

Development and Validation of a Machine Learning-Based Clinical Prediction Model for Liver Metastasis in Stage III–IV Non-Small Cell Lung Cancer

Jin Gao

Beijing University of Chinese Medicine

Yan-Yuan Du

China Academy of Chinese Medical Sciences

Yu-Ming Liu

Beijing University of Chinese Medicine

Hui-Bo Yu

Beijing University of Chinese Medicine

Hong-Gang Zheng

China Academy of Chinese Medical Sciences

Wei Hou

houw2025@126.com

China Academy of Chinese Medical Sciences

Article

Keywords: non-small cell lung cancer, liver metastasis, machine learning, prediction model

Posted Date: July 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9815563/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on August 28th, 2026.
See the published version at <https://doi.org/10.1038/s41598-026-69223-w>.

Development and Validation of a Machine Learning-Based Clinical Prediction Model for Liver Metastasis in Stage III–IV Non-Small Cell Lung Cancer

Jin Gao^{1,2}, Yan-Yuan Du², Yu-Ming Liu^{1,2}, Hui-Bo Yu^{1,2}, Hong-Gang Zheng², Wei Hou^{2*}

Co-first authors: Jin Gao, Yan-Yuan Du, Yu-Ming Liu.

Corresponding author: Wei Hou.

¹Beijing University of Chinese Medicine, Beijing 100029, China.

²Department of Oncology, Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing 100053, China.

Email: houw2025@126.com

Abstract

Liver metastasis is among the major determinants of prognosis in non-small cell lung cancer (NSCLC), and early identification of high-risk patients has direct implications for clinical management. We retrospectively enrolled 849 patients with stage III–IV NSCLC and split them 8:2 into a training set ($n = 679$) and a validation set ($n = 170$). Feature selection followed a Venn intersection approach combining Least Absolute Shrinkage and Selection Operator (LASSO) regression, random forest (RF), and decision tree (DT) algorithms, which identified six shared predictors: lactate dehydrogenase (LDH), multi-organ metastasis, hemoglobin (HGB), thrombin time (TT), platelet count (PLT), and alkaline phosphatase (ALP). Ten machine learning (ML) classifiers were trained on these variables, and two with complementary strengths were assembled into a Stacking ensemble. On the validation set, the Stacking model outperformed all candidate models (AUC = 0.852, 95% CI: 0.769–0.921; AUPRC = 0.565; sensitivity = 76.0%; NPV = 0.952), with well-calibrated probability

estimates confirmed by calibration curves and decision curve analysis. SHAP analysis identified elevated LDH (roughly ≥ 250 U/L) and multi-organ metastasis as the two strongest contributors to individual predictions, while lower HGB, shortened TT, and abnormal PLT or ALP each added smaller but independent predictive weight. These findings demonstrate that a Stacking ML approach can predict liver metastasis risk in stage III–IV NSCLC with strong discrimination and reliable calibration. An online calculator has been deployed for point-of-care use; external validation in prospective cohorts is needed before broader clinical application.

Keywords: non-small cell lung cancer; liver metastasis; machine learning; prediction model

INTRODUCTION

Among all cancers, lung cancer leads worldwide in both new diagnoses and cancer-related deaths. GLOBOCAN 2022 estimated roughly 2.48 million new diagnoses and 1.82 million deaths annually—12.4% of all incident cancers and 18.7% of cancer deaths [1]. Non-small cell lung cancer (NSCLC) accounts for roughly 85% of cases and is the predominant histological type [2]. Molecularly targeted agents and immune checkpoint inhibitors have improved survival for certain patients, yet 40%–50% of patients already have distant metastases when first seen. In these patients, the 5-year survival rate hovers around 10% [3]. The outlook remains poor.

Among metastatic sites, liver metastasis carries an especially poor prognosis in NSCLC. It is reported in roughly 15%–20% of patients with metastatic NSCLC [4,5]; once it develops, median overall survival drops to about 4 months—among the shortest of any metastatic site [6]. A large SEER database study found liver metastasis to carry a 53% higher mortality hazard than brain metastasis [7]. The liver's immunotolerant microenvironment may also suppress systemic antitumor immunity [8]. This could possibly explain why immunotherapy tends to

underperform in these patients. Therefore, predicting liver metastasis risk before it becomes apparent is clinically meaningful.

In practice, predicting liver metastasis risk in individual NSCLC patients remains unsolved. Most published work has relied on Cox regression or nomogram approaches [9,10], which can lose accuracy once predictor relationships turn nonlinear or variable interactions multiply. Machine learning (ML) models for bone [11] and brain [12] metastasis in lung cancer have since been built and validated. In colorectal cancer, XGBoost- and random-forest-based models using SHapley Additive exPlanations (SHAP) to explain predictions have been applied to liver metastasis [13,14]. Yet few studies in NSCLC specifically have compared multiple ML classifiers head-to-head, assembled the strongest into an ensemble, and then used SHAP to explain what drove each prediction.

Interpretability poses a further obstacle: high-performing ML models are often opaque, with outputs that can be inscrutable even to the engineers who built them [15]. Clinicians are hesitant to act on predictions they cannot examine. SHAP, rooted in cooperative game theory, quantifies how much and in which direction each variable shifts a given prediction at both the population and individual patient levels [16]. Pairing SHAP with an ensemble model retains predictive accuracy while showing clinicians which variables drove each prediction.

In this study, ten ML classifiers were trained and compared in stage III–IV NSCLC patients, and the two strongest were assembled into a Stacking ensemble. SHAP values were then computed to explain what each prediction rested on, and an online calculator was deployed for direct use at the point of care.

Materials and Methods

Design and participants

We retrospectively enrolled consecutive inpatients admitted to the Department of Oncology,

Guang'anmen Hospital, China Academy of Chinese Medical Sciences, between April 2016 and August 2025. Inclusion criteria: (1) histopathologically or cytologically confirmed NSCLC; (2) clinical stage III or IV by the 8th-edition AJCC TNM staging system; (3) complete baseline clinical data available at first admission, covering demographics, laboratory results, and imaging findings. Exclusion criteria: (1) concurrent or prior history of other primary malignancies; (2) concurrent primary hepatic diseases (e.g., liver cirrhosis or active viral hepatitis) that might interfere with identifying liver metastasis; (3) severely incomplete clinical records (missing rate >30% for core variables) or records of questionable quality. Figure 1 shows the study design and workflow.

Institutional review board statement

This study was conducted in accordance with the principles established by the Declaration of Helsinki and was approved by the Ethics Committee of Guang'anmen Hospital, China Academy of Chinese Medical Sciences. Given the retrospective nature of the study, the fact that most participants had either died or been lost to follow-up, and that all data were fully anonymised, the Ethics Committee of Guang'anmen Hospital, China Academy of Chinese Medical Sciences granted a waiver of informed consent.

Outcome, Candidate Predictors, and Definitions

Outcome:

The primary outcome was liver metastasis, defined as hepatic lesions confirmed by contrast-enhanced CT, MRI, or PET-CT, or by pathological biopsy, and recorded as a binary variable.

Candidate predictors:

Candidate predictors were drawn from NSCLC clinical guidelines, expert consensus, published reviews, and clinical experience, subject to data availability, and totaled 38 variables.

We extracted clinical data from the electronic medical record system of Guang'anmen Hospital, covering five categories:

(1) Demographic and general characteristics: sex, age, smoking history, alcohol consumption history, body mass index (BMI), and Karnofsky Performance Status (KPS) score.

(2) Comorbidities: hypertension, diabetes, and fatty liver disease.

(3) Pathological features: histological subtype, classified as adenocarcinoma, squamous cell carcinoma, or other types.

(4) Laboratory parameters — complete blood count: neutrophil count (NEUT), lymphocyte count (LYMPH), monocyte count (MONO), platelet count (PLT), and hemoglobin (HGB); liver function tests: albumin (ALB), total bilirubin (TBIL), alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), and lactate dehydrogenase (LDH); coagulation function: D-dimer (D-Dimer), international normalized ratio (INR), thrombin time (TT), activated partial thromboplastin time (APTT), and fibrinogen (FIB); inflammatory marker: C-reactive protein (CRP); tumor markers: carcinoembryonic antigen (CEA), cytokeratin 19 fragment (CYFRA21-1), neuron-specific enolase (NSE), and squamous cell carcinoma antigen (SCC).

(5) Sites of distant metastasis: bone, brain, adrenal gland, pleural metastasis/malignant pleural effusion, and pulmonary metastasis. Multi-organ metastasis (Multi_organ_mets) was defined as the simultaneous presence of two or more distant metastatic sites (excluding liver metastasis itself) and was coded as a binary variable.

Variable types and measurement:

Directly quantifiable indicators, such as age and KPS score, were retained in continuous form to avoid information loss from discretization. Naturally categorical variables (sex,

comorbidities, histological subtype, and metastatic site status) were binary-coded. All laboratory values reflect the first measurement taken during the index admission. Demographic data were extracted using a standardized template. Since all candidate predictors were drawn from objective test results and electronic records, no subjective assessment was involved and predictor blinding was not required.

Data Preprocessing

All candidate variables were first screened for missingness; those with more than 30% missing data were excluded from further analysis. For variables falling below this threshold, Multiple Imputation by Chained Equations (MICE) was applied (maximum iterations = 10, random seed = 0). After MICE imputation, the dataset contained no remaining missing values, requiring no further case deletion. (Supplementary Material 1). Continuous variables were then standardized via Z-score normalization, while binary variables were carried through without transformation. To prevent data leakage, both the imputation model and standardization parameters were fit solely on the training set before being applied to the validation set.

Sample Size Calculation

Sample size directly affects how reliable and transferable the model will be [17]. The calculation here combined the classical “10 events-per-variable (EPV)” rule with the four-criterion framework from Riley et al., taking the larger estimate as the minimum requirement. We computed three sub-criteria separately and retained the largest value:

Criterion 1 — Precision of overall risk estimation:

$$n_1 = \left\lceil \left(\frac{1.96}{\delta} \right)^2 \cdot \varphi (1 - \varphi) \right\rceil$$

Setting the absolute margin of error $\delta = 0.05$, this yields $n_1 = \lceil (1.96 / 0.05)^2 \times 0.146 \times$

$$0.854] = 192.$$

Criterion 2 — Overall shrinkage factor $S = 0.9$, anticipated Cox–Snell $R^2_{CS} = 0.15$:

$$n_2 = \left\lceil \frac{p}{(S - 1) \ln \left(1 - \frac{R^2_{CS}}{S} \right)} \right\rceil$$

This yields $n_2 = 330$.

Criterion 3 — Root mean square prediction error (RMSPE = 0.05):

$$n_3 = \left\lceil \frac{p}{S \cdot \text{RMSPE}^2} \right\rceil$$

This yields $n_3 = \lceil 6 / (0.9 \times 0.0025) \rceil = 2667$.

The maximum across the three criteria is $n_{\text{Riley}} = \max(192, 330, 2667) = 2667$, driven by the RMSPE precision requirement of Criterion 3. The available sample ($n = 849$) satisfies the EPV criterion and Riley Criteria 1 and 2, consistent with current recommendations for prediction model development at moderate discriminative ability.

Predictor Selection

Considering the data characteristics, research objectives, and complementary strengths of each method, we applied three predictor selection methods: Least Absolute Shrinkage and Selection Operator (LASSO) regression, random forest (RF), and decision tree (DT). LASSO adds an L1 penalty to the cost function, pushing coefficients of uninformative features toward zero; this simultaneously reduces multicollinearity and filters out noise—particularly useful when candidate variables are internally correlated [18]. RF ranks feature importance by aggregating many decision trees and handles nonlinear relationships between laboratory and clinical variables, while being relatively robust to individual outliers and data noise [19]. DT maps the feature–outcome relationship as a tree structure, keeping the decision path transparent [20].

Each method operated on a distinct mathematical basis (L1 regularization, ensemble averaging, and information gain) and the final predictor set was the intersection of all three. Only variables that survived all three filters entered the model.

Model Development and Validation

Model development strategy: We trained ten machine learning algorithms: logistic regression (LR), DT, RF, extreme gradient boosting (XGBoost), support vector machine (SVM), adaptive boosting (AdaBoost), gradient boosting machine (GBM), K-nearest neighbors (KNN), artificial neural network (ANN), and extremely randomized trees (ET). All models underwent 10-fold cross-validation on the training set to maximize data use and limit random variance. Performance was then measured on the held-out validation set [21,22].

Model performance evaluation: Area Under the Curve (AUC) and Area Under the Precision-Recall Curve (AUPRC) were the primary discrimination metrics. AUC captures overall discrimination for binary outcomes; AUPRC is better suited to class-imbalanced data and more directly reflects performance on low-frequency events [23,24]. We also calculated accuracy, sensitivity, specificity, F1 score, Positive Predictive Value (PPV), and Negative Predictive Value (NPV). Calibration curves assessed the match between predicted probabilities and observed event rates. Final model selection was driven by validation-set performance, with training metrics checked to rule out both underfitting and overfitting.

Clinical utility assessment: Decision Curve Analysis (DCA) was used to evaluate the clinical utility of each model. Based on decision theory, DCA quantifies net benefit across a spectrum of threshold probabilities, incorporating the trade-off between the consequences of under- and over-treatment, thereby offering a more clinically meaningful assessment than discrimination metrics alone. [25,26].

Stacking Ensemble Model

To further improve predictive performance, we also tested a stacking ensemble strategy alongside the individual models. Stacking is a two-tier heterogeneous ensemble that integrates the complementary strengths of multiple algorithms: base learners (Level 0) generate predicted probabilities that feed into a meta-learner (Level 1) for a second round of learning [27]. Out-of-fold (OOF) probabilities generated through K-fold stratified cross-validation on the training set served as meta-features, ensuring that the meta-learner was trained on data unseen by the base learners and thereby preventing data leakage. A grid search covered base learner combinations, the passthrough option (whether raw features were also forwarded to the meta-learner), and the meta-learner's regularization parameter. The final model configuration was selected based on high prediction diversity among base learners, strong inter-model complementarity, and superior overall validation-set performance. The ensemble model was evaluated using the same metrics and as the individual models.

Model Interpretability Analysis

We applied SHAP to the optimal model to examine its prediction logic. SHAP, grounded in cooperative game theory, assigns each feature a marginal contribution to individual predictions, enabling both global and local interpretation [28]. Globally, a mean |SHAP| bar chart and a beeswarm plot displayed the direction and magnitude of each feature's effect on model output. For individual cases, force and waterfall plots showed how each feature shifted the predicted probability away from the baseline. Dependence plots were also generated to examine how SHAP contributions varied with feature values across the observed range.

Statistical Analysis

Continuous variables are reported as median (Q1, Q3) according to their distribution; group comparisons used the Mann–Whitney U test. Categorical variables are reported as counts and percentages and compared by the chi-square (χ^2) test; Fisher's exact test replaced it when any

expected cell count fell below 5. All analyses were run in Python 3.13.9. Statistical significance was set at a two-tailed $P < 0.05$.

Results

Baseline Characteristics of Study Participants

A total of 849 patients were included in the final analysis, of whom 124 had liver metastases and 725 did not, corresponding to a positive event rate of 14.6%. Thirty-eight candidate variables were collected in total. Three were excluded during data cleaning due to missingness $\geq 30\%$: GGT (51.8%), ALT (48.7%), and AST (48.7%). The remaining 35 variables were retained; among these, NEUT had the highest missingness rate (25.3%), followed by Diabetes (15.1%), BMI (11.7%), and KPS (11.3%). Most laboratory indicators had missingness rates below 8%, and binary variables had no or negligible missing data. All missing values were handled via MICE imputation. Regarding pathological subtype, the "other" category (originally comprising adenocarcinoma, squamous cell carcinoma, and other) contained fewer than 10 patients and was therefore removed as too small for stable estimation, leaving two subtypes for analysis. The dataset was split into training ($n=679$) and validation ($n=170$) sets at an 8:2 ratio using stratified random sampling. Baseline characteristics were comparable between the two groups with no significant differences ($P>0.05$), except for sex ($P=0.014$); details are presented in Table 1.

Feature Selection

Three ML methods (LASSO regression, RF, and DT) were used to screen candidate predictor variables; the final set was determined by their intersection, visualized as a Venn diagram. For LASSO (Figure 2A), 10-fold cross-validation AUC served as the evaluation metric. At $\lambda = 9.3260$ (optimal value), the model reached peak AUC with 19 variables retaining non-zero coefficients. Ranked by descending coefficient magnitude (Figure 2B): multi-organ metastasis,

brain metastasis, LDH, NSE, pathological subtypes, PLT, MONO, D-dimer, CEA, alcohol consumption history, HGB, ALP, FIB, TT, CYFRA21-1, KPS, NEUT, bone metastasis, and smoking history. With RF(Figure 2C), the mean importance score (Mean = 0.0286) was used as the cutoff; 20 variables exceeded this threshold. LDH ranked highest, followed by CEA, NSE, CYFRA21-1, D-dimer, multi-organ metastasis, HGB, ALP, NEUT, BMI, TT, PLT, age, CRP, SCC, MONO, ALB, LYMPH, APTT, and TBIL. With DT(Figure 2D), the same threshold (Mean = 0.0286) yielded 12 above-average variables. LDH again ranked first, followed by multi-organ metastasis, CRP, PLT, BMI, SCC, ALB, HGB, TT, FIB, age, and ALP.

To ensure methodological rigor, three feature selection methods were integrated. A Venn diagram was drawn to identify variables selected consistently across all three algorithms (Figure 2E). Six predictors emerged from this intersection: LDH, multi-organ metastasis, PLT, HGB, ALP, and TT. Six variables were ultimately retained for model development: LDH, multi-organ metastasis, PLT, HGB, ALP, and TT. All three algorithms, despite being grounded in different mathematical principles, independently converged on the same set of predictors. This cross-method agreement suggests that these variables carry stable predictive value across algorithmic frameworks, lending credibility to the model's reliability and reproducibility.

Model Development and Performance Comparison

Using the six variables from feature selection, ten ML classification models (LR, KNN, SVM, DT, RF, GBM, XGBoost, AdaBoost, ET, and ANN) were built to predict liver metastasis risk in stage III/IV NSCLC. Full performance metrics are in Table 2 and Figure 3. Among the individual models, KNN (validation-set AUC: 0.839) and LR (validation-set AUC: 0.829) stood out. PR curves showed a similar pattern: ANN (AP = 0.576), LR (AP = 0.540), and

KNN (AP = 0.526) recorded the three highest AP values on the validation set, all well above the baseline level, suggesting reasonable discrimination for the minority class (liver metastasis-positive cases).

We then evaluated a Stacking ensemble strategy. KNN and AdaBoost were selected as base learners for their contrasting prediction approaches (instance-based vs. tree-boosting), complementary error profiles, and strong validation-set performance; logistic regression served as the meta-learner. The meta-learner assigned coefficients of 2.462 to AdaBoost and 2.031 to KNN, reflecting substantive input from both base models. On the validation set, the Stacking model achieved the best AUC (0.852, 95% CI: 0.769 – 0.921) and AUPRC (0.565), outperforming both base learners and all other individual models. Sensitivity held at 76.0% and NPV at 95.2%, giving the model strong negative-exclusion capability (Table 3, Figure 4A). The PR curve (Figure 4B) confirmed adequate discrimination for minority-class events. The calibration curve (Figure 4C) showed the Stacking model's predicted probabilities tracking the ideal line closely in the low-risk range, better than most individual models. DCA (Figure 4D) showed consistently positive net benefit over both the treat-all and treat-none reference lines across threshold probabilities of roughly 0.07 – 0.48, indicating net clinical benefit across a wide decision range. The normalized feature importance heatmap (Figure 5) showed LDH and multi-organ metastasis leading the rankings in most models; HGB, TT, PLT, and ALP showed consistent contributions despite some variation across models. Among all tested configurations, the Stacking ensemble achieved the best discrimination and calibration. DCA confirmed clinical utility across a useful threshold range. It was selected as the final model for liver metastasis prediction in stage III/IV NSCLC.

Feature Contribution Analysis via SHAP Values

We applied SHAP to the Stacking model to examine its prediction logic at both global and

individual levels. Globally, multi-organ metastasis and LDH emerged as the two strongest predictors, with mean |SHAP| values well above those of the other four variables; HGB, TT, PLT, and ALP followed in descending order (Figure 6A). The beeswarm plot clarified the directional effects: LDH showed a clear positive-driving pattern (high values pushed risk up, low values pulled it down), with the widest SHAP distribution of all variables. HGB and TT ran in the opposite direction: low values mapped to positive SHAP (higher risk), high values to negative SHAP (protective) (Figure 6C). For the binary variable multi-organ metastasis, presence carried a SHAP of +0.055 versus -0.024 for absence (Figure 6B). Dependence plots added finer detail. LDH contributions climbed sharply above roughly 250 U/L (Figure 7A). HGB showed a clear inflection near its cohort mean of 122.66 g/L—below-mean values pushed risk up, above-mean values pulled it down (Figure 7B). TT displayed a step-like shift at 15.83 s (Figure 7C). PLT followed an approximate inverted-U pattern, with positive SHAP values below the mean and negative ones above (Figure 7D). ALP contributions climbed progressively past 99.42 U/L, reaching their strongest positive values at the upper end of the distribution (Figure 7E). Waterfall plots detailed the prediction pathway for three representative patients. In a high-risk patient, an LDH of 3,600 U/L contributed +0.22, alongside contributions from multi-organ metastasis (+0.06) and low HGB (+0.03); the combination pushed the predicted probability from a baseline of 0.15 to 0.467 (Figure 8A). In a low-risk patient, a normal-to-low LDH (100 U/L, -0.05) combined with elevated HGB (-0.02) and absent multi-organ metastasis (-0.02) held the predicted probability down to 0.057 (Figure 8C). In a borderline case, the positive push from multi-organ metastasis (+0.05) was partly offset by a normal LDH (-0.02), leaving the predicted probability just above baseline (Figure 8B).

Discussion

Six predictors—LDH, multi-organ metastasis, HGB, TT, PLT, and ALP—emerged from the three-algorithm feature selection step. A Stacking ensemble was then built on this predictor set. SHAP was applied to unpack the model’s reasoning at both cohort and patient level. ML models built specifically to predict the occurrence of liver metastasis in NSCLC are scarce. Most prior work has focused on survival after liver metastasis is already confirmed, using Cox regression or nomograms [10,29]—methods that handle nonlinear interactions poorly once predictor count grows. A practical tool for prospectively identifying patients at elevated liver metastasis risk does not yet exist for this disease context. This study attempts to provide an interpretable, quantitative solution for this clinical need.

On the validation set, the Stacking model reached an AUC of 0.852 (95% CI: 0.769–0.921), AUPRC of 0.565, sensitivity of 76.0%, and NPV of 95.2%—outperforming all ten individual ML models. Among comparable studies targeting liver metastasis occurrence in NSCLC, Zhang et al. [30] built a nomogram using multivariable logistic regression and reported a validation-set AUC of 0.799 (95% CI = 0.683–0.915); the present model’s AUC of 0.852 is somewhat higher. All six predictors come from routine admission bloodwork. No extra molecular testing is required. ML has been applied to liver metastasis risk prediction in colorectal cancer [13,14]. The present work extends this approach to NSCLC using Stacking ensemble learning, with a three-algorithm intersection for feature selection and OOF cross-validation for meta-feature generation—an approach that limits leakage while holding performance gains. An NPV of 95.2% means a low-risk classification makes liver metastasis unlikely. For that subgroup, less frequent abdominal imaging may be warranted. DCA showed positive net benefit over both reference lines across thresholds of 0.07–0.48. This range covers a clinically useful spread of decision thresholds. SHAP analysis identified elevated LDH and multi-organ metastasis as the two strongest predictors; HGB, TT, PLT, and ALP each added independent predictive weight. Dependence plots showed LDH contributions climbing sharply

above roughly 250 U/L. LDH is a key effector enzyme of the Warburg effect. LDHA-catalyzed conversion of pyruvate to lactate acidifies the tumor microenvironment (TME), and that acidity promotes epithelial-mesenchymal transition (EMT), tumor cell invasion, and immune evasion [31]. Multi-organ metastasis (≥ 2 organs involved) reflects a higher overall tumor burden and wider systemic spread. Xu et al. [32] analyzed metastatic site combination patterns in a large NSCLC cohort and found that the liver had the lowest single-site incidence, while bone–liver–lung was the most frequent triple-site combination—suggesting that hepatic involvement typically arises with high tumor burden. The high SHAP weight for multi-organ metastasis fits this observation. The WHO defines anemia as HGB <120 g/L in females and <130 g/L in males; positive SHAP values below the mean HGB of 122.66 g/L indicate that anemia correlates with higher liver metastasis risk. In advanced NSCLC, anemia is common. It signals diminished systemic reserve and ongoing tumor-related consumption, and may also worsen local tumor microenvironmental hypoxia by reducing blood oxygen-carrying capacity—activating HIF-1 α . HIF-1 α upregulation follows, promoting angiogenesis via VEGF and driving EMT through Snail, β -catenin, and Notch signaling—both pathways heightening invasiveness and metastatic spread [33]. Shortened TT points to a systemic hypercoagulable state, closely linked to higher metastatic propensity. High tissue factor (TF) expression in tumor cells sustains activation of the extrinsic coagulation cascade, and the resulting thrombin promotes invasion, distant colonization, and angiogenesis via PAR-1/PAR-4 receptors [34]. PLT showed an inverse relationship with liver metastasis risk—lower PLT values mapped to higher risk. In stage III–IV disease, tumor-related systemic wasting is common. Persistent inflammatory activation sustains elevated TNF- α levels, suppressing bone marrow hematopoiesis and accelerating platelet destruction [35,36]. Advanced lung cancer cells can also infiltrate bone marrow via hematogenous routes, impairing megakaryocyte proliferation and reducing platelet output [37]. Bone marrow invasion marks advanced systemic dissemination and tends to co-occur with

liver metastasis. In this context, a low platelet count is not just an isolated lab finding. It signals bone marrow reserve depletion and widening systemic spread—both directly relevant to liver metastasis risk. Elevated ALP may reflect hepatic or skeletal tumor involvement. Tumor infiltration of the hepatic parenchyma or biliary obstruction drives large-scale hepatic ALP release; bone metastasis-associated skeletal destruction can similarly raise total serum ALP. In NSCLC, elevated ALP has been linked to both liver and bone metastasis risk and to reduced ICI response [38,39].

The model runs on six standard laboratory values from routine admissions. It has been deployed as an online risk prediction platform (<https://heileg-hz3shhgtznvsj5a4tjdusflung-liver-metastasis-predictor.streamlit.app/>). For patients flagged as low risk, imaging follow-up could reasonably be spaced out. Intermediate- or high-risk patients warrant closer and more frequent abdominal surveillance. For each case, the SHAP breakdown shows which variables drove the prediction—clinicians can inspect this output directly rather than treating the score as a black box.

Conclusion

Among the ten ML classifiers evaluated, the Stacking model performed best on the validation set (AUC 0.852, NPV 95.2%). It draws on six routine laboratory values and pairs each risk estimate with SHAP-based variable-level explanations. Prospective external validation is still required before this approach can be adopted in routine clinical care.

Limitations

Several constraints deserve mention. The data are drawn from a single center—Guang'anmen Hospital, a traditional Chinese medicine institution—so the patient profile may not match general oncology populations; generalization requires caution. The cohort ($n = 849$) satisfies Riley criteria 1 and 2 but falls short of criterion 3, limiting the precision of individual-level

probability estimates. The predictor set is also limited by data availability: driver mutation status (EGFR, ALK, KRAS), immunotherapy history, and radiomic features were not captured in these records, and including them might improve discrimination or recalibrate the risk estimates. The model should be tested in prospective multicenter cohorts before it can be used outside this institution.

Data sharing statement

The datasets used and analysed during the current study are available from the corresponding author on reasonable request. The source code used in this study is available at <https://github.com/heilegehei/lung-liver-metastasis-predictor>.

References

1. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024).
2. Molina, J. R. et al. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin. Proc.* **83**, 584–594 (2008).
3. Siegel, R. L., Kratzer, T. B., Wagle, N. S., Sung, H. & Jemal, A. Cancer statistics, 2026. *CA Cancer J.* **76**, e70043 (2026).
4. Kang, I. et al. Different prognostic implications of hepatic metastasis according to front-line treatment in non-small cell lung cancer: a real-world retrospective study. *Transl. Lung Cancer Res.* **10**, 2745–2755 (2021).
5. Čeriman Krstić, V. et al. Treatment options for patients with non-small cell lung cancer and liver metastases. *Curr. Issues Mol. Biol.* **46**, 13443–13455 (2024).
6. Qu, F. J., Zhou, Y. & Wu, S. Progress of immune checkpoint inhibitors therapy for

- non-small cell lung cancer with liver metastases. *Br. J. Cancer* **130**, 165–175 (2024).
7. Riihimäki, M. et al. Metastatic sites and survival in lung cancer. *Lung Cancer* **86**, 78–84 (2014).
 8. Corrao, G. et al. Stereotactic radiotherapy in metastatic non-small cell lung cancer: combining immunotherapy and radiotherapy with a focus on liver metastases. *Lung Cancer* **142**, 70–79 (2020).
 9. Wang, Y. et al. Clinical characteristics and prognosis of non-small cell lung cancer patients with liver metastasis: a population-based study. *World J. Gastrointest. Oncol.* **14**, 1982–1994 (2022).
 10. Xu, T. et al. Development and validation of a nomogram for predicting the overall survival in non-small cell lung cancer patients with liver metastasis. *Transl. Cancer Res.* **12**, 3061–3073 (2023).
 11. So, K. W. L. et al. Machine learning models to predict bone metastasis risk in patients with lung cancer. *Cancer Med.* **13**, e70383 (2024).
 12. Visonà, G. et al. Machine-learning-aided prediction of brain metastases development in non-small-cell lung cancers. *Clin. Lung Cancer* **24**, e311–e322 (2023).
 13. Xing, Y. et al. Development of prediction models for liver metastasis in colorectal cancer based on machine learning: a population-level study. *Transl. Cancer Res.* **13**, 5943–5952 (2024).
 14. Li, Y. et al. An interpretable deep learning framework for predicting liver metastases in postoperative colorectal cancer patients using natural language processing and clinical data integration. *Cancer Med.* **12**, 19785–19798 (2023).
 15. Cabitza, F., Rasoini, R. & Gensini, G. F. Unintended consequences of machine learning in

- medicine. *JAMA* **318**, 517–518 (2017).
16. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
 17. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. & Feinstein, A. R. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **49**, 1373–1379 (1996).
 18. Ranstam, J. & Cook, J. A. LASSO regression. *Br. J. Surg.* **105**, 1348 (2018).
 19. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
 20. Podgorelec, V., Kokol, P., Stiglic, B. & Rozman, I. Decision trees: an overview and their use in medicine. *J. Med. Syst.* **26**, 445–463 (2002).
 21. Efthimiou, O. et al. Developing clinical prediction models: a step-by-step guide. *BMJ* **386**, e078276 (2024).
 22. Wong, T. & Yeh, P. Reliable accuracy estimates from k-fold cross validation. *IEEE Trans. Knowl. Data Eng.* **32**, 1586–1594 (2020).
 23. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
 24. Cao, Y. & Shen, Y. Bayesian active learning for optimization and uncertainty quantification in protein docking. *J. Chem. Theory Comput.* **16**, 5334–5347 (2020).
 25. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26**, 565–574 (2006).
 26. Vickers, A. J., van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn. Progn. Res.* **3**, 18 (2019).
 27. Wolpert, D. H. Stacked generalization. *Neural Netw.* **5**, 241–259 (1992).

28. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
29. Zhao, R., Dai, Y., Li, X. & Zhu, C. Construction and validation of a nomogram for non-small cell lung cancer patients with liver metastases based on a population analysis. *Sci. Rep.* **12**, 4011 (2022).
30. Zhang, T. et al. Construction of a nomogram model based on biomarkers for liver metastasis in non-small cell lung cancer. *Thorac. Cancer* **15**, 1897–1911 (2024).
31. Claps, G. et al. The multiple roles of LDH in cancer. *Nat. Rev. Clin. Oncol.* **19**, 749–762 (2022).
32. Xu, Z. et al. Clinical associations and prognostic value of site-specific metastases in non-small cell lung cancer: a population-based study. *Oncol. Lett.* **17**, 5590–5600 (2019).
33. Wu, Q. et al. Hypoxia-inducible factors: master regulators of hypoxic tumor immune escape. *J. Hematol. Oncol.* **15**, 77; (2022).
34. Wojtukiewicz, M. Z. et al. Thrombin—unique coagulation system protein with multifaceted impacts on cancer and metastasis. *Cancer Metastasis Rev.* **35**, 213–233 (2016).
35. Zhu, R. et al. Updates on the pathogenesis of advanced lung cancer-induced cachexia. *Thorac. Cancer* **10**, 8–16 (2019).
36. Chu, T. et al. Bifunctional effect of the inflammatory cytokine tumor necrosis factor α on megakaryopoiesis and platelet production. *J. Thromb. Haemost.* **20**, 2998–3010 (2022).
37. Yang, R., Jia, L. & Cui, J. Mechanism and clinical progression of solid tumors bone marrow metastasis. *Front. Pharmacol.* **15**, 1390361 (2024).
38. Jiang, X. et al. Do alkaline phosphatases have great potential in the diagnosis, prognosis,

and treatment of tumors? *Transl. Cancer Res.* **12**, 2760–2772 (2023).

39. Hu, J. et al. Pretreatment levels of serum alkaline phosphatase are associated with the prognosis of patients with non-small cell lung cancer receiving immune checkpoint inhibitors. *Oncol. Lett.* **25**, 147 (2023).

Acknowledgements

We would like to express our gratitude to each patient who participated in this study.

Authors' contributions

Jin Gao and Yu-Ming Liu were responsible for data curation, drafting the original manuscript, and reviewing and editing; Yan-Yuan Du and Hong-Gang Zheng were responsible for software development and data analysis; Hui-Bo Yu contributed to data curation; Wei Hou conceived and designed the study, conducted the investigation, and acquired funding.

Funding Statement

This study was supported by the National Key Research and Development Program of China (NO.2023YFC3503301) and (NO.Z210-014).

Additional information

Competing interests

The authors declare no competing interests.

Open Access

This article is an open-access article that was selected by an in-house editor and fully peer-reviewed by external reviewers. It is distributed in accordance with the Creative Commons Attribution NonCommercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative

works on different terms, provided the original work is properly cited and the use is non-commercial. See: <https://creativecommons.org/licenses/by-nc/4.0/>

Consent for publication

All authors agree to publish the final manuscript.

Figure

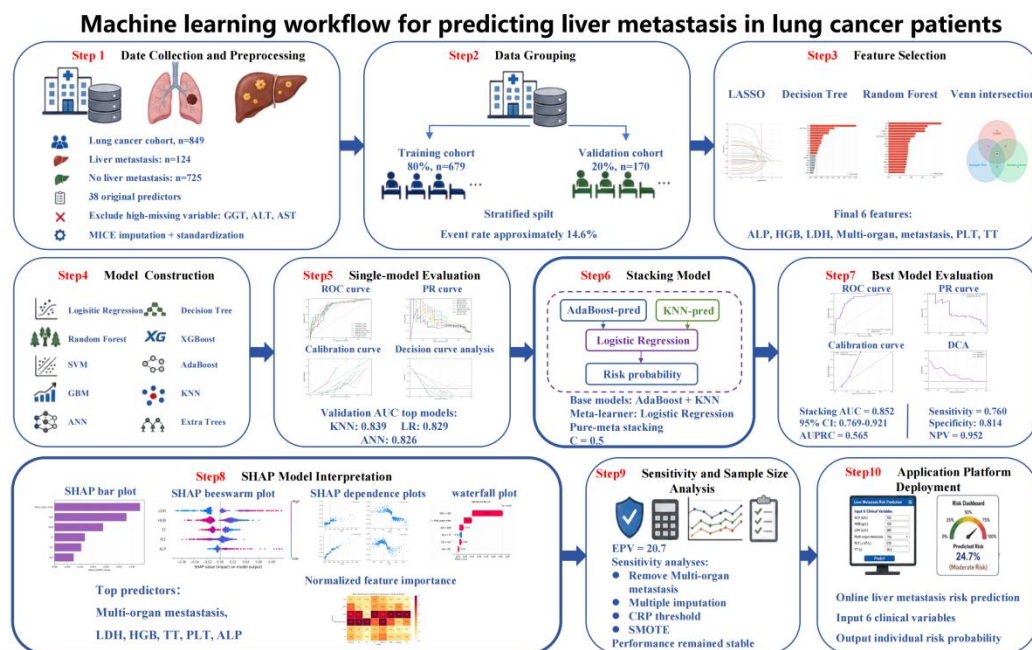


Figure 1. Flowchart of the study protocol. GGT: gamma-glutamyl transferase; ALT: alanine aminotransferase; AST: aspartate aminotransferase; MICE: Multiple Imputation by Chained Equations; LASSO: Least Absolute Shrinkage and Selection Operator; ALP: alkaline phosphatase; HGB: hemoglobin; LDH: lactate dehydrogenase; PLT: platelet count; TT: thrombin time; LR: Logistic Regression; XGBoost: Extreme Gradient Boosting; SVM: Support Vector Machine; GBM: Gradient Boosting Machine; ANN: Artificial Neural Network; AdaBoost: Adaptive Boosting; KNN: K-Nearest Neighbors; ROC: Receiver Operating Characteristic; PR: Precision-Recall; DCA: Decision Curve Analysis; AUC: Area Under the Curve; CI: Confidence Interval; NPV: Negative Predictive Value; AUPRC: Area Under the Precision-Recall Curve; SHAP: SHapley Additive exPlanations; CRP: C-reactive protein;

EPV: Events Per Variable.

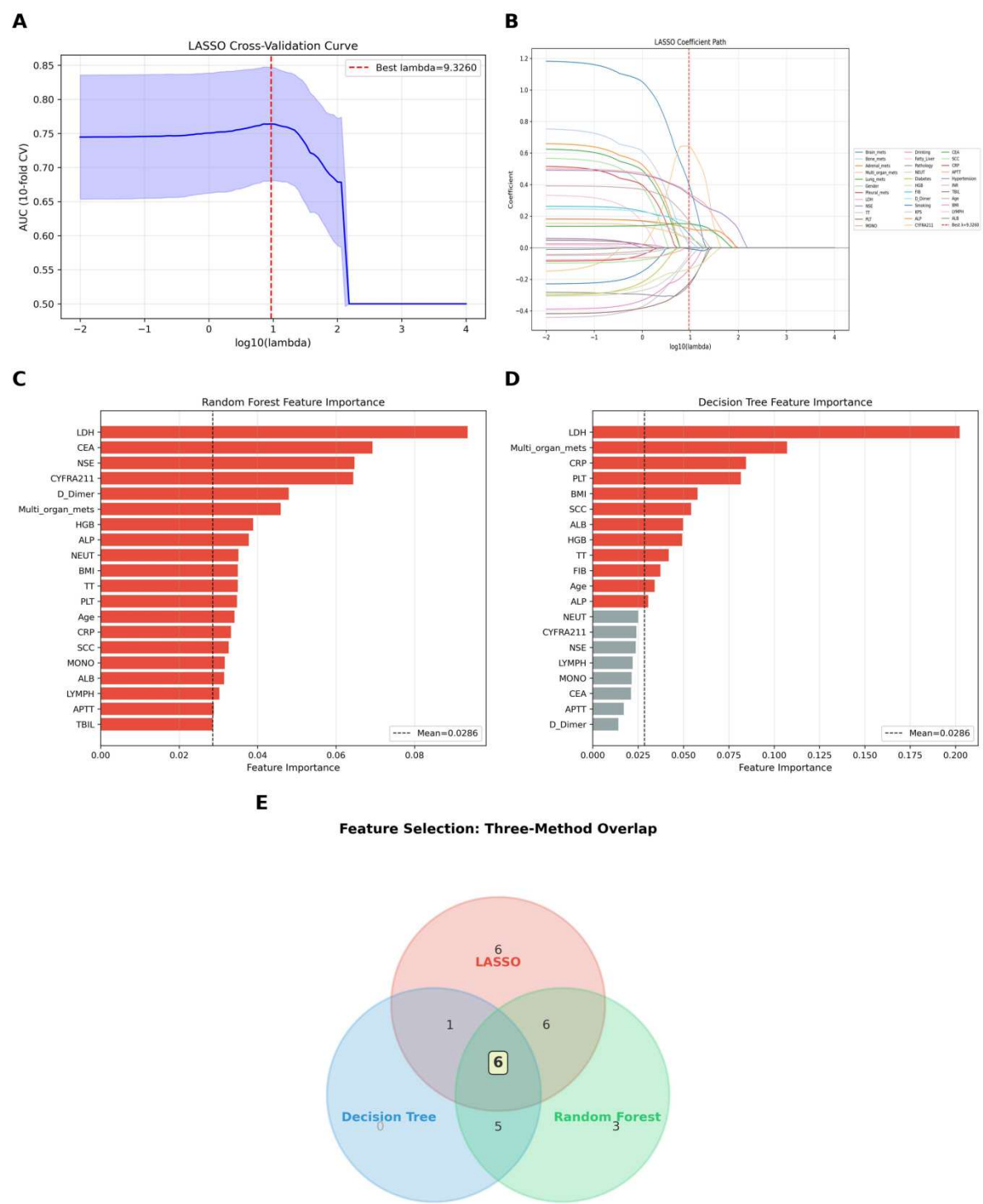


Figure 2. Feature selection and identification of consensus predictors. (A) Cross-validation curve for the LASSO method. (B) LASSO regression coefficient path plot for candidate

predictor variables. (C) Feature importance ranking derived from the random forest algorithm. (D) Feature importance ranking derived from the decision tree algorithm. (E) Venn diagram illustrating the intersection of the three algorithms, identifying six consensus predictors for model development.

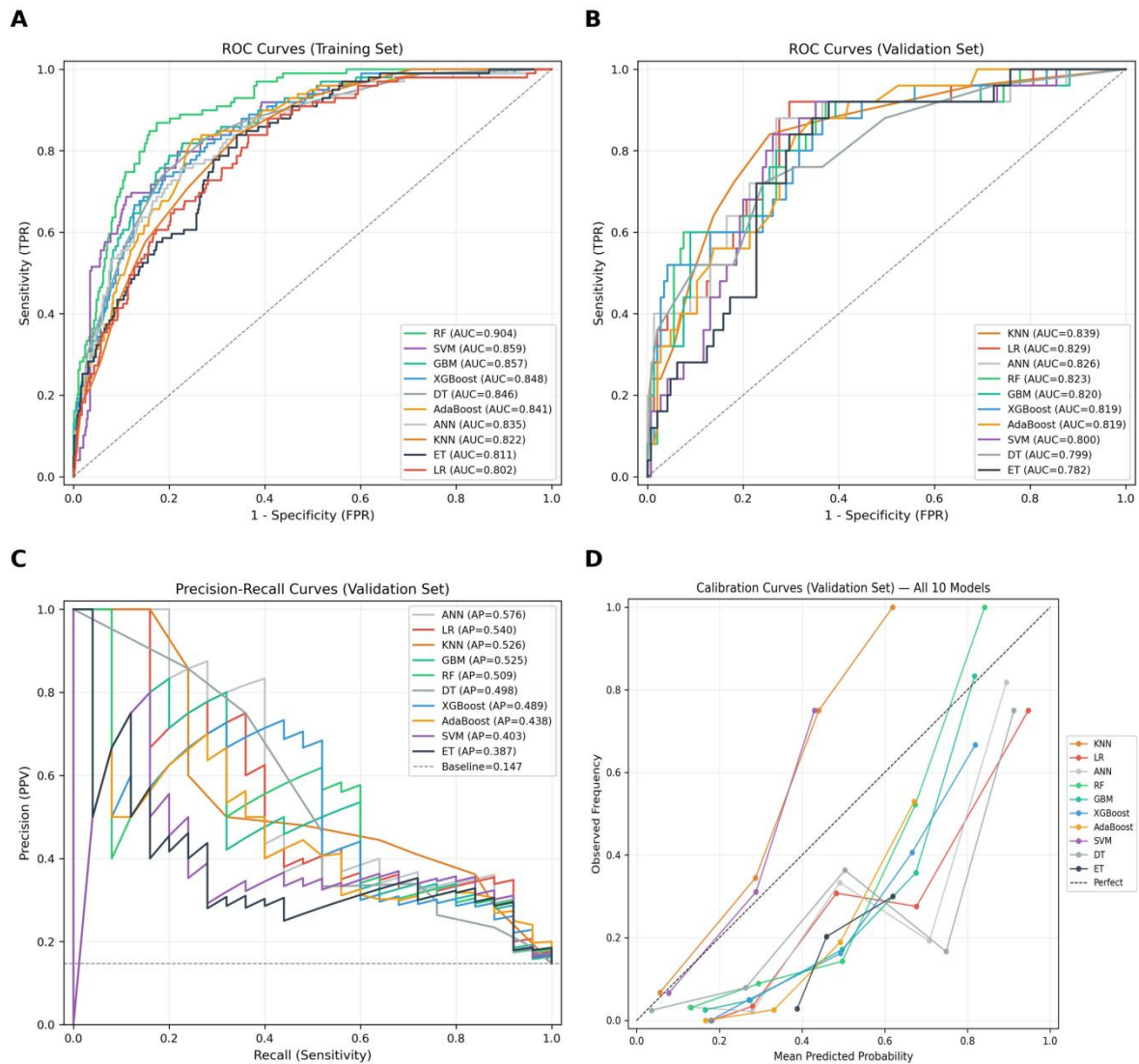


Figure 3. Performance curves of the ten machine learning models. (A) ROC curves for the training set. (B) ROC curves for the validation set. (C) PR curves for the validation set. (D) Calibration curves for the validation set. DT: Decision Tree; RF: Random Forest; ET: Extra Trees.

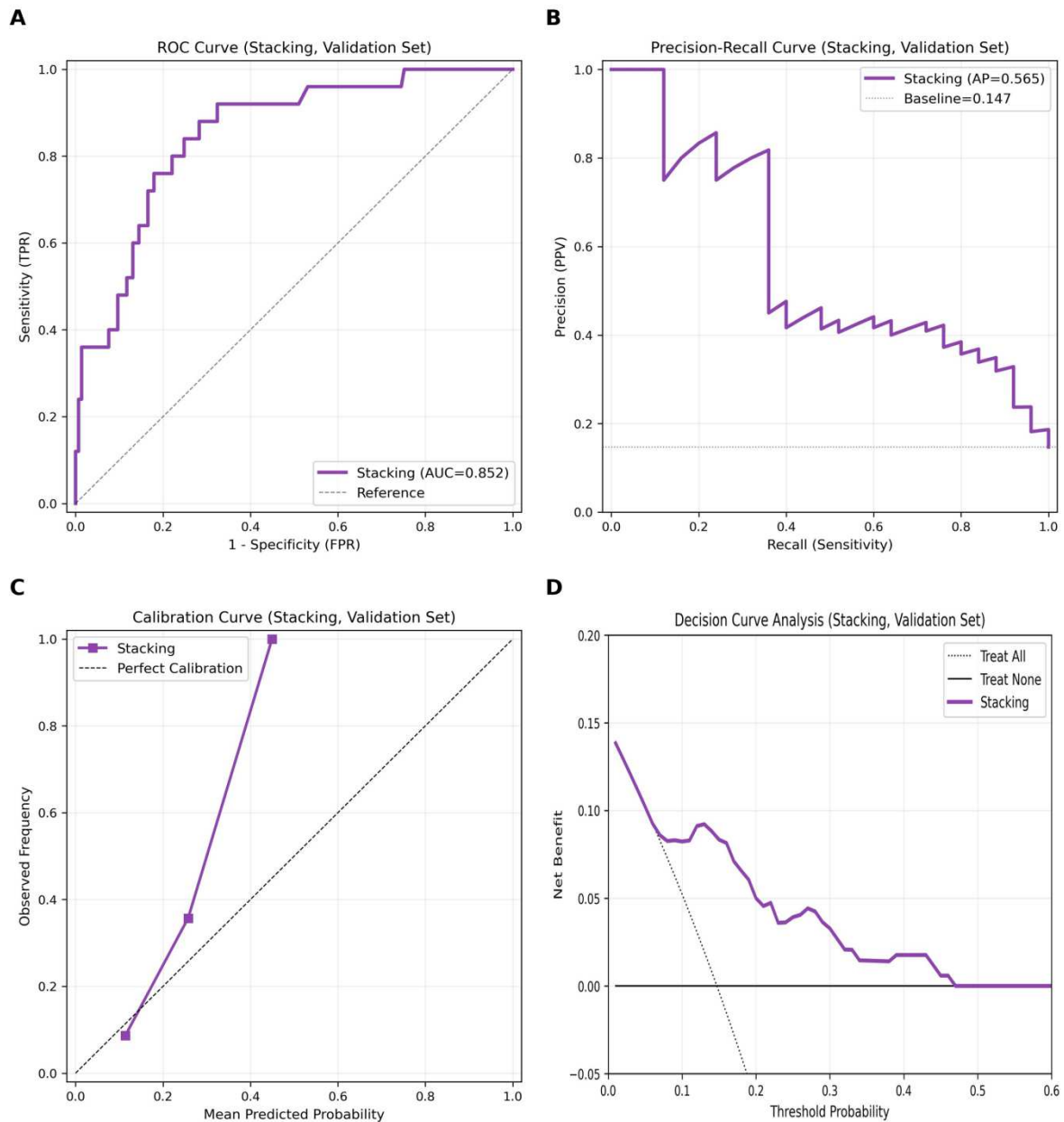


Figure 4. Performance curves of the Stacking ensemble model. (A) ROC curve. (B) PR curve. (C) Calibration curve. (D) DCA .

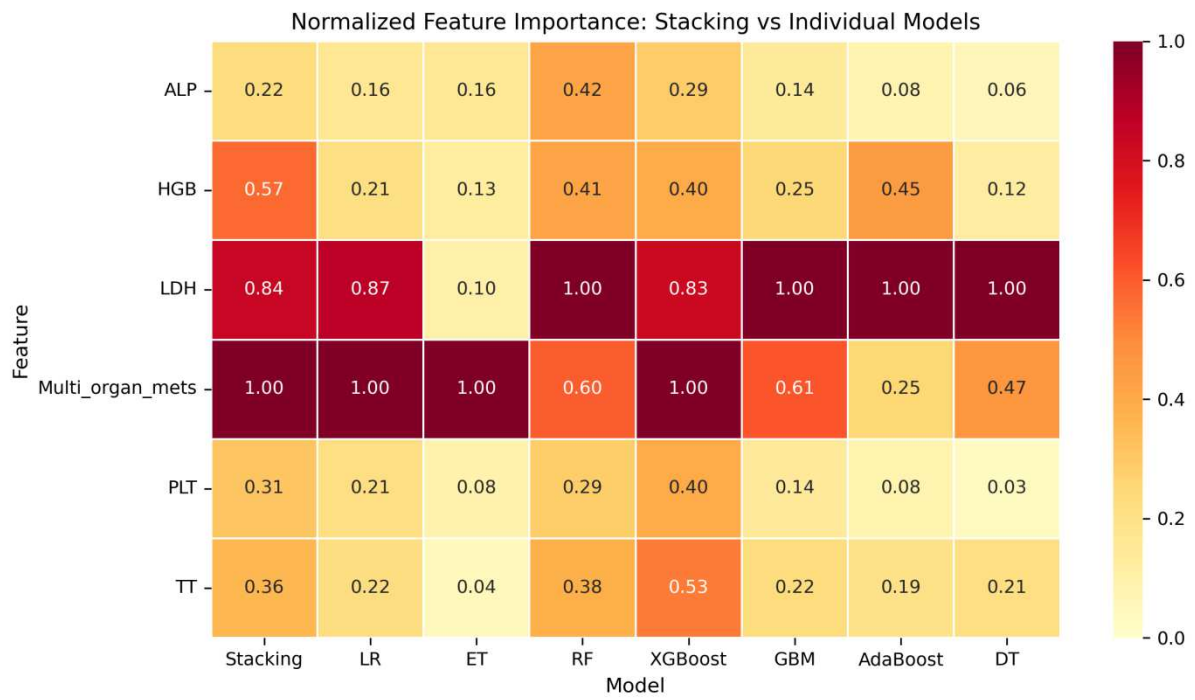


Figure 5. Normalized feature importance heatmap across all machine learning models.

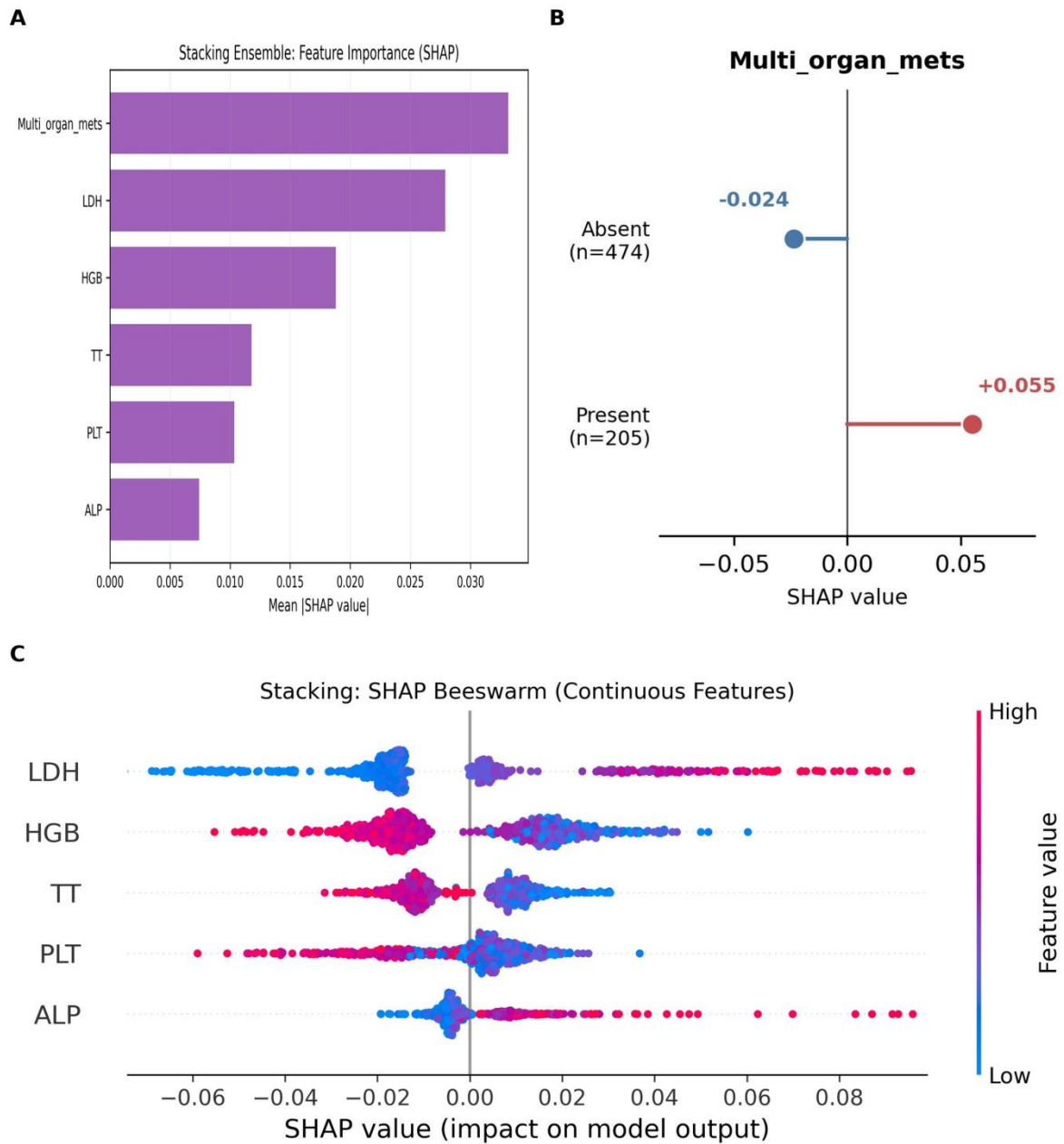


Figure 6. SHAP analysis of the Stacking ensemble model. (A) Global feature importance ranking based on mean absolute SHAP values. (B) SHAP value analysis for the binary feature multi-organ metastasis (Multi_organ_mets). (C) Beeswarm plot of SHAP values for continuous features.

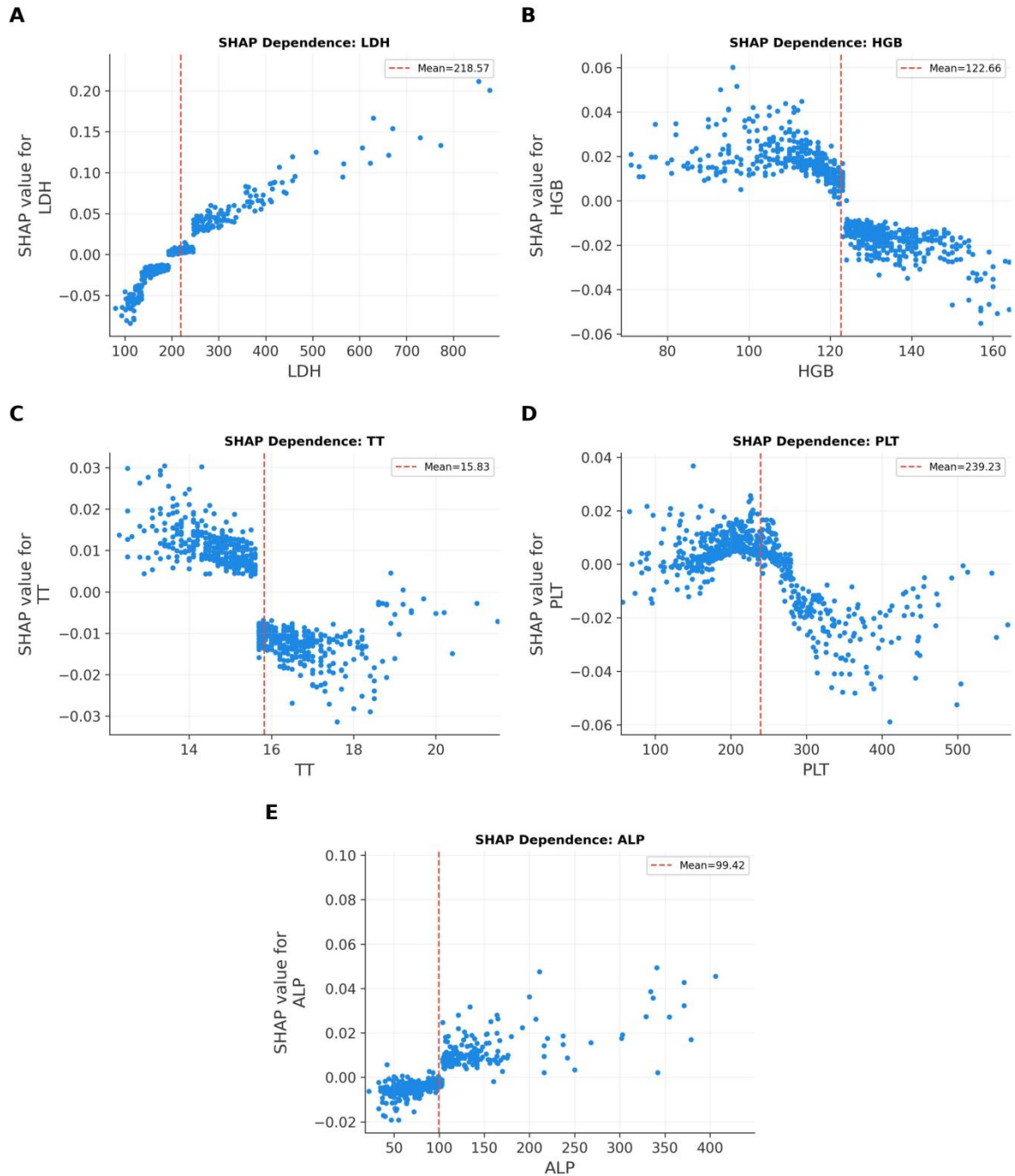


Figure 7. SHAP dependence plots for the Stacking ensemble model. Panels A–E show each predictor variable (x-axis) against its SHAP value (y-axis), reflecting each feature's contribution to model output. The red dashed line indicates the mean value of each feature in the study cohort.

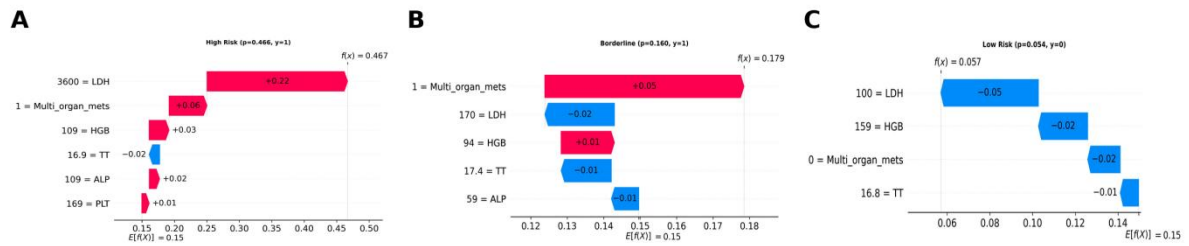


Figure 8. Individual- and cohort-level SHAP waterfall plots for the Stacking ensemble model.

Panels A–C show individual-level waterfall plots for three representative patients: (A) high-risk case, (B) borderline case, and (C) low-risk case. Each plot shows how individual feature values cumulatively push the prediction from the baseline $E[f(X)]$ to the final output $f(x)$. Red features raise the predicted probability of liver metastasis; blue features lower it.

Tables

Table 1. Baseline characteristics of patients with non-small cell lung cancer

Characteristic	Overall (n=849)	Training set (n=679)	Validation set (n=170)	P value
Age, median (Q1, Q3)	65.00 (59.00–72.00)	65.00 (59.00–71.00)	65.00 (58.00–72.00)	0.793
BMI, median (Q1, Q3)	23.31 (21.10–25.65)	23.31 (20.96–25.64)	23.66 (21.47–25.78)	0.501
KPS, median (Q1, Q3)	90.00 (80.00–90.00)	90.00 (80.00–90.00)	90.00 (80.00–90.00)	0.718
NEUT, median (Q1, Q3)	4.02 (2.86–5.61)	3.99 (2.91–5.56)	4.29 (2.71–5.71)	0.649
LYMPH, median (Q1, Q3)	1.40 (1.02–1.80)	1.39 (1.02–1.77)	1.42 (1.06–1.90)	0.274
MONO, median (Q1, Q3)	0.43 (0.34–0.56)	0.43 (0.34–0.56)	0.41 (0.32–0.54)	0.335
PLT, median (Q1, Q3)	227.00 (180.00–285.00)	225.00 (178.00–283.50)	237.50 (188.00–289.50)	0.221
HGB, median (Q1, Q3)	124.00 (112.00–134.00)	124.00 (112.00–134.00)	123.00 (112.00–134.00)	0.881
ALB, median (Q1, Q3)	37.00 (34.06–39.70)	36.95 (33.90–39.69)	37.10 (34.80–39.87)	0.348
TBIL, median (Q1, Q3)	11.30 (9.10–14.30)	11.30 (9.00–14.30)	11.30 (9.40–14.10)	0.579
ALP, median (Q1, Q3)	83.00 (66.00–105.00)	84.00 (66.00–105.00)	79.00 (67.00–100.00)	0.489
CEA, median (Q1, Q3)	5.36 (2.31–21.92)	5.44 (2.42–24.67)	4.54 (2.15–14.13)	0.122
CYFRA21-1, median (Q1,Q3)	5.01 (2.80–12.41)	5.21 (2.83–12.41)	4.24 (2.72–12.11)	0.189
NSE, median (Q1,Q3)	13.29 (10.90–18.90)	13.34 (11.00–18.88)	12.80 (10.50–18.94)	0.204
SCC, median (Q1,Q3)	1.21 (0.80–2.24)	1.24 (0.82–2.33)	1.20 (0.80–2.15)	0.656
D-dimer, median (Q1,Q3)	0.73 (0.38–1.64)	0.73 (0.39–1.64)	0.73 (0.33–1.63)	0.419
INR, median (Q1,Q3)	1.03 (0.98–1.09)	1.03 (0.98–1.10)	1.03 (0.98–1.08)	0.759
TT, median (Q1,Q3)	15.60 (14.70–16.50)	15.60 (14.70–16.60)	15.50 (14.60–16.40)	0.126

APTT, median (Q1,Q3)	26.90 (25.10–29.10)	26.80 (25.10–29.00)	27.10 (25.00–30.10)	0.213
FIB, median (Q1,Q3)	4.03 (3.05–4.82)	4.07 (3.09–4.90)	3.92 (2.89–4.70)	0.165
CRP, median (Q1,Q3)	11.28 (3.76–31.04)	8.38 (1.86–28.09)	8.73 (1.37–23.46)	0.592
LDH, median (Q1,Q3)	180.00 (153.00–225.00)	182.00 (154.00–226.25)	171.00 (145.50–216.00)	0.051
Gender				0.014
Male	541 (63.7%)	447 (65.8%)	94 (55.3%)	
Female	308 (36.3%)	232 (34.2%)	76 (44.7%)	
Smoking history				0.169
Yes	516 (60.8%)	421 (62.0%)	95 (55.9%)	
No	329 (38.8%)	258 (38.0%)	75 (44.1%)	
Alcohol consumption history				0.837
Yes	368 (43.3%)	296 (43.6%)	72 (42.4%)	
No	471 (55.5%)	383 (56.4%)	98 (57.6%)	
Hypertension				1
Yes	348 (41.0%)	278 (40.9%)	70 (41.2%)	
No	493 (58.1%)	401 (59.1%)	100 (58.8%)	
Diabetes				0.563
Yes	164 (19.3%)	128 (18.9%)	36 (21.2%)	
No	557 (65.6%)	551 (81.1%)	134 (78.8%)	
Fatty liver				0.537
Yes	63 (7.4%)	48 (7.1%)	15 (8.8%)	
No	786 (92.6%)	631 (92.9%)	155 (91.2%)	
Pathological classification				0.996
Adenocarcinoma	552 (65.0%)	442 (65.1%)	110 (64.7%)	
Squamous cell carcinoma	297 (35.0%)	237 (34.9%)	60 (35.3%)	
Liver metastasis				1
Yes	124 (14.6%)	99 (14.6%)	25 (14.7%)	
No	725 (85.4%)	580 (85.4%)	145 (85.3%)	
Bone metastasis				0.092
Yes	231 (27.2%)	194 (28.6%)	37 (21.8%)	
No	618 (72.8%)	485 (71.4%)	133 (78.2%)	
Brain metastasis				0.897
Yes	120 (14.1%)	97 (14.3%)	23 (13.5%)	
No	729 (85.9%)	582 (85.7%)	147 (86.5%)	
Adrenal metastasis				0.131

Yes	60 (7.1%)	53 (7.8%)	7 (4.1%)	
No	789 (92.9%)	626 (92.2%)	163 (95.9%)	
Pleural metastasis / malignant pleural effusion				0.251
Yes	169 (19.9%)	141 (20.8%)	28 (16.5%)	
No	680 (80.1%)	538 (79.2%)	142 (83.5%)	
Pulmonary metastasis				0.675
Yes	223 (26.3%)	181 (26.7%)	42 (24.7%)	
No	626 (73.7%)	498 (73.3%)	128 (75.3%)	
Multi-organ metastasis				0.391
Yes	250 (29.4%)	205 (30.2%)	45 (26.5%)	
No	599 (70.6%)	474 (69.8%)	125 (73.5%)	

Table 2. Performance of ten machine learning models in predicting liver metastasis in patients with stage III/IV non-small cell lung cancer

MLs	AUC (95% CI)	AUPRC (95% CI)	Accuracy	Sensitivity	Specificity	F1 score	PPV	NPV
Training set								
RF	0.904 (0.877–0.930)	0.632 (0.542–0.720)	0.683 (0.648–0.717)	0.929 (0.878–0.978)	0.641 (0.601–0.679)	0.461 (0.400–0.519)	0.307 (0.255–0.357)	0.982 (0.968–0.995)
SVM	0.859 (0.818–0.896)	0.497 (0.403–0.608)	0.676 (0.641–0.710)	0.859 (0.787–0.925)	0.645 (0.605–0.684)	0.436 (0.374–0.495)	0.292 (0.241–0.344)	0.964 (0.945–0.982)
GBM	0.857 (0.818–0.892)	0.520 (0.427–0.625)	0.714 (0.680–0.747)	0.848 (0.778–0.913)	0.691 (0.652–0.729)	0.464 (0.404–0.525)	0.319 (0.267–0.375)	0.964 (0.947–0.981)
XGBoost	0.848 (0.810–0.884)	0.520 (0.423–0.612)	0.713 (0.677–0.744)	0.828 (0.759–0.900)	0.693 (0.655–0.730)	0.457 (0.392–0.518)	0.315 (0.260–0.371)	0.959 (0.941–0.977)
DT	0.846 (0.805–0.883)	0.467 (0.372–0.560)	0.666 (0.630–0.701)	0.879 (0.816–0.939)	0.629 (0.590–0.667)	0.434 (0.374–0.495)	0.288 (0.239–0.339)	0.968 (0.951–0.984)
AdaBoost	0.841 (0.802–0.876)	0.459 (0.369–0.562)	0.728 (0.695–0.757)	0.838 (0.768–0.904)	0.709 (0.673–0.742)	0.473 (0.409–0.534)	0.329 (0.275–0.384)	0.963 (0.945–0.979)
ANN	0.835 (0.793–0.876)	0.514 (0.422–0.616)	0.767 (0.733–0.797)	0.758 (0.674–0.842)	0.769 (0.734–0.804)	0.487 (0.418–0.550)	0.359 (0.296–0.423)	0.949 (0.929–0.970)
KNN	0.822 (0.782–0.858)	0.417 (0.329–0.521)	0.754 (0.722–0.785)	0.707 (0.620–0.796)	0.762 (0.728–0.795)	0.456 (0.390–0.519)	0.337 (0.276–0.402)	0.938 (0.917–0.959)
ET	0.811 (0.766–0.852)	0.469 (0.371–0.571)	0.714 (0.677–0.747)	0.778 (0.691–0.856)	0.703 (0.664–0.739)	0.443 (0.374–0.504)	0.309 (0.251–0.366)	0.949 (0.927–0.969)
LR	0.802 (0.757–0.845)	0.427 (0.338–0.541)	0.635 (0.598–0.670)	0.869 (0.804–0.936)	0.595 (0.555–0.632)	0.410 (0.351–0.468)	0.268 (0.222–0.318)	0.964 (0.945–0.983)
Validation set								

KNN	0.839 (0.751–0.913)	0.526 (0.358–0.704)	0.806 (0.747–0.871)	0.720 (0.542–0.893)	0.821 (0.753–0.884)	0.522 (0.373–0.658)	0.409 (0.270–0.561)	0.944 (0.905–0.983)
LR	0.829 (0.744–0.911)	0.540 (0.369–0.730)	0.688 (0.618–0.759)	0.920 (0.806–1.000)	0.648 (0.569–0.725)	0.465 (0.340–0.585)	0.311 (0.211–0.423)	0.979 (0.948–1.000)
ANN	0.826 (0.735–0.912)	0.576 (0.406–0.750)	0.771 (0.706–0.835)	0.720 (0.542–0.895)	0.779 (0.710–0.849)	0.480 (0.348–0.615)	0.360 (0.236–0.500)	0.942 (0.896–0.982)
RF	0.823 (0.730–0.909)	0.509 (0.335–0.707)	0.694 (0.624–0.765)	0.800 (0.640–0.950)	0.676 (0.595–0.757)	0.435 (0.313–0.562)	0.299 (0.200–0.412)	0.951 (0.906–0.990)
GBM	0.820 (0.727–0.905)	0.525 (0.353–0.717)	0.735 (0.671–0.806)	0.760 (0.583–0.923)	0.731 (0.662–0.810)	0.458 (0.325–0.588)	0.328 (0.217–0.455)	0.946 (0.900–0.983)
AdaBoost	0.819 (0.740–0.893)	0.438 (0.287–0.657)	0.724 (0.659–0.794)	0.720 (0.542–0.900)	0.724 (0.649–0.801)	0.434 (0.293–0.564)	0.310 (0.196–0.434)	0.938 (0.889–0.981)
XGBoost	0.819 (0.726–0.901)	0.489 (0.331–0.714)	0.712 (0.647–0.782)	0.680 (0.517–0.857)	0.717 (0.647–0.793)	0.410 (0.278–0.539)	0.293 (0.185–0.413)	0.929 (0.881–0.972)
SVM	0.800 (0.707–0.884)	0.403 (0.259–0.601)	0.724 (0.653–0.788)	0.840 (0.690–0.964)	0.703 (0.622–0.780)	0.472 (0.333–0.589)	0.328 (0.219–0.446)	0.962 (0.922–0.991)
DT	0.799 (0.700–0.896)	0.498 (0.318–0.700)	0.653 (0.576–0.729)	0.760 (0.600–0.923)	0.634 (0.553–0.714)	0.392 (0.273–0.509)	0.264 (0.172–0.366)	0.939 (0.888–0.981)
ET	0.782 (0.698–0.865)	0.387 (0.253–0.580)	0.712 (0.641–0.782)	0.720 (0.545–0.880)	0.710 (0.633–0.784)	0.424 (0.289–0.546)	0.300 (0.192–0.419)	0.936 (0.892–0.975)

Table 3. Performance comparison of the Stacking ensemble model, KNN, and AdaBoost in the validation set

Model	AUC (95% CI)	AUPRC (95% CI)	Accuracy	Sensitivity	Specificity	F1 score	PPV	NPV
Stacking	0.852 (0.769–0.921)	0.565 (0.399–0.745)	0.806 (0.747–0.871)	0.760 (0.591–0.929)	0.814 (0.750–0.882)	0.535 (0.387–0.675)	0.413 (0.276–0.561)	0.952 (0.912–0.984)
KNN	0.839 (0.751–0.913)	0.526 (0.358–0.704)	0.806 (0.747–0.871)	0.720 (0.542–0.893)	0.821 (0.753–0.884)	0.522 (0.373–0.658)	0.409 (0.270–0.561)	0.944 (0.905–0.983)
AdaBoost	0.819 (0.740–0.893)	0.438 (0.287–0.657)	0.724 (0.659–0.794)	0.720 (0.542–0.900)	0.724 (0.649–0.801)	0.434 (0.293–0.564)	0.310 (0.196–0.434)	0.938 (0.889–0.981)

Supplementary Files

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

- [SupplementaryTable1.docx](#)