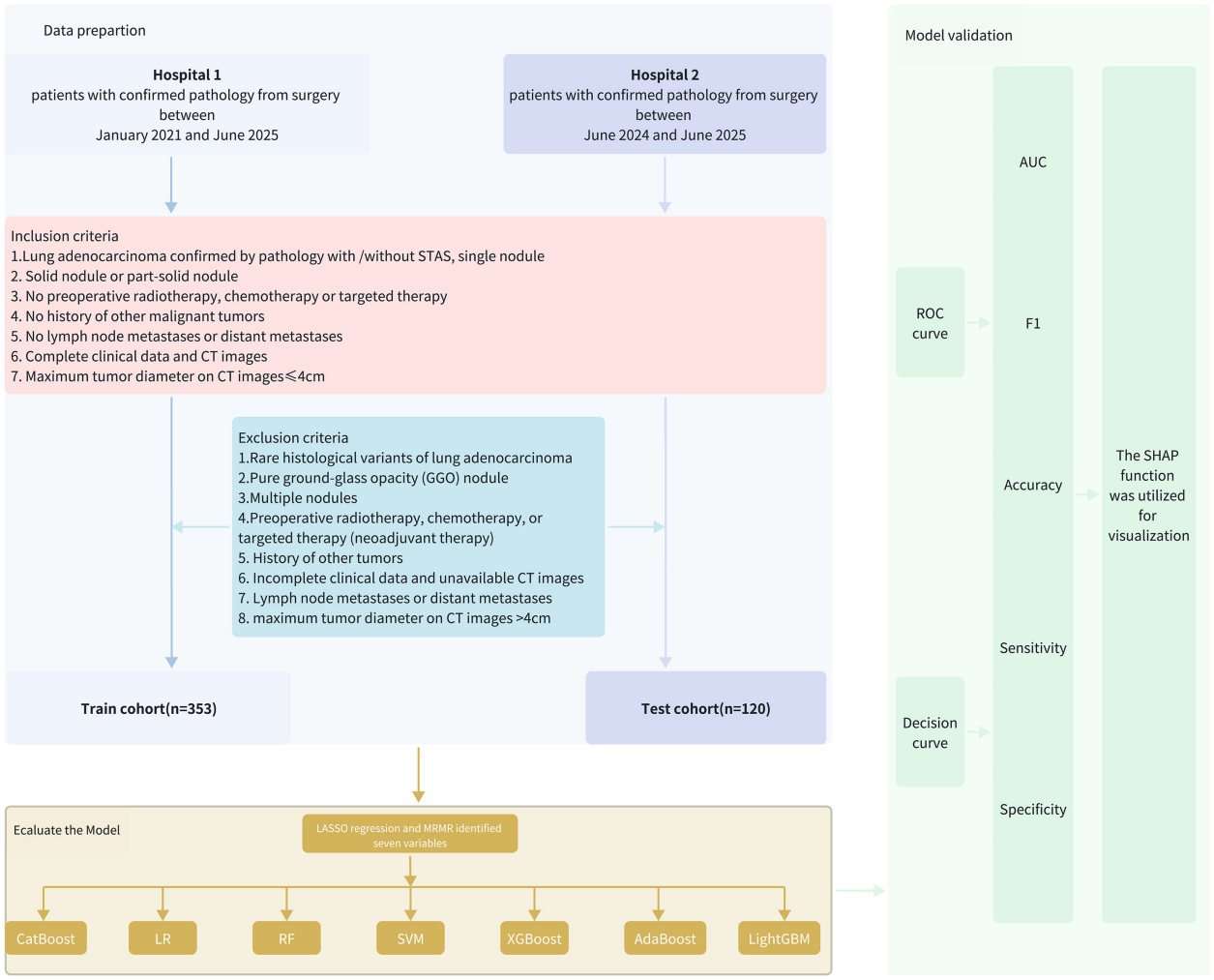

FIGURE 1
The flow diagram of the study.

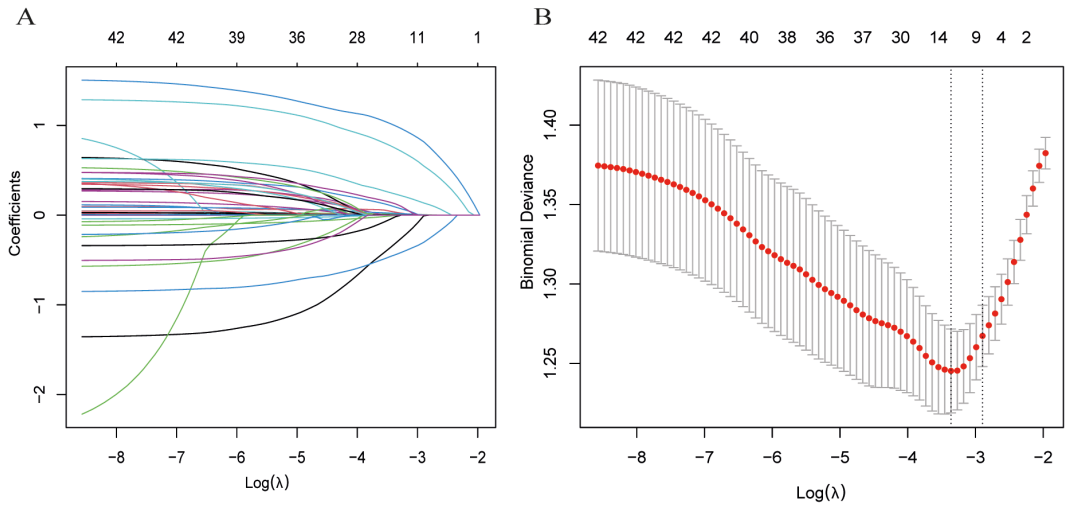

FIGURE 2
Process of selecting variables through LASSO regression. **(A)** LASSO coefficient profiles across a sequence of log(λ) values. The optimal penalty parameter λ was determined via ten-fold cross-validation. **(B)** Validation of the optimal λ, with dotted vertical lines indicating the chosen value. Seven variables with nonzero coefficients were selected by λ. min.

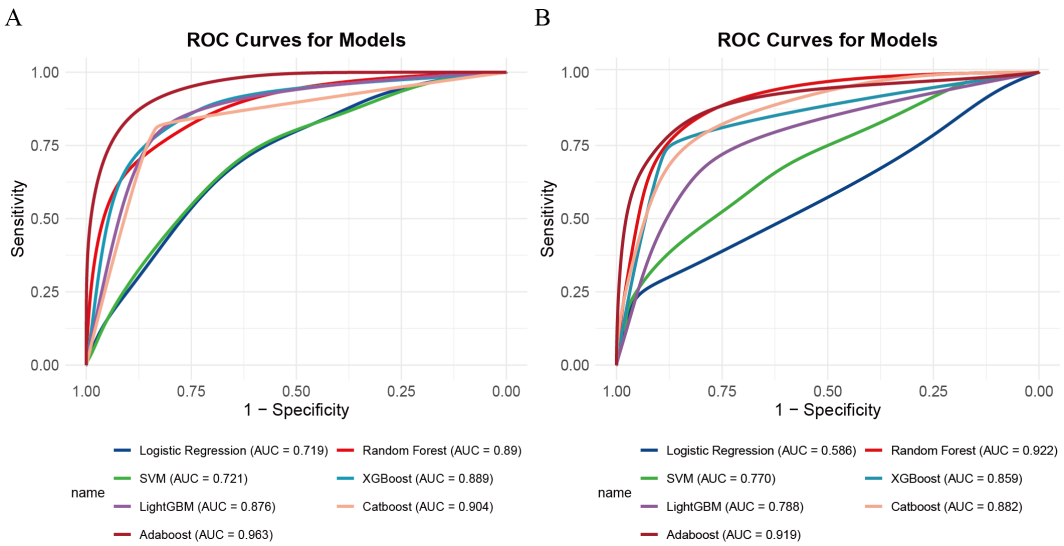

FIGURE 3
ROC curves for seven machine learning models. **(A)** Training set, **(B)** validation set. ROC, receiver operating characteristic.

never) shifts predictions toward higher risk. Age shows a smaller but directionally consistent effect.

To facilitate clinical utility, a final prediction nomogram was constructed using the seven predictive variables (CEA, vascular convergence, proGRP, age, AFP, smoking history, CTR) and implemented in a web-based application for clinical use. The web application is accessible online at: https://cuncun.shinyapps.io/DynNomapp/.

Discussion

This study presents a retrospective analysis of clinical features associated with STAS in stage I lung adenocarcinoma, utilizing seven risk factors (CEA, vascular convergence, proGRP, age, AFP, smoking history, and CTR) to develop predictive models. We compared seven machine learning models, with AdaBoost demonstrating the best diagnostic performance. However, XGBoost exhibited superior discriminatory power and predictive accuracy, We employed the SHAP function to visualize model interpretability, thereby enhancing the model's transparency.

Our study has several advantages (1). Multi-center design: This is a multi-center retrospective study leveraging preoperative clinical indicators to predict STAS in stage I lung adenocarcinoma, aiming to develop a predictive model for assessing STAS risk. (2) Clinical variables: Unlike radiomics-based studies, the clinical variables (27, 28), used in this model are easily accessible, significantly improving the model's generalizability. Additionally, we developed a web-

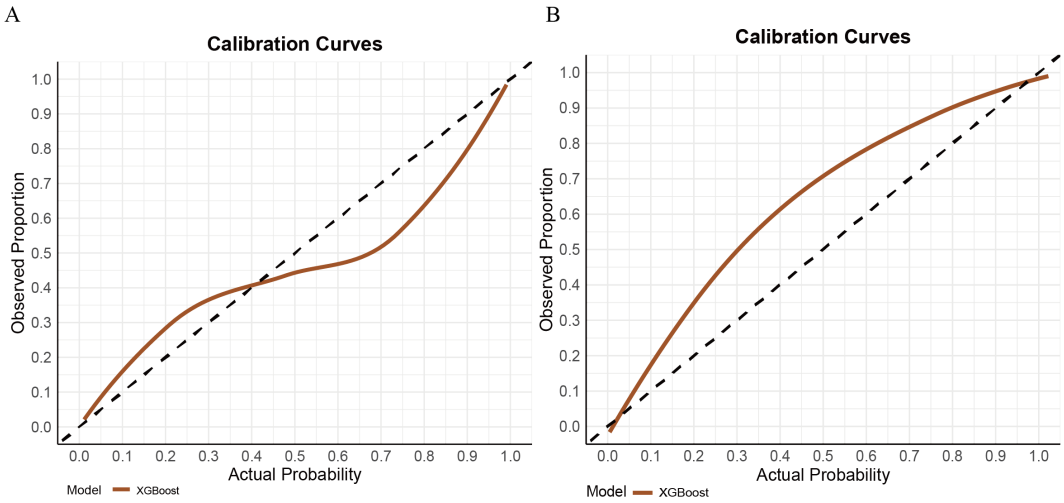

FIGURE 4
Calibration curves for XGBoost model in the train set **(A)** and validation **(B)** set.

TABLE 2 Performance of seven models.

Model	Group	AUC (95%CI)	Accuracy	Specificity	Sensitivity	NPV	PPV	F1
logistic regression	train	0.719 (0.667 - 0.772)	0.677	0.703	0.646	0.710	0.638	0.642
	validation	0.586 (0.483 - 0.689)	0.567	0.814	0.328	0.541	0.645	0.435
random forest	train	0.890 (0.857 - 0.923)	0.793	0.821	0.759	0.805	0.774	0.767
	validation	0.922 (0.886-0.97)	0.867	0.881	0.852	0.865	0.881	0.867
SVM	train	0.721 (0.667 - 0.774)	0.669	0.672	0.665	0.726	0.621	0.642
	validation	0.770 (0.686 - 0.853)	0.633	0.898	0.377	0.582	0.793	0.511
XGBoost	train	0.889 (0.852- 0.926)	0.807	0.805	0.81	0.840	0.771	0.79
	validation	0.859 (0.789 - 0.928)	0.808	0.881	0.738	0.765	0.865	0.796
AdaBoost	train	0.963 (0.947 - 0.980)	0.898	0.908	0.886	0.907	0.886	0.886
	validation	0.919 (0.866 - 0.972)	0.817	0.915	0.721	0.761	0.898	0.8
LightGBM	train	0.876 (0.838 - 0.913)	0.822	0.831	0.81	0.853	0.795	0.803
	validation	0.788 (0.705 - 0.872)	0.725	0.814	0.639	0.685	0.78	0.703
CatBoost	train	0.904 (0.872 - 0.935)	0.819	0.836	0.797	0.836	0.797	0.797
	validation	0.882 (0.8226 - 0.942)	0.792	0.864	0.721	0.750	0.846	0.779

CI, Confidence interval; AUC, Area under the receiver operating characteristics curve; F1, F1 score; SVM, support vector machines; XGBoost, Extreme Gradient Boosting; AdaBoost, adaptive boosting; LightGBM, light gradient boosting machine; CatBoost, category boosting; NPV, negative predictive value; PPV, positive predictive value.

based nomogram, which facilitates practical use in clinical settings. (3) Superior interpretability: Compared to other ML models, our nomogram shows similar diagnostic performance but offers superior interpretability and operational ease.

STAS has emerged as a critical pathological feature linked to local recurrence and poor prognosis in lung cancer (10, 29, 30). It is a powerful independent predictor of recurrence and prognosis in stage I lung adenocarcinoma. STAS-positive patients exhibit significantly worse postoperative outcomes than STAS-negative patients in stage IA, with prognosis approaching that of stage IB patients (31, 32). STAS also predicts recurrence and prognosis in stage I lung squamous cell carcinoma, although this association does not extend to stages II–III (8). Radiologically, the presence of solid components correlates with higher STAS risk, with a threefold increase in STAS risk for every 1% increase in the solid component of the tumor (33).

Surgical outcomes are significantly influenced by STAS. In T1N0M0 lung adenocarcinoma, patients with STAS who undergo sublobar resection have higher recurrence rates and lung cancer–specific mortality than those treated with lobectomy (34).

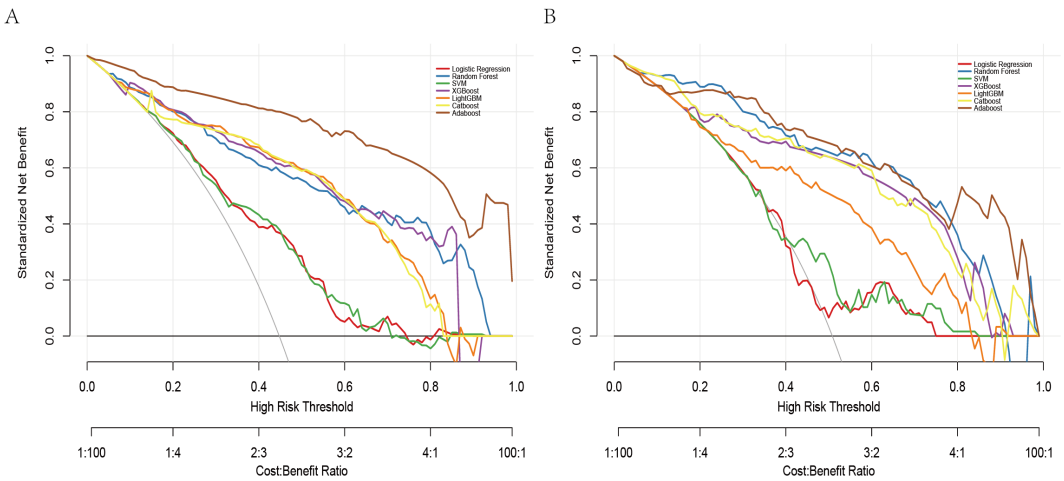

FIGURE 5
Decision curves for seven machine learning models: **(A)** training set, **(B)** validation set.

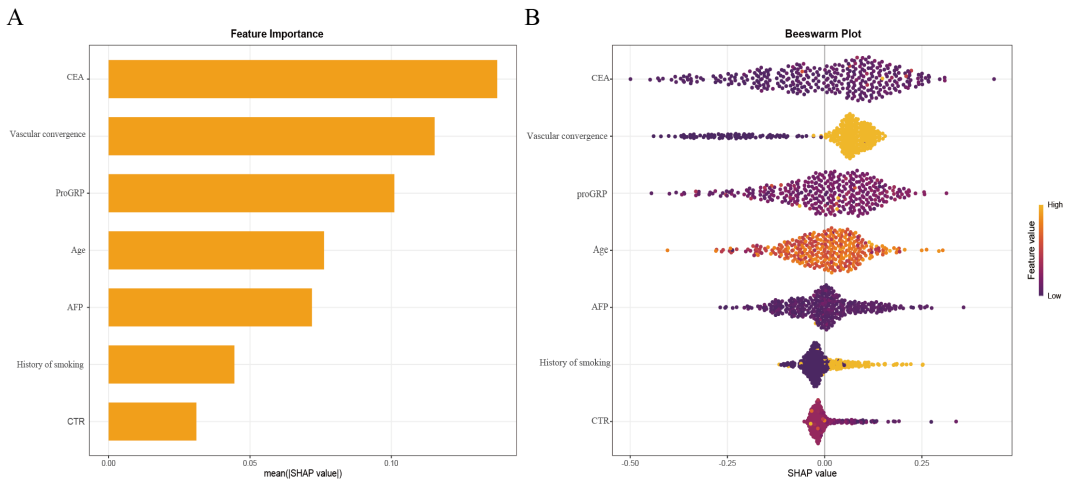

FIGURE 6
XGBoost model interpretability using SHAP.SHAP summary bar plot ranking feature importance **(A)**. SHAP summary dot plot illustrating feature contributions **(B)**. Each dot represents an individual patient's SHAP value for a given feature, with color intensity indicating the feature's actual value. Higher SHAP values reflect an increased probability of STAS positivity.

Moreover, among sublobar approaches, wedge resection is associated with inferior OS and RFS compared with segmentectomy or lobectomy (11). Therefore, accurate preoperative identification of STAS-positive patients is crucial for surgical decision-making—particularly when considering sublobar resection—as failure to recognize STAS preoperatively may lead to undertreatment and a substantially increased risk of recurrence. Patients at high predicted risk may be considered for lobectomy to mitigate recurrence risk.

Our study identified several predictive factors, including baseline characteristics, CT imaging features and tumor markers, Smoking, the most significant risk factor for lung cancer, has been linked to increased STAS risk, particularly in older patients (10, 32). However, the exact relationship between smoking, age, and STAS positivity remains unclear and warrants further investigation.

Tumor markers, including CEA, proGRP, AFP reflect tumor biology and systemic disease burden (35). CEA, a glycoprotein associated with cell adhesion, is typically absent in healthy adult blood (36). Our feature interpretability analysis revealed CEA as the most significant factor in the XGBoost model, which is consistent with findings from other studies (37, 38). High CEA expression can promote epithelial-mesenchymal transition (EMT) by modulating various signaling molecules within the EMT pathway. During EMT, tumor cells lose epithelial adhesion markers and gain mesenchymal markers, enhancing motility and invasiveness, which in turn increases STAS likelihood (31, 39).

ProGRP plays a vital role in diagnosing and subtyping lung cancer (40), particularly in small cell lung cancer(SCLC). It has been widely applied as a biomarker for SCLC diagnosis, monitoring, and evaluation of treatment response (41, 42), and is also considered an effective marker for diagnosing lung neuroendocrine neoplasms (43). Recent studies further suggest that ProGRP, when combined with artificial intelligence approaches, can accurately predict lung cancer risk (44). Nevertheless, the role of ProGRP in lung adenocarcinoma remains insufficiently understood, and further research is warranted to clarify its potential diagnostic and prognostic value.

AFP is a glycoprotein originally identified as the first oncoprotein and is now widely used as a biomarker in hepatocellular carcinoma screening (45, 46), Elevated serum AFP levels have also been reported in some patients with primary lung cancer (47, 48), and extremely high concentrations are a distinguishing feature of hepatoid adenocarcinoma of the lung (49), However, the intrinsic relationship between AFP and lung adenocarcinoma remains poorly understood and warrants further investigation.

Vascular convergence has been identified as a strong indicator of STAS, appearing frequently in STAS-positive patients (50, 51).

The aggressiveness of lung cancer is also linked to the proportion of solid tumor components observed on CT, a higher solid component indicates a more significant, CTR have a positive correlation with STAS (32, 52) and as the most accurate CT characteristic for forecasting STAS in lung adenocarcinomas measuring ≤2 cm (53). Our research shows that an increase in solid components is an independent predictor of STAS, significantly heightening the risk, consistent with previous studies.

In this study, we utilized clinical baseline characteristics, imaging characteristics and tumor markers to develop various machine learning models to preoperatively predict the presence of STAS preoperatively. the XGBoost model, which effectively manages high-dimensional data and complex interactions, showed superior performance, with the predicted values aligning closely with actual results. The SHAP algorithm was used to enhance model interpretability, making the results more accessible to clinicians.

Previous studies based on CT radiomics models have faced challenges due to the low incidence of STAS and the typically single-center design of such studies, limiting their generalizability. Multi-center studies have also struggled with robustness (28, 54).

Additionally, radiomics models often suffer from a lack of interpretability, creating a "black box" effect that reduces clinical confidence. In contrast, the clinical variables in our model are derived from preoperative data and CT images, which are easily accessible. The use of the SHAP algorithm further enhances interpretability, and the model can be accessed through a web-based platform, improving clinical applicability.

Nonetheless, this study has several limitations. First, the retrospective design introduces potential selection bias, highlighting the need for prospective validation. Second, although external validation was performed, it was derived from a single-center cohort, which limits the generalizability of the findings. Third, the relatively small sample size raises concerns about potential overfitting of the model, and the lack of long-term follow-up further restricts the strength of the conclusions. In addition, pure GGO nodules and patients with multiple nodules were excluded; future investigations should develop strategies to better evaluate STAS in these subgroups. Overall, larger multi-center prospective studies with extended follow-up are required to confirm and extend our findings. However, the web-based tool developed in this study has not yet undergone prospective, multi-center validation or formal clinical impact assessment, and thus its clinical applicability remains preliminary. At this stage, it should be regarded as a research prototype rather than a tool to guide individual patient care.

The predictive models based on XGBoost regression demonstrated significant preoperative predictive accuracy for STAS in stage I LUAD solid and part-solid nodules. The application of SHAP analysis augmented the model's interpretability by establishing associations between predictions and relevant clinical variables, thereby enhancing its clinical applicability. This interpretable model offers a promising tool for personalized preoperative surgical planning and tailored postoperative management.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The Ethics Committee of Northern Jiangsu People's Hospital reviewed and approved this retrospective study involving human participants. A waiver of informed consent was granted in accordance with national regulations and institutional policies, as the research design utilized anonymized archival data without new interventions.

Author contributions

Q-LR: Software, Writing – review & editing, Writing – original draft, Investigation, Resources, Formal analysis, Methodology, Conceptualization, Visualization, Validation. LL: Writing – review & editing, Investigation, Writing – original draft, Data curation, Validation, Visualization, Conceptualization, Methodology. KC: Data curation, Software, Conceptualization, Writing – review & editing, Resources, Visualization. X-RX: Methodology, Software, Conceptualization, Visualization, Validation, Writing – review & editing. H-JW: Writing – review & editing, Investigation, Methodology, Formal analysis. JW: Writing – review & editing, Conceptualization, Methodology. J-ZY: Writing – review & editing, Conceptualization, Validation, Visualization, Methodology. J-XH: Methodology, Writing – review & editing, Investigation, Formal analysis. X-LW: Validation, Project administration, Funding acquisition, Writing – review & editing, Formal analysis, Conceptualization, Methodology, Supervision. Y-SS: Formal analysis, Resources, Project administration, Investigation, Data curation, Methodology, Conceptualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. This study was supported by the Jiangsu Provincial Health Commission Research Project on Elderly Health (LKZ2022019), the Suqian Sci&Tech Program (Grant No. Z2023119), the Yangzhou Social Development and Clinical Frontier Technology Project (YZ2023084 and YZ2021078),and the Yangzhou Innovation Capability Building Design Plan Project (YZ2022168).

Acknowledgments

We extend our gratitude to all authors for their valuable contributions to this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. Generative AI was used solely for language polishing and grammar refinement during manuscript preparation. No AI tools were used for data analysis, interpretation, or generation of scientific content. The authors take full responsibility for the integrity and accuracy of the manuscript.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2025.1682633/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Calibration curves of seven machine learning models in the training set **(A)** and validation set **(B)**. Closer alignment of the curves with the diagonal line indicates better agreement between predicted and observed probabilities.

References

1. Han B, Zheng R, Zeng H, Wang S, Sun K, Chen R, et al. Cancer incidence and mortality in China, 2022. *J Natl Cancer Center*. (2024) 4:47–53. doi: 10.1016/j.jncc.2024.01.006

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J Clin*. (2021) 71:209–49. doi: 10.3322/caac.21660

3. Travis WD, Brambilla E, Riely GJ. New pathologic classification of lung cancer: relevance for clinical practice and clinical trials. *J Clin Oncol*. (2013) 31:992–1001. doi: 10.1200/JCO.2012.46.9270

4. Altorki N, Wang X, Damman B, Mentlick J, Landreneau R, Wigle D, et al. Lobectomy, segmentectomy, or wedge resection for peripheral clinical T1an0 non-small cell lung cancer: A *post hoc* analysis of calgb 140503 (Alliance). *J Thorac Cardiovasc Surg*. (2024) 167:338–47.e1. doi: 10.1016/j.jtcvs.2023.07.008

5. Warth A, Beasley MB, Mino-Kenudson M. Breaking new ground: the evolving concept of spread through air spaces (Stas). *J Thorac Oncol*. (2017) 12:176–8. doi: 10.1016/j.jtho.2016.10.020

6. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 world health organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol*. (2015) 10:1243–60. doi: 10.1097/JTO.0000000000000630

7. Ding Y, Chen Y, Wen H, Li J, Chen J, Xu M, et al. Pretreatment prediction of tumour spread through air spaces in clinical stage I non-small-cell lung cancer. *Eur J Cardio-thoracic Surg*. (2022) 62:ezac248. doi: 10.1093/ejcts/ezac248

8. Yanagawa N, Shiono S, Endo M, Ogata S-Y. Tumor Spread through Air Spaces Is a Useful Predictor of Recurrence and Prognosis in Stage I Lung Squamous Cell Carcinoma, but Not in Stage Ii and Iii. *Lung Cancer (amsterdam Netherlands)*. (2018) 120:14–21. doi: 10.1016/j.lungcan.2018.03.018

9. Kadota K, Nitadori J-I, Sima CS, Ujiie H, Rizk NP, Jones DR, et al. Tumor Spread through Air Spaces Is an Important Pattern of Invasion and Impacts the Frequency and Location of Recurrences after Limited Resection for Small Stage I Lung Adenocarcinomas. *J Thorac Oncol*. (2015) 10:806–14. doi: 10.1097/JTO.0000000000000486

10. Khalil HA, Shi W, Mazzola E, Lee DN, Norton-Hughes E, Dolan D, et al. Analysis of recurrence in lung adenocarcinoma with spread through air spaces. *J Thorac Cardiovasc Surg*. (2023) 166:1317–28.e4. doi: 10.1016/j.jtcvs.2023.01.030

11. Ikeda T, Kadota K, Go T, Misaki N, Haba R, Yokomise H. Segmentectomy provides comparable outcomes to lobectomy for stage ia non-small cell lung cancer with spread through air spaces. *Semin Thorac Cardiovasc Surg*. (2023) 35:156–63. doi: 10.1053/j.semtcvs.2022.02.001

12. Eguchi T, Kameda K, Lu S, Bott MJ, Tan KS, Montecalvo J, et al. Lobectomy Is Associated with Better Outcomes Than Sublobar Resection in Spread through Air Spaces (Stas)-Positive T1 Lung Adenocarcinoma: A propensity Score-Matched Analysis. *J Thorac Oncol*. (2019) 14:87–98. doi: 10.1016/j.jtho.2018.09.005

13. Dai Z-Y, Shen C, Wang X, Wang F-Q, Wang Y. Could less be enough: sublobar resection vs lobectomy for clinical stage ia non-small cell lung cancer patients with visceral pleural invasion or spread through air spaces. *Int J Surg (london England)*. (2025) 111:2675–85. doi: 10.1097/JS9.0000000000002249

14. Deo RC. Machine learning in medicine. *Circulation*. (2015) 132:1920–30. doi: 10.1161/CIRCULATIONAHA.115.001593

15. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol*. (2022) 23:40–55. doi: 10.1038/s41580-021-00407-0

16. Hasegawa M, Sakai F, Ishikawa R, Kimura F, Ishida H, Kobayashi K. Ct features of epidermal growth factor receptor-mutated adenocarcinoma of the lung: comparison with nonmutated adenocarcinoma. *J Thorac Oncol*. (2016) 11:819–26. doi: 10.1016/j.jtho.2016.02.010

17. Toyokawa G, Takada K, Okamoto T, Shimokawa M, Kozuma Y, Matsubara T, et al. Computed tomography features of lung adenocarcinomas with programmed death ligand 1 expression. *Clin Lung Cancer*. (2017) 18:e375–e83. doi: 10.1016/j.cllc.2017.03.008

18. Zhou JY, Zheng J, Yu ZF, Xiao WB, Zhao J, Sun K, et al. Comparative analysis of clinicoradiologic characteristics of lung adenocarcinomas with alk rearrangements or egfr mutations. *Eur Radiol*. (2015) 25:1257–66. doi: 10.1007/s00330-014-3516-z

19. Ma X, He W, Chen C, Tan F, Chen J, Yang L, et al. A ct-based deep learning model for preoperative prediction of spread through air spaces in clinical stage I lung adenocarcinoma. *Front Oncol*. (2024) 14:1482965. doi: 10.3389/fonc.2024.1482965

20. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The eighth edition ajcc cancer staging manual: continuing to build a bridge from a population-based to a more "Personalized" Approach to cancer staging. *CA: A Cancer J Clin*. (2017) 67:93–9. doi: 10.3322/caac.21388

21. Zhang Y, Ding C, Li T. Gene selection algorithm by combining relieff and mrmr. *BMC Genomics*. (2008) 9 Suppl 2:S27. doi: 10.1186/1471-2164-9-S2-S27

22. Li S, Liu S, Sun X, Hao L, Gao Q. Identification of endocrine-disrupting chemicals targeting key dcm-associated genes via bioinformatics and machine learning. *Ecotoxicol Environ Saf*. (2024) 274:116168. doi: 10.1016/j.ecoenv.2024.116168

23. LaValley MP. Logistic regression. *Circulation*. (2008) 117:2395–9. doi: 10.1161/CIRCULATIONAHA.106.682658

24. Gohari MR, Doggett A, Patte KA, Ferro MA, Dubin JA, Hilario C, et al. Using random forest to identify correlates of depression symptoms among adolescents. *Soc Psychiatry Psychiatr Epidemiol*. (2024) 59:2063–71. doi: 10.1007/s00127-024-02695-1

25. Hao N, Sun P, Zhao W, Li X. Application of a developed triple-classification machine learning model for carcinogenic prediction of hazardous organic chemicals to the us, eu, and who based on chinese database. *Ecotoxicol Environ Saf*. (2023) 255:114806. doi: 10.1016/j.ecoenv.2023.114806

26. Wang J, Chen H, Wang H, Liu W, Peng D, Zhao Q, et al. A risk prediction model for physical restraints among older chinese adults in long-term care facilities: machine learning study. *J Med Internet Res*. (2023) 25:e43815. doi: 10.2196/43815

27. Lin M-W, Chen L-W, Yang S-M, Hsieh M-S, Ou D-X, Lee Y-H, et al. Ct-based deep-learning model for spread-through-air-spaces prediction in ground glass-predominant lung adenocarcinoma. *Ann Surg Oncol*. (2024) 31:1536–45. doi: 10.1245/s10434-023-14565-2

28. Liao G, Huang L, Wu S, Zhang P, Xie D, Yao L, et al. Preoperative ct-based peritumoral and tumoral radiomic features prediction for tumor spread through air spaces in clinical stage I lung adenocarcinoma. *Lung Cancer (amsterdam Netherlands)*. (2022) 163. doi: 10.1016/j.lungcan.2021.11.017

29. Chae M, Jeon JH, Chung J-H, Lee SY, Hwang WJ, Jung W, et al. Prognostic Significance of Tumor Spread through Air Spaces in Patients with Stage Ia Part-Solid Lung Adenocarcinoma after Sublobar Resection. *Lung Cancer (amsterdam Netherlands)*. (2021) 152:21–6. doi: 10.1016/j.lungcan.2020.12.001

30. Laville D, Désage A-L, Fournel P, Bayle-Bleuez S, Neifer C, Picot T, et al. Spread through Air Spaces in Stage I to Iii Resected Lung Adenocarcinomas: Should the Presence of Spread through Air Spaces Lead to an Upstaging? *Am J Surg Pathol*. (2024) 48:596–604. doi: 10.1097/PAS.0000000000002188

31. Luo Y, Ding W, Yang X, Bai H, Jiao F, Guo Y, et al. Construction and validation of a predictive model for meningoencephalitis in pediatric scrub typhus based on machine learning algorithms. *Emerging Microbes Infections*. (2025) 14:2469651. doi: 10.1080/22221751.2025.2469651

32. Toyokawa G, Yamada Y, Tagawa T, Kozuma Y, Matsubara T, Haratake N, et al. Significance of spread through air spaces in resected pathological stage I lung adenocarcinoma. *Ann Thorac Surg*. (2018) 105:1655–63. doi: 10.1016/j.athoracsur.2018.01.037

33. Gao Z, An P, Li R, Wu F, Sun Y, Wu J, et al. Development and validation of a clinic-radiological model to predict tumor spread through air spaces in stage I lung adenocarcinoma. *Cancer Imaging*. (2024) 24:25. doi: 10.1186/s40644-024-00668-w

34. David EA, Atay SM, McFadden PM, Kim AW. Sublobar or suboptimal: does tumor spread through air spaces signify the end of sublobar resections for T1n0 adenocarcinomas? *J Thorac Oncol*. (2019) 14:11–2. doi: 10.1016/j.jtho.2018.11.001

35. Varadhachary GR, Abbruzzese JL, Lenzi R. Diagnostic strategies for unknown primary cancer. *Cancer*. (2004) 100:1776–85. doi: 10.1002/cncr.20202

36. Konstantopoulos K, Thomas SN. Cancer cells in transit: the vascular interactions of tumor cells. *Annu Rev Biomed Eng*. (2009) 11:177–202. doi: 10.1146/annurev-bioeng-061008-124949

37. Wang J, Yao Y, Tang D, Gao W. An individualized nomogram for predicting and validating spread through air space (Stas) in surgically resected lung adenocarcinoma: A single center retrospective analysis. *J Cardiothor Surg*. (2023) 18:337. doi: 10.1186/s13019-023-02458-0

38. Wang Y, Lyu D, Zhang D, Hu L, Wu J, Tu W, et al. Nomogram based on clinical characteristics and radiological features for the preoperative prediction of spread through air spaces in patients with clinical stage ia non-small cell lung cancer: A multicenter study. *Diagn Interventional Radiol (ankara Turkey)*. (2023) 29:771–85. doi: 10.4274/dir.2023.232404

39. Yoshikawa M, Morine Y, Ikemoto T, Imura S, Higashijima J, Iwahashi S, et al. Elevated preoperative serum cea level is associated with poor prognosis in patients with hepatocellular carcinoma through the epithelial-mesenchymal transition. *Anticancer Res*. (2017) 37:1169–75. doi: 10.21873/anticanres.11430

40. Korkmaz ET, Koksal D, Aksu F, Dikmen ZG, Icen D, Maden E, et al. Triple test with tumor markers cyfra 21.1, he4, and progrp might contribute to diagnosis and subtyping of lung cancer. *Clin Biochem*. (2018) 58:15–9. doi: 10.1016/j.clinbiochem.2018.05.001

41. Zeng Q, Liu M, Zhou N, Liu L, Song X. Serum human epididymis protein 4 (He4) may be a better tumor marker in early lung cancer. *Clin Chim Acta; Int J Clin Chem*. (2016) 455:102–6. doi: 10.1016/j.cca.2016.02.002

42. Wojcik E, Kulpa JK. Pro-gastrin-releasing peptide (Progrp) as a biomarker in small-cell lung cancer diagnosis, monitoring and evaluation of treatment response. *Lung Cancer (Auckl)*. (2017) 8:231–40. doi: 10.2147/LCTT.S149516

43. Rosiek V, Kogut A, Kos-Kudła B. Pro-gastrin-releasing peptide as a biomarker in lung neuroendocrine neoplasm. *Cancers (Basel)*. (2023) 15. doi: 10.3390/cancers15133282

44. Hu W, Zhang X, Saber A, Cai Q, Wei M, Wang M, et al. Development and validation of a nomogram model for lung cancer based on radiomics artificial intelligence score and clinical blood test data. *Front In Oncol*. (2023) 13:1132514. doi: 10.3389/fonc.2023.1132514

45. Yeo YH, Lee Y-T, Tseng H-R, Zhu Y, You S, Agopian VG, et al. Alpha-fetoprotein: past, present, and future. *Hepatol Commun*. (2024) 8. doi: 10.1097/HC9.0000000000000422

46. Force M, Park G, Chalikonda D, Roth C, Cohen M, Halegoua-DeMarzio D, et al. Alpha-fetoprotein (Afp) and afp-L3 is most useful in detection of recurrence of hepatocellular carcinoma in patients after tumor ablation and with low afp level. *Viruses*. (2022) 14. doi: 10.3390/v14040775

47. Gavrancic T, Park Y-HA. A novel approach using sorafenib in alpha fetoprotein-producing hepatoid adenocarcinoma of the lung. *J Natl Compr Canc Netw*. (2015) 13. doi: 10.6004/jnccn.2015.0054

48. Yamagata T, Yamagata Y, Nakanishi M, Matsunaga K, Minakata Y, Ichinose M. A case of primary lung cancer producing alpha-fetoprotein. *Can Respir J*. (2004) 11:504–6. doi: 10.1155/2004/510350

49. Grossman K, Beasley MB, Braman SS. Hepatoid adenocarcinoma of the lung: review of a rare form of lung cancer. *Respir Med*. (2016) 119:175–9. doi: 10.1016/j.rmed.2016.09.003

50. Yang Y, Li L, Hu H, Zhou C, Huang Q, Zhao J, et al. A nomogram integrating the clinical and ct imaging characteristics for assessing spread through air spaces in clinical stage ia lung adenocarcinoma. *Front Immunol*. (2025) 16:1519766. doi: 10.3389/fimmu.2025.1519766

51. Gu Y, Zheng B, Zhao T, Fan Y. Computed tomography features and tumor spread through air spaces in lung adenocarcinoma: A meta-analysis. *J Thorac Imaging*. (2023) 38:W19–29. doi: 10.1097/RTI.0000000000000693

52. Qin L, Sun Y, Zhu R, Hu B, Wu J. Clinicopathological and ct features of tumor spread through air space in invasive lung adenocarcinoma. *Front Oncol*. (2022) 12:959113. doi: 10.3389/fonc.2022.959113

53. Qi L, Xue K, Cai Y, Lu J, Li X, Li M. Predictors of ct morphologic features to identify spread through air spaces preoperatively in small-sized lung adenocarcinoma. *Front Oncol*. (2020) 10:548430. doi: 10.3389/fonc.2020.548430

54. Yu W, Yun D, Xin L, Xin L, Xiap J, Jiu L, et al. Preoperative ct-based radiomics combined with tumour spread through air spaces can accurately predict early recurrence of stage I lung adenocarcinoma: A multicentre retrospective cohort study. *Cancer Imaging*. (2023) 23. doi: 10.1186/s40644-023-00605-3