



OPEN Comparative study on predicting postoperative distant metastasis of lung cancer based on machine learning models

Xi Guo^{1,10}, Tingting Xu^{2,10}, Yu Luo^{1,10}, Siyuan Liu³, Xiaoqing Tang¹, Rui Han⁴, Shiqin Li⁵, Yujie Du⁶, Jin Chen¹, Feng Jun⁷✉ & Hongjiang Pu^{8,9}✉

Lung cancer remains the leading cause of cancer-related incidence and mortality worldwide. Its tendency for postoperative distant metastasis significantly compromises long-term prognosis and survival. Accurately predicting the metastatic potential in a timely manner is crucial for formulating optimal treatment strategies. This study aimed to comprehensively compare the predictive performance of nine machine learning (ML) models and to enhance interpretability through SHAP (Shapley Additive Explanations), with the goal of developing a practical and transparent risk stratification tool for postoperative lung cancer management. Clinical data from 3,120 patients with stage I–III lung cancer who underwent radical surgery were retrospectively collected and randomly divided into training and testing cohorts. A total of 52 clinical, pathological, imaging, and laboratory variables were analyzed. Nine ML models—including eXtreme Gradient Boosting (XGBoost), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Adaptive Boosting (AdaBoost), Decision Tree (DT), Gradient Boosting Decision Tree (GBDT), Gaussian Naive Bayes (GNB), Complement Naive Bayes (CNB), and Multilayer Perceptron classifier (MLP)—were developed and evaluated. Model performance was assessed using accuracy, precision, recall, F1 score, ROC–AUC, PR–AUC, calibration, and decision curve analysis (DCA). All models were evaluated using nested cross-validation (outer stratified 70/30 splits repeated 10 times; inner fivefold tuning), with decision thresholds prespecified in the inner loop and applied unchanged to held-out tests. Given the approximately 4:1 class imbalance, cost-sensitive learning was primarily adopted, and PR–AUC was reported in addition to ROC–AUC. Among the nine models, GBDT demonstrated the highest predictive performance, achieving an AUC of 0.810 (95% CI: 0.748–0.872), accuracy of 0.766, sensitivity of 0.698, and specificity of 0.786 in the test set. SHAP analysis revealed that adjuvant chemotherapy, adjuvant radiotherapy, pathological N stage, age, body mass index (BMI), and preoperative neutrophil count (Pre-ANC) were the most influential predictors of distant metastasis. The combination of model performance and interpretability supported the model’s potential for integration into clinical workflows to assist in real-time decision-making. In this work, we carried out a systematic comparison of nine machine learning algorithms in a large postoperative cohort under a coherent and interpretable framework. By jointly considering discrimination, calibration, clinical benefit (via decision curve analysis), and SHAP-based explanations, we constructed a practical prognostic tool to guide personalized treatment strategies and follow-up care. This methodology offers a data-driven basis for precision management. Ultimately, our findings

provide an internally validated reference framework that warrants external and multicenter validation prior to clinical deployment.

Keywords Lung cancer, Distant metastasis, Machine learning, GBDT, SHAP, Risk prediction, Clinical decision support

Abbreviations

ML	Machine learning
SHAP	Shapley additive explanations
XGBoost	EXtreme gradient boosting
LightGBM	Light gradient boosting machine
AdaBoost	Adaptive boosting
DT	Decision tree
GBDT	Gradient boosting decision tree
GNB	Gaussian naive bayes
CNB	Complement naive bayes
MLP	Multilayer perceptron
NSCLC	Non-small cell lung cancer
PD-L1	Programmed death-ligand 1
TNM	Tumor, node, metastasis
STROBE	Strengthening the reporting of observational studies in epidemiology
PET-CT	Positron emission tomography–computed tomography
MRI	Magnetic resonance imaging
CEA	Carcinoembryonic antigen
DCA	Decision curve analysis
ROC	Receiver operating characteristic
AUC	Area under the curve
PR	Precision–recall
AP	Average precision
Pre-ANC	Preoperative neutrophil count
RECIST	Response evaluation criteria in solid tumors
AJCC	American joint committee on cancer

¹Department of Oncology, The Third People's Hospital of Kunming, Kunming 650041, China. ²Department of Oncology, The People's Hospital of Chu Xiong Yi Autonomous Prefecture, Chu Xiong 675000, China. ³Department of Surgery, The Third People's Hospital of Kunming, Kunming 650041, China. ⁴Department of Ophthalmology, The First People's Hospital of Kunming, Kunming 650034, China. ⁵Department of Intensive Care Unit, The People's Hospital of Chu Xiong Yi Autonomous Prefecture, Chu Xiong 675000, China. ⁶Jiangxi Vocational College of Industry & Engineering, Pingxiang 337000, China. ⁷Department of Radiology, Yunnan Cancer Hospital, The Third Affiliated of Kunming Medical University, Xishan District, Number 519 Kunzhou Road, Kunming 650118, China. ⁸The Third Affiliated Hospital of Kunming Medical University, Peking University Cancer Hospital Yunnan Hospital, Yunnan Cancer Hospital, Yunnan Cancer Centre, No. 519 Kunzhou Road, Xishan District, Kunming 650118, China. ⁹Department of Oncology, Dazhou Central Hospital, Dazhou 635000, Sichuan, China. ¹⁰Xi Guo, Tingting Xu and Yu Luo have contributed equally to this work. ✉email: 20859180@qq.com; puhongjiang@qq.com

Lung cancer remains the leading cause of cancer-related morbidity and mortality worldwide¹. Despite advances in screening, minimally invasive surgery, targeted therapy, and immunotherapy, long-term survival after curative resection is still curtailed by postoperative distant metastasis^{2,3}. Clinical series suggest that ~30–50% of non-small cell lung cancer (NSCLC) patients ultimately develop metastases—most frequently in the brain, bone, adrenal glands, and liver^{4,5}. Timely, accurate risk stratification is therefore crucial to tailor surveillance, optimize adjuvant therapy, and allocate resources efficiently.

Multiple prognostic approaches—TNM staging, clinicopathologic nomograms, imaging-derived features, and blood-based biomarkers—have been explored^{6–8}. Yet many reported models show only modest sensitivity and specificity, limited generalizability across institutions, and suboptimal calibration, hindering clinical uptake^{9–13}. Moreover, conventional regression may not capture the non-linear, high-dimensional interactions among heterogeneous clinical, pathological, imaging, and laboratory variables seen in practice^{14,15}.

Machine learning (ML) offers principled tools to model such complexity and uncover latent patterns linked to metastatic risk^{16–18}. Prior work often applied single algorithms with narrow scope, small cohorts, and limited interpretability^{19–21}. Because opaque predictions are difficult to trust, transparent models whose reasoning can be interrogated and aligned with clinical knowledge are needed^{22–26}.

To address these gaps, we performed a head-to-head comparison of nine ML algorithms—XGBoost, RF, LightGBM, AdaBoost, DT, GBDT, GNB, CNB, and MLP— in a single-center cohort of 3,120 stage I–III lung cancer patients treated with radical resection. Fifty-two candidate variables spanning clinical features, pathology, imaging-derived metrics, and laboratory indices were assessed. Performance was evaluated using accuracy, precision, recall, F1 score, ROC–AUC, precision–recall metrics, calibration, and decision curve analysis (DCA). We further applied SHAP to provide global and individualized interpretations of the best-performing model, with top contributors aligning with clinical knowledge—for example, nodal status, adjuvant therapy use, and systemic inflammatory indices.

We hypothesized that gradient-boosting approaches would outperform linear baselines and that SHAP would surface clinically plausible predictors. Our aim is to deliver accurate, explainable risk estimates to support postoperative stratification, follow-up planning, and shared decision-making. In summary, we contribute: (i) a large single-center cohort focused on distant metastasis; (ii) a unified, nested cross-validation pipeline for fair benchmarking of nine algorithms; and (iii) an integrated methodology combining LASSO-based feature selection, probability calibration, DCA, and SHAP. These results position our model as a methodological reference and a practical decision-support prototype warranting external, multicenter validation.

Data and methods

Data

Ethics statement

This study's methodology was thoroughly reviewed and approved by the Ethics Committee of Yunnan Oncology Center (Authorization Code: KY2019141) and conducted in full compliance with the principles of the Helsinki Declaration. All participants provided written informed consent prior to enrollment.

Study population

In accordance with the guidelines outlined in the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement²⁷, this study retrospectively analyzed 3,120 patients with stage I-III primary lung cancer confirmed by pathology. All patients underwent radical resection at Yunnan Oncology Center between January 2013 and December 2018.

Inclusion and exclusion criteria

Inclusion Criteria:

1. Pathologically confirmed primary lung cancer diagnosis.
2. Underwent radical surgical treatment.
3. Postoperative follow-up duration of at least 2 years.

Exclusion Criteria:

1. Presence of other malignant tumors.
2. Received preoperative radiotherapy or chemotherapy.
3. Incomplete clinical data Fig. 1.

Postoperative monitoring and metastasis

Postoperative monitoring involved chest and abdominal CT scans and blood tumor marker tests every 3 to 6 months during the first 3 years, followed by tests every 6 to 12 months in years 4 and 5. Additional imaging tests, such as PET-CT or bone scans for bone pain, head MRI for headaches, and enhanced CT, abdominal ultrasound, or gastrointestinal endoscopy for abdominal pain, were performed as needed based on patient symptoms. Upon confirmation of metastasis, a multidisciplinary team was consulted to determine the appropriate treatment plan.

Classification and definition of first organ metastases

An analysis was conducted on the post-relapse survival (PRS) of patients with metastases. PRS is defined as the time from the first detection of metastasis to the last follow-up or all-cause death in patients without event review. The metastatic sites were classified into seven categories: (i) lungs, (ii) brain, (iii) bone, (iv) abdomen (including liver and adrenal glands), (v) pleura, (vi) lymph nodes, and (vii) multiple sites (involving two or more organs).

Research variables

The study included 52 variables across the following categories:

1. General Information: Age, gender, smoking history, number of cigarettes smoked, alcohol consumption history, family history of cancer, previous cancer history, emphysema status, and BMI.
2. Imaging Data: Preoperative whole-body bone scan, lesion type, preoperative pathological biopsy, imaging-based TNM staging, Skeletal muscle area at the T4 level, CT attenuation of skeletal muscle at the T4 level, splenic area at the level of the splenic hilum, splenic hilum CT value, L3 skeletal muscle area, L3 skeletal muscle CT value, L3 visceral fat area, L3 visceral fat CT value, L3 subcutaneous fat area, L3 subcutaneous fat CT value, maximum lesion diameter on CT imaging, and the count of metastatic lymph nodes.
3. Treatment Status: Surgical method, type of resection, lesion location, and number of lymph nodes removed.
4. Pathological Findings: Pathological type, differentiation grade, vascular invasion, bronchial stump involvement, pleural invasion, EGFR and ALK mutation status, and pathological TNM staging.
5. Laboratory Data: Preoperative tumor markers (CEA, CA125, CA199, CA724, SCC, neuron-specific enolase, ferritin, malignant tumor markers), preoperative absolute neutrophil count, and preoperative absolute lymphocyte count (including CD3, CD4, CD8, NK cells).

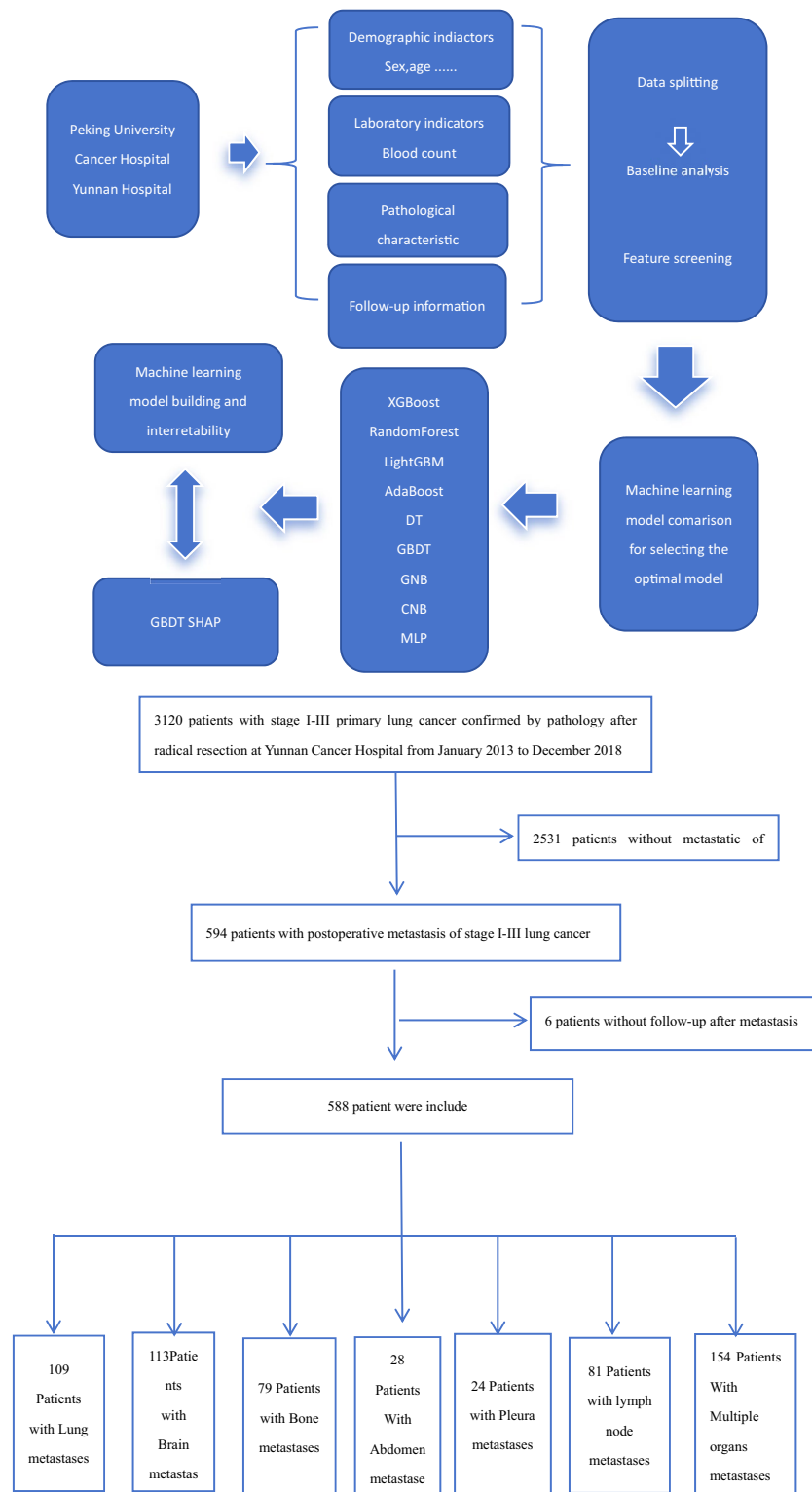


Fig. 1. The flow chart of patient inclusion and exclusion.

Imaging feature extraction and reproducibility

ROI delineation and HU thresholds. Body-composition metrics, including skeletal muscle and adipose tissue areas at the T4 and L3 levels, were derived using semi-automated segmentation with independent manual corrections by two radiologists; discrepancies were resolved by consensus. Hounsfield Unit thresholds followed conventional ranges—muscle: ~ -29 to $+150$ HU; subcutaneous/visceral fat: ~ -190 to -30 HU—with scanner-specific calibration verified monthly²⁸.

Reproducibility and harmonization. Inter-observer agreement was evaluated on an independently annotated subset, and intra-observer agreement on a separate subset with repeat annotations after a washout interval. For continuous features (areas and HU attenuation), agreement was quantified using two-way random-effects, absolute-agreement intraclass correlation coefficients, ICC(2,1), with 95% CIs obtained via bootstrap resampling, alongside Bland–Altman analysis (mean difference and 95% limits of agreement) and the coefficient of variation ($CV = SD/mean$)²⁹. For mask overlap, the Dice similarity coefficient was computed for paired annotations ($Dice = 2|A \cap B|/(|A| + |B|)$). HU stability was maintained through routine phantom-based QA, and sensitivity analyses included scanner type as a covariate. Detailed ICC/Bland–Altman/CV/Dice results are provided in the Supplement³⁰.

Methods

Overview of the evaluation workflow

All models were trained and evaluated within a nested cross-validation framework with fold-confined preprocessing, class-imbalance handling, calibration, and thresholding, as detailed in Sects. “Feature Selection (LASSO)” Sects. “Thresholding and Performance Reporting”. Statistical comparisons and descriptive analyses are described in Sect. “Statistical Analysis”.

Handling class imbalance

The cohort was moderately imbalanced (metastasis ~19.1%, approximately a 4:1 ratio). To mitigate potential training bias, we primarily adopted cost-sensitive learning as the main strategy, using `class_weight = “balanced”` for scikit-learn models and `scale_pos_weight = N_neg/N_pos` for XGBoost and LightGBM. All weights were determined within the inner cross-validation and fixed for the corresponding outer test split to avoid information leakage.

We additionally reported PR-AUC, which is more sensitive to minority-class performance under class imbalance. Oversampling methods (SMOTE and ADASYN) were explored during preliminary model development but were not adopted in the final pipeline due to inferior or unstable performance. Class-imbalance handling was applied only within training folds; the held-out test folds were kept untouched to preserve the real-world prevalence. This strategy has been widely adopted in previous studies addressing class imbalance in medical prediction models³¹.

Feature selection (LASSO)

We applied LASSO-based feature selection to reduce dimensionality and mitigate collinearity prior to model training. Candidate predictors included clinical, pathological, imaging-derived, and laboratory variables. Selected features from the training data were passed forward to all downstream models within the nested cross-validation (CV) pipeline to ensure a consistent predictor set and to avoid information leakage.

Overfitting mitigation and parameterization

Nine classifiers were trained within a single nested CV pipeline. The outer loop used stratified 70/30 (7:3) splits repeated 10 times to obtain unbiased test estimates; the inner loop used stratified fivefold CV for hyperparameter tuning. Tuning primarily optimized ROC–AUC, with PR–AUC and the Brier score as secondary criteria.

Given class imbalance (19.1% metastasis), imbalance handling was performed inside the inner loop to prevent leakage: class weighting (or `scale_pos_weight` for boosting models) and. For tree-based learners (including the final GBDT), the search space covered the number of estimators, learning rate, maximum depth, subsample and column-sample fractions, and L2 regularization; early stopping on inner-CV validation folds was used to curb overfitting. Linear models were trained with L1/L2 penalties, with regularization strength selected by CV. The configuration with the highest mean inner-CV ROC–AUC advanced to the corresponding outer held-out fold. Predicted probabilities were calibrated on inner-CV predictions and then applied unchanged to the outer-test evaluation^{32–34}.

Thresholding and performance reporting

Decision thresholds were pre-specified in the inner CV using Youden’s J and then fixed for the outer held-out sets. In addition to AUROC and PR-AUC, we reported sensitivity, specificity, PPV, NPV, accuracy, and F1 score at the fixed operating threshold. Together, these metrics fully characterize the 2×2 classification outcome and are functionally equivalent to a confusion matrix. Test-set prevalence is additionally provided to contextualize PPV and NPV. Global and individualized interpretability analyses for the best-performing model were conducted using SHAP, as detailed in the corresponding Results subsection^{35,36}.

Statistical analysis

The training and testing datasets were compared across all variables. Continuous variables were summarized as medians with interquartile ranges (IQRs) and compared using the Mann–Whitney U test, whereas categorical variables were expressed as counts and percentages and analyzed with the chi-square test. Effect sizes were reported as standardized mean differences (SMD) for continuous variables and Cramér’s V for categorical variables. To account for multiple testing, false discovery rate control with the Benjamini–Hochberg procedure was applied³⁷. A two-tailed p-value < 0.05 was considered statistically significant.

All performance estimates were derived from internal validation only (repeated outer splits with nested hyperparameter tuning); no external or multicenter validation was available at the time of submission. Analyses were conducted using R (v4.2.3), Python (v3.11.4), and SPSS (v25.0).

Results

Patient characteristics

Baseline characteristics were comparable between the training and testing cohorts

Of the 3,120 eligible patients, 2,184 were assigned to the training cohort and 936 to the testing cohort via stratified random split. There were no statistically significant differences between cohorts across demographic, clinical, pathological, imaging, and laboratory variables (all $P > 0.05$), indicating that the random split achieved a balanced distribution of baseline characteristics and minimized allocation bias for subsequent model development (Table 1).

Baseline characteristics by metastasis status are summarized in Table 2.

Compared with the non-metastasis group ($n = 2,524$), the metastasis group ($n = 596$) was more likely to be male, have a history of smoking and alcohol use, and present with solid or part-solid pulmonary nodules (all $P < 0.05$). They also exhibited larger tumor size on CT and more advanced clinical and pathological T/N stages. Pathology features—poor differentiation, vascular tumor thrombus, and positive bronchial margins—were significantly enriched among patients who developed metastasis. In terms of treatment, the proportions receiving adjuvant chemotherapy and adjuvant radiotherapy were higher in the metastasis group (both $P < 0.001$). Regarding laboratory and body-composition indices, patients with metastasis had higher BMI, elevated Pre-ANC, and higher tumor markers (e.g., CEA, CA-125, CA-199, ferritin) compared with those without metastasis (most comparisons $P < 0.05$). Age was comparable between groups (median 59 vs 58 years, $P = 0.439$).

Feature selection for postoperative distant metastasis

LASSO regression analysis was conducted on the remaining independent variables, with lung cancer metastasis as the dependent variable³⁸. LASSO reduced 52 candidates to 9 predictors at the λ minimizing inner-CV MSE (Fig. 2; Table S1). Of these, age, BMI, pathological N stage, adjuvant chemotherapy, adjuvant radiotherapy, and Pre-ANC remained significant ($p < 0.05$).

Multivariate logistic regression analysis

Multivariate logistic regression analysis was conducted to identify independent risk factors associated with postoperative distant metastasis (Table 3). Age (OR = 1.08, 95% CI: 1.05–1.11, $P < 0.001$), history of smoking (OR = 0.24, 95% CI: 0.12–0.46, $P < 0.001$), BMI (OR = 1.11, 95% CI: 1.03–1.24, $P = 0.032$), wedge resection (OR = 2.62, 95% CI: 1.08–6.92, $P = 0.029$), pathological N2 stage (OR = 3.87, 95% CI: 2.17–16.92, $P < 0.001$), adjuvant chemotherapy (OR = 9.34, 95% CI: 4.05–21.91, $P < 0.001$), and adjuvant radiotherapy (OR = 49.34, 95% CI: 16.18–199.29, $P < 0.001$) were identified as independent predictors of distant metastasis. These large effect sizes for adjuvant therapies likely reflect confounding by indication, as patients with more advanced or aggressive disease were preferentially selected for additional treatment, rather than a direct causal effect. These findings highlight that both clinicopathological features and treatment-related factors significantly contribute to metastasis risk.

Model performance comparison

These features were subsequently used to train nine machine-learning models (GBDT, XGBoost, RF, LightGBM, AdaBoost, DT, GNB, CNB, and MLP). Each model was trained and compared within the nested-CV pipeline (outer stratified 70/30 splits repeated 10 times; inner fivefold tuning). GBDT demonstrated the highest performance (mean AUROC ≈ 0.81), followed by XGBoost and RF (AUROC ≈ 0.79 – 0.80), whereas CNB and GNB performed lower (AUROC ≈ 0.74 – 0.75). Calibration curves indicated that GBDT exhibited the best agreement between predicted and observed probabilities, and decision-curve analysis (DCA) showed the greatest clinical net benefit across a broad threshold range (Fig. 3). Detailed operating-point metrics are provided in Sect. “Optimal Model Selection” and Table S2; representative ROC/PR/calibration/DCA curves are shown in Fig. 3. In the test set, GBDT achieved accuracy 0.766, sensitivity 0.698, specificity 0.786, and F1 0.564 (Fig. 3e–f; Sect. “Optimal Model Selection”), confirming its selection as the optimal model for this dataset.

Optimal model selection

Based on the head-to-head comparison, GBDT was selected as the final model. Held-out performance and the fixed operating point are detailed in Sect. “Thresholded performance (confusion-matrix-equivalent)”.

Thresholded performance (confusion-matrix-equivalent)

All reported results were obtained under the imbalance-aware, cost-sensitive training framework described above, with class weights fixed within the inner cross-validation and applied unchanged to the outer test folds. On the held-out set, the final GBDT achieved AUC 0.810 (95% CI 0.748–0.872) at the pre-specified threshold 0.199 (0.185–0.212). At this operating point, the confusion-matrix-equivalent metrics were: sensitivity 0.698 (0.672–0.724), specificity 0.786 (0.759–0.814), PPV 0.478 (0.436–0.520), NPV 0.904 (0.894–0.914), accuracy 0.766 (0.748–0.785), and F1 0.564 (0.537–0.590) (in Table S2). These metrics together describe the model's classification behavior at the selected operating point, allowing clinicians to understand how often metastasis is correctly identified and how often it may be missed in routine practice. Given sensitivity < specificity, false negatives are relatively more frequent at the observed prevalence, consistent with our DCA-based threshold-benefit trade-offs (Fig. 4).

SHAP explanation of the optimal model

SHAP analysis was applied to enhance model transparency and interpretability. The summary plot identified adjuvant chemotherapy, adjuvant radiotherapy, pathological N stage, BMI, age, and Pre-ANC as the leading

Variable	total (n = 3120)	Training set (n = 2184)	Testing set (n = 936)	Z	p
Sex, n(%)					
Male	1683(53.942)	1186(54.304)	497(53.098)	0.383	0.536
Female	1437(46.058)	998(45.696)	439(46.902)		
Smoking history, n(%)					
no	1822(59.194)	1276(59.211)	546(59.155)	0.001	0.977
yes	1256(40.80612556)	879(40.789)	377(40.845)		
Family history of tumors, n(%)					
no	2815(90.631)	1961(90.161)	854(91.729)	1.889	0.169
yes	291(9.369)	214(9.839)	77(8.271)		
Alcohol history, n(%)					
no	2481(80.657)	1739(80.659)	742(80.652)	0.000	0.997
yes	595(19.343)	417(19.341)	178(19.348)		
History of malignant tumors, n(%)					
no	2915(93.851)	2039(93.747)	876(94.092)	0.135	0.714
yes	191(6.149)	136(6.253)	55(5.908)		
Type of lesion, n(%)					
Ground-glass nodules	384(12.375)	269(12.385)	115(12.352)	0.145	0.930
Partial-solid nodules	475(15.308)	329(15.147)	146(15.682)		
Solid nodules	2244(72.317)	1574(72.468)	670(71.966)		
Preoperative biopsy, n(%)					
no	633(22.203)	453(22.695)	180(21.053)	0.935	0.334
yes	2218(77.797)	1543(77.305)	675(78.947)		
Pathological type, n(%)					
adenocarcinoma	2416(77.436)	1676(76.740)	740(79.060)	2.055	0.358
squamous cell carcinoma	553(17.724)	400(18.315)	153(16.346)		
other types	151(4.840)	108(4.945)	43(4.594)		
Degree of differentiation, n(%)					
Unknown	2264(72.611)	1580(72.377)	684(73.155)	1.659	0.894
High differentiation	9(0.289)	7(0.321)	2(0.214)		
Medium differentiation	447(14.336)	317(14.521)	130(13.904)		
Low differentiation	348(11.161)	247(11.315)	101(10.802)		
Undifferentiated	26(0.834)	16(0.733)	10(1.070)		
Others	24(0.770)	16(0.733)	8(0.856)		
Vascular cancer thrombus, n(%)					
no	3067(98.301)	2146(98.260)	921(98.397)	0.074	0.786
yes	53(1.699)	38(1.740)	15(1.603)		
Bronchial stump, n(%)					
no	3036(97.308)	2126(97.344)	910(97.222)	0.037	0.847
yes	84(2.692)	58(2.656)	26(2.778)		
Pleural invasion, n(%)					
no	2560(83.415)	1776(82.490)	784(85.590)	nan	nan
yes	508(16.553)	376(17.464)	132(14.410)		
Unknown	1(0.033)	1(0.046)	0(0.000)		
Pathological N staging, n(%)					
N0	2270(72.756)	1584(72.527)	686(73.291)	3.831	0.429
N1	185(5.929)	135(6.181)	50(5.342)		
N2	537(17.212)	383(17.537)	154(16.453)		
N3	23(0.737)	16(0.733)	7(0.748)		
Unknown	105(3.365)	66(3.022)	39(4.167)		
Adjuvant chemotherapy, n(%)					
no	1306(46.100)	885(44.629)	421(49.529)	6.085	0.048
yes	1364(48.147)	984(49.622)	380(44.706)		
Chemotherapy after metastasis, n(%)					
	163(5.754)	114(5.749)	49(5.765)		
Adjuvant radiotherapy, n(%)					
no	2369(91.045)	1653(91.125)	716(90.863)	0.121	0.941
Continued					

Variable	total (n = 3120)	Training set (n = 2184)	Testing set (n = 936)	Z	p
yes	209(8.032)	145(7.993)	64(8.122)		
Unknown	24(0.922)	16(0.882)	8(1.015)		
Metastasis, n(%)					
no	2524(80.897)	1764(80.769)	760(81.197)	0.077	0.781
yes	596(19.103)	420(19.231)	176(18.803)		
Age(years), Media(IQR)	59.000[52.000,66.000]	58.000[52.000,66.000]	59.000[51.000,66.000]	-1.048	0.294
BMI(kg/m2), Media(IQR)	22.313[20.761,24.567]	22.376[20.761,24.567]	22.266[20.761,24.539]	0.417	0.677
T4 pectoral muscle skeletal muscle area, median[IQR]	27.790[22.190,34.410]	27.900[22.200,34.420]	27.380[22.110,34.360]	0.103	0.918
T4 chest muscle CT value, median[IQR]	34.690[28.050,40.330]	34.540[27.770,40.200]	35.200[28.460,40.610]	-1.718	0.086
Spleen area at the level of splenic hilum, median[IQR]	27.370[22.020,33.730]	27.490[22.320,33.790]	27.120[21.230,33.560]	1.417	0.157
CT value of splenic hilum and spleen, median[IQR]	48.170[45.810,50.490]	48.080[45.730,50.470]	48.290[46.020,50.520]	-1.048	0.295
L3 skeletal muscle area, median[IQR]	115.700[97.280,139.300]	115.600[97.160,140.100]	115.800[97.930,137.800]	0.288	0.774
L3 skeletal muscle CT value, median[IQR]	38.660[34.440,42.430]	38.880[34.700,42.650]	38.390[34.080,42.170]	1.258	0.208
L3 visceral fat area, median[IQR]	82.430[41.590,130.900]	81.720[41.590,130.200]	86.080[42.140,131.700]	-0.434	0.664
L3 subcutaneous fat area, median [IQR]	105.500[70.200,150.300]	106.700[71.340,151.400]	103.100[67.160,147.900]	1.174	0.240
CT image shows the maximum diameter of the lesion, median[IQR]	2.500[1.600,3.600]	2.500[1.700,3.700]	2.400[1.600,3.600]	0.934	0.350
Pre-CEA, median[IQR]	3.310[2.040,5.760]	3.310[2.040,5.720]	3.310[2.070,5.780]	0.435	0.663
Pre-CA125, median[IQR]	13.720[9.860,20.510]	13.760[9.920,20.620]	13.610[9.730,20.240]	1.020	0.308
Pre-CA-99, median[IQR]	11.410[7.340,18.930]	11.500[7.500,19.170]	10.960[7.030,17.650]	2.047	0.041
Pre-CA724, median[IQR]	2.410[1.240,5.750]	2.300[1.200,5.420]	2.650[1.370,5.950]	-1.764	0.078
Pre-NSE, median[IQR]	13.490[11.530,16.010]	13.510[11.530,16.050]	13.450[11.510,15.950]	0.395	0.693
Pre-SCC, median[IQR]	0.800[0.600,1.160]	0.800[0.600,1.200]	0.800[0.600,1.100]	2.187	0.028
Pre-Ferritin, median[IQR]	266.900[143.500,459.500]	264.800[141.100,466.200]	274.200[150.700,443.000]	-0.523	0.601
Pre-ANC, median[IQR]	3.870[2.960,5.170]	3.880[2.970,5.130]	3.870[2.950,5.220]	-0.402	0.687
Pre-LYM, median[IQR]	1.880[1.500,2.290]	1.900[1.510,2.300]	1.850[1.470,2.250]	1.910	0.056
Pre_CD3, median[IQR]	1177.000[896.000,1539.000]	1166.000[902.000,1529.000]	1209.000[882.000,1546.000]	-0.602	0.547
Pre_CD4, median[IQR]	649.000[486.000,860.000]	649.000[490.000,857.000]	656.000[475.000,867.000]	0.163	0.870
Pre_CD8, median[IQR]	381.000[262.000,540.000]	386.000[263.000,549.000]	372.000[262.000,525.000]	0.830	0.407
Pre-_NK, median[IQR]	300.000[186.000,458.000]	304.000[186.000,465.000]	287.000[189.000,450.000]	0.441	0.659
Total number of metastatic lymph nodes, median[IQR]	0.000[0.000,0.000]	0.000[0.000,0.000]	0.000[0.000,0.000]	0.680	0.356
Total number of lymph nodes cleared, median[IQR]	10.000[5.000,15.000]	10.000[6.000,15.000]	9.000[5.000,15.000]	1.137	0.255

Table 1. Baseline characteristics in training set and testing set.

contributors to metastasis risk (Fig. 5a). Although age did not show a statistically significant difference in the univariate comparison (Table 3, $P=0.439$), this does not contradict its contribution in the SHAP analysis. Univariate tests evaluate marginal group differences, whereas SHAP quantifies the conditional contribution of a feature within a multivariable, nonlinear model. In our model, age mainly contributed through interaction patterns with other clinical variables (e.g., pathological N stage, BMI, and treatment-related variables), which may not be detectable in univariate comparisons. Higher SHAP values for adjuvant therapy and advanced N stage reflected a stronger association with metastasis probability. However, these associations likely reflect treatment indication, whereby patients at higher baseline risk are more likely to receive adjuvant therapies, rather than a direct causal effect.

Dependence plots showed that low BMI and younger age were linked to elevated predicted risk, potentially reflecting more aggressive tumor biology or host frailty (Fig. 5b). Case-specific force plots further illustrated SHAP-based individualized predictions: in a high-risk case ($f(x)=0.92$), multiple nodal metastases and poor differentiation were dominant risk-enhancing features (Fig. 5c); in contrast, in a low-risk case ($f(x)=0.07$), a high proportion of ground-glass opacity and EGFR mutation positivity contributed to a protective profile (Fig. 5d).

Overall, the GBDT model achieved strong predictive performance and provided clinically interpretable insights when combined with SHAP analysis. By capturing adjuvant therapies as proxies for baseline severity rather than independent causal drivers. Importantly, this pattern should not be interpreted as a detrimental effect of adjuvant treatment itself, but rather as a reflection of baseline disease severity and clinical decision-making pathways. The model's explanations therefore align with established clinical understanding and provide a credible foundation for potential clinical integration.

Discussion

Recent advances in machine learning have enabled improved risk stratification across multiple domains of lung cancer management, including diagnosis, treatment response prediction, and prognostic assessment. However,

Characteristic	Total		p-value
	Non-metastasis N = 2,524	Metastasis N = 596	
Age, median [IQR]	59 (52, 66)	58 (52, 66)	0.439
Number of cigarettes smoked per day, median [IQR]	0 (0, 20)	0 (0, 20)	0.040
Sex, n (%)			< 0.001
Male	1,320 (52.3%)	363 (60.9%)	
Female	1,204 (47.7%)	233 (39.1%)	
Smoking history, n(%)			0.017
no	1,499 (60.2%)	323 (54.8%)	
yes	990 (39.8%)	266 (45.2%)	
Alcohol history, n(%)			0.002
no	2,033 (81.7%)	448 (76.2%)	
yes	455 (18.3%)	140 (23.8%)	
Family history of tumors, n(%)			0.361
no	2,269 (90.4%)	546 (91.6%)	
yes	241 (9.6%)	50 (8.4%)	
History of malignant tumors, n(%)			0.623
no	2,354 (93.7%)	561 (94.3%)	
yes	157 (6.3%)	34 (5.7%)	
Emphysema, n(%)			0.855
Absent	2,328 (92.6%)	545 (92.4%)	
Present	182 (7.2%)	45 (7.6%)	
Unknown	3 (0.1%)	0 (0.0%)	
Preoperative whole-body bone scan, n(%)			0.475
no	332 (13.2%)	85 (14.3%)	
yes	2,192 (86.8%)	511 (85.7%)	
Type of lesion, n(%)			< 0.001
Ground-glass nodules	363 (14.4%)	21 (3.6%)	
Partial-solid nodules	405 (16.1%)	70 (11.9%)	
Solid nodules	1,745 (69.4%)	499 (84.6%)	
Preoperative biopsy, n(%)			0.007
no	534 (23.2%)	99 (17.9%)	
yes	1,765 (76.8%)	453 (82.1%)	
Clinical T Stage, n(%)			
Tis	1 (0.0%)	0 (0.0%)	
T1A	255 (10.1%)	19 (3.2%)	
T1B	792 (31.4%)	122 (20.5%)	
T1C	584 (23.1%)	141 (23.7%)	
T2A	452 (17.9%)	142 (23.8%)	
T2B	181 (7.2%)	62 (10.4%)	
T3	138 (5.5%)	58 (9.7%)	
T4	121 (4.8%)	52 (8.7%)	
Clinical N Stage, n(%)			< 0.001
N0	1,705 (67.6%)	287 (48.2%)	
N1	225 (8.9%)	100 (16.8%)	
N2	408 (16.2%)	157 (26.3%)	
N3	87 (3.4%)	30 (5.0%)	
Unknown	99 (3.9%)	22 (3.7%)	
Surgical approach, n(%)			< 0.001
Thoracotomy;	860 (34.1%)	331 (55.5%)	
Thoracoscopy	1,664 (65.9%)	265 (44.5%)	
Old Version of Surgical Resection Methods, n(%)			0.215
Wedge Resection	389 (15.6%)	91 (15.5%)	
Segmentectomy	34 (1.4%)	5 (0.9%)	
Lobectomy	2,018 (81.0%)	469 (80.0%)	
Combined Lobectomy	5 (0.2%)	1 (0.2%)	
Continued			

Characteristic	Total		p-value
	Non-metastasis N = 2,524	Metastasis N = 596	
Bronchial Sleeve Lobectomy	22 (0.9%)	9 (1.5%)	
Pneumonectomy	23 (0.9%)	11 (1.9%)	
Lesion Location, n(%)			
Right Upper Lobe of Lung	706 (28.0%)	160 (26.8%)	
Right Middle Lobe of Lung	151 (6.0%)	45 (7.6%)	
Right Lower Lobe of Lung	571 (22.6%)	124 (20.8%)	
Left Upper Lobe of Lung	575 (22.8%)	132 (22.1%)	
Left Lower Lobe of Lung	429 (17.0%)	106 (17.8%)	
Right Lung	71 (2.8%)	18 (3.0%)	
Left Lung	20 (0.8%)	11 (1.8%)	
Bilateral Lungs	1 (0.0%)	0 (0.0%)	
New Version of Surgical Methods, n(%)			< 0.001
Lobectomy	2,199 (87.1%)	514 (86.2%)	
Segmentectomy	98 (3.9%)	6 (1.0%)	
Wedge Resection	179 (7.1%)	55 (9.2%)	
Pneumonectomy	48 (1.9%)	21 (3.5%)	
Pathological Type, n(%)			< 0.001
Adenocarcinoma	1,983 (78.6%)	433 (72.7%)	
Squamous Cell Carcinoma	435 (17.2%)	118 (19.8%)	
Others	106 (4.2%)	45 (7.6%)	
Degree of Differentiation, n(%)			
Unknown	1,882 (74.6%)	382 (64.1%)	
Well-differentiated	9 (0.4%)	0 (0.0%)	
Moderately Differentiated	346 (13.7%)	101 (16.9%)	
Poorly Differentiated	250 (9.9%)	98 (16.4%)	
Undifferentiated	16 (0.6%)	10 (1.7%)	
Other	19 (0.8%)	5 (0.8%)	
Vascular Tumor Thrombus, n(%)			< 0.001
Absent	2,494 (98.8%)	573 (96.1%)	
Present	30 (1.2%)	23 (3.9%)	
Bronchial Resection Margin, n(%)			< 0.001
Negative	2,478 (98.2%)	558 (93.6%)	
Positive	46 (1.8%)	38 (6.4%)	
Pleural Invasion, n(%)			0.290
Absent	2,083 (83.9%)	477 (81.3%)	
Present	398 (16.0%)	110 (18.7%)	
Unknown	1 (0.0%)	0 (0.0%)	
Pathological T Stage, n(%)			< 0.001
T1A	215 (8.5%)	13 (2.2%)	
T1B	586 (23.2%)	95 (15.9%)	
T1C	335 (13.3%)	77 (12.9%)	
T2A	752 (29.8%)	191 (32.0%)	
T2B	197 (7.8%)	65 (10.9%)	
T3	263 (10.4%)	101 (16.9%)	
T4	175 (6.9%)	54 (9.1%)	
Pathological N Stage, n(%)			
N0	1,987 (78.7%)	283 (47.5%)	
N1	120 (4.8%)	65 (10.9%)	
N2	336 (13.3%)	201 (33.7%)	
N3	8 (0.3%)	15 (2.5%)	
Unknown	73 (2.9%)	32 (5.4%)	
Adjuvant Chemotherapy, n(%)			< 0.001
None	1,217 (53.8%)	89 (15.6%)	
Administered	960 (42.4%)	404 (70.9%)	
Continued			

Characteristic	Total		p-value
	Non-metastasis N = 2,524	Metastasis N = 596	
Chemotherapy after Recurrence/Metastasis	86 (3.8%)	77 (13.5%)	
Adjuvant Radiotherapy, n(%)			<0.001
None	2,015 (97.3%)	354 (66.7%)	
Administered	34 (1.6%)	175 (33.0%)	
Unknown	22 (1.1%)	2 (0.4%)	
T4 pectoral muscle skeletal muscle area, median [IQR]	27 (22, 34)	29 (23, 35)	0.003
CT value of the T4 pectoralis muscle, median [IQR]	35 (28, 40)	35 (29, 41)	0.201
Spleen area at the splenic hilum level, median [IQR]	27 (22, 34)	28 (22, 35)	0.019
CT value of the spleen at the splenic hilum, median [IQR]	48.1 (45.8, 50.5)	48.5 (45.8, 50.8)	0.158
Skeletal muscle area at the L3 level, median [IQR]	115 (97, 138)	120 (99, 145)	0.010
CT value of skeletal muscle at the L3 level, median [IQR]	38.5 (34.1, 42.2)	39.4 (35.7, 43.1)	0.023
Visceral fat area at the L3 level, median [IQR]	82 (41, 128)	83 (43, 141)	0.100
Subcutaneous fat area at the L3 level, median [IQR]	106 (70, 151)	105 (71, 148)	0.828
BMI, median [IQR]	22.23 (20.76, 24.45)	22.65 (20.76, 25.00)	0.014
Maximum diameter of the lesion on CT image (cm), median [IQR]	2.30 (1.60, 3.50)	3.00 (2.10, 4.30)	<0.001
Preoperative CEA, median [IQR]	3 (2, 5)	4 (3, 10)	<0.001
Preoperative CA-125, median [IQR]	13 (10, 20)	16 (11, 27)	<0.001
Preoperative CA-199, median [IQR]	11 (7, 18)	13 (7, 22)	0.005
Preoperative CA-724, median [IQR]	2 (1, 6)	3 (1, 6)	0.057
Preoperative NSE, median [IQR]	13.5 (11.5, 15.9)	13.6 (11.6, 16.5)	0.234
Preoperative SCC, median [IQR]	0.80 (0.60, 1.10)	0.80 (0.60, 1.20)	0.315
Preoperative Ferritin, median [IQR]	261 (141, 445)	286 (160, 517)	0.004
Malignant tumor factor, median [IQR]	50 (41, 56)	51 (45, 57)	0.535
Preoperative ANC, median [IQR]	3.82 (2.89, 5.13)	4.07 (3.26, 5.34)	<0.001
Preoperative ALC, median [IQR]	1.88 (1.50, 2.28)	1.88 (1.50, 2.32)	0.594
Preoperative Absolute T-cell Count, median [IQR]	1,189 (907, 1,542)	1,112 (833, 1,512)	0.107
Preoperative Absolute Th Cell Count, median [IQR]	656 (496, 865)	632 (464, 848)	0.200
Preoperative Absolute Ts Cell Count, median [IQR]	384 (268, 546)	366 (234, 527)	0.133
Preoperative Absolute NK Cell Count, median [IQR]	301 (186, 464)	288 (185, 444)	0.398
Total number of metastatic lymph nodes, median [IQR]	0.00 (0.00, 0.00)	0.00 (0.00, 3.00)	<0.001
Total number of dissected lymph nodes, median [IQR]	9 (5, 14)	11 (6, 16)	0.004

Table 2. Baseline characteristics by metastasis status (Non-metastasis vs Metastasis).

most existing models focus on imaging-based or molecular endpoints, with limited attention to postoperative distant metastasis and insufficient emphasis on interpretability^{39–45}. This gap motivates the present study.

Performance of different machine learning models in predicting distant metastasis after lung cancer surgery

We systematically compared nine machine-learning algorithms for predicting postoperative distant metastasis. Average AUROCs were obtained within the nested-CV pipeline (outer stratified 70/30 splits repeated 10 times; inner fivefold tuning), rather than a single tenfold CV (Fig. 3). GBDT achieved the best held-out performance (AUROC 0.810, accuracy 0.766, sensitivity 0.698, specificity 0.786; Sect. “Optimal Model Selection”, Table S2). This profile of moderate sensitivity and higher specificity implies relatively more false negatives at the observed prevalence. Clinically, if the priority is to minimize missed metastasis, a left-shift of the deployment threshold (favoring higher sensitivity and NPV at the expense of more false positives) is appropriate; if avoiding unnecessary interventions is prioritized, the current threshold—or a right-shift after recalibration—may be preferable. Consistent with prior oncology studies, gradient boosting has been reported to outperform linear baselines^{46,47}, and to maintain stable performance across internal validation^{48,49}, although external transportability still requires confirmation in multicenter cohorts⁵⁰.

Clinical application value of the optimal model

We evaluated the clinical utility of the GBDT model from multiple complementary perspectives, including discrimination, calibration, decision-curve analysis, and learning-curve stability (Figs. 3b, 3c, 3f). The GBDT achieved an AUROC of 0.810 (95% CI 0.748–0.872) at the prespecified operating threshold derived from inner cross-validation. Thresholded metrics (sensitivity, specificity, PPV, NPV, accuracy, F1) are reported in Sect. “Optimal Model Selection” and Table S2, alongside PR-AUC, calibration, and DCA summaries (Fig. 3c). In addition, Learning-curve and repeated-split analyses suggested stable internal generalization and efficient

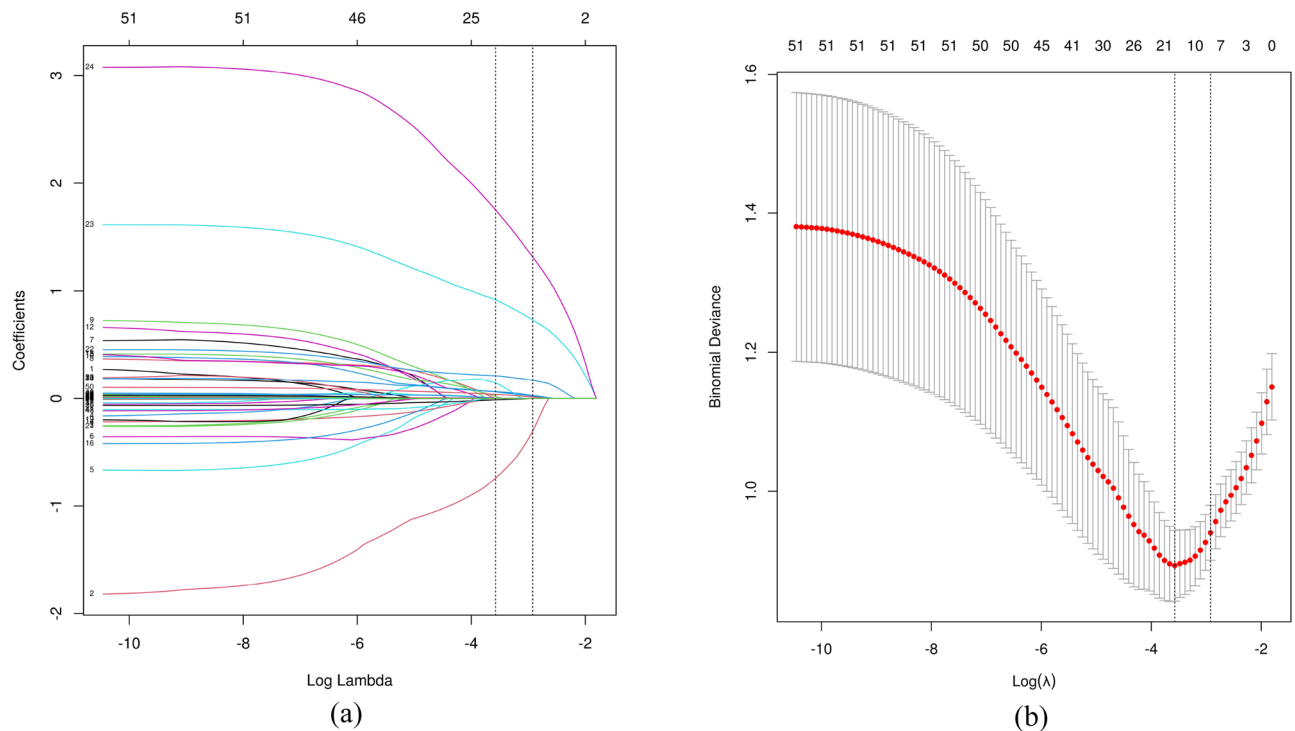


Fig. 2. LASSO regression analysis was employed to select key factors. **(a)** Ten-fold cross-validation was used to plot vertical lines at the chosen values, with the optimal lambda yielding nine non-zero coefficients. **(b)** In the LASSO model, the coefficient profiles for the 52 candidate features were plotted based on the $\log(\lambda)$ sequence. Vertical dashed lines indicate the points at the minimum mean square error ($\lambda = 0.028$) and the standard error for the minimum distance ($\lambda = 0.054$).

data utilization for GBDT, which is particularly valuable in real-world settings where sample sizes may be limited (Fig. 4d). External transportability remains to be established. According to prior literature, GBDT offers several practical advantages. (i) Dynamic risk stratification enables tailored follow-up schedules across low-, intermediate-, and high-risk groups. (ii) Resource optimization is supported by decision-curve analysis, which indicates net clinical benefit across broad thresholds (Fig. 3c). (iii) Actionable interpretability is achievable when predictions are paired with SHAP summaries and case-level explanations, facilitating clinician review rather than replacing clinical judgment. Notably, these findings are associational rather than causal, and prospective multicenter validation is required before model-guided interventions can be recommended^{51,52}.

Misclassification profile and operating-point strategy

At the reported operating point, sensitivity is lower than specificity, implying relatively more false negatives at the observed prevalence. This aligns with our DCA: if the clinical priority is to minimize missed metastasis, the deployment threshold can be shifted left to gain sensitivity (with more false positives); if avoiding unnecessary interventions is prioritized, the current threshold—or a right-shift after recalibration—may be preferable. We report the fixed-threshold baseline here and outline a clear path for external validation and recalibration⁵³. Together, these findings indicate that model performance, error profile, and interpretability must be considered jointly when translating predictive tools into clinical workflows.

Clinical significance of SHAP interpretation results

SHAP-based interpretation enhances the transparency of the GBDT model by providing clinically meaningful explanations for individual predictions. This may improve clinicians' understanding of model predictions and facilitate more informed decision-making. Previous studies have suggested that SHAP-based interpretability tools may enhance clinicians' confidence in model-assisted decision-making and facilitate more rational treatment adjustments^{45–55}. The SHAP analysis based on the GBDT model in this study reveals the significant impact of clinical features such as adjuvant chemotherapy, adjuvant radiotherapy, pathological N stage, BMI, age, and Pre-ANC on distant metastasis after lung cancer surgery. Notably, age appeared among the top contributors despite a non-significant univariate difference. This is expected because SHAP reflects model-based conditional contribution rather than marginal group separation. Clinically, age may modulate risk in conjunction with nodal status, body composition and treatment decisions, highlighting the importance of considering patient factors jointly rather than in isolation. This finding is consistent with the CALGB 9633 study, which reported a survival benefit from adjuvant chemotherapy, and with the 8th edition of the AJCC staging system, which emphasizes pathological N stage as a core prognostic factor. Using LASSO preserved variable-level semantics and facilitated alignment^{56,57}. Selection frequencies across outer repetitions suggested a core subset of predictors was repeatedly

Variable	R	SE	Z	p	OR(95%CI)
Age	0.077	0.015	5.182	0.0	1.08(1.05–1.113)
History of smoking	–1.421	0.334	–4.257	0.0	0.241(0.124–0.46)
T4 pectoral muscle skeletal muscle area	–0.03	0.017	–1.777	0.076	0.97 (0.937–1.003)
CT value of splenic hilum and spleen	0.029	0.029	0.998	0.319	1.029(0.937–1.09)
Subcutaneous fat area of L3	–0.002	0.003	–0.762	0.446	0.998(0.991–1.003)
BMI	0.108	0.051	2.144	0.032	1.114(1.028–1.242)
CT image shows the maximum diameter of the lesion	0.299	0.162	1.841	0.066	1.348(0.971–1.851)
pre-CA-125	0.001	0.001	0.988	0.323	1.001(0.999–1.003)
pre-Absolute value of Th cells	0.0	0.0	–0.67	0.503	1.0(0.999–1.0)
Clinical T staging(T1B)	1.146	0.729	1.572	0.116	3.145(0.851–15.587)
Clinical T staging(T1C)	0.892	0.774	1.152	0.249	2.441(0.589–12.8855)
Clinical T staging(T2A)	1.087	0.834	1.303	0.193	2.965(0.625–17.151)
Clinical T staging(T2B)	0.997	0.975	1.023	0.306	2.711(0.423–20.005)
Clinical T staging(T3)	0.443	1.086	0.408	0.683	1.557(0.189–13.754)
Clinical T staging(T4)	0.383	1.399	0.274	0.784	1.467(0.096–24.123)
New version of surgical methods(Segmentectomy of lung)	0.253	0.702	0.36	0.719	1.287(0.269–4.563)
New version of surgical methods(wedge resection)	0.962	0.439	2.19	0.029	2.618(1.076–6.92)
New version of surgical methods(pneumonectomy)	–0.443	1.059	–0.419	0.676	0.642(0.071–4.486)
Degree of differentiation(moderately differentiated)	–0.12	0.778	–0.155	0.877	0.887(0.18–3.968)
Degree of differentiation(Poorly differentiated)	–0.384	0.771	–0.499	0.618	0.681(0.144–3.114)
Degree of differentiation(undifferentiation)	1.347	1.407	0.958	0.338	3.846(0.146–63.96)
Pathological N staging(N1)	0.397	0.472	0.84	0.401	1.487(0.56–83.662)
Pathological N staging(N2)	1.352	0.295	4.584	0.0	3.866(2.17–16.92)
Pathological N staging(N3)	1.834	1.065	1.722	0.085	6.258(0.765–60.72)
Pathological N staging(unknown)	0.252	0.654	0.385	0.7	1.286(0.348–4.582)
adjuvant chemotherapy	2.234	0.429	5.207	0.0	9.335(4.047–21.912)
Adjuvant radiotherapy	3.899	0.63	6.19	0.0	49.338(16.18–199.292)
(Intercept)	–10.737	2.122	–5.061	0.0	0.0 (-)

Table 3. Multivariate logistic regression analysis.

retained, indicating robustness to resampling. Complementary selectors (e.g., elastic-net for grouped correlation) could be considered as sensitivity analyses while keeping reporting focused on clinically interpretable variables. This pattern is clinically plausible, as younger patients with advanced nodal disease or aggressive pathology may experience higher metastatic risk despite similar chronological age distributions.

SHAP-based interpretation made the model's predictions transparent and clinically interpretable. Summary plots identified adjuvant chemotherapy, adjuvant radiotherapy, pathological N stage, BMI, age, and Pre-ANC as the leading contributors to predicted metastasis risk (Fig. 5a–b).

These explanations are associational rather than causal; in particular, attributions involving adjuvant therapies likely reflect confounding by indication, as patients at higher baseline risk are more likely to receive additional treatment (Fig. 5c–d). This observation is compatible with established clinicopathologic knowledge summarized in the WHO classification of lung tumors⁵⁸.

Comparison with prior work

Compared with prior studies that typically assessed a single algorithm, used narrow feature sets, or emphasized discrimination alone, our work adopts a broader algorithmic scope, a richer predictor set, and a more clinically oriented evaluation. We conducted a unified, head-to-head comparison of nine ML models; complemented ROC-AUC/PR-AUC with calibration and decision curve analysis (DCA) to quantify clinical utility; and used SHAP to connect predictions with clinical reasoning at both global and patient levels⁵⁹. These design choices enhance reproducibility and interpretability, and may partly explain why specific variables emerge as dominant predictors in both traditional between-group analyses and the ML framework.

Recent radiomics-based studies have reported higher discrimination for metastasis-related endpoints. For example, a CT radiomics study published in BMC Cancer reported an AUC approaching 0.90 under an internal split validation scheme⁶⁰. Importantly, the higher AUC reported in radiomics-based studies is often obtained under single-split internal validation without fold-confined preprocessing or probability calibration, which may inflate apparent performance. In contrast, our repeated nested cross-validation framework yields more conservative but less optimistically biased estimates. However, direct numerical comparison is not straightforward because the predicted endpoint (e.g., brain metastasis versus overall distant metastasis), feature modality (radiomics versus routinely available clinicopathologic and treatment variables), and validation strategy differ. In the present study, we intentionally restricted predictors to readily available perioperative

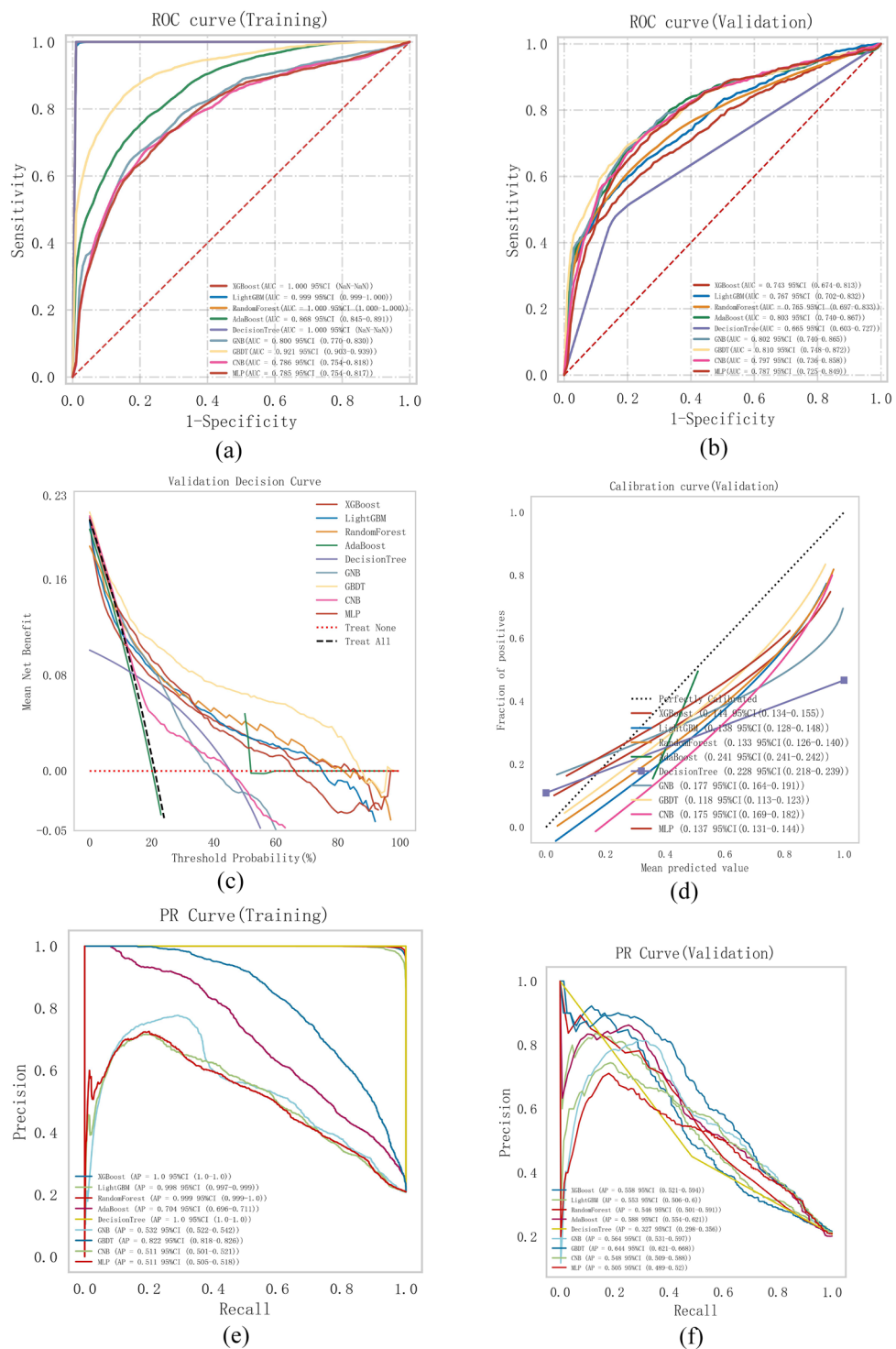


Fig. 3. Comprehensive Evaluation of ML Models. **(a)** ROC curves and AUC for the training set. **(b)** ROC curves and AUC for the test set. Lung cancer patients were split within the nested-CV outer loop (stratified 70/30, repeated 10 times). **(c)** Decision curve analysis (DCA) for the test set. The black dashed line assumes all patients have distant metastases, while the red dashed line and thin black line assume none have distant metastases. Solid lines represent various models. **(d)** Calibration curves for the test set. Coordinates indicate average predicted probabilities, with the slanted line representing actual probabilities. The diagonal dashed line serves as a reference, while the smoothed solid lines reflect the fit of different models. The closer the fit to the reference line and the smaller the values in parentheses, the more accurate the model predictions. **(e)** PR curves and AP for the training set. **(f)** PR curves and AP for the test set. Precision is shown on the vertical axis, and recall on the horizontal axis. If one model's PR curve entirely encompasses another's, it demonstrates superior performance. Higher AP values indicate stronger model performance. Colors in the charts denote different models, results are shown as means with 95% confidence intervals. (CI).

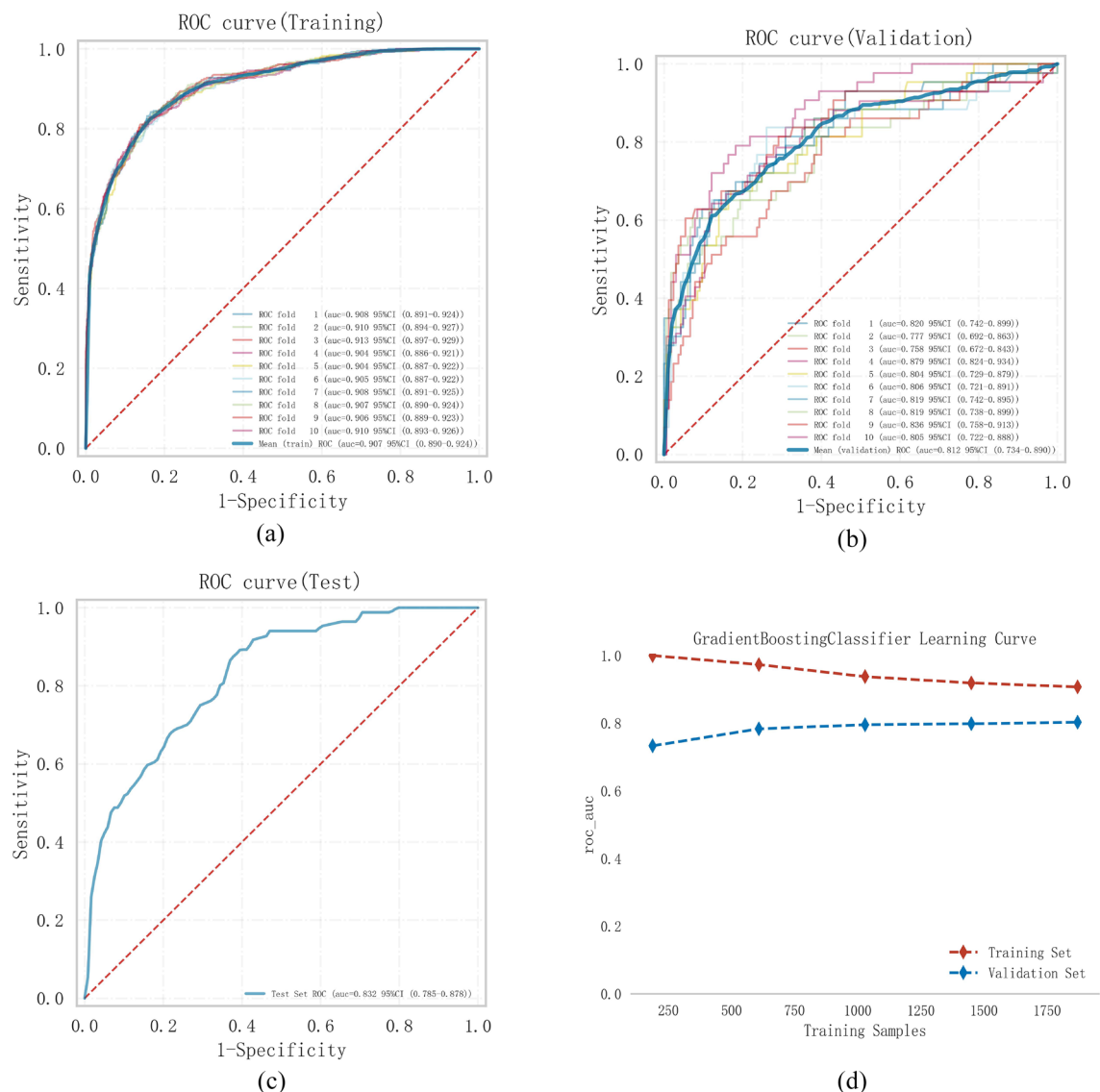


Fig. 4. Learning curves and stability analysis of the GBDT model under the nested cross-validation framework. **(a)** ROC and AUC for the training set, and **(b)** ROC and AUC for the validation set. The results reflect tenfold cross-validation, with each solid line representing a different run. **(c)** ROC and AUC for the test set, showing outcomes for 30% of the patients. **(d)** Learning curves illustrating model performance stability across increasing training set sizes within the nested-CV framework. Curves show mean AUROC with 95% confidence intervals over 10 outer-loop repetitions.

variables and adopted repeated nested cross-validation, which yields more conservative but less optimistically biased estimates (Table S4).

Limitations of the research and future prospects

This study delivers a head-to-head, multi-algorithm comparison on a large, clinically rich cohort; complements discrimination with calibration, decision-curve, and precision-recall evaluation; and provides transparent explanations via SHAP at both global and individual levels. These choices enhance reproducibility, interpretability, and clinical relevance, supporting workflow integration for risk-stratified postoperative management. Nevertheless, several limitations merit discussion.

First, this was a single-center retrospective study, which may limit generalizability and transportability across different clinical settings. Although we used repeated nested cross-validation with fold-confined preprocessing, tuning, and calibration to mitigate overfitting, the absence of external validation may still leave residual optimism, and multicenter validation is therefore warranted before clinical implementation. Although repeated outer splits with nested tuning and systematic internal evaluation (ROC-/PR-AUC, calibration, and DCA) reduce optimism, they do not substitute for external or multicenter validation. Generalizability may be constrained by center-specific case-mix, scanner/reconstruction variability, and local practice patterns. GBDT demonstrated

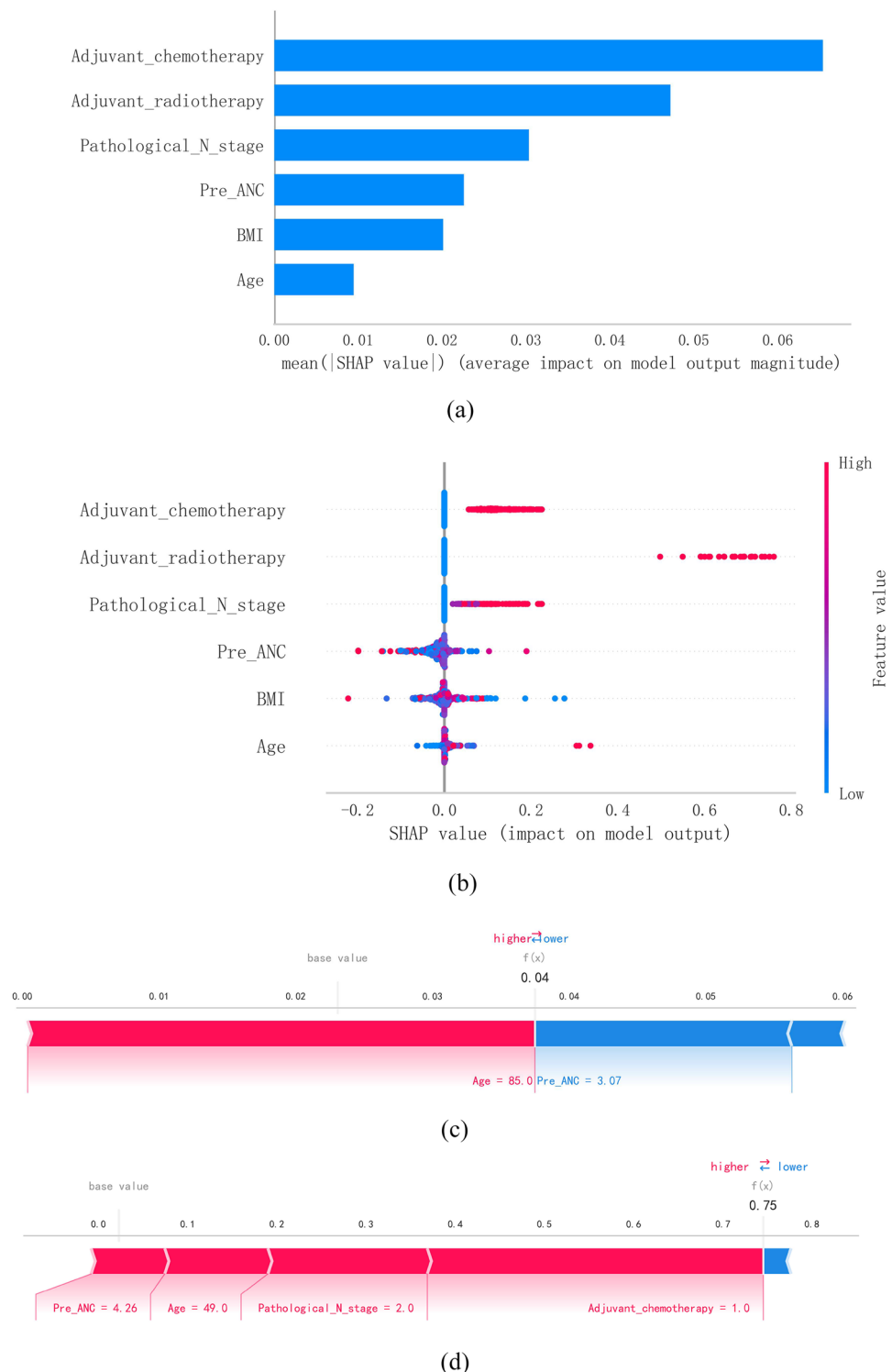


Fig. 5. SHAP model interpretation. (a) SHAP attribute visualization: each line represents a feature, with the x-axis showing the SHAP values. Red dots indicate higher feature values, while blue dots represent lower values. (b) Feature importance ranking by SHAP: the matrix diagram outlines the contribution of each covariate to the final prediction model. (c) Individual SHAP values for patients without metastasis and (d) for those with metastasis: SHAP values depict the predictive characteristics for each patient, showing how each feature contributes to the predicted metastasis risk. The bold number is the forecast probability ($f(x)$), while the base value represents the prediction without input data. $F(x)$ is the logarithmic ratio for each observation. Red features highlight an increased risk of metastasis, while blue features show a reduced risk. The length of the arrows illustrates the magnitude of each feature's impact—the longer the arrow, the more significant the effect on the prediction.

stable convergence across increasing training set sizes (Fig. 4d), suggesting robust internal generalization. This aligns with Lee et al.⁶¹, who noted that single-center datasets are prone to selection bias affecting performance across diverse populations.

Second, the mechanistic linkage of selected features remains limited. LASSO screening (Fig. 2; e.g., 9 non-zero coefficients identified at $\lambda = 0.028$ via tenfold cross-validation) successfully recognized imaging-derived and clinicopathologic features with predictive value. However, the biological mechanisms underlying these features have not been deeply linked to pathological indicators or molecular markers (e.g., EGFR). The coefficient distribution plot (Fig. 3) only reflects the statistical significance of the features, rather than providing insights into the tumor microenvironmental context. This limitation aligns with a common gap in the field of radiomics: statistically significant features often lack mechanistic validation—a challenge also highlighted in the study by Hatt et al.⁶².

Third, imaging-feature variability. Imaging-derived features can vary with slice selection, segmentation, and vendor/kernel differences. We mitigated and quantified these via semi-automated HU-thresholded segmentation for body composition extraction, isotropic resampling, and inter/intra-observer assessments (ICC(2,1), Bland–Altman, Dice, CV%). Sensitivity analyses ($\pm$ HU, morphological operations) suggested minimal impact on conclusions (Table Sx, Figure Sx), but multicenter, multi-vendor data and automated segmentation are needed to further reduce measurement error⁶³.

Fourth, static postoperative inputs. The present model uses cross-sectional postoperative variables and does not incorporate time-varying information (e.g., treatment response, surveillance trajectories), which may understate risk in evolving clinical courses⁶⁴.

Future work will include temporal and geographic external validation across multiple centers equipped with imaging devices from different vendors, reconstruction kernels, and case-prevalence profiles to rigorously assess model generalizability. In parallel, we will evaluate model updating strategies, incorporate SHAP interaction values to clarify synergistic feature effects, and explore dynamic prediction approaches together with multi-omics integration. Through these efforts, our goal is to transition the model from an “internally validated” construct to a deployable decision-support tool with verified performance in out-of-sample populations.

In this context, the current findings should be regarded as an internal benchmark pending external validation and head-to-head multicenter evaluation. To enable rigorous out-of-sample assessment and prospective clinical integration, we have provided experimental protocols, containerized tools, and a scoping dossier. These findings should be interpreted in the context of existing machine learning-based prognostic studies in lung cancer⁶⁵.

Conclusion

This study established and compared nine machine learning models to predict distant metastasis after lung cancer surgery, with the GBDT model demonstrating the most robust performance (76.6% accuracy at the prespecified operating threshold; AUROC 0.810, 95% CI 0.748–0.872). SHAP-based interpretation revealed adjuvant chemotherapy, adjuvant radiotherapy, pathological N stage, BMI, age, and Pre-ANC as the principal contributors to metastatic risk. By enabling postoperative risk stratification, the model may support individualized diagnostic and therapeutic planning. Although constrained by a single-center cohort and the absence of external validation, the findings highlight the potential role of machine learning for lung cancer prognosis. Importantly, they lay a foundation for the next stage of research—integrating multi-omics data, leveraging multicenter cohorts, and moving toward a clinically deployable, transparent decision-support tool capable of enhancing personalized lung cancer management.

Data availability

Data are available from the corresponding author, Jun F, upon reasonable request.

Received: 9 November 2024; Accepted: 19 January 2026

Published online: 28 January 2026

References

1. Torre, L. A. et al. Lung cancer incidence and mortality trends in 50 countries, 1990–2020: A joinpoint regression analysis. *J Thorac. Oncol.* **17**(2), 237–248 (2022).
2. Sung, H. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**(3), 229–263 (2024).
3. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**(5), 395–409 (2011).
4. Howlader, N. et al. The effect of advances in lung-cancer treatment on population mortality. *N. Engl. J. Med.* **383**(7), 640–649 (2020).
5. Chen, W. et al. Cancer statistics in China, 2015. *CA Cancer J Clin.* **66**(2), 115–132 (2016).
6. Wang, Y. et al. Prognostic factors for patients with non-small cell lung cancer and distant metastasis after surgery: A meta-analysis. *BMC Cancer* **21**(1), 1234 (2021).
7. Thiery, J. P. et al. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**(5), 871–890 (2009).
8. Chen, L. et al. Metastasis: From dissemination to organ-specific colonization. *Nat. Rev. Cancer* **11**(4), 274–284 (2011).
9. Topalian, S. L., Drake, C. G. & Pardoll, D. M. Immune checkpoint blockade: A common denominator approach to cancer therapy. *Cancer Cell* **27**(4), 450–461 (2015).
10. Zhang, Y. et al. Prognostic significance of epithelial-mesenchymal transition markers in non-small cell lung cancer: A systematic review and meta-analysis. *Oncotarget* **7**(45), 72539–72550 (2016).
11. Goldstraw, P. et al. The IASLC lung cancer staging project: Proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. *J. Thorac. Oncol.* **11**(1), 39–51 (2016).
12. Owonikoko, T. K. et al. Lung cancer in elderly patients: An analysis of the surveillance, epidemiology, and end results database. *J. Clin. Oncol.* **25**(35), 5570–5577 (2007).

13. Mok, T. S. et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N. Engl. J. Med.* **361**(10), 947–957 (2009).
14. Kong, X. F. et al. A nomogram to predict distant metastasis in resected non-small-cell lung cancer. *J. Thorac. Cardiovasc. Surg.* **156**(1), 359–367.e3 (2018).
15. Abbosh, C. et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**(7655), 446–451 (2017).
16. Li, Y. et al. Exosomal miR-21 and miR-106b as biomarkers for predicting distant metastasis in non-small cell lung cancer. *J. Cell Biochem.* **120**(9), 14679–14688 (2019).
17. Yang, X. et al. Prognostic models for predicting distant metastasis in resected non-small cell lung cancer: A systematic review and meta-analysis. *BMC Cancer* **22**(1), 567 (2022).
18. Amin MB, et al. AJCC cancer staging manual (9th Edition) Springer (2023).
19. Abbosh, C. et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **617**(7961), 553–560 (2023).
20. Rusch, V. W. et al. The IASLC lung cancer staging project: Analysis of racial disparities in stage IB non-small cell lung cancer. *J. Thorac. Oncol.* **17**(3), 401–412 (2022).
21. Wang, X. et al. Challenges in integrating multi-omic data for precision oncology. *Nat Rev Clin Oncol.* **18**(12), 737–751 (2021).
22. Jordan, M. I. et al. Machine learning: Trends, perspectives, and prospects. *Science* **374**(6566), 603–608 (2021).
23. Collins, F. S. et al. Precision medicine initiative: Leveraging data for better health. *N. Engl. J. Med.* **387**(19), 1773–1782 (2022).
24. Lehman, C. D. et al. Deep learning for breast cancer screening: A population-based study. *Radiology* **303**(3), 531–541 (2022).
25. Zhang, Y. et al. Machine learning outperforms TNM staging for colorectal cancer metastasis prediction: A multicenter cohort study. *J. Clin. Oncol.* **41**(12), 2134–2143 (2023).
26. Zhang, Y. et al. Meta-analysis of ML in cancer prognosis. *Lancet Digit. Health* **5**(2), e105–e114 (2023).
27. von Elm, E. et al. The strengthening of reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Int. J. Surg.* **12**(12), 1495–1499 (2014).
28. Peng, X. et al. Repeatability and reproducibility of computed tomography radiomic features: A multicenter phantom study. *Invest. Radiol.* **57**(4), 242–253 (2022).
29. Ackermans, L. L. G. C. et al. Deep learning automated segmentation for muscle and adipose tissue from abdominal computed tomography in polytrauma patients. *Sensors* **21**(6), 2083 (2021).
30. Whybra, P. et al. The image biomarker standardization initiative: Standardized convolutional filters for reproducible radiomics and enhanced clinical insights. *Radiology* **310**, e231319 (2024).
31. Szymanski, J. J. & Kajdanowicz, T. A scikit-learn compatible framework for imbalanced learning. *J. Mach. Learn. Res.* **22**(166), 1–5 (2021).
32. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLoS ONE* **14**(11), e0224365 (2019).
33. Chicco, D., Warrens, M. J. & Jurman, G. The coefficient of determination R^2 , precision–recall AUC, and Brier score: Loss functions for binary classification. *Sci. Rep.* **11**, 218 (2021).
34. Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
35. Fluss, R., Faraggi, D. & Reiser, B. Estimation of the Youden Index and its associated cutoff point. *Biom. J.* **47**(4), 458–472 (2005).
36. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 252–259 (2020).
37. Austin, P. C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **28**(25), 3083–3107 (2009).
38. Wang, H. et al. Machine learning with LASSO for early detection of lung cancer recurrence: Addressing overfitting in imbalanced datasets. *Comput. Biol. Med.* **145**, 105432 (2023).
39. LiuHuangChen, Y. J. J. C. et al. Predicting treatment response in multicenter non-small cell lung cancer patients based on federated learning. *BMC Cancer* **24**(1), 688 (2024).
40. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**(6), 954–961 (2019).
41. Wang, X. et al. Deep learning for low-dose CT lung cancer screening: A multi-center validation study. *Nat. Med.* **28**(3), 571–579 (2022).
42. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**(8), 1301–1309 (2019).
43. Jiang, Y. et al. Dynamic prediction of immunotherapy response in lung cancer using Bayesian networks. *Lancet Digit. Health* **4**(3), e186–e195 (2022).
44. Goldstein, B. A. et al. Predicting lung cancer recurrence: A comparison of machine learning approaches. *J. Thorac. Oncol.* **16**(5), 832–842 (2021).
45. Zhang, Y. et al. DeepVariant 2.0: Improved variant calling with deep learning. *Nat. Biotechnol.* **40**(5), 613–622 (2022).
46. Smeets, D. M. et al. AI-based diabetic retinopathy screening in primary care: A multi-center randomized trial. *Lancet Digit. Health* **5**(2), e105–e114 (2023).
47. Yan, Z. et al. XGBoost algorithm and logistic regression to predict the postoperative 5-year outcome in patients with glioma. *Ann. Transl. Med.* **10**(16), 98928 (2022).
48. Hassan, M. M. et al. A comparative assessment of machine learning algorithms with the least absolute shrinkage and selection operator for breast cancer detection and prediction. *Decis. Anal. J.* **7**, 100245 (2023).
49. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **7**, 91 (2006).
50. Wilimitis, D. et al. Practical considerations and applied examples of cross-validation in clinical ML. *Patterns* **4**(11), 100866 (2023).
51. Collins, G. S. et al. TRIPOD+AI: Updated guidance for reporting ML and regression prediction models. *BMJ* **385**, e078378 (2024).
52. Miller, K. D. et al. Machine learning-guided postoperative surveillance in lung cancer. *J. Clin. Oncol.* **41**(10), 1120–1130 (2023).
53. Huang, L. et al. Federated learning for multicenter lung cancer prediction. *IEEE J. Biomed. Health Inform.* **27**(2), 789–800 (2023).
54. Vickers, A. J., Van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn. Progn. Res.* **3**, 18 (2019).
55. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**(1), 56–67 (2020).
56. Bates, D. W. et al. Evaluating the clinical impact of explainable AI in diagnostic decision support. *NPJ Digit. Med.* **5**(1), 1–9 (2022).
57. Strauss, G. M. et al. Adjuvant paclitaxel plus carboplatin in resected stage IB non-small cell lung cancer: CALGB 9633. *J. Clin. Oncol.* **26**(31), 5043–5051 (2008).
58. Amin MB, et al. AJCC Cancer staging manual (8th Edition) Springer (2017).
59. Travis, W. D. et al. The 2015 World health organization classification of lung tumors. *J. Thorac. Oncol.* **10**(9), 1243–1260 (2015).
60. Collins, G. S. et al. TRIPOD+AI: Updated guidance for reporting clinical prediction models using regression or machine learning. *BMJ* **385**, e078378 (2024).
61. Peng, Z. et al. Application of prediction model based on CT radiomics in prognosis of patients with non-small cell lung cancer. *BMC Cancer* **25**, 1273 (2025).
62. Lee, S. et al. External validation of predictive models for lung cancer outcomes: A multi-center study. *Ann. Oncol.* **32**(5), 678–685 (2021).
63. Aerts, H. J. W. L. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).

64. Zwanenburg, A. et al. The Image Biomarker Standardisation Initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**(2), 328–338 (2020).
65. Baghirzada, L. et al. Incorporating time-varying covariates in clinical prediction models: A review. *Stat. Methods Med. Res.* **31**(4), 635–651 (2022).

Author contributions

Conception and design: Pu HJ, Jun F; Acquisition, analysis, or interpretation of data: Guo X, Pu HJ, Du YJ; Drafting of the manuscript: Guo X, Tang XQ, Chen J; Critical revision of the manuscript for important intellectual content: Guo X, Xu TT, Han R; Statistical analysis: Guo X, Pu HJ, Liu SY, Li SQ; Administrative, technical, or material support: Jun F, Luo Y; Study supervision: Guo X, Pu HJ, Jun F; Final approval of the manuscript: all authors.

Funding

This study was supported by the Yunnan Provincial Department of Science and Technology Research Fund Project (202201AY070001-165), the Yunnan Medical Association Imaging Research Fund Project (YN-2022-S-01), and The Zheng jiasheng Expert Workstation Project (2018IC115).

Declarations

Competing interests

The authors declare no competing interests.

Ethics

This study was approved by the Ethics Committee of Yunnan Oncology Center (KY2019141) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-026-37113-w>.

Correspondence and requests for materials should be addressed to F.J. or H.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026