scientific reports

Check for updates

OPEN

Development and validation of a machine learning model to predict early recurrence after surgery in NSCLC patients

Guo-Yong Lin[1,2], Run-Nan Chen[1,2], Shun Wu[1,2], Zhi-Sen Gao[1,2], Xiang-Qiong Guo[1,2], Xiao-Hong Zheng[1,2] & Shu-Zhen Chen[1,2]✉

To develop and validate a machine learning (ML) model for predicting early recurrence (ER) within two years post-surgery in non-small cell lung cancer (NSCLC) patients. This multicenter cohort study included 3,171 NSCLC patients who underwent radical surgery between January 2015 and January 2021. Patients were randomly allocated to training and testing cohorts, with nine machine learning algorithms employed to construct prediction models for early recurrence, and a stacking method utilized to combine the three best-performing models. Furthermore, external validation was performed from a single institution (n = 619). Model performance was evaluated using various metrics, including Area Under the Receiver Operating Characteristic Curve (AUC). SHapley Additive exPlanations (SHAP) methodology was used to interpret predictions. Among the cohort, 553 patients (17.4%) experienced ER, with common recurrence sites including the lung (35.2%), brain (24.1%), and bone (18.5%). Significant predictors identified included pathological T stage (pT3-4: 77% vs. 50%, p < 0.001), pathological N stage (pN1-2: 51% vs. 31%, p < 0.001), and tumor differentiation grade (poorly differentiated: 68% vs. 54%, p < 0.001). The stacking model achieved superior predictive performance (AUC = 0.81, accuracy = 0.83, Brier score = 0.03), outperforming individual ML models, whose AUC values ranged from 0.72 to 0.79. SHAP analysis revealed pT stage, maximum tumor diameter, and tumor markers as key determinants of ER risk. An online computing platform (https://nsclc-risk.shinyapps.io/NSCLC_early_recurrence/) for this stacking model is publicly available and free-to-use by doctors and patients. This study developed an interpretable ML model with high predictive performance for ER in post-operative NSCLC patients. The model, based on readily available clinical data, offers a valuable tool for personalized treatment decisions and follow-up strategies. Future prospective studies across multiple centers are needed to further validate the model's generalizability and accuracy.

Keywords Non-small cell lung cancer (NSCLC), Early recurrence prediction, Machine learning, Stacking model, Postoperative prognosis

Lung cancer remains the most frequently diagnosed malignancy worldwide, with its incidence continuing to rise. Recent epidemiological data indicate approximately 2.5 million new cases and 1.8 million lung cancer-related deaths annually on a global scale[1]. Despite surgical resection being the cornerstone of curative treatment for non-small cell lung cancer (NSCLC), postoperative survival outcomes remain suboptimal due to the substantial risk of recurrence[2]. Early recurrence (ER), particularly within two years post-operation, is strongly associated with poor prognosis, significantly diminishing both survival rates and quality of life[3].

The TNM staging system is the primary factor in predicting prognosis and determining treatment options for NSCLC patients[4]. However, even patients with the same TNM stage may have different prognoses due to tumor heterogeneity[5]. Additional prognostic factors beyond staging alone play a crucial role. These factors include patient-related variables such as age, sex, body mass index (BMI), and comorbid conditions, as well as surgical factors like the extent of resection, type of surgical procedure, and postoperative complications[6–8]. Furthermore, the management of postoperative infections, inflammatory responses, and nutritional status have been identified

[1]Department of Respiratory and Critical Illness Medicine, The First Hospital of Putian, Putian City 351100, Fujian Province, China. [2]Department of Respiratory and Critical Illness Medicine, The School of Clinical Medicine, Fujian Medical University, Putian City 351100, Fujian Province, China. ✉email: 13959551815@163.com

as contributing to recurrence risk[9]. Accurately predicting ER is crucial for optimizing postoperative care, allowing clinicians to deliver timely and tailored interventions that can enhance patient outcomes.

As a core branch of artificial intelligence, machine learning fundamentally transforms oncological clinical practice into a core branch of AI. It can automatically extract key features from large volumes of heterogeneous medical data, identify complex patterns, and make valuable decisions with minimal human intervention[10]. This capability holds immense promise for early detection of neoplasms, precision medicine, and health surveillance[11]. For instance, Famularo et al. developed a machine learning model that effectively guided the treatment allocation for hepatocellular carcinoma recurrence following surgical intervention[12]. The model proposes optimal treatment strategies based on patient-specific characteristics and tumor burden, thereby significantly enhancing the projected survival rates. Machine-learning algorithms can elucidate patterns in complex medical data that may elude human perception and offer novel insights into molecular states, prognostic prediction, and treatment sensitivity assessment[13,14].

This multicenter cohort study aims to develop and validate a robust ML-based model for predicting the risk of postoperative recurrence within two years in patients with NSCLC. By utilizing standardized and comprehensive datasets from multiple centers, this study seeks to enhance the accuracy of recurrence risk assessment and support personalized clinical decision-making.

Methods

Patient selection

This multi-center population-based cohort study utilized data from three major tertiary regional medical centers in China (Fig. 1). Data from a separate teaching hospital were used for external validation. Ethical approval for this study was obtained from the Ethics Committee of the First Hospital of Putian (Ethics approval ID: 2022-004). The study was conducted in accordance with the Declaration of Helsinki, and the need for informed consent was waived by the committee due to the retrospective analysis of anonymized data. All patients diagnosed with NSCLC who underwent surgical resection between January 2015 and January 2021 were eligible for inclusion. Exclusion criteria comprised: (1) presence of metastasis at other sites prior to surgery; (2) mortality within 30 days post-surgery; and (3) insufficient clinical follow-up data. The primary endpoint was defined as ER, occurring within two years after resection. This study adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline[15].

Data collection

All clinical, pathological, and laboratory parameters were collected and reviewed from patient records. Prior to data collection, all relevant personnel at participating centers received training on data extraction forms. To ensure data relevance and consistency, only laboratory examinations conducted within 14 days before definitive surgery were considered. To standardize laboratory measurements across different centers, units for all variables were normalized. Extreme outliers were flagged and reassessed by the principal investigator or designated physicians to verify their validity and exclude data entry errors. Additionally, data distribution was adjusted to address imbalances between clinical categories. TNM stage was determined in accordance with the 8th edition International Association for the Study of Lung Cancer (IASLC) lung cancer staging system[16]. All patients had follow-up data for a minimum of 24 months to ensure comprehensive and accurate outcomes. The diagnosis

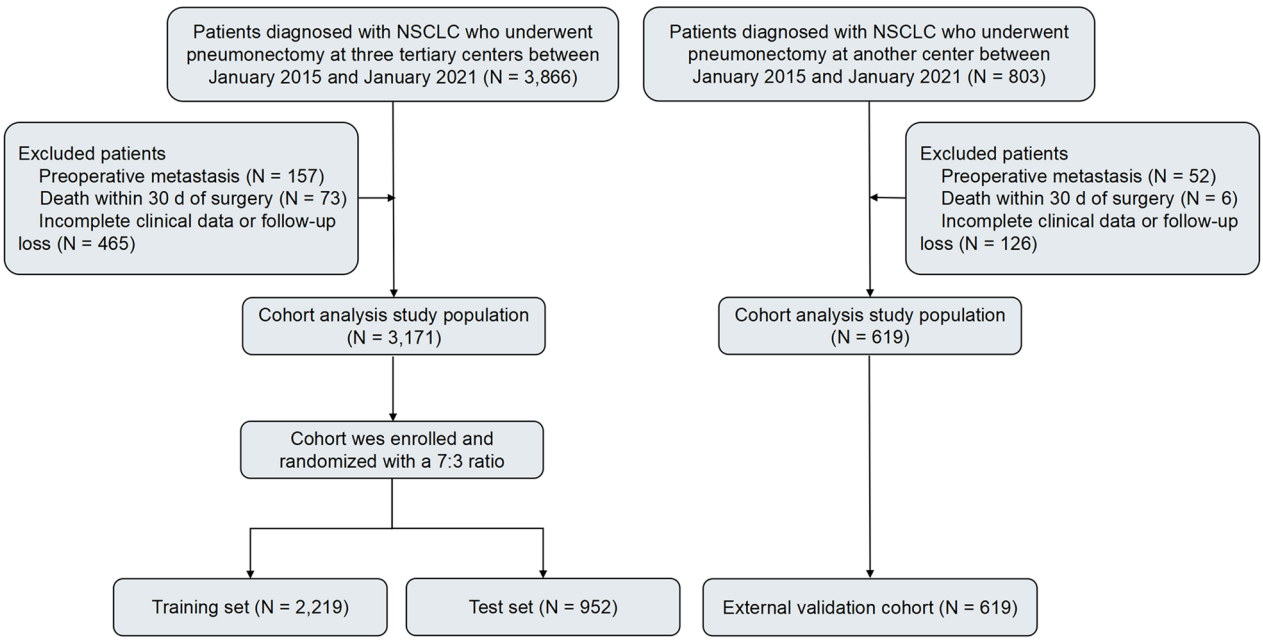

Fig. 1. Study flowchart: patient selection and cohort assignment.

of NSCLC recurrence primarily relied on imaging examinations and pathological assessment, particularly CT scans and biopsy.

Variable selection

Predictor variables were selected based on a comprehensive review of the literature and their clinical relevance. These variables are routinely assessed in clinical practice and have a well-established association with patient survival. To ensure data quality, variables with more than 20% missing values were excluded from further analysis. The overall rate of missing data was 1.17% (eFigure 1). For missing clinical data, we employed multiple imputation using the random forest method (50 cycles) from the "mice" package. The imputation quality was assessed through convergence plots of multiple imputation results and density plots of imputed data (eFigure 2). Correlations between variables were quantified using Spearman correlation coefficients (r) (eFigure 3), with $|r| \geq 0.9$ indicating strong correlations (eFigure 4). Strongly correlated variables were considered redundant and selectively removed from the analysis. Additionally, we calculated the following immune-nutritional scores based on laboratory data at admission: prognostic nutritional index (PNI)[17], lung Immune prognostic index

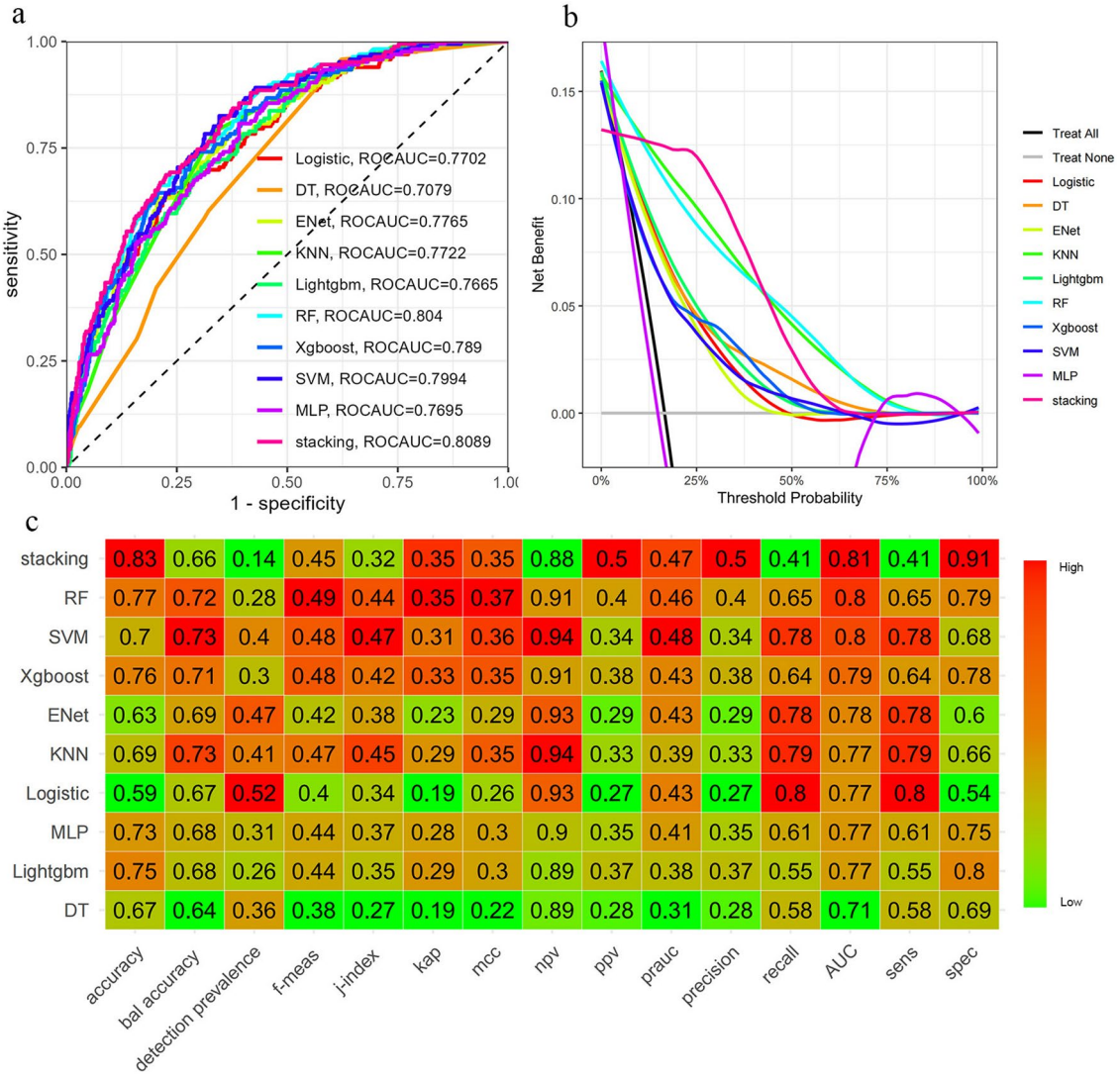

Fig. 2. Performance comparison of 10 machine learning models in the testing cohort. (**a**) ROC curves for each model in the testing cohort; (**b**) decision curve analysis (DCA) curves for each model in the testing cohort; (**c**) heatmap comparing performance metrics of each model in the testing cohort. Abbreviations: Logistic: Logistic regression model; MLP: Multilayer perceptron model; RF: Random forest model; SVM: Support vector machine model; Xgboost: Extreme gradient boosting model; ENet: Elastic net model; DT: Decision tree model; KNN: K-nearest neighbors model; Lightgbm: Light gradient boosting machine model; Stacking: Stacking ensemble model; F-means: F1 score; J-index: Youden index; Kap: Kappa coefficient; MCC: Matthews correlation coefficient; NPV: Negative Predictive Value; PPV: Positive Predictive Value; PRAUC: Precision-recall area under the curve; AUC: Area under the receiver operating characteristic curve; Sens: Sensitivity; Spec: Specificity.

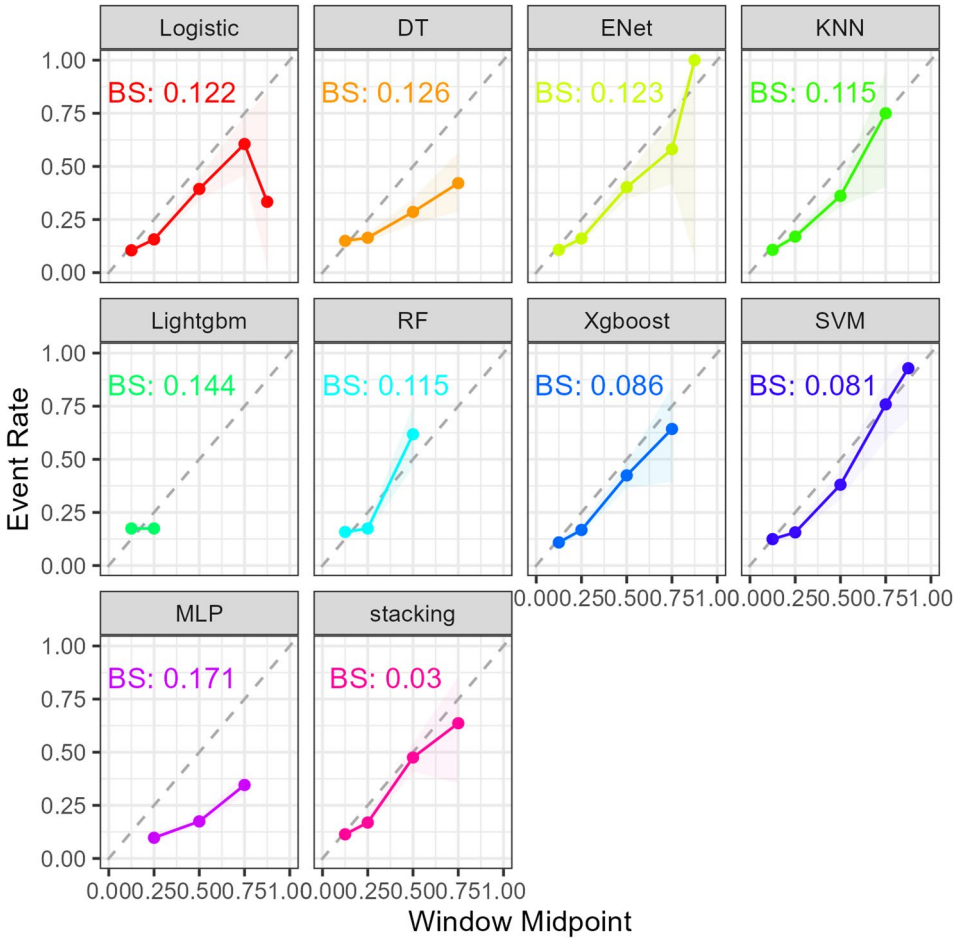

Fig. 3. Calibration plots (reliability curve, dashed line represents perfectly calibrated) comparing the predictive performance of ten models using testing data.

(LIPI)[18], and systemic immune-inflammation index (SII)[19]. The PNI which is an easily calculated nutritional index, is significantly associated with patient outcomes in various solid malignancies[17]. LIPI is calculated based on the derived neutrophil-to-lymphocyte ratio (NLR) and lactate dehydrogenase levels, reflecting the patient's systemic inflammatory status and tumor burden[18]. SII is a composite inflammatory marker based on peripheral NLR and platelet count, reflecting the integrated level of systemic inflammation and immune status[19].

Model construction and validation

The study cohort was stratified into a training cohort comprising 70% of the data and a testing cohort comprising the remaining 30%. Model development incorporated nine machine learning algorithms and one stacking algorithm: decision tree (DT), random forest (RF), extreme gradient boosting (XGBoost), LASSO-ridge regression elastic net (ENET), support vector machine (SVM), artificial neural network (ANN), light gradient boosting machine (LightGBM), k-nearest neighbors (KNN), and logistic regression (LR). Each classification algorithm underwent hyperparameter tuning through 5-fold cross-validation internally. After selecting the optimal hyperparameters, the model was retrained on the complete training subset to finalize the weighting and generate a locked model. These locked models were then assessed on the internal validation cohort.

Model performance was evaluated using multiple metrics, including the Area Under the Receiver Operating Characteristic Curve (AUC), sensitivity, specificity, positive predictive value, negative predictive value, accuracy, and F1 score. Calibration plots and the Brier score (BS) were used to assess the accuracy of predicted probabilities. In the model ensemble stage, we employed a stacking method to combine the three best-performing models (RF, XGBoost, and SVM) to construct the final ensemble model. During the stacking process, logistic regression was used as a meta-learner to linearly weight the predictions from each base model, thereby generating the final prediction probabilities.

Statistical methods

Data analysis was performed using R statistical computing software version 4.4.1 (R Foundation for Statistical Computing, Vienna, Austria). Patient data were presented as continuous or categorical variables. For normally distributed continuous variables, data were described as mean ± standard deviation and compared using t-tests. For continuous variables not meeting normality assumptions, Mann-Whitney U tests were employed, with

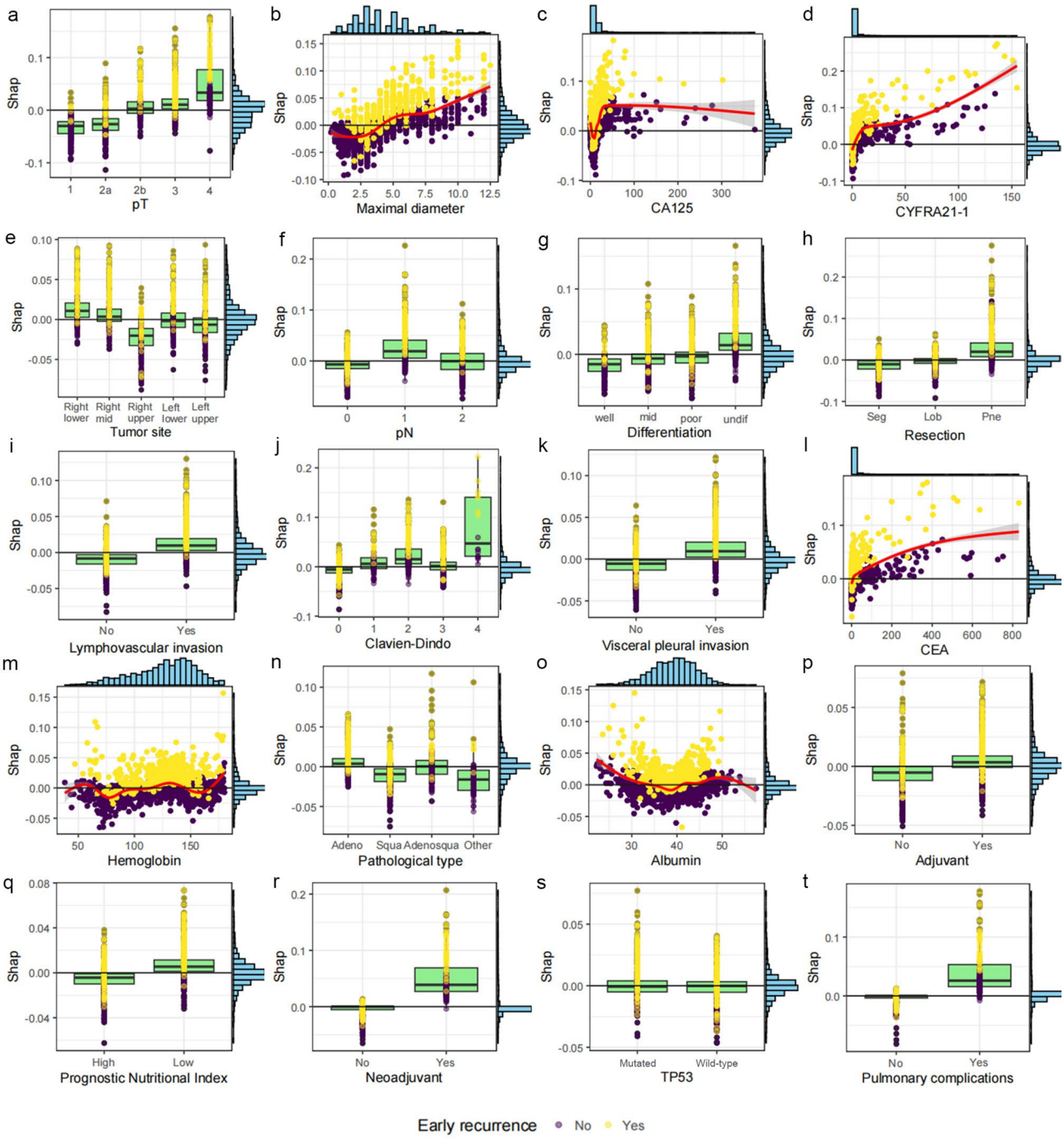

Fig. 4. SHAP dependence plot for early recurrence prediction.

results presented as median (interquartile range). Categorical data were presented as numbers and frequencies and compared using chi-square tests or Fisher's exact tests where appropriate. All statistical tests were two-sided, with P values < 0.05 considered statistically significant.

Model training was conducted using the tidymodels workflow framework (version 1.2.0)[20], with multiple models undergoing tuning and performance evaluation. Variable importance analysis was performed with the DALEX package (version 2.4.3) to assess each feature's contribution to model performance. Cross-validation strategies were employed throughout to ensure result robustness. Additionally, the SHapley Additive exPlanations (SHAP) methodology was applied to rank input feature importance and enhance interpretability at both global and local levels using the iBreakDown package (version 2.1.2).

Results

Patient characteristics

During the study period, 3,866 patients who underwent surgical treatment for NSCLC were reviewed. Exclusions included 157 patients with pre-existing metastases, 73 who died within 30 days post-surgery, and 465 with incomplete clinical or follow-up data (Fig. 1). Ultimately, 3,171 patients who underwent radical surgery for NSCLC were included, with a median age of 62 years (interquartile range: 54–68 years). During follow-up, 553 patients (17.4%) experienced recurrence within two years post-surgery. Among these, 126 (22.9%) had local recurrences, while 366 (66.2%) developed distant metastases, commonly affecting the brain (13.9%), mediastinum (16.6%), adrenal glands (7.2%), liver (11.4%), and bone (5.6%). Additionally, 106 patients (19.2%) presented with both local and distant recurrence, while data were missing for 13 cases (2.4%) (Table 1). The mean time to ER for all patients who experienced relapse was 379.7 ± 173.4 days.

Comparison of clinical and tumor characteristics between ER and Non-ER groups

Patients in the ER group were slightly older than those in the non-ER group (62 vs. 61 years, $p = 0.030$), though the gender distribution did not differ significantly (male proportion: 77% vs. 74%, $p = 0.13$). In terms of body mass index (BMI), the ER group exhibited marginally lower values compared to the non-ER group (21.5 vs. 21.9, $p = 0.007$). No statistically significant differences were observed in smoking history (43% vs. 40%, $p = 0.2$) or rates of neoadjuvant therapy (27% vs. 25%, $p = 0.2$) between the two groups. With respect to tumor characteristics, patients in the ER group were more likely to present with advanced pathological tumor stage (pT3-4: 77% vs. 50%, $p < 0.001$) and advanced nodal involvement (pN1-2: 51% vs. 31%, $p < 0.001$). Additionally, the ER group demonstrated a significantly higher prevalence of poorly differentiated tumors, with poorly differentiated and undifferentiated/anaplastic tumors collectively comprising 68% of cases, compared to 54% in the non-ER group ($p < 0.001$). Additional clinical and pathological characteristics are detailed in Table 2.

Factor importance rankings for ER using the Boruta algorithm

Through predetermined stratified random sampling, 70% ($n = 2,219$) of patients were allocated to the training cohort and 30% ($n = 952$) to the testing cohort. Both cohorts maintained identical ER rates of 17% and demonstrated overall balance in baseline characteristics with no statistically significant differences ($p > 0.05$). The absence of significant differences in features indicated that both groups were drawn from the same population, ensuring good comparability (eTable 1). The external validation cohort ($n = 619$) exhibited comparable early recurrence rate (18%, $p = 0.8$), but showed statistically significant differences in age, BMI, smoking history, neoadjuvant therapy rate, tumor stage, tumor site, pathological type, and several inflammation- and nutrition-related indexes including LIPI, PNI, and SII (all $p < 0.05$), indicating a moderately heterogeneous validation population.

To enhance the predictive accuracy of our machine learning model, we employed the Boruta algorithm in R[21] for feature selection in the training cohort. This approach assessed the importance of 42 features, encompassing gene mutations, pathological characteristics, and clinical parameters, through 500 simulation runs. Variables categorized in the green zone were identified as crucial factors exerting significant influence on the model. Notably, pT stage emerged as a key predictor for ER in post-surgical NSCLC patients (eFigure 5). Variables in the yellow zone were classified as tentative factors, potentially associated with adverse outcomes, while those in the red zone were deemed non-significant.

Validation and calibration of the ML prediction model

In the model training process, the positive class represented NSCLC with recurrence within two years, while the negative class represented absence of recurrence. Following variable selection, the input data for model training comprised 20 indicators closely associated with ER. Utilizing these features, we developed a total of 10 machine learning models, including a stacking model. Upon completion of model training, evaluation was conducted on the independent testing dataset, which was not included in the training dataset. The findings of this study indicate that the stacking model exhibited significantly higher AUC values (Fig. 2a). Decision curve analysis (DCA), a direct method for evaluating the clinical utility of disease diagnostic models, was employed. The DCA

Type of early recurrence	Overall ($N = 553$)
Locoregional recurrence, No. (%)	126 (22.9%)
Distant metastasis, No. (%)	366 (66.2%)
Brain	77 (13.9%)
Mediastinum	92 (16.6%)
Adrenal glands	40 (7.2%)
Liver	63 (11.4%)
Bone	31 (5.6%)
Other sites	17 (3.1%)
Locoregional and distant recurrence, No. (%)	106 (19.2%)
Missing data, No. (%)	13 (2.4%)
Time, mean SD, days	379.7 ± 173.4

Table 1. Pattern of initial early recurrence.

Variable	Total $N=3,171$[1]	Non early recurrence $N=2,618$[1]	Early recurrence $N=553$[1]	p-value[2]
Age, median (IQR), years	62 (54, 68)	61 (54, 68)	62 (54, 69)	0.030
Sex, n (%)				0.13
Female	821 (26%)	692 (26%)	129 (23%)	
Male	2,350 (74%)	1,926 (74%)	424 (77%)	
BMI, median (IQR), kg/m^2	21.8 (20.0, 24.2)	21.9 (20.0, 24.2)	21.5 (19.6, 23.9)	0.007
Smoking history, n (%)	1,280 (40%)	1,043 (40%)	237 (43%)	0.2
Neoadjuvant Therapy, n (%)	802 (25%)	651 (25%)	151 (27%)	0.2
Charlson Comorbidity Index, n (%)				>0.9
≤2	3,006 (95%)	2,482 (95%)	524 (95%)	
>2	165 (5.2%)	136 (5.2%)	29 (5.2%)	
COPD, n (%)	280 (8.8%)	233 (8.9%)	47 (8.5%)	0.8
pT, n (%)				<0.001
1	573 (18%)	560 (21%)	13 (2.4%)	
2	869 (27%)	759 (29%)	110 (20%)	
3	1,537 (48%)	1,181 (45%)	356 (64%)	
4	192 (6.1%)	118 (4.5%)	74 (13%)	
pN, n (%)				<0.001
0	2,079 (66%)	1,810 (69%)	269 (49%)	
1	604 (19%)	430 (16%)	174 (31%)	
2	488 (15%)	378 (14%)	110 (20%)	
Tumor site, n (%)				<0.001
Right lower lobe	604 (19%)	475 (18%)	129 (23%)	
Right middle lobe	745 (23%)	595 (23%)	150 (27%)	
Right upper lobe	638 (20%)	556 (21%)	82 (15%)	
Left lower lobe	497 (16%)	412 (16%)	85 (15%)	
Left upper lobe	687 (22%)	580 (22%)	107 (19%)	
Degree of differentiation, n (%)				<0.001
Well differentiated	361 (11%)	326 (12%)	35 (6.3%)	
Moderately differentiated	1,011 (32%)	869 (33%)	142 (26%)	
Poorly differentiated	1,193 (38%)	974 (37%)	219 (40%)	
Undifferentiated/anaplastic	606 (19%)	449 (17%)	157 (28%)	
Pathological type, n (%)				0.4
Adenocarcinoma	1,733 (55%)	1,414 (54%)	319 (58%)	
Squamous	1,099 (35%)	922 (35%)	177 (32%)	
Adenosquamous	252 (7.9%)	208 (7.9%)	44 (8.0%)	
Other	87 (2.7%)	74 (2.8%)	13 (2.4%)	
LIPI, n (%)				0.9
Low	2,440 (77%)	2,013 (77%)	427 (77%)	
Mid	658 (21%)	546 (21%)	112 (20%)	
High	73 (2.3%)	59 (2.3%)	14 (2.5%)	
PNI, n (%)				<0.001
Low	1,779 (56%)	1,512 (58%)	267 (48%)	
High	1,392 (44%)	1,106 (42%)	286 (52%)	
SII, n (%)				<0.001
Low	1,063 (34%)	914 (35%)	149 (27%)	
High	2,108 (66%)	1,704 (65%)	404 (73%)	
[1]Median (Q1, Q3); n (%)				
[2]Wilcoxon rank sum test; Pearson's Chi-squared test				

Table 2. Overall summary of variables. *BMI* Body mass index, *COPD* Chronic obstructive pulmonary disease, *LIPI* Lung immune prognostic index, *PNI* Prognostic nutritional index, *SII* Systemic immune-inflammation index.

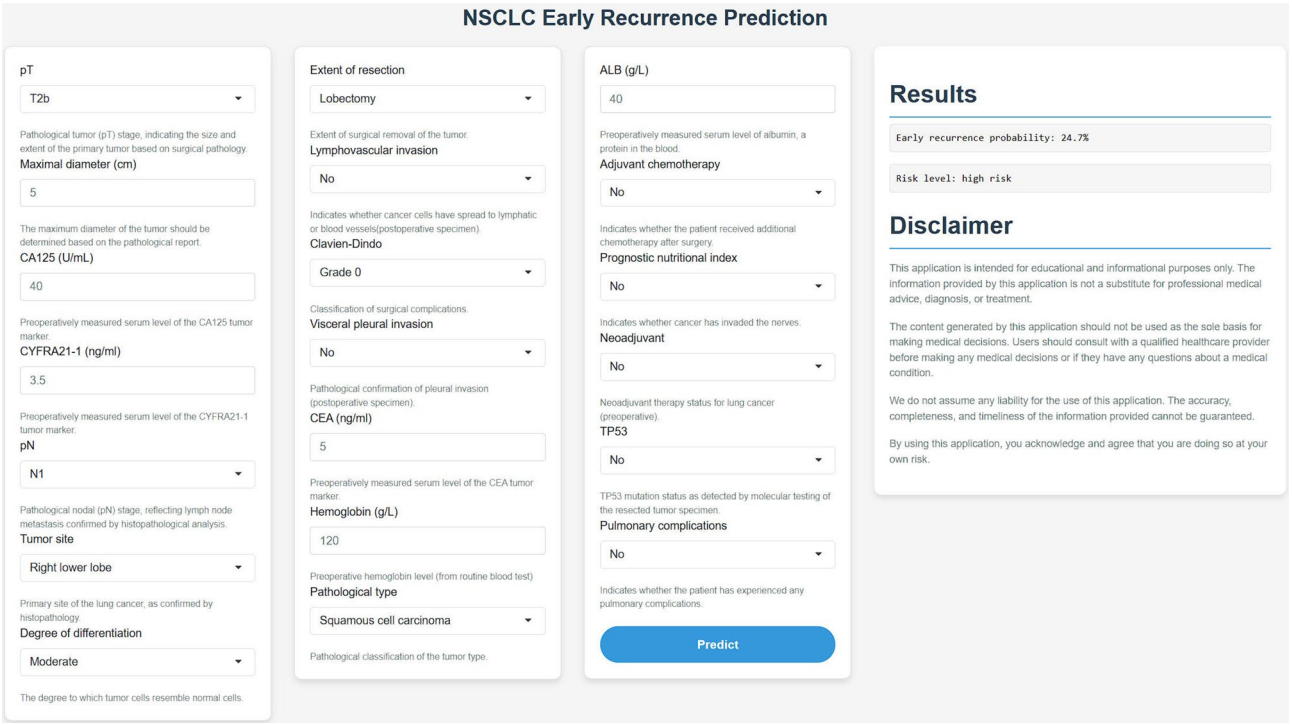

Fig. 5. Online computing platform presentation of the optimal stacking model.

curves depicted in Fig. 2b further demonstrated that the RF model possessed the highest clinical utility. Further examination of the data in the testing cohort revealed that the stacking method demonstrated superior predictive capabilities across multiple metrics, achieving higher accuracy (0.83), AUC (0.81), and negative predictive value (NPV = 0.88) compared to individual models (Fig. 2c). However, the relatively low detection prevalence (0.14) suggested potential limitations in identifying positive cases (Fig. 2c). The performance of the models was evaluated on the external validation cohort. The stacking model exhibited the highest AUC value of 0.8043. The decision curve analysis indicated the stacking model's superior clinical utility, with the highest net benefit across various threshold probabilities. Additionally, the stacking model achieved the highest accuracy (0.84), AUC (0.80), and sensitivity (0.94), while maintaining a balanced precision and recall. However, a relatively low detection prevalence (0.11) was observed (eFigure 6).

The calibration plot illustrated the degree of alignment between predicted probabilities and actual events (Fig. 3). Overall, there was strong concordance between actual ER outcomes and predictions by the stacking model for all NSCLC patients in our database, with most points falling nearly directly on the 45-degree line. The stacking model achieved the lowest Brier score (BS = 0.03), indicating optimal calibration performance, while MLP (BS = 0.171) and LightGBM (BS = 0.144) showed relatively poorer performance. The calibration plot on the external validation cohort demonstrated that the stacking model maintained strong calibration, with most points lying close to the 45-degree line, indicating good alignment between predicted probabilities and observed outcomes. The stacking model achieved the lowest Brier score (BS = 0.122) among all models, while other models such as MLP (BS = 0.185) and LightGBM (BS = 0.142) exhibited relatively poorer calibration performance (eFigure 7).

Model interpretations with SHAP

To ensure a comprehensive understanding of the selected variables, we employed the SHAP algorithm to elucidate their predictive importance within the optimal stacking model for ER. This approach provided both global feature-level and local individual-level interpretations. In the global analysis, the variables pT stage, maximal tumor diameter, CA125, and CYFRA21-1 were identified as the most significant predictors, as evidenced by their elevated SHAP values (eFigure 8). The magnitude of SHAP value variation showed a positive association with increasing pT stages (Fig. 4a). Variable-specific SHAP dependence plots demonstrated progressively increased predictive contributions to ER risk with higher maximum tumor diameter measurements (Fig. 4b), elevated serum concentrations of CA125 (Fig. 4c), and CYFRA21-1 (Fig. 4d). Tumor site emerged as a significant factor, with middle and lower lobe tumors demonstrating higher SHAP values, which correlated with elevated recurrence risk (Fig. 4e). Higher pN stages were significantly associated with increased model prediction probabilities for ER (Fig. 4f). Tumor differentiation also emerged as a critical determinant, with poorly differentiated tumors demonstrating higher SHAP values (Fig. 4g). The model assigned greater predictive weights to extended resection margins (Fig. 4h), lymphovascular invasion (Fig. 4i), and postoperative complications, particularly Clavien-Dindo grade IV complications (Fig. 4j). Visceral pleural invasion was consistently associated with enhanced ER prediction (Fig. 4k). CEA levels demonstrated progressively increased predictive contributions with elevated

concentrations (Fig. 4l). Hemoglobin levels showed variable SHAP contributions (Fig. 4m), while pathological type (Fig. 4n) and albumin levels (Fig. 4o) demonstrated distinct patterns in their predictive importance. Both adjuvant (Fig. 4p) and lower PNI (Fig. 4q) showed comparable effect magnitudes, with neoadjuvant therapies (Fig. 4r) also demonstrating positive associations with increased SHAP values for ER prediction. Conversely, variables such as TP53 status (Fig. 4s) and pulmonary complications (Fig. 4t) exhibited lower SHAP values.

At the local level, our analysis demonstrated how individualized input data were integrated to generate predictions for specific cases. For instance, a force plot illustrated a patient with a 57.1% predicted probability of ER within two years post-radical surgery (eFigure 9a), while another patient was successfully classified into the non-ER group with only a 19.3% probability (eFigure 9b). Factors such as lymphovascular invasion and pleural invasion increased the predicted probability of ER, whereas tumor location in the upper lobe or well-differentiated histology reduced it. This comprehensive interpretative approach enhances the model's transparency and clinical applicability, facilitating more informed decision-making in patient management and follow-up strategies. To further support clinical implementation, we developed an online computational platform (https://nsclc-risk.shinyapps.io/NSCLC_early_recurrence/) based on the optimal stacking model (Fig. 5). This platform facilitates convenient online calculations for both physicians and patients.

Discussion

This study aimed to investigate and compare ten ML prediction models for ER analysis in NSCLC populations. We employed the Boruta algorithm to identify critical variables associated with ER in NSCLC and subsequently developed a robust classification algorithm for patients undergoing radical lung resection surgery. Utilizing SHAP methodology, we provided individualized risk assessments based on specific patient characteristics, enabling the identification of high-risk patients for ER. This approach has the potential to assist clinicians in formulating more aggressive and rational treatment strategies, potentially reducing mortality risks.

The majority of studies on risk factors associated with NSCLC prognosis have focused on survival outcomes[22,23]. Our research addresses the gap in short-term recurrence prediction and management, providing data support for more comprehensive disease management. Compared to traditional single biomarker approaches, our developed final model demonstrates superior capability in predicting early postoperative recurrence in NSCLC[24]. The complex biological characteristics of NSCLC and its multifaceted influencing factors have made single analyte approaches challenging to directly apply in clinical practice[25]. Several previously published studies have also developed machine learning models to predict recurrence in NSCLC patients. Kwon et al. developed a CT-based deep learning model for segmentectomy in clinical stage IA NSCLC, which demonstrated robust prognostic performance, with an AUC of 0.86 for 2-year freedom from recurrence. This model outperformed the Japan Clinical Oncology Group (JCOG) criteria, offering a more accurate risk stratification tool for clinical decision-making[26]. Hindocha et al. developed and validated a machine learning model for predicting recurrence after curative-intent radiotherapy for NSCLC. The model achieved an AUC of 0.687 in the validation set and 0.722 in the external test set[27]. However, the model relies on specific clinical data, which limits its generalizability.

Recurrence timing exhibits multi-peak characteristics, with recurrence risk demonstrating dynamic changes across different postoperative time periods. The 6–9 month and 18–24 month postoperative intervals represent high-risk periods for NSCLC recurrence[2,28]. Accurate prediction of recurrence within two years post-surgery is particularly crucial. ML techniques offer a powerful computational approach for handling complex and voluminous data, processing highly variable datasets and understanding intricate relationships among variables in a flexible and trainable manner. The combination of extensive patient data with sophisticated ML algorithms facilitates innovation in clinical prediction models. After evaluating the performance of these models across various metrics, the stacking model demonstrated optimal AUC values and calibration performance.

Stacking is an ensemble learning method that enhances predictive performance by combining the strengths of multiple base models[29]. This technique has demonstrated exceptional predictive value in the medical field[30–32]. In our study, we employed a stacking approach to develop the final model, incorporating 20 features. The majority of these features can be readily obtained or evaluated during the patient's first follow-up visit post-discharge, making this model a promising tool for guiding postoperative follow-up care. Furthermore, we collected data from a large cohort of patients across various institutions, potentially enhancing the generalizability and robustness of our findings. This multi-institutional approach adds credibility to our results and increases the likelihood of the model's applicability across diverse clinical settings. By leveraging the power of ML and stacking techniques, our model not only addresses the complexities inherent in NSCLC recurrence prediction but also offers a practical solution that can be integrated into existing clinical workflows. The predictions from this model can assist clinicians in developing personalized follow-up and treatment plans based on the risk of recurrence. For patients with a higher recurrence risk, more aggressive treatment regimens or more frequent follow-up may be required.

ML techniques are often described as a "black box," offering little explanation for how predictions are derived. This can lead to clinicians' reluctance to use them, as they are unwilling to make medical decisions based on opaque information. Our study addresses this limitation by utilizing SHAP methodology to elucidate the ML model's decision-making process. This approach revealed several potentially important features, providing direction for future research. As anticipated, pathological factors remain the primary predictors of recurrence[24,33,34]. Key variables in this category include pleural invasion, lymphovascular invasion, pathological type, and differentiation grade. Furthermore, lung cancer-associated tumor markers have demonstrated significant predictive value. Clinical variables, such as postoperative complications, have also emerged as crucial predictors of ER. This multifaceted approach to prediction, incorporating both pathological and clinical factors, provides a more comprehensive assessment of recurrence risk in NSCLC patients.

Recurrence patterns among different NSCLC pathological types remain controversial. Squamous cell carcinoma exhibits a higher risk of local recurrence after stereotactic body radiation therapy compared to

adenocarcinoma[35], while adenocarcinoma is more prone to lymph node metastasis[36] and distant metastasis[37]. Our study supports the notion that tumors with adenocarcinoma components have a higher tendency for ER. The occurrence of complications, particularly infectious complications[38], has been found to promote tumor metastasis[38]. Postoperative complications may facilitate cancer metastasis primarily due to alterations in the immune system and inflammatory responses. Specifically, postoperative infectious complications can trigger systemic inflammatory response syndrome, potentially transitioning to an immunosuppressive state that may facilitate cancer progression and metastasis[39–41].

Despite the potential for richer information provided by incorporating additional features such as radiomics, genomics, and metabolomics, an excess of features may limit the model's clinical applicability. Moreover, including non-causal features could potentially reduce prediction accuracy[42,43]. The SHAP method was employed to assist in analyzing feature contributions, enhancing model transparency and credibility. As a simple and convenient ML prediction tool, the model developed in this study shows promise in facilitating the clinical management of NSCLC patients.

However, the findings of this study should be interpreted within the context of certain limitations. Firstly, this is a multi-center retrospective study with inherent data integrity issues. Secondly, the model was constructed based on a Chinese population, and its applicability to other global populations remains uncertain. Nevertheless, the inclusion of large-scale data from multiple centers, encompassing NSCLC patients at various stages, lends some support to the generalizability of the results. Future research should focus on validating the model using prospective samples from additional centers to encompass greater sample diversity. In predicting the risk of recurrence, sensitivity is of paramount importance, as it reflects the model's ability to correctly identify patients at risk. While the stacked model demonstrates high specificity, its lower sensitivity may limit its utility in this context, and further improvements may be needed to enhance sensitivity. Furthermore, the current model can only predict the occurrence of ER, not the specific timing. Further investigations are needed to explore the prediction of recurrence timing, particularly within the 2-year and 2–5 year time windows.

Conclusion

In summary, we have successfully developed an interpretable ML prediction model based on readily available clinical data to predict ER in post-operative NSCLC patients. The final stacking model demonstrated superior predictive performance, with pT stage, maximum tumor diameter, tumor markers, and pN stage identified as the main contributing factors. This study provides clinicians with an effective tool for making personalized treatment decisions. However, to further validate the model's generalizability and accuracy, future prospective clinical studies across multiple centers are necessary to provide more robust supporting evidence.

Data availability

The datasets and code used and/or analyzed during the current study are available from the corresponding author upon reasonable request. Additionally, the code for model construction and analysis can be accessed at the following GitHub repository: https://github.com/13959551815/NSCLC-Early-Recurrence-Prediction.

Received: 17 March 2025; Accepted: 24 October 2025
Published online: 25 November 2025

References

1. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J. Clin.* **74** (3), 229–263 (2024).
2. Nicole, F. N. et al. Hazards of recurrence, second primary, or other tumor at ten years after surgery for non-small-cell lung cancer. *Clin. Lung Cancer* **21**(4), 333 (2020).
3. Katsuya, W. et al. Postoperative follow-up strategy based on recurrence dynamics for non-small-cell lung cancer. *Eur. J. Cardiothorac. Surg.* **49**(6), 1624 (2016).
4. Jae Kwang, Y. et al. Various recurrence dynamics for non-small cell lung cancer depending on pathological stage and histology after surgical resection. *Transl. Lung Cancer Res.* **11**(7), 1327 (2022).
5. Jian, W. et al. Nomogram integrating gene expression signatures with clinicopathological features to predict survival in operable NSCLC: a pooled analysis of 2164 patients. *J. Exp. Clin. Cancer Res.* **36**(1), 4 (2017).
6. Dariusz Adam, D. et al. Risk factors for local and distant recurrence after surgical treatment in patients with non-small-cell lung cancer. *Clin. Lung Cancer* **17**(5), e157 (2016).
7. Wong, M. L. et al. Impact of age and comorbidity on treatment of non-small cell lung cancer recurrence following complete resection: A nationally representative cohort study. *Lung Cancer*. **102**, 108–117 (2016).
8. Sepesi, B. et al. The influence of body mass index on overall survival following surgical resection of Non-Small cell lung cancer. *J. Thorac. Oncol.* **12** (8), 1280–1287 (2017).
9. Choi, H. & Hwang, W. Perioperative inflammatory response and cancer recurrence in lung cancer surgery: A narrative review. *Front. Surg.* **9**, 888630 (2022).
10. Jeremy, G. et al. How machine learning will transform biomedicine. *Cell* **181**(1), 92 (2020).
11. Kyle, S. et al. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell* **186**(8), 1772 (2023).
12. Simone, F. et al. Machine learning predictive model to guide treatment allocation for recurrent hepatocellular carcinoma after surgery. *JAMA Surg.* **158**(2), 192 (2022).
13. Yangzi, C. et al. Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer. *Nat. Commun.* **15**(1), 1657 (2024).
14. Yuming, J. et al. Predicting peritoneal recurrence and disease-free survival from CT images in gastric cancer with multitask deep learning: a retrospective study. *Lancet Digit. Health* **4**(5), e340 (2022).
15. Collins, G. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ (Clinical Res. ed)*. **350**, pg7594 (2015).

16. Rami-Porta, R. et al. The international association for the study of lung cancer lung cancer staging project: proposals for revision of the TNM stage groups in the forthcoming (Ninth) edition of the TNM classification for lung cancer. *J. Thorac. Oncology: Official Publication Int. Association Study Lung Cancer*. **19** (7), 1007–1027 (2024).
17. Buzby, G. P. Prognostic nutritional index in gastrointestinal surgery. *Am. J. Surg.* **139**(1), 160 (1980).
18. Laura, M. et al. Association of the lung immune prognostic index with immune checkpoint inhibitor outcomes in patients with advanced non-small cell lung cancer. *JAMA Oncol.* **4**(3), 351 (2018).
19. Bo, H. et al. Systemic immune-inflammation index predicts prognosis of patients after curative resection for hepatocellular carcinoma. *Clin. Cancer Res.* **20**(23), 6212 (2014).
20. Vaughan, D. C.S. *workflows: Modeling Workflows*. [cited 2025 2025-05-12];. Available from: https://workflows.tidymodels.org(2025)
21. Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36** (11), 1–13 (2010).
22. She, Y. et al. Development and validation of a deep learning model for Non-Small cell lung cancer survival. *JAMA Netw. open.* **3** (6), e205842 (2020).
23. Guo, Y. et al. Development and validation of a survival prediction model for patients with advanced non-small cell lung cancer based on LASSO regression. *Front. Immunol.* **15**, 1431150 (2024).
24. Zhang, Y. et al. Development and validation of Web-Based nomograms to precisely predict conditional risk of Site-Specific recurrence for patients with completely resected Non-small cell lung cancer: A multiinstitutional study. *Chest* **154** (3), 501–511 (2018).
25. Lucas, W. et al. Challenges in Predicting Recurrence After Resection of Node-Negative Non-Small Cell Lung Cancer. *Ann. Thorac. Surg.* **106**(5), 1460 (2018).
26. Christopher, G., Garo, H. & Misako, N. Neoadjuvant therapy in non-small cell lung cancer. *Crit. Rev. Oncol. Hematol.* **190**(0), 104080 (2023).
27. Sumeet, H. et al. A comparison of machine learning methods for predicting recurrence and death after curative-intent radiotherapy for non-small cell lung cancer: Development and validation of multivariable clinical prediction models. *EBioMedicine* **77**, 103911 (2022).
28. Romano, D. et al. *Recurrence dynamics for non-small-cell lung cancer: effect of surgery on the development of metastases. J. Thorac. Oncol.* **7**(4), 723 (2012).
29. Kevin, M. et al. Investigation of model stacking for drug sensitivity prediction. *BMC Bioinform.* **19**, 71 (2018).
30. Satish Kumar, K., Suryakanth, G. & Kanuri, G. A novel stacking technique for prediction of diabetes. *Comput. Biol. Med.* **135**, 104554 (2021).
31. Yuji, H. et al. Fabrication of functional three-dimensional tissues by stacking cell sheets in vitro. *Nat. Protoc.* **7**(5), 850 (2012).
32. Xing-Qi, Z. et al. Machine learning prediction of early recurrence in gastric cancer: a nationwide real-world study. *Ann. Surg. Oncol.* **32**, 2637 (2025).
33. John, M. et al. Nodal stage of surgically resected non-small cell lung cancer and its effect on recurrence patterns and overall survival. *Int. J. Radiat. Oncol. Biol. Phys.* **91**(4), 765 (2015).
34. Mitsuhiro, I. et al. Risk factors for local recurrence after lobectomy and lymph node dissection in patients with non-small cell lung cancer: Implications for adjuvant therapy. *Lung Cancer* **115**(0), 28 (2018).
35. Nozomi, K. et al. Comparison of recurrence patterns between adenocarcinoma and squamous cell carcinoma after stereotactic body radiotherapy for early-stage lung cancer. *Cancers (Basel)* **15**(3), 887 (2023).
36. Han-Yu, D. et al. Lung adenocarcinoma has a higher risk of lymph node metastasis than squamous cell carcinoma: a propensity score-matched analysis. *World J. Surg.* **43**(3), 955 (2018).
37. Yu, K. et al. Differences in patterns of recurrence of squamous cell carcinoma and adenocarcinoma after radiotherapy for stage III non-small cell lung cancer. *Jpn J. Radiol.* **39**(6), 611 (2021).
38. Kazuhiro, U. et al. Postoperative complications and cancer recurrence: impact on poor prognosis of lower lobe cancer. *Ann. Thorac. Surg.* **109**(6), 1750 (2020).
39. Hironori, T. et al. Potential mechanisms of tumor progression associated with postoperative infectious complications. *Cancer Metastasis Rev.* **40**(1), 285 (2021).
40. Stephen, T., Paul, M. H. & Donald, M. The impact of the type and severity of postoperative complications on long-term outcomes following surgery for colorectal cancer: a systematic review and meta-analysis. *Crit. Rev. Oncol. Hematol.* **97**(0), 168 (2015).
41. Takashi, Y. et al. Impact of postoperative complications on the long-term outcome in lung cancer surgery. *Surg. Today* **52**(9), 1254 (2022).
42. Ranjana, M. et al. Prediction of postoperative recurrence-free survival in non-small cell lung cancer by using an internationally validated gene expression model. *Clin. Cancer Res.* **17**(9), 2934 (2011).
43. Peiwen, W. et al. Application of a comprehensive model based on CT radiomics and clinical features for postoperative recurrence risk prediction in non-small cell lung cancer. *Acad. Radiol.* **31**(6), 2579 (2024).

Author contributions

GYL and RNC contributed to the concept and design of the research; XQG was responsible for the acquisition of data; SW and XHZ performed the analysis and interpretation of data; ZSG conducted the statistical analysis; GYL drafted the manuscript; GYL and RNC revised the manuscript for important intellectual content; SZC provided supervision. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-25775-x.

Correspondence and requests for materials should be addressed to S.-Z.C.

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.