scientific reports

Check for updates

OPEN

Interpretable multiclassification machine learning models in predicting the extent of small cell lung cancer metastasis

Jijing Xu[1]✉, Hai Shen[2], Ronghua Cao[1], Jiaxuan Li[1], Qiang Wang[2], Qiaoming Gao[1], Yue Liu[1], Guangyao Liu[3], Kai Xiao[4], Linghua Wu[5] & Yonggao Huang[1]✉

Small cell lung cancer (SCLC) is known for its rapid disease progression and high metastasis, significantly reducing patient survival. The current lack of clinical models that can accurately predict the metastatic stage of SCLC limits the timeliness and effectiveness of treatment strategies. Machine learning techniques offer the possibility to improve the accuracy and operationalisation of predictions by utilising complex algorithmic pattern recognition. This study aimed to develop and interpret a predictive model for SCLC metastatic staging using real-world data. The Surveillance, Epidemiology, and End Results database (SEER, 2004–2017) of the National Cancer Institute was selected as the data source for this study, and multiple machine learning algorithms were used to predict the metastatic stage of SCLC. Lasso regression was used for feature selection and model parameters were optimised by cross-validation. Model performance was assessed by area under the receiver operating characteristic curve (AUC) and other statistical metrics. To enhance the interpretability of the model, SHAP (SHapley Additive exPlanations) was used to analyse the contribution of key predictor variables to model decisions. After comparing different machine learning models, the XGBoost model demonstrated the highest accuracy in predicting SCLC metastatic stages, especially in advanced metastatic stages (M1b and M1c), with significantly better AUC values than the other models. SHAP analyses revealed that age, tumour size, and treatment regimen were the main variables influencing the model predictions. LIME analyses further confirmed these variables' specific influence in the prediction of individual cases. This study successfully applied advanced machine learning methods in the prediction of SCLC metastatic stage, which significantly improved the accuracy of prediction and the interpretability of the model. These findings can help clinicians make more accurate treatment decisions in SCLC management, especially in strategy selection and early intervention.

Keywords Small cell lung cancer, Metastasis stage, SEER, Machine learning, Shap, Model interpretability tools

Small cell lung cancer (SCLC) is a highly malignant subtype of lung cancer. Although its incidence is lower than that of non-small cell lung cancer, it accounts for a significant proportion of lung cancer-related deaths due to its rapid growth and high propensity to metastasise[1]. Epidemiological studies have shown that SCLC accounts for approximately 10–15% of all lung cancer cases, and its five-year survival rate is significantly lower than that of other types, especially in patients diagnosed at advanced stages[2,3]. Distant metastases of SCLC mainly involve vital organs such as brain, liver, bone and adrenal glands[4,5]. This extensive metastatic ability stems in part from the rapid proliferation of tumour cells, their invasiveness, and their high degree of anti-apoptotic adaptability[6,7]. The From a molecular biological perspective, SCLC cells aberrantly express a variety of growth factors and receptors that promote tumour growth and metastasis[8]. Metastasis of SSCLC significantly worsens patient prognosis and greatly limits treatment options[9]. Due to the rapid progression of SCLC, patients often miss the optimal time for treatment, and even aggressive chemotherapy and radiotherapy are difficult to eradicate the disease[9]. Therefore, early identification and prediction of the metastatic stage of SCLC is crucial for improving

[1]Thoracic Surgery, Taizhou Hospital of Traditional Chinese Medicine, Jiangsu, China. [2]Thoracic Surgery, Taizhou Fourth People's Hospital, Jiangsu, China. [3]Pathology Department, Taizhou Hospital of Traditional Chinese Medicine, Jiangsu, China. [4]Medical Oncology, Taizhou Hospital of Traditional Chinese Medicine, Jiangsu, China. [5]Internal Medicine, Taizhou Fifth Peoples Hospital, Jiangsu, China. ✉email: 7259191@qq.com; wsying0303@163.com

patient survival and quality of life. Early diagnosis not only provides patients with more treatment options, but also slows down disease progression through targeted therapies or surgical interventions, prolonging disease-free survival and overall survival[10].

Predicting the pattern of metastasis at initial diagnosis in SCLC has become a key factor in clinical decision-making[11]. Traditional predictive models rely heavily on statistical methods such as logistic regression and Cox proportional risk models, which are trained on historical datasets to identify potential biomarkers and clinical parameters that influence metastasis[12]. While these methods initially provide valuable insights, they are typically weak in handling high-dimensional data and model prediction accuracy is often affected by sample size and variable selection bias[13]. With the development of machine learning techniques and the proliferation of clinical databases, more studies have employed these advanced algorithms to predict SCLC metastasis[14]. For example, methods such as Random Forest, Support Vector Machines (SVM) and Artificial Neural Networks have been used for modelling to improve the accuracy and reliability of predictions by learning complex patterns from large amounts of patient data[15]. These machine learning models can handle more input variables and theoretically provide higher predictive performance. However, machine learning methods also face challenges in practice, as their "black box" nature makes the decision-making process difficult to interpret, which is particularly problematic in clinical applications, where doctors and patients often require explicit explanations to support treatment decisions[16].

The aim of this study was to optimise the prediction of SCLC metastatic stages by utilising the SEER (Surveillance, Epidemiology and End Results Database) database with advanced machine learning techniques, in particular integrated learning methods and model interpretation tools.The SEER database provides a rich resource of clinical data, including extensive case follow-up and detailed treatment records, which provides a reliable basis for constructing efficient and accurate prediction models provides a reliable foundation[17]. In addition, this study will focus on the use of model interpretation tools such as SHAP (SHapley Additive exPlanations)[18–22]. The introduction of these tools aims to increase the transparency and interpretability of models, so that machine learning models are not just a "black box" but can provide clinicians with interpretable data to support more informed treatment decisions[23]. Through this study, we expect not only to improve the accuracy of SCLC metastatic stage prediction, but also to enhance the transparency and credibility of clinical decision-making through the explanatory tools of the model, which will significantly improve the effectiveness of the clinical decision support system, and at the same time, promote the application and popularisation of machine learning technology in the medical field.

Materials and methods

Data sources and data sets

Patients with an ICD-O3 standard pathological diagnosis of small cell lung cancer (SCLC) in the SEER database between 2004 and 2017 were selected for this study. Specific diagnostic codes included: malignancy, small cell type (8002/3); small cell carcinoma, not otherwise specified (NOS) (8041/3); small cell carcinoma, spindle cell (8043/3); and small cell carcinoma, intermediate cell (8044/3);Combined small cell carcinoma (8045/3); squamous cell carcinoma, small cell type (8073/3).

Inclusion Criteria: 1. All study subjects were histopathologically confirmed as SCLC patients. 2. Age and gender were not restricted to ensure the generalisability of the study results. 3. All patients with known TNM staging were included to provide comprehensive coverage of the different stages of the disease. 4. Patients with any form of treatment were included. Exclusion criteria: 1. Patient records with incomplete information were excluded. 2. Duplicate records in the database were excluded to ensure sample independence. 3. Non-SCLC patients were excluded to ensure study specificity.

To address potential concerns regarding multiple primary cancers, we included all patients with pathologically confirmed SCLC, regardless of the SEER sequence number. According to SEER definitions, the sequence number represents the chronological order of reportable tumors diagnosed in a patient's lifetime but does not indicate metastatic spread between different tumors. SCLC cases with a sequence number > 1 thus still represent distinct primary lung cancers. Furthermore, to assess the potential confounding effects of prior malignancies, we conducted a sensitivity analysis excluding patients with Sequence Number ≥ 2 (i.e., retaining only those with Sequence Number = 1).

To ensure consistency with SEER data definitions and maintain external comparability, race/ethnicity was categorized into seven predefined groups: White, Black, Chinese, Japanese, Filipino, Hawaiian, and Others. These classifications are based on SEER's coding standards and have been widely adopted in prior population-based studies to capture potential genetic, socio-cultural, and care-access heterogeneity.

Data pre-processing and descriptive analysis

In this study, records missing key information (e.g., tumour stage, treatment type, or key clinical outcomes) were excluded during data pre-processing to ensure data quality and enhance the reliability of model predictions. Because M-staging may vary across years, a uniform scheme was applied: M0 indicates no distant metastases; M1a indicates limited distant metastases (e.g., ipsilateral pleural or regional lymph node involvement as defined in our coding); M1b indicates single-organ distant metastasis; and M1c indicates multiple-organ metastases or substantial tumour burden. This classification standardized entries prior to analysis to ensure consistent criteria across years and sources and to avoid sparsity caused by staging drift. The dataset was split into a 70% training set and a 30% test set with fixed random seeds for reproducibility. Normality of continuous variables (e.g., age, tumor size) was assessed using the Shapiro–Wilk test. Non-normal variables ($p < 0.05$) were summarized as median (IQR), whereas normal variables were summarized as mean (SD). Group differences across M-stages were evaluated using ANOVA (normal) or Kruskal–Wallis (non-normal) tests; categorical variables were compared with chi-square tests. A two-sided p-value < 0.05 was considered statistically significant. The prediction target

was the metastasis category at initial diagnosis (M0/M1a/M1b/M1c); no time-to-event endpoints were modeled. Predictors were limited to information available at or before the diagnostic encounter; no post-diagnostic outcomes or follow-up variables were used.

Variable selection

Lasso regression is a linear model that uses L1 regularisation to automatically select and shrink variables. The method effectively addresses multicollinearity in the data and selects variables by reducing the coefficients to zero. The regularisation parameter λ is crucial in determining the strength of the L1 penalty in the model, as well as influencing model complexity and variable selection. In this study, the optimal value of λ was determined through cross-validation (CV). Specifically, a K-fold cross-validation method was used to calculate the validation error of the model for different values of λ. The value of λ that minimises the cross-validation error was selected. Lasso regressions performed at this value of λ will automatically filter out the variables that contribute significantly to the predicted outcomes (M0, M1a, M1b, M1c) and reduce the coefficients of the insignificant variables to zero.

In order to visualise the overlap and differences between the variables selected for different predictive outcomes, we plotted Wayne and UpSet diagrams. These plots effectively show the intersection and peculiar sets of variables selected between different models, helping researchers to understand which variables are consistently selected across multiple models and which variables are specific to certain predictive outcomes. Based on the results of the Wayne and UpSet plots, we selected the intersection of variables in the three predictive outcome models as predictor variables for the final model. This step ensured that the selected variables had statistical significance and predictive power for all predicted outcomes.

Model development

All data analysis and model development were carried out in R (version 4.3.2), primarily using the tidymodels framework and its associated packages. Tidymodels integrates several modular components—such as parsnip, recipes, tune, workflows, stacks, and yardstick—to provide a unified pipeline for model development and evaluation. A variety of base learners were constructed using the parsnip package, including decision trees, random forests, XGBoost, support vector machines (SVMs), LightGBM, multilayer perceptrons (MLPs), k-nearest neighbors (KNN), elastic net, ridge regression, and multinomial logistic regression. Each model was configured with an appropriate engine and hyperparameters tailored to its algorithmic structure. The recipes package was used for preprocessing tasks, including variable transformation, missing data imputation, and feature engineering, to ensure the dataset was appropriately structured for machine learning analysis.

To address the notable class imbalance among the M-stage categories—particularly the underrepresentation of M1c—we applied SMOTE (Synthetic Minority Oversampling Technique) to the training set during resampling. SMOTE was implemented using the themis package, which integrates seamlessly into the tidymodels workflow. Oversampling was conducted within each resample split (i.e., inside the resampling loop) to avoid data leakage. This technique was applied consistently across all model pipelines to improve the sensitivity and robustness of classifiers toward minority classes. Model hyperparameter spaces were defined using the dials package, and tuning was conducted via grid search or random search with k-fold cross-validation using the tune package. The stacks package was used to implement stacked generalization, combining predictions from base learners into a meta-model trained via lasso regression, thus leveraging model diversity to enhance predictive performance. Model evaluation was performed with the yardstick package using multiple metrics: accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC (Precision-Recall Area Under the Curve), which are particularly informative under imbalanced classification settings. The final evaluation results were visualized using ggplot2 for performance comparison across models, folds, and metrics.

For multiclass model evaluation, we computed ROC-AUC values using a One-vs-Rest (OvR) decomposition strategy, in which each metastatic stage (M0, M1a, M1b, M1c) was treated as the positive class in turn and compared against all remaining stages. This approach allowed us to derive class-specific ROC and PR curves, enabling a more granular analysis of model discrimination across metastasis stages. The OvR-based AUC computation was implemented using the yardstick::roc_auc() function with the estimator = "macro" or "hand_till" argument as appropriate.

The choice of the 10 candidate algorithms was based on their distinct methodological strengths and popularity in structured medical data classification tasks. For example, KNN and decision trees offer strong interpretability, tree-based ensembles (RF, XGBoost, LightGBM) have high non-linear modeling capacity, linear models (ridge, elastic net) provide regularization benefits, while MLP and SVM are known for handling high-dimensional interactions[24–26]. This diversity enabled a comprehensive comparison across classical and modern machine learning strategies.

Analysis of model performance and interpretability

The area under the receiver operating characteristic curve (ROC-AUC) was used as the main evaluation metric to assess the performance of different predictions on the training and test sets, respectively. The AUC value quantifies the ability of the model to correctly classify under different thresholds, and is an ideal metric for assessing the performance of medical diagnostic tests. Generating ROC curves for each model on the test set, showing the changes in false positive and true positive rates of the models under different classification thresholds, helps to visually assess and compare the diagnostic efficacy of the models. An analysis of variance (ANOVA) was used to compare the AUC values of the different models on the test set to determine if statistically significant differences existed. This step identifies the model with optimal performance and provides a scientific basis for model selection. Performance metrics (ROC-AUC, PR-AUC, accuracy, F1) were selected for a cross-sectional, multiclass classification setting.

The models with the best performance on the test set are selected for SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) analyses, both of which are advanced model interpretation techniques based on different theories. SHAP is a game theory-based approach for explaining the contribution of individual features to model predictions, while LIME provides local interpretability of model predictions. SHAP values and LIME results are computed for the data on the test set, and feature importance graphs and local contribution graphs are plotted, which show the average contribution of different features to the model prediction results and the impact of specific prediction instances, helping to identify the variables that have the most impact on the predictions. This approach enhances the transparency and understandability of the modelling decision-making process.

Reproducibility and code availability

To ensure transparency and reproducibility, all procedures for data preprocessing, model training, performance evaluation, and interpretability analysis were conducted using R (version 4.3.2) and the tidymodels framework. The modeling pipeline includes key components such as data cleaning, LASSO-based feature selection, SMOTE resampling, cross-validation tuning, and ensemble stacking.

A de-identified version of the source code, along with example datasets and detailed documentation, has been prepared and will be made available upon manuscript acceptance. For peer-review purposes, these materials are provided as supplementary files. A permanent open-access link to the GitHub or Zenodo repository will be published with the final version of the manuscript.

Results

Results of descriptive statistics

A total of 12,128 patients diagnosed with small cell lung cancer (SCLC) from 2004 to 2017 were included in the final analysis, stratified by metastatic stage: M0 (N = 4439), M1a (N = 1138), M1b (N = 6054), and M1c (N = 497). The mean age at diagnosis was 68 ± 10 years, with modest differences across groups (M0: 69 ± 10; M1a: 70 ± 10; M1b: 68 ± 10; M1c: 68 ± 9; $p < 0.001$). Tumor survival time decreased significantly with disease progression (overall: 12 ± 17 months; M0: 20 ± 21; M1a: 11 ± 16; M1b: 8 ± 10; M1c: 3 ± 3; $p < 0.001$). Similarly, tumor size increased across metastatic stages (overall: 50 ± 34 mm; M0: 44 ± 30; M1a: 58 ± 33; M1b: 52 ± 34; M1c: 58 ± 44; $p < 0.001$). The cohort was predominantly White (83.9%), with a balanced gender distribution (51.4% female) and 50.0% of patients being married. Notably, the gender ratio varied by metastasis status, with more females in M0 (55.3%) and M1a (56.9%), while M1b showed a slight male predominance (52.4%) ($p < 0.001$). Married patients comprised the majority in all subgroups, ranging from 45.3 to 51.3%.

Tumor characteristics revealed stage-dependent trends. In M0, early T stage (T1: 29.0%) and absent lymph node involvement (N0: 26.3%) were most common, whereas M1b had the highest T4 tumor proportion (44.6%) and M1c the highest N3 node involvement (30.6%) ($p < 0.001$ for both T and N stage distributions). The upper lobe remained the most frequent tumor site across groups (~ 50%). Regarding treatment, systemic therapy after surgery was most common in the M1a group (31.1%), while no systemic therapy or surgery was recorded for 76.2% overall, peaking at 81.2% in M1b. This reflects the shift toward palliative treatment in later-stage disease. Detailed distributions of all demographic and clinical variables are presented in Table 1.

Variable selection and model optimisation

We applied Lasso regression for variable selection in each of the four metastatic stage models (M0, M1a, M1b, M1c) to identify features with optimal discriminatory power. The regularization parameter λ was tuned via tenfold cross-validation, and the value that minimized misclassification error was selected (Fig. 1E). The log(λ) values for the optimal models were approximately − 5.43 for M0, − 5.12 for M1a, − 5.66 for M1b, and − 6.03 for M1c, respectively, as indicated by the minimal cross-validation error points in the λ path plots (Fig. 1A–D). The number of features retained at these optimal λ values varied by metastatic stage: 39 variables were selected for M0, 19 for M1a, 27 for M1b, and 15 for M1c. The path plots clearly demonstrate how most coefficients shrink toward zero as λ increases, with only the most informative variables retaining non-zero weights, indicating their contribution to metastatic classification.

To evaluate the overlap and specificity of variables selected across metastatic stages, we constructed an UpSet plot and a Venn diagram (Fig. 2A, B). The UpSet plot revealed that 10 variables were consistently retained across all four models, representing core features with robust predictive value. These overlapping features—such as age at diagnosis, tumor size, N stage, and marital status—were also supported by clinical relevance and prior literature.

Based on these results, we retained the intersecting variables from all four Lasso models as the final feature set for downstream model training. This unified feature selection strategy ensured that only those predictors demonstrating consistent significance across metastatic subtypes were used, thereby enhancing the generalizability and interpretability of the final predictive models. A complete listing of all included variables, distributions, and inter-group comparisons is provided in Table 1. Notably, the selection process also reflected clinical plausibility: for instance, tumor size and N stage were prominent in later stages (M1b, M1c), while demographic and treatment-related variables such as marital status, laterality, and chemotherapy showed greater contributions at earlier stages.

Evaluation of model performance

To assess the discriminatory power of each model across different metastatic stages of small cell lung cancer (SCLC), we compared the performance of ten machine learning algorithms under both training and testing conditions. The evaluation focused on four M-stage categories (M0, M1a, M1b, and M1c), using a range of performance metrics including ROC-AUC, PR-AUC, F1-score, accuracy, and balanced accuracy.The entire

Characteristic	Tumor metastasis					*p*-Value
	Overall N = 12,128	M0 N = 4,439	M1a N = 1,138	M1b N = 6,054	M1c N = 497	
Age of diagnosis						< 0.001[1]
Mean ± SD	68 ± 10	69 ± 10	70 ± 10	68 ± 10	68 ± 9	
Median (Q1, Q3)	68 (62, 76)	69 (62, 76)	70 (62, 78)	68 (61, 75)	68 (63, 75)	
Min, Max	24, 100	29, 96	37, 98	24, 100	36, 91	
Survival times						< 0.001[1]
Mean ± SD	12 ± 17	20 ± 21	11 ± 16	8 ± 10	3 ± 3	
Median (Q1, Q3)	7 (2, 15)	13 (6, 25)	7 (1, 14)	6 (1, 11)	2 (1, 5)	
Min, Max	0, 107	0, 107	0, 107	0, 107	0, 11	
Tumor size						< 0.001[1]
Mean ± SD	50 ± 34	44 ± 30	58 ± 33	52 ± 34	58 ± 44	
Median (Q1, Q3)	44 (25, 69)	36 (21, 60)	53 (32, 80)	48 (28, 70)	50 (31, 76)	
Min, Max	0, 718	0, 500	0, 190	0, 690	0, 718	
Marital status at diagnosis, n (%)						
Single (never married)	1,776 (14.6%)	625 (14.1%)	187 (16.4%)	874 (14.4%)	90 (18.1%)	
Widowed	2,157 (17.8%)	806 (18.2%)	251 (22.1%)	1,020 (16.8%)	80 (16.1%)	
Married (including common law)	6,059 (50.0%)	2,199 (49.5%)	516 (45.3%)	3,106 (51.3%)	238 (47.9%)	
Divorced	1,970 (16.2%)	743 (16.7%)	169 (14.9%)	978 (16.2%)	80 (16.1%)	
Unmarried or domestic partner	45 (0.4%)	17 (0.4%)	2 (0.2%)	22 (0.4%)	4 (0.8%)	
Separated	121 (1.0%)	49 (1.1%)	13 (1.1%)	54 (0.9%)	5 (1.0%)	
Race ethnicity, n (%)						
Black	1,185 (9.8%)	413 (9.3%)	145 (12.7%)	580 (9.6%)	47 (9.5%)	
White	10,170 (83.9%)	3,724 (83.9%)	909 (79.9%)	5,109 (84.4%)	428 (86.1%)	
Filipino	122 (1.0%)	41 (0.9%)	11 (1.0%)	68 (1.1%)	2 (0.4%)	
Japanese	149 (1.2%)	66 (1.5%)	21 (1.8%)	60 (1.0%)	2 (0.4%)	
Others	247 (2.0%)	95 (2.1%)	27 (2.4%)	111 (1.8%)	14 (2.8%)	
Chinese	114 (0.9%)	43 (1.0%)	12 (1.1%)	57 (0.9%)	2 (0.4%)	
Hawaiian	141 (1.2%)	57 (1.3%)	13 (1.1%)	69 (1.1%)	2 (0.4%)	
Gender, n (%)						< 0.001[2]
Male	5,890 (48.6%)	1,986 (44.7%)	491 (43.1%)	3,172 (52.4%)	241 (48.5%)	
Female	6,238 (51.4%)	2,453 (55.3%)	647 (56.9%)	2,882 (47.6%)	256 (51.5%)	
Primary site labeled, n (%)						< 0.001[2]
Main bronchus	1,134 (9.4%)	337 (7.6%)	127 (11.2%)	621 (10.3%)	49 (9.9%)	
Lower lobe, lung	2,964 (24.4%)	1,109 (25.0%)	265 (23.3%)	1,479 (24.4%)	111 (22.3%)	
Upper lobe, lung	6,323 (52.1%)	2,449 (55.2%)	562 (49.4%)	3,054 (50.4%)	258 (51.9%)	
Lung, NOS	1,023 (8.4%)	267 (6.0%)	110 (9.7%)	598 (9.9%)	48 (9.7%)	
Middle lobe lung	522 (4.3%)	217 (4.9%)	50 (4.4%)	233 (3.8%)	22 (4.4%)	
Overlapping lesion of lung	162 (1.3%)	60 (1.4%)	24 (2.1%)	69 (1.1%)	9 (1.8%)	
Laterality, n (%)						
Left	5,046 (41.6%)	1,833 (41.3%)	479 (42.1%)	2,542 (42.0%)	192 (38.6%)	
Right	6,831 (56.3%)	2,556 (57.6%)	636 (55.9%)	3,345 (55.3%)	294 (59.2%)	
Paired site, but no information concerning laterality	162 (1.3%)	37 (0.8%)	8 (0.7%)	108 (1.8%)	9 (1.8%)	
Bilateral, single primary	51 (0.4%)	0 (0.0%)	10 (0.9%)	40 (0.7%)	1 (0.2%)	
Only one side. side unspecified	15 (0.1%)	5 (0.1%)	2 (0.2%)	8 (0.1%)	0 (0.0%)	
Not a paired site	23 (0.2%)	8 (0.2%)	3 (0.3%)	11 (0.2%)	1 (0.2%)	
T, n (%)						< 0.001[2]
T0	154 (1.3%)	41 (0.9%)	3 (0.3%)	104 (1.7%)	6 (1.2%)	
T1	2,205 (18.2%)	1,287 (29.0%)	79 (6.9%)	784 (13.0%)	55 (11.1%)	
T2	3,027 (25.0%)	1,233 (27.8%)	236 (20.7%)	1,477 (24.4%)	81 (16.3%)	
T3	2,540 (20.9%)	829 (18.7%)	284 (25.0%)	1,338 (22.1%)	89 (17.9%)	
T4	3,904 (32.2%)	991 (22.3%)	507 (44.6%)	2,157 (35.6%)	249 (50.1%)	
TX	298 (2.5%)	58 (1.3%)	29 (2.5%)	194 (3.2%)	17 (3.4%)	
RX Summ. Systemic Sur. Seq, n (%)						
No systemic therapy and or surgical procedures	9,237 (76.2%)	3,039 (68.5%)	903 (79.3%)	4,916 (81.2%)	379 (76.3%)	
Systemic therapy after surgery	2,846 (23.5%)	1,382 (31.1%)	234 (20.6%)	1,114 (18.4%)	116 (23.3%)	
Other	45 (0.4%)	18 (0.4%)	1 (0.1%)	24 (0.4%)	2 (0.4%)	
Continued						

Characteristic	Tumor metastasis					*p*-Value
	Overall N = 12,128	M0 N = 4,439	M1a N = 1,138	M1b N = 6,054	M1c N = 497	
Visceral and parietal pleural invasion recode, n (%)						
Not documented; No resection of primary; Not assessed or if assessed	11,208 (92.4%)	3,885 (87.5%)	1,067 (93.8%)	5,779 (95.5%)	477 (96.0%)	
PL0; No evidence; Tumor does not completely traverse the elastic layer of pleura	678 (5.6%)	421 (9.5%)	37 (3.3%)	211 (3.5%)	9 (1.8%)	
Tumor extends to pleura, NOS; not stated if visceral or parietal	157 (1.3%)	63 (1.4%)	29 (2.5%)	55 (0.9%)	10 (2.0%)	
PL1 or PL2; Invasion of visceral pleura present, NOS	67 (0.6%)	60 (1.4%)	2 (0.2%)	4 (0.1%)	1 (0.2%)	
PL3; Tumor invades into or through the parietal pleura OR chest wall	18 (0.1%)	10 (0.2%)	3 (0.3%)	5 (0.1%)	0 (0.0%)	
N, n (%)						<0.001[2]
N0	2,167 (17.9%)	1,167 (26.3%)	162 (14.2%)	768 (12.7%)	70 (14.1%)	
N1	1,044 (8.6%)	516 (11.6%)	65 (5.7%)	429 (7.1%)	34 (6.8%)	
N2	6,332 (52.2%)	2,126 (47.9%)	633 (55.6%)	3,332 (55.0%)	241 (48.5%)	
N3	2,585 (21.3%)	630 (14.2%)	278 (24.4%)	1,525 (25.2%)	152 (30.6%)	

Table 1. Patient demographics and baseline characteristics. RX. Summ. Systemic. Sur. Seq refers to the treatment sequence of systemic therapy and/or surgery: No systemic therapy and/or surgical procedures – patients received neither systemic treatment nor surgery; Systemic therapy after surgery – systemic therapy was administered following surgical procedures; Other – includes all other or unspecified treatment sequences. [1]Kruskal-Wallis rank sum test. [2]Pearson's Chi-squared test.

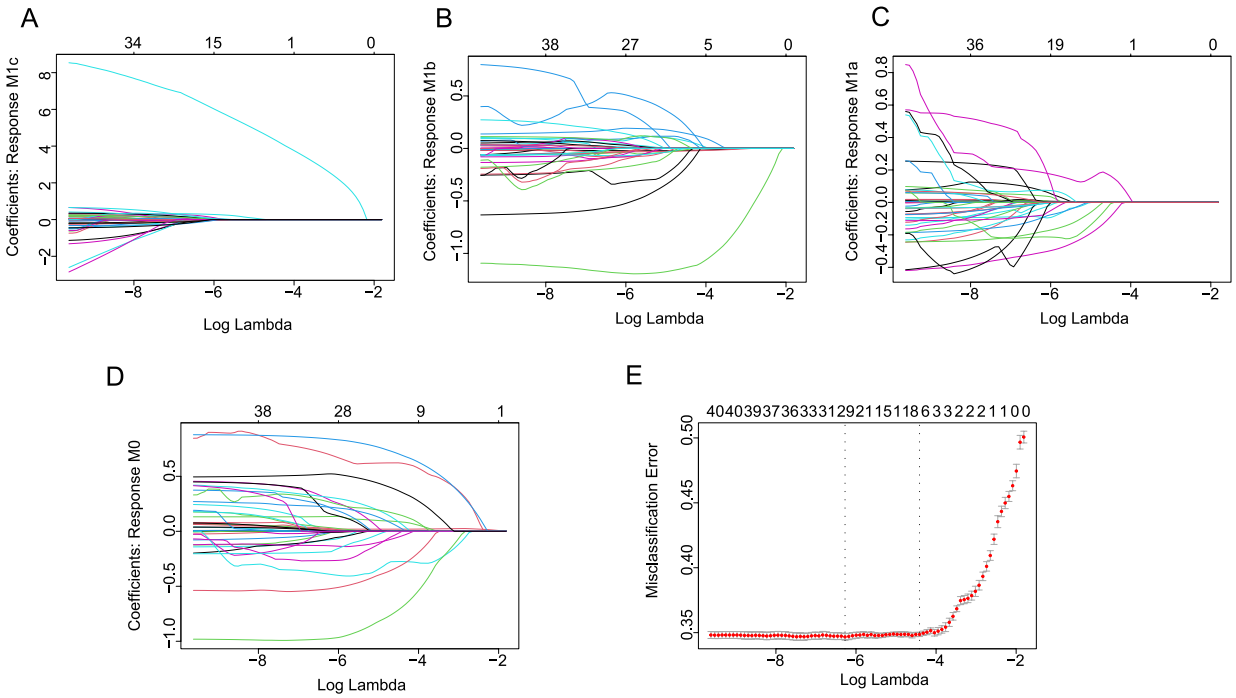

Fig. 1. Lasso coefficient path plots and tuning error curve for each metastatic stage (M0–M1c). (**A**–**D**) Each colored line represents the coefficient trajectory of a predictor variable as a function of log(λ). Vertical dashed lines indicate the optimal log(λ) corresponding to minimum cross-validation error. (**E**) Misclassification error vs. log(λ) showing optimal λ selection (minimum error) used to determine feature subset.

cohort of 12,128 patients was randomly split into a training set (70%) and a test set (30%) using stratified sampling to preserve the proportional distribution of M-stage categories. Baseline demographic and clinical characteristics of the two subsets are presented in Supplementary Table S1. Model development was conducted on the training set, where we employed tenfold cross-validation, and to address the pronounced class imbalance (especially for M1a and M1c), SMOTE (Synthetic Minority Over-sampling Technique) was applied within each fold to generate synthetic minority class samples.

As shown in Fig. 3A, B, KNN consistently outperformed other models across cross-validation folds, achieving the highest and most stable PR-AUC and ROC-AUC values. This suggests strong class discrimination and reliable calibration across M-stage categories. Other nonlinear models such as XGBoost, SVM, and MLP)

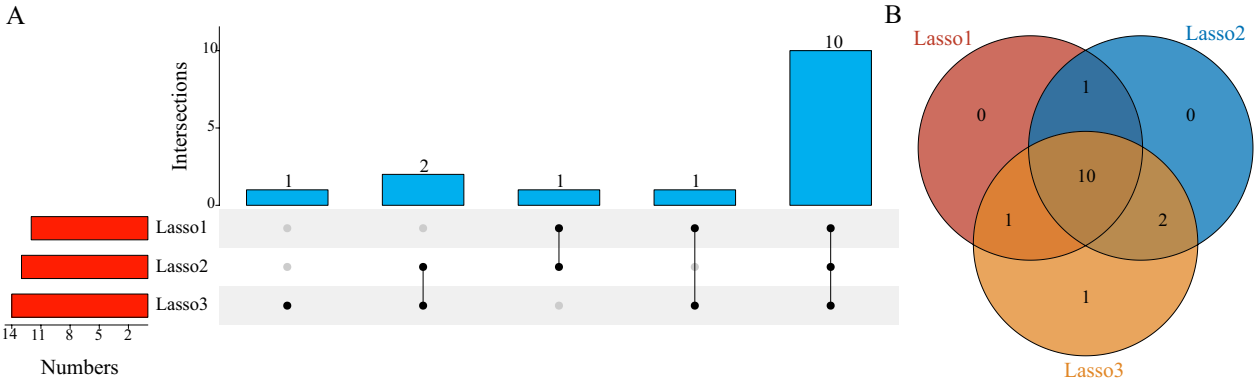

Fig. 2. (**A**) UpSet plot showing intersections of Lasso-selected variables across stage-specific models. The bar on the right indicates 10 variables shared across all stages. (**B**) Venn diagram illustrating variable overlaps between M0 (Lasso1), M1a (Lasso2), and M1b (Lasso3) models.

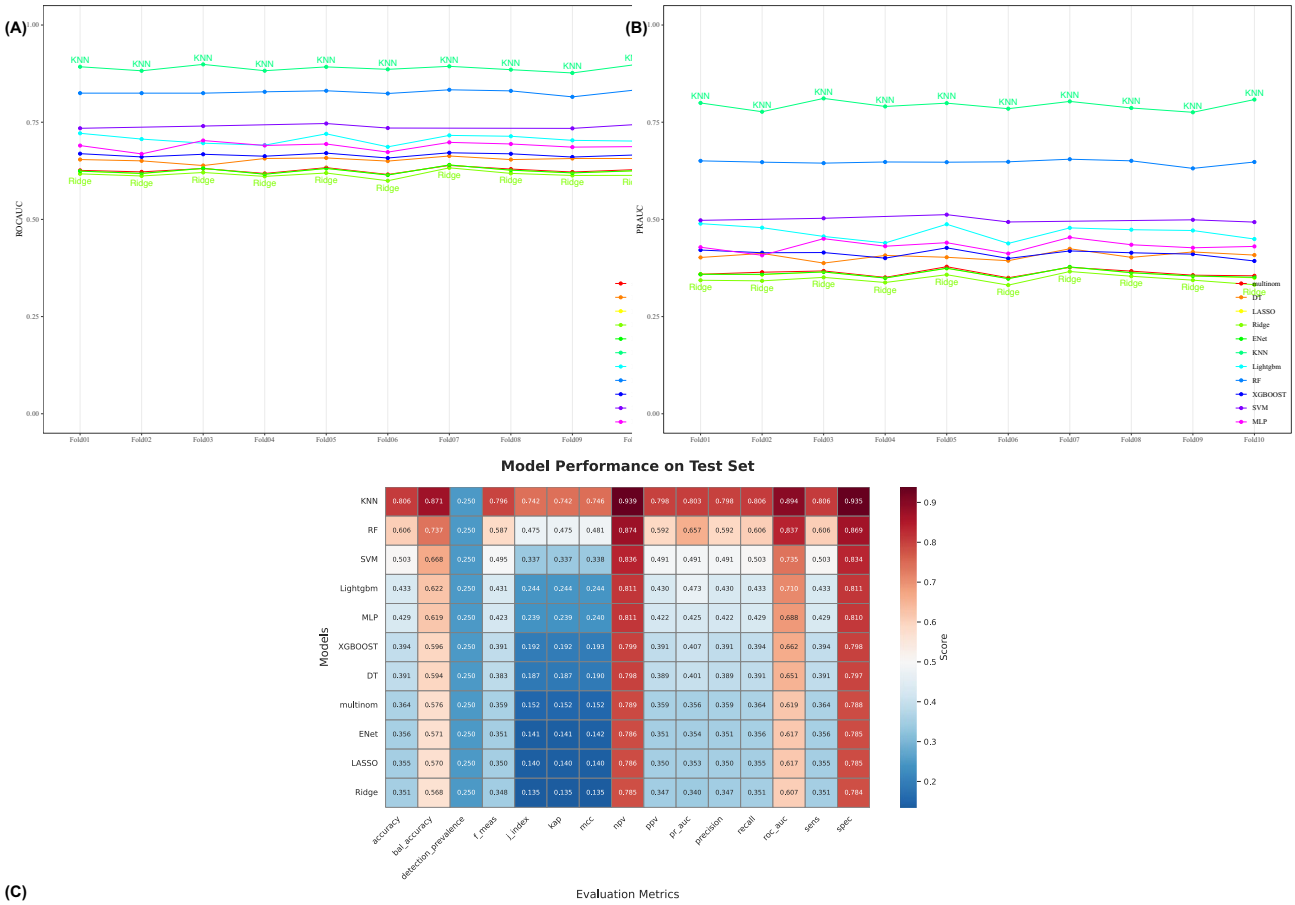

Fig. 3. Model performance in tenfold cross-validation and test set evaluation. (**A**) Mean PR-AUC values of 10 models across tenfold cross-validation on the training set. K-nearest neighbors (KNN) consistently achieved the highest PR-AUC scores across all folds, indicating superior sensitivity to minority classes. (**B**) Mean ROC-AUC values of 10 models across tenfold cross-validation. KNN also maintained the highest ROC-AUC, followed by XGBoost and SVM. (**C**) Heatmap summarizing test set performance of each model across 12 metrics, including accuracy, balanced accuracy, F1-score, ROC-AUC, PR-AUC, precision, recall, specificity, and others. Abbreviations: PR-AUC = Area Under the Precision-Recall Curve; ROC-AUC = Area Under the Receiver Operating Characteristic Curve; MCC = Matthews Correlation Coefficient; J_index = Youden index; PPV = Positive Predictive Value; NPV = Negative Predictive Value.

also demonstrated favorable performance, whereas linear models (Ridge, LASSO, Elastic Net) and decision trees exhibited relatively lower discriminative ability.

On the test set, model performance was visualized using a comprehensive heatmap of 12 metrics (Fig. 3C). KNN demonstrated outstanding overall performance, achieving the highest scores in accuracy (0.806), balanced accuracy (0.871), F1-score (0.796), ROC-AUC (0.894), PR-AUC (0.803), and specificity (0.935). These results highlight its superior generalizability and robustness in unseen data, even under class imbalance. In contrast, XGBoost and RF performed moderately well with ROC-AUC values around 0.83–0.66 but underperformed in recall and PR-AUC. Linear models and decision trees were consistently inferior in most metrics.

To better understand class-wise performance, we plotted multiclass ROC and PR curves for each metastatic stage on the test set (Figs. 4 and 5). KNN demonstrated clear dominance across stages M0, M1a, and M1c, with steep ROC curves and elevated PR-AUCs. In M1c—a clinically significant yet underrepresented group—KNN, XGBoost, and SVM showed relatively better PR-AUC values, though with notable drops in precision at higher recall levels. Ridge, LASSO, and multinomial regression showed limited discriminative power, with flatter ROC curves and degraded PR performance.

Overall, KNN emerged as the top-performing model in both cross-validation and external testing, exhibiting robust, interpretable, and stage-consistent performance across all evaluation metrics. Its effectiveness under class imbalance, particularly in detecting rare metastatic subtypes like M1c, underscores its potential utility in clinical decision-making for metastatic staging in SCLC. Given class imbalance, per-class PR-AUC provides a complementary view to ROC-AUC; detailed per-class results and diagnostic plots are provided alongside calibration and confusion matrices.

Interpretability analysis of the optimal model

To elucidate the decision process of the optimal classifier, we computed SHAP values for the KNN model using a one-vs-rest (OvR) decomposition for each metastatic stage (M0, M1a, M1b, M1c). Global impact was summarized with stage-specific SHAP beeswarm plots, and cross-stage importance was aggregated with stacked mean|SHAP| bars (Fig. 6A). All explanations were computed on the finalized modeling feature set. To complement global beeswarm plots, we report class-conditional SHAP dependence plots and patient-level force

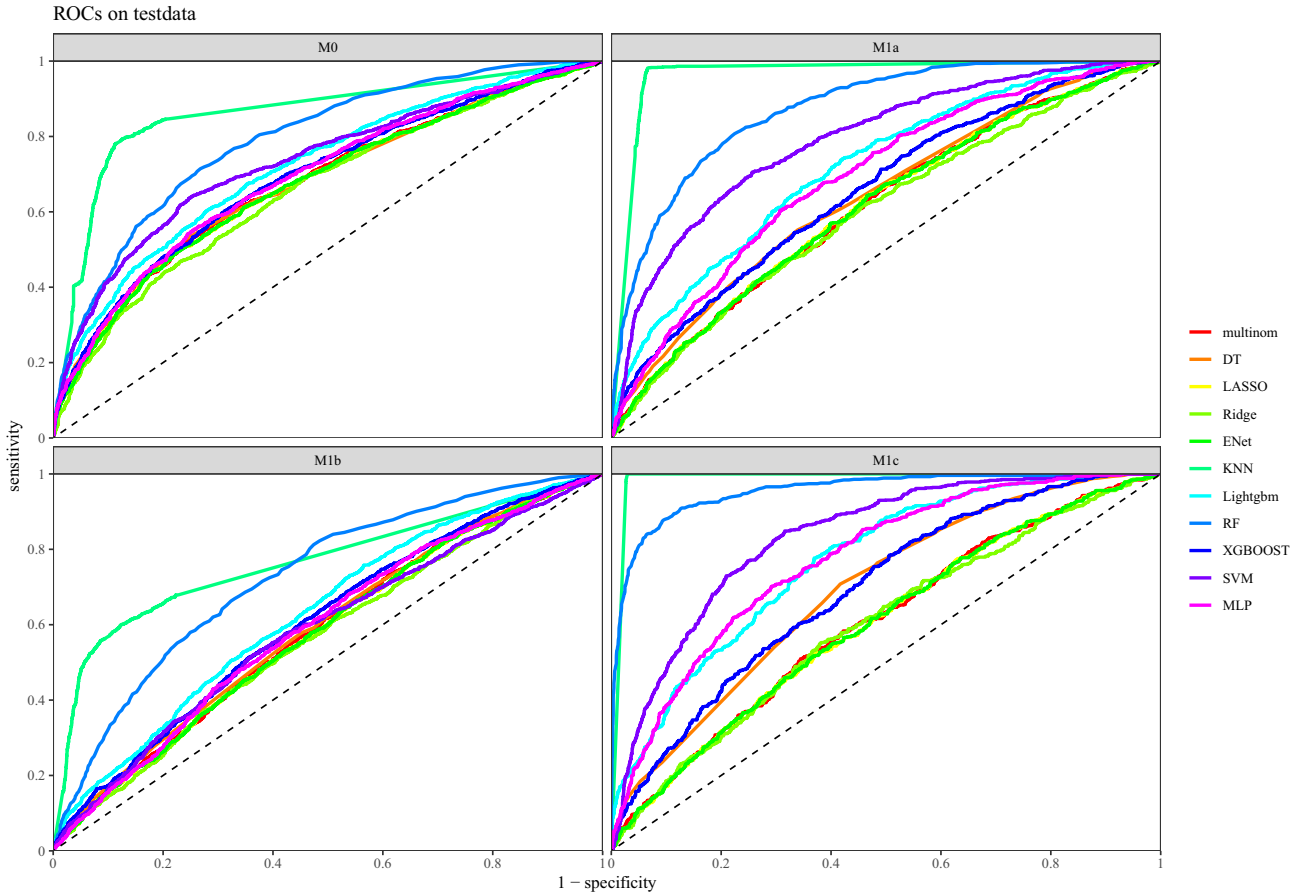

Fig. 4. Multiclass ROC curves for different M-stage predictions on the test set.One-vs-Rest ROC curves for each metastatic stage (M0, M1a, M1b, M1c) across 10 machine learning models. KNN demonstrated dominant performance in M0 and M1a stages, with superior true positive rates. In M1c, KNN, XGBoost, and SVM performed better than others, showing meaningful sensitivity to advanced metastatic cases. ROC-AUC = Area Under the Receiver Operating Characteristic Curve. OvR = One-vs-Rest decomposition.

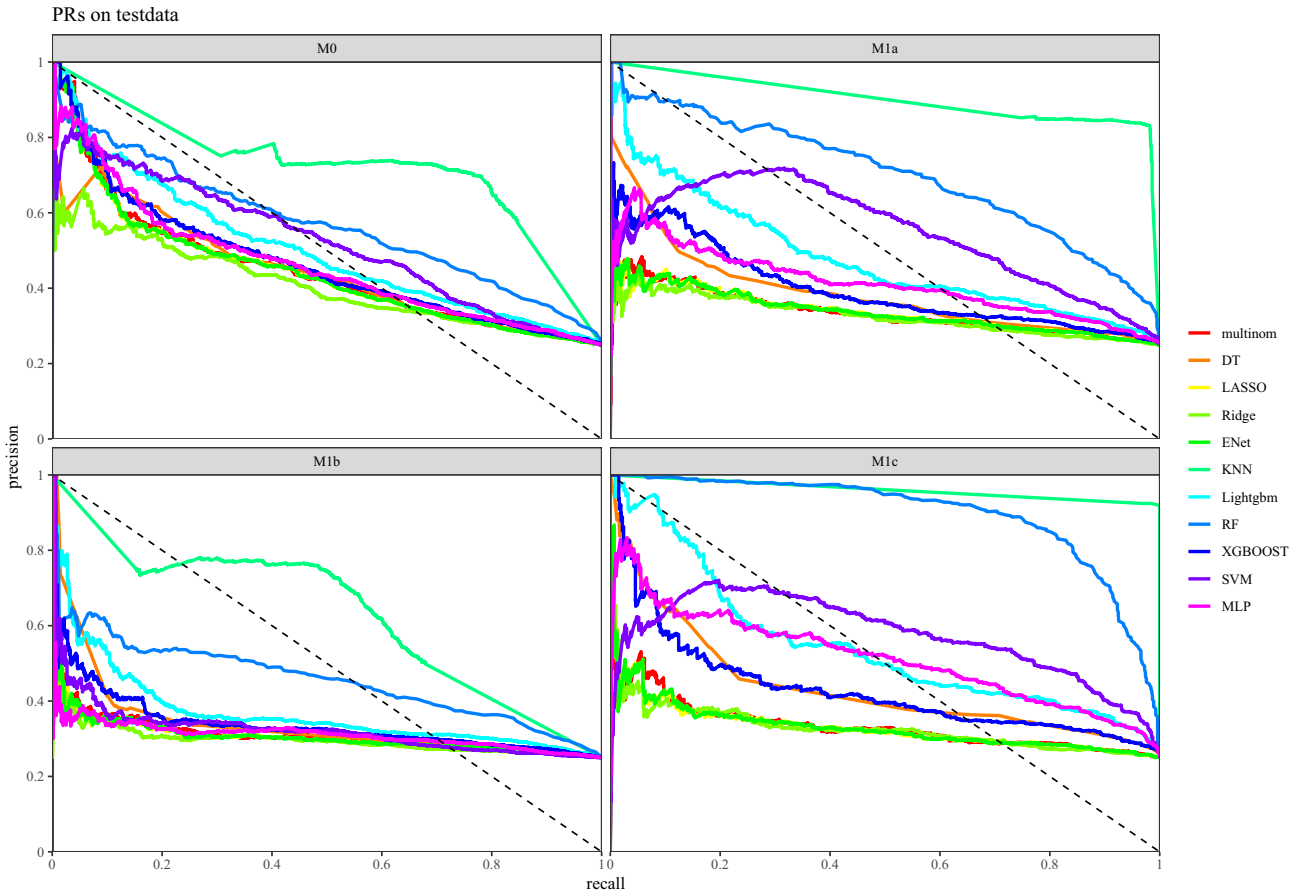

Fig. 5. Multiclass Precision-Recall curves for different M-stage predictions on the test set.One-vs-Rest PR curves for each metastatic stage (M0, M1a, M1b, M1c). KNN achieved the highest PR-AUC in M0 and M1a, with consistently high precision and recall. In M1c, KNN, XGBoost, and SVM showed competitive PR performance. Linear models and decision trees displayed lower PR curves overall, especially in minority classes. PR-AUC = Area Under the Precision-Recall Curve. OvR = One-vs-Rest decomposition.

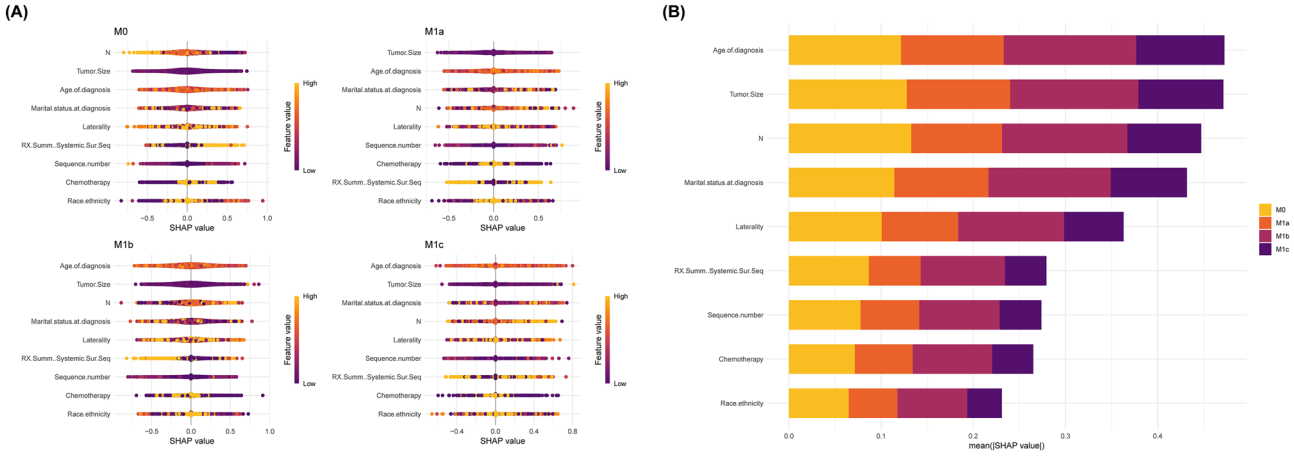

Fig. 6. SHAP summary plot for the KNN model across four metastatic stages (M0, M1a, M1b, M1c). The SHAP (SHapley Additive exPlanations) summary plot illustrates the global interpretability of the KNN model under a one-vs-rest multiclass decomposition. Each subplot corresponds to a specific M-stage classification (e.g., M0 vs. others). Features are ordered by mean absolute SHAP value, with longer bars indicating higher global importance. Each point represents one patient; the color indicates the actual feature value (red = low, blue = high).

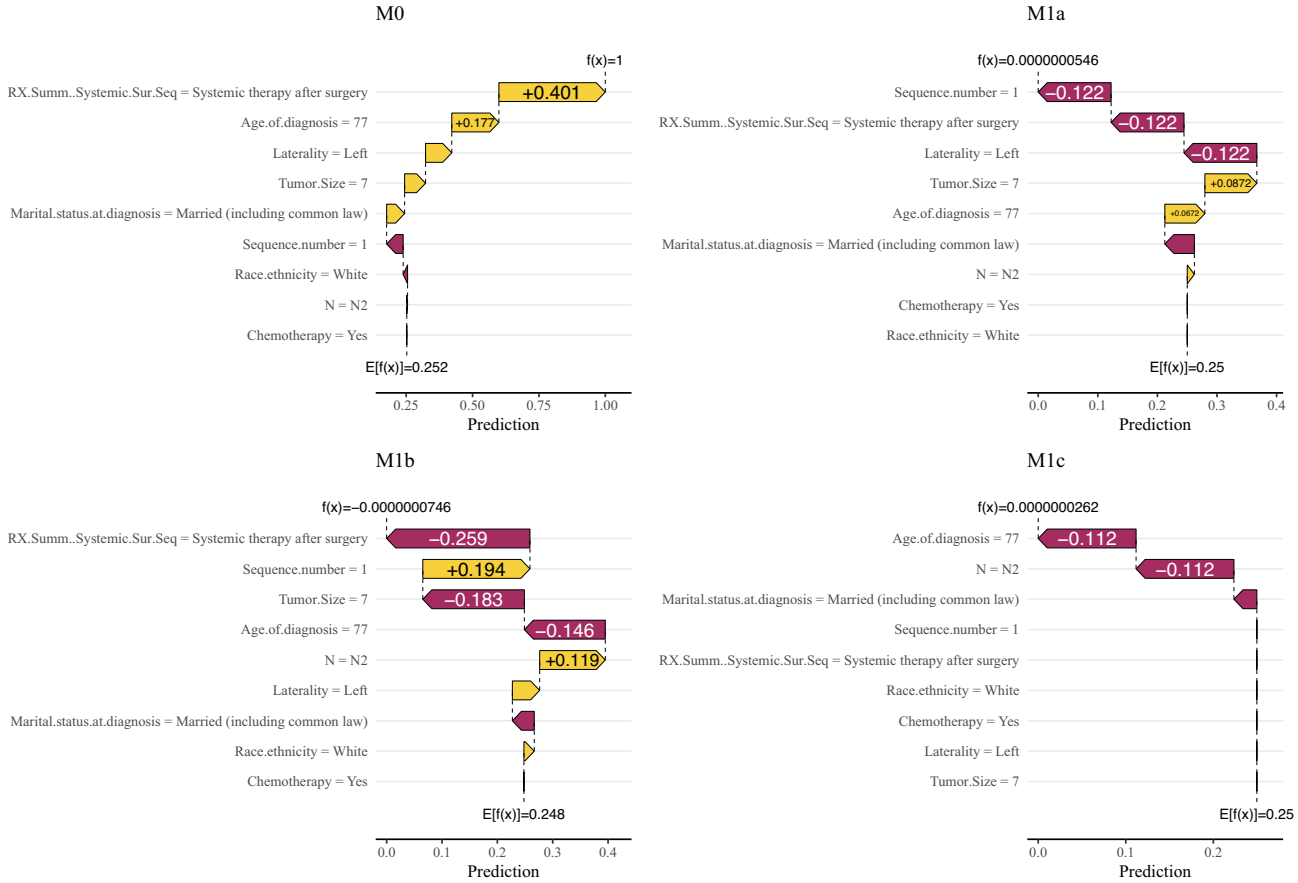

Fig. 7. SHAP force plots for representative test-set patients across OvR tasks.Each panel displays the additive decomposition of the predicted probability for one metastatic stage (M0, M1a, M1b, M1c). The dashed vertical line shows the baseline probability E[f(x)] for the corresponding OvR classifier. Bars pointing right increase the probability of the displayed class; bars pointing left decrease it. Variables such as N stage, tumor size, age at diagnosis, marital status, laterality, systemic-therapy sequence, sequence number, chemotherapy, and race/ethnicity contribute with stage-dependent signs and magnitudes, illustrating how case-specific combinations of tumor burden, nodal status, and contextual factors shape the final decision.

plots to clarify feature directionality. Discrimination is further summarized with per-class PR-AUC, calibration, and confusion matrices.

Across stages, the most influential predictors were N stage, tumor size, and age at diagnosis, followed by marital status at diagnosis, laterality, systemic-therapy sequence, SEER sequence number, chemotherapy, and race/ethnicity. Directionality was stage-dependent. In M0, lower nodal burden (N0–N1) and smaller tumors increased the probability of M0, whereas higher N and larger size reduced it. In M1a, larger tumors and older age showed modest positive contributions. In M1b/M1c, higher age, larger tumor size, and higher N generally shifted predictions toward advanced classes, while demographic and treatment variables exerted smaller, mixed effects.

The beeswarm color scale encodes actual feature values (yellow/orange = higher, purple = lower), enabling visual assessment of monotonic and non-monotonic patterns. Patient-level SHAP force plots (Fig. 6B) illustrate how features combine additively to move the prediction toward or away from a given stage; for example, low N with small tumor size favors M0, whereas higher N with larger tumors and older age favors M1b–M1c.

We visualized four representative test-set patients (one per OvR task) using SHAP force plots (Fig. 7). The dashed vertical line denotes the baseline probability E[f(x)] for the OvR classifier; colored bars show additive feature contributions that move the prediction toward (right) or away from (left) the displayed class. In the M0 example, contributions from systemic-therapy sequence, age at diagnosis, laterality, and tumor size shifted the probability toward M0, whereas N stage and chemotherapy opposed it. In the M1a example, sequence number, systemic-therapy sequence, and laterality decreased the probability, with a smaller positive shift from tumor size. In the M1b example, systemic-therapy sequence and tumor size reduced the probability, while sequence number, age, and N stage provided countervailing positive contributions. In the M1c example, negative shifts from age and N stage yielded a low final probability. These force plots demonstrate how a small set of clinically familiar variables combine to determine the stage-specific decision at the individual level.

Discussion

The KNN model achieved the best overall discrimination for metastatic-stage classification in our dataset, particularly for advanced stages (M1b/M1c). KNN is a non-parametric, instance-based classifier; its discrimination here arises from local neighborhood structure rather than tree ensembles or boosting. By leveraging distances in the feature space, KNN can adapt to nonlinear decision boundaries without explicit parametric fitting, which may be advantageous when the signal is local and heterogeneous.

Other machine-learning models (e.g., SVMs, Random Forests) also performed competitively. In our setting, KNN yielded the highest average metrics, but differences were modest for some classes. Random Forests can be robust to noise and interactions; however, under class imbalance and partially overlapping feature distributions, neighborhood-based methods may capture local structure more effectively. Performance variability across cohorts is expected given heterogeneity and potential feature correlations; we therefore report per-class metrics and calibration to contextualize discrimination.

Consequently, further research is warranted to bolster the model's generalization ability and robustness through technical refinements like data balancing procedures, feature engineering, and deeper model tuning. Such endeavors aim to fortify the model's reliability and applicability in clinical settings, thus facilitating more informed decision-making in SCLC management. Although our cohort included some patients with a SEER sequence number greater than one, indicating a history of prior primary malignancies, these cases were not excluded from the main analysis. This decision aligns with SEER's coding guidelines, which treat each tumor as a distinct primary rather than a metastasis of a previous cancer. Additionally, we performed a sensitivity analysis restricting the cohort to first primary SCLC cases (Sequence Number = 1), and found consistent model performance. This supports the conclusion that prior malignancies did not significantly confound metastatic stage prediction in our model.

The utilization of SHAP as advanced model interpretation tools has played a pivotal role in enhancing the transparency and interpretability of predictive models for the transfer stage of SCLC[27]. By employing these tools, we've not only bolstered the confidence of healthcare professionals in the modeling decision-making process but also facilitated the integration and application of these models within clinical settings[28].

SHAP offer healthcare professionals an intuitive glimpse into how each predictor variable influences the model output, thereby demystifying the modeling decision-making process[29]. SHAP values, for instance, quantitatively elucidate the contribution of each feature to the model's predictions, allowing clinicians to discern critical factors influencing a patient's transfer stage prediction[30]. This comprehensive understanding empowers clinicians to make more informed decisions when devising treatment plans, as they grasp which variables play pivotal roles in shaping the predicted transfer stage[31].

The application of SHAP in our study has underscored the significant influence of key variables like age, tumor size, and treatment regimen in predicting the metastatic stage of SCLC. SHAP analysis, in particular, has highlighted that larger tumor size and advanced age are commonly associated with heightened metastatic risk, aligning with the known biological behavior of SCLC[32]. Furthermore, differences in treatment regimens, such as chemotherapy or surgery, have been demonstrated to exert specific impacts on metastatic risk prediction through the insightful analyses provided by these tools.

The machine learning models developed in this study, particularly the KNN model, offer a promising avenue for enhancing the clinical management of small cell lung cancer (SCLC) by providing accurate predictions of metastatic stage. These models serve as valuable diagnostic tools, enabling clinicians to identify high-risk patients who may experience rapid metastasis, thereby facilitating timely adjustments in treatment plans.

In clinical practice, the deployment of these predictive models can guide clinicians in prioritizing more aggressive treatment strategies for patients in advanced metastatic stages (M1b and M1c). This may involve considering early intervention or the adoption of intensified chemotherapy regimens, potentially improving patient outcomes and quality of life[33]. Moreover, the implementation of such models supports the development of personalized medicine, allowing for the tailoring of treatment strategies based on individual patient data[34].

However, the application of these models in real-world clinical settings necessitates careful consideration of several factors. Firstly, seamless integration with existing clinical workflows and healthcare information systems is essential to ensure timely access to patient data and the delivery of predictive results. Additionally, the predictive accuracy of these models may vary across different patient populations, highlighting the importance of validation studies and potential model retraining to maintain reliability across diverse demographics.

Future research endeavors should focus on enhancing the explanatory power and usability of these models to foster trust and acceptance among clinicians. Furthermore, continuous updates and optimization of the models are imperative to incorporate emerging medical research and evolving treatment strategies. As clinical data accumulation and biomarker discovery progress, ongoing refinement of these predictive models will be crucial to their continued efficacy in improving patient care and outcomes.

Indeed, despite the promising outcomes achieved in predicting the metastatic stage of SCLC, there are noteworthy limitations that warrant consideration. Firstly, the reliance on the SEER) database for both model training and testing introduces potential biases inherent to the dataset. As the SEER database primarily encompasses data from specific regions within the United States, it may not fully represent the global or diverse ethnic backgrounds of the SCLC patient population. Moreover, the presence of data entry errors or incompleteness within the database could introduce inaccuracies that may impact the reliability and applicability of the predictive model.

Furthermore, while advanced machine learning techniques were employed in model development, challenges persist regarding the generalizability of the model across diverse populations. Variations in genetic predispositions, lifestyle factors, and treatment approaches among different demographic groups may influence the performance of the predictive model in real-world clinical settings. As such, efforts to validate and refine

the model across diverse patient populations are imperative to ensure its robustness and reliability in clinical practice.

Addressing these limitations will require incorporating additional data sources to enhance representativeness and conducting rigorous validation across diverse patient cohorts. Importantly, the present model is intended for diagnostic-time staging and data-quality harmonization rather than forecasting subsequent metastasis; prospective, longitudinal time-to-event modeling will be pursued in future work. Ongoing monitoring and model updating as new data accrue will further improve performance and clinical relevance. Ultimately, a comprehensive approach that reflects the complexities and nuances of SCLC populations is essential to maximize the utility and impact of predictive models in clinical practice.

Conclusion

This study optimises the prediction of SCLC metastatic stages through the machine learning methods and explanatory tools employed, which are of great significance in improving the accuracy of clinical diagnosis and treatment. In particular, the use of the KNN model significantly improves the accuracy of predicting the metastatic stage of SCLC, which is extremely important for early detection and therapeutic intervention, especially in the advanced metastatic stage of the disease, for clinical applications.

Data availability

The data that support the findings of this study are available from the SEER database (http://seer.cancer.gov/about/), but restrictions apply to the availability of these data, which were used under license for the current study and are therefore not publicly available. Data are, however, available from the authors upon reasonable request by contacting the corresponding author. Additionally, the full codebase used for data preprocessing, model development, and evaluation has been included in the supplementary materials to facilitate reproducibility. Additionally, the full codebase used for data preprocessing, model development, and evaluation has been included in the supplementary materials to facilitate reproducibility.

Received: 1 June 2024; Accepted: 24 November 2025
Published online: 01 December 2025

References

1. Franco, F. et al. Epidemiology, treatment, and survival in small cell lung cancer in Spain: Data from the Thoracic Tumor Registry. *PLoS ONE* **16**(6), e0251761 (2021).
2. Dumoulin, D. W. et al. Trends in the epidemiology of small-cell lung cancer: A Dutch nationwide population-based study over 1989–2020. *Eur. J. Cancer* **191**, 112985 (2023).
3. Rudin, C. M. et al. Small-cell lung cancer. *Nat. Rev. Dis. Primers.* **7**(1), 3 (2021).
4. Saltos, A., Shafique, M. & Chiappori, A. Update on the biology, management, and treatment of small cell lung cancer (SCLC). *Front. Oncol.* **10**, 1074 (2020).
5. Bunn, P. A. Jr. et al. Small cell lung cancer: can recent advances in biology and molecular biology be translated into improved outcomes?. *J. Thorac. Oncol.* **11**(4), 453–474 (2016).
6. Ko, J., Winslow, M. M. & Sage, J. Mechanisms of small cell lung cancer metastasis. *EMBO Mol. Med.* **13**(1), e13122 (2021).
7. Gazdar, A. F., Bunn, P. A. & Minna, J. D. Small-cell lung cancer: What we know, what we need to know and the path forward. *Nat. Rev. Cancer* **17**(12), 725–737 (2017).
8. Ding, X. L. et al. Clinical characteristics and patient outcomes of molecular subtypes of small cell lung cancer (SCLC). *World J. Surg. Oncol.* **20**(1), 54 (2022).
9. Früh, M. et al. Small-cell lung cancer (SCLC): ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **24**, 199–1105 (2013).
10. Rossi, A. et al. Vascular endothelial growth factor receptor as target for advanced non-small cell lung cancer therapy. *Curr. Drug Targets* **11**(7), 865–8784 (2010).
11. Wang, P. P. et al. Circulating tumor cells as a new predictive and prognostic factor in patients with small cell lung cancer. *J. Cancer* **11**(8), 2113–2122 (2020).
12. Lu, Y. J. et al. A novel nomogram based on SEER database for the prediction of liver metastasis in patients with small-cell lung cancer. *Ann. Palliat. Med.* **9**(5), 3123–3137 (2020).
13. Buil-Bruna, N. et al. Early prediction of disease progression in small cell lung cancer: Toward model-based personalized medicine in oncology. *Cancer Res.* **75**(12), 2416–2425 (2015).
14. Kriegsmann, M. et al. Deep learning for the classification of small-cell and non-small-cell lung cancer. *Cancers (Basel)* **12**(6), 1604 (2020).
15. Zhao, H. et al. The machine learning model for distinguishing pathological subtypes of non-small cell lung cancer. *Front Oncol* **12**, 875761 (2022).
16. Sanchez-Martinez, S. et al. Machine learning for clinical decision-making: Challenges and opportunities in cardiovascular imaging. *Front Cardiovasc Med* **8**, 765693 (2021).
17. Zhang, D. et al. Interpretable deep learning survival predictive tool for small cell lung cancer. *Front Oncol* **13**, 1162181 (2023).
18. Knapi, S., et al. Explainable artificial intelligence for human decision-support system in medical domain. 2021.
19. Akter, S. et al. Brain tumor classification utilizing pixel distribution and spatial dependencies higher-order statistical measurements through explainable ML models. *Sci. Rep.* **14**(1), 25800 (2024).
20. Khater, T., et al. Explainable AI for breast cancer detection: A LIME-driven approach. in *2023 16th International conference on developments in eSystems engineering (DeSE)*.
21. Venkata Krishnamoorthy, T. et al. A novel NASNet model with LIME explanability for lung disease classification. *Biomed. Signal Process. Control* **93**, 106114 (2024).
22. Nguyen, H. V. & Byeon, H. LIME-based ensemble machine for predicting performance status of patients with liver cancer. *Digital Health* **9**, 20552076231211636 (2023).
23. Aldughayfiq, B. et al. Explainable AI for retinoblastoma diagnosis: Interpreting deep learning models with LIME and SHAP. *Diagnostics (Basel)* **13**(11), 1932 (2023).
24. Li, W. et al. Development of a machine learning-based predictive model for lung metastasis in patients with ewing sarcoma. *Front. Med. (Lausanne)* **9**, 807382 (2022).

25. Shi, Y. et al. Application and interpretation of machine learning models in predicting the risk of severe obstructive sleep apnea in adults. *BMC Med. Inform. Decis. Mak.* **23**(1), 230 (2023).
26. Wang, K. et al. A machine learning model for visualization and dynamic clinical prediction of stroke recurrence in acute ischemic stroke patients: A real-world retrospective study. *Front. Neurosci.* **17**, 1130831 (2023).
27. Rikta, S. T. et al. XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer. *J. Pathol. Inform.* **14**, 100307 (2023).
28. Cui, Y. et al. Machine learning approaches for prediction of early death among lung cancer patients with bone metastases using routine clinical characteristics: An analysis of 19,887 patients. *Front. Public Health* **10**, 1019168 (2022).
29. Prabhu, S. et al. Multi-organ squamous cell carcinoma classification using feature interpretation technique for explainability. *Biocybernet. Biomed. Eng.* **44**(2), 312–326 (2024).
30. Astley, J. R. et al. Explainable deep learning-based survival prediction for non-small cell lung cancer patients undergoing radical radiotherapy. *Radiother. Oncol.* **193**, 110084 (2024).
31. Gupta, P., Application of explainable artificial intelligence in the identification of non-small cell lung cancer biomarkers. 2023.
32. Favali, M., Explainable machine learning and deep learning models to predict immunotherapy response in NSCLC patients using CT scans. 2022.
33. Shi, Y. et al. Machine learning prediction models for different stages of non-small cell lung cancer based on tongue and tumor marker: A pilot study. *BMC Med. Inform. Decis. Mak.* **23**(1), 197 (2023).
34. Guo, J. et al. Improving differential diagnosis of pulmonary large cell neuroendocrine carcinoma and small cell lung cancer via a transcriptomic, biological pathway-based machine learning model. *Translat. Oncol.* **14**(12), 101222 (2021).

Author contributions

YGH, JJX, HS and RHC conceived the project. YGH, LHW and KX designed most of the experiments. JXL, QW, QMG and YL analyzed data. GYL and KX provided significant intellectual input. QW and LHW wrote the manuscript with input from all other authors. All authors reviewed the final version of the 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/s41598-025-30314-9.

Correspondence and requests for materials should be addressed to J.X. or Y.H.