



OPEN

Development and validation of machine learning models for diagnosis and prognosis of lung adenocarcinoma, and immune infiltration analysis

Lin Lin & Yongxia Bao

The aim of our study was to develop robust diagnostic and prognostic models for lung adenocarcinoma (LUAD) using machine learning (ML) techniques, focusing on early immune infiltration. Feature selection was performed on The Cancer Genome Atlas (TCGA) data using least absolute shrinkage and selection Operator (LASSO), random forest (RF), and support vector machine (SVM) algorithms. Six ML algorithms were employed to construct the diagnostic models, which were evaluated through receiver operating characteristic (ROC) curves, precision-recall curves (PRC), and classification error (CE), and validated on the GSE7670 dataset. Additionally, a lasso cox prognostic model was built on the TCGA-LUAD dataset and externally validated using independent Gene Expression Omnibus datasets (GSE30219, GSE31210, GSE50081, and GSE37745). Single-sample gene set enrichment analysis (ssGSEA) was performed to assess immune cell infiltration in stage I LUAD samples, revealing significant differences in immune cell types. These findings demonstrate a positive correlation between immune infiltration in stage I LUAD and Th2 cells, Tcm cells, and T helper cells, while a negative correlation was observed with Macrophages, Eosinophils, and Tem cells. These insights provide novel perspectives for clinical diagnosis and treatment of LUAD.

Keywords Lung adenocarcinoma, Machine learning, Diagnosis, Immune infiltration

Lung cancer, characterized by its high incidence and mortality rates, has imposed a significant disease burden worldwide¹. Non-small cell lung cancer accounts for 85% of lung cancer cases, with lung adenocarcinoma (LUAD) being the predominant subtype². Given that computed tomography often detects uncertain nodules, the identification of biomarkers that can accurately distinguish between malignant and benign nodules could lead to more effective diagnosis of LUAD³. Furthermore, lung cancer is acknowledged for its high heterogeneity. Its molecular classification precisely categorizes tumors into informative subtypes, informing prognosis and enabling more targeted, personalized treatments⁴. Hence, novel biomarkers are urgently needed to enhance LUAD diagnosis and prognosis, surpassing current diagnostic and therapeutic methods.

Machine learning (ML) constructs predictive models through deep analysis and learning of large-scale datasets, thereby uncovering underlying patterns and regularities⁵. In the biomedical domain, its robust data mining and computational capabilities offer solutions to complex problems, particularly when dealing with large-scale and high-dimensional patient data⁶. ML models can predict disease progression based on such data, providing decision support for clinicians and facilitating the development of personalized treatment plans⁷. Furthermore, their application in areas like medical image analysis and genetic sequencing data processing has significantly improved diagnostic accuracy and research efficiency⁸. ML in biomedicine holds great potential for medical innovation. However, ML models for precise early lung nodule classification are scarce, with few studies using ≥ 6 ML methods. In LUAD early diagnosis, developing a highly sensitive and specific model is challenging. More research is needed to explore ML algorithms for lung nodule classification and early LUAD diagnosis, enhancing diagnostic accuracy for better patient outcomes.

Department of Respiratory and Critical Care Medicine, The Second Affiliated Hospital of Harbin Medical University, No. 148, Health Care Road, Nangang District, Harbin 150086, Heilongjiang, People's Republic of China. email: hydbaoxyongxia@163.com

In this study, we utilized LUAD data from The Cancer Genome Atlas (TCGA) to identify stage I markers via least absolute shrinkage and selection operator (LASSO), random forest (RF), and support vector machine (SVM), constructing diagnostic models with six ML algorithms. Model performances were evaluated using receiver operating characteristic (ROC) curves, precision-recall curves (PRC), and classification error (CE), validated on the GSE6768 dataset. A lasso cox prognosis model was developed and assessed through K-M, time-dependent ROC, and gene expression analysis, validated on four Gene Expression Omnibus datasets. Single sample gene set enrichment analysis (ssGSEA) evaluated immune cell infiltration in stage I LUAD, with significant immune cell types identified by Wilcoxon and LASSO tests. Our aim is to enhance clinical decision-making accuracy and efficiency through precise LUAD diagnostic and prognostic models, while exploring early-stage immune infiltration mechanisms. The study flowchart is in Fig. 1.

Results

Feature selection

Regarding the mRNA dataset, the RF model's performance was evaluated using OOB estimation, yielding a low error rate of 1.44%, indicating high classification accuracy on the training set. The confusion matrix showed a class error rate of 5.56% for the normal class and 0.58% for the stage I class, emphasizing effective differentiation. To identify influential features, we used varImpPlot to visualize the top 30 variables by mean decrease gini (Fig. 2A). Figure 2B illustrates the change in CE as the number of trees varies, showing the effect of tree count on model performance. In LASSO cross-validation (Fig. 2C), increasing λ initially reduced error before rising, balancing underfitting and overfitting. Figure 2D shows how λ 's logarithmic increase decreases model complexity, shrinks

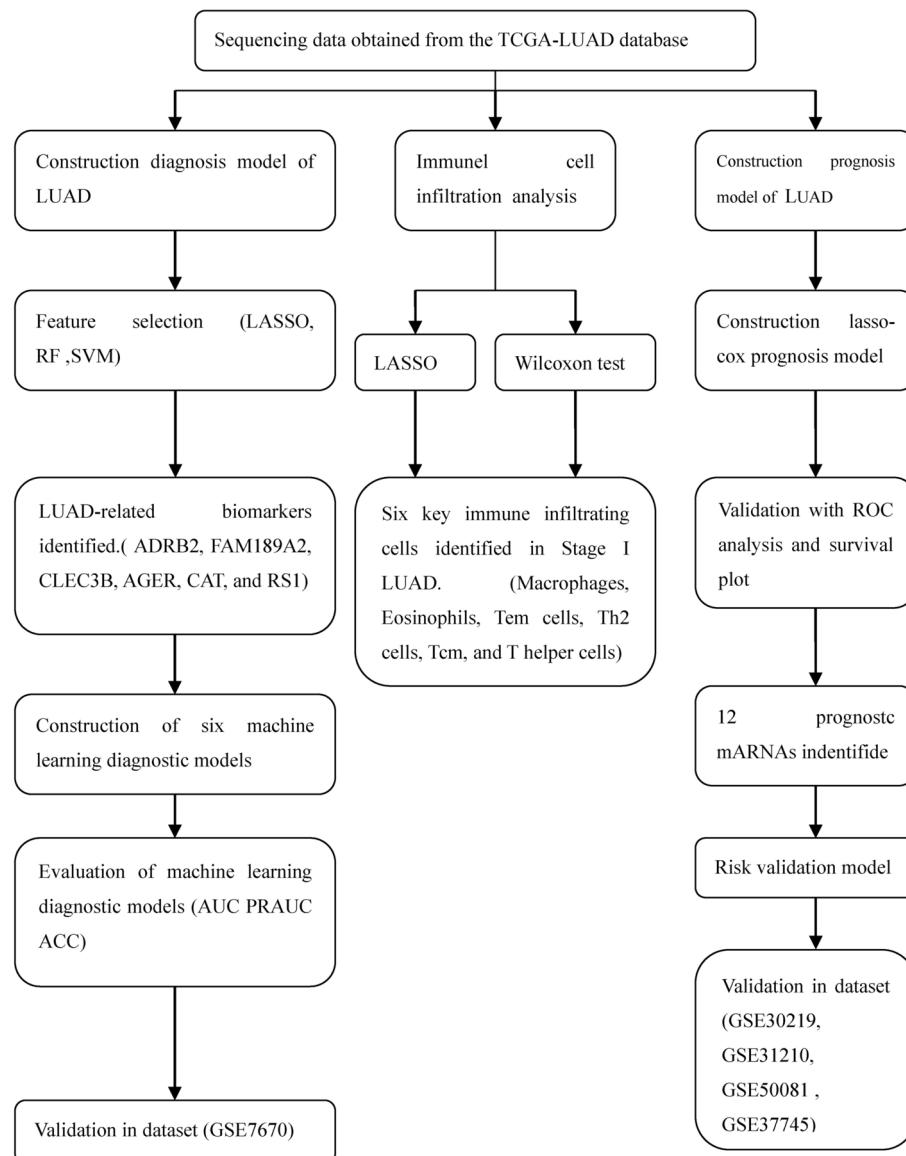


Fig. 1. Flow chart of the study.

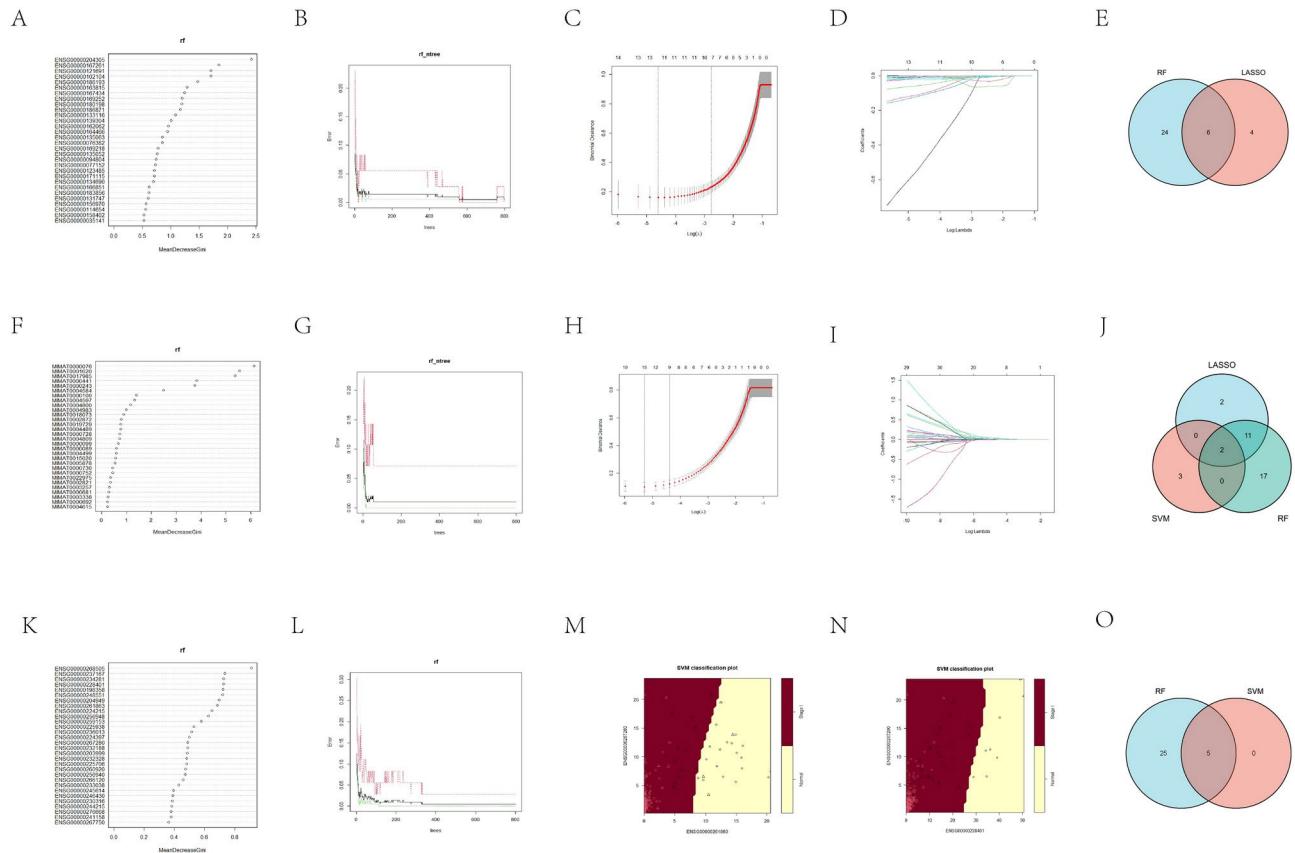


Fig. 2. Feature selection. **(A)** Feature selection of mRNAs using RF. RF analysis classifies the levels of importance of mRNAs. The X-axis represents RF mean decrease in accuracy and gini. The Y-axis represents a ranking of variables was obtained from RF on the basis of mean decrease in accuracy and gini coefficients. **(B)** Performance (error rate) per number of tree generated by the RF algorithm. The x-axis shows the number of trees, Y-axis is the error rate given by RF (out of bag error estimation from 800 trees). Black, red and green lines correspond to the gross distribution, stage I LUAD distribution and adjacent normal tissue distribution, respectively. **(C)** Lasso analysis results of mRNAs. The lower horizontal axis represents lambda value, and the upper horizontal axis scale represents the number of variables in the lasso model, the regression coefficient (x) of which is not 0. **(D)** The trajectory of each independent variable, the horizontal axis represents the log value of the independent variable lambda, and the vertical axis represents the coefficient of the independent variable. **(E)** A Venn diagram of two feature selection methods. **(F)** Feature selection of miRNAs using RF. RF analysis classifies the levels of importance of miRNAs. **(G)** Performance (error rate) per number of tree generated by the RF algorithm. **(H)** Lasso analysis results of miRNAs. **(I)** The trajectory of each independent variable, the horizontal axis represents the log value of the independent variable lambda, and the vertical axis represents the coefficient of the independent variable. **(J)** A Venn diagram of three feature selection methods. **(K)** Feature selection of lncRNAs using RF. RF analysis classifies the levels of importance of lncRNAs. **(L)** Performance (error rate) per number of tree generated by the RF algorithm. **(M, N)** Two-dimensional projections using four of the SVM features. The support vectors and non-support vectors are denoted with triangles and circles, respectively. Red areas represent predicted positive regions, yellow areas represent predicted negative regions. **(O)** A Venn diagram of two feature selection methods.

coefficients, and may converge some to zero. `lambda.min` (0.01005025) offers optimal training but with high complexity, while `lambda.1se` (0.06281407) provides a concise model with near-optimal performance, reducing overfitting risk. Thus, we selected the `lambda.1se` model as final. By intersecting the two predictions, six optimal mRNAs (ADRB2, FAM189A2, CLEC3B, AGER, CAT, and RS1) were obtained. The Venn diagram is presented in Fig. 2E. Figure 2F,G displays the results of miRNA variable selection using RF, while Fig. 2H,I presents the miRNA variable selection outcomes from the LASSO algorithm. After intersecting, miR-21-5p and miR-29c-3p were selected as the final features. The Venn diagram is provided in Fig. 2J. Figure 2K,L show cases the lncRNA variable selection results from RF, and Fig. 2M,N demonstrates the lncRNA variable selection outcomes using the SVM algorithm. After intersecting, five lncRNAs (BX2-AS1, PTPRD-AS1, LINC01996, AL590226.1 and AC084030.1) were obtained as the results of feature selection. The corresponding Venn diagram is presented in Fig. 2O.

Construction and validation of six machine learning diagnostic models

To classify lung nodules, we employed six traditional machine learning algorithms, including k-nearest neighbor (KNN), naive bayes (NB), RF, decision tree (DT), SVM, and eXtreme Gradient Boosting (XGBoost), on the training set and evaluated their performance using ten-fold cross-validation (Table 1). For mRNA data, the area under the ROC curve AUC ranged from 0.9844472 to 1 (Fig. 3A), the AUC values of the precision-recall curve fell between 0.9942892 and 1 (Fig. 3B), and the average prediction accuracy exceeded 95% (Fig. 3C). When applied to the test set, the AUC values of these algorithms remained between 0.9337719 and 1, with PRC AUC values ranging from 0.9881881 to 1, and the prediction accuracy also exceeded 95%, see Table 2 for detailed results. For miRNA data, the AUC values of the ROC curve for the six algorithms on the training set ranged from 0.9469609 to 1 (Fig. 3D), the AUC values of the PRC fell between 0.9824365 and 1 (Fig. 3E), and the average prediction accuracy exceeded 90% (Fig. 3F). On the test set, the AUC values of the ROC curve for these algorithms were between 0.9665007 and 0.9859375, the AUC values of the PRC were between 0.9806327 and 0.9952889, and the prediction accuracy also exceeded 90%, see Table 3 for detailed results. For lncRNA data, the AUC values of the ROC curve for the six algorithms on the training set ranged from 0.9860368 to 1 (Fig. 3G), the AUC values of the PRC fell between 0.9942586 and 1 (Fig. 3H), and the average prediction accuracy exceeded 95% (Fig. 3I). On the test set, the AUC values of the ROC curve for these algorithms were between 0.9833333 and 1, the AUC values of the PRC were between 0.9947368 and 1, and the prediction accuracy also exceeded 95%, see Table 4 for detailed results. In summary, these six classifiers exhibited excellent performance in the task of lung nodule classification.

Comparison of six ML diagnostic models

This study evaluated various mRNA ML models and found that KNN, RF and SVM performed exceptionally well, achieving AUC and PRAUC scores of 1.000000 on both training and testing sets, emphasizing their advantages in data fitting and generalization. The NB model, while slightly lower with AUC (0.9973684) and PRAUC (0.9997300) on the testing set, still showed commendable performance, indicating its potential for classification tasks. In contrast, XGBoost and DT maintained high AUC scores of 0.9364035 and 0.9337719, respectively, and PRAUC scores of 0.9884611 and 0.9881881, respectively, showcasing their predictive capabilities. In summary, the study recommends prioritizing KNN, RF, and SVM, while recognizing NB as a viable alternative.

Ablation studies for variable importance assessment

To assess feature contributions, we conducted ablation studies by removing each of six mRNA features (ADRB2, AGER, CLEC3, FAM189A2, CAT, RS1) and retraining models. We evaluated model performance using AUC on training and test sets. The baseline model, with all features, achieved near-perfect AUCs on the training set and high AUCs on the test set, indicating excellent generalization.

Removing ADRB2 had minimal impact, except for XGBoost, which improved from 0.9364035 to 0.9716667 AUC on the test set. AGER ablation caused minimal changes, except for DT, which saw a decrease from 0.9337719 to 0.9131356 AUC on the test set. XGBoost improved to 0.9822442 AUC.

Removing CLEC3 moderately affected models, with XGBoost showing the most significant improvement (from 0.9364035 to 0.9802083 AUC). DT still showed a decrease. FAM189A2 ablation had minimal effects on learners with perfect training AUCs but caused slight decreases in XGBoost and NB test AUCs. DT also showed a moderate decrease. Deleting CAT had a pronounced positive effect, especially on DT and XGBoost, with improvements to 0.9800000 and 0.9877381 AUC, respectively. Ablating RS1 resulted in mixed effects, with some learners showing minor changes and others, like XGBoost, experiencing moderate improvements. DT showed a relatively large decrease, indicating RS1's importance. Detailed ablation study data and model error rates (mostly < 5%) are in Supplementary Table S1 and Supplementary Figure S2, respectively.

Variable	Description	Categories/Values
Data Source		TCGA, GEO
Number of Samples in TCGA	Total	594 (LUAD: 535, Normal: 59)
LUAD Stages in TCGA	Stage I 294	
	Stage II 123	
	Stage III 84	
	Stage IV 26	
GEO Dataset Accession Numbers	GSE7670	27 paired samples
GSE30219	14 normal, 293 cancer	
GSE31210	226 Stage I-II cancers	
GSE50081	127 cancers	
GSE37745	106 cancers	
RNA-seq Platform	TCGA	Level 3 data
	GEO	Affymetrix Human Genome U133 Plus 2.0 Array

Table 1. Baseline demographic and clinical information.

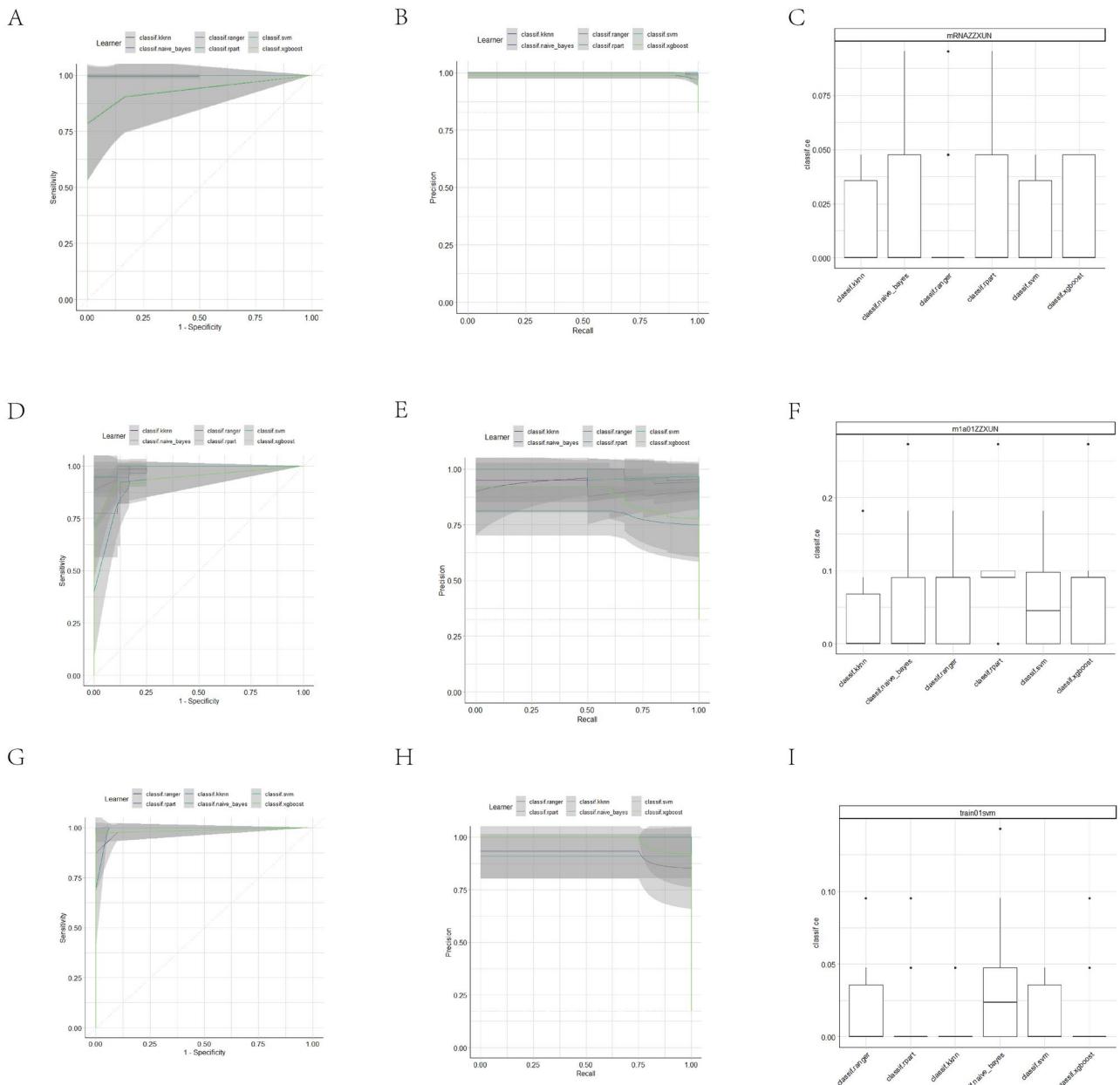


Fig. 3. Performance of machine learning models was evaluated by AUC–ROC and AUC–PRC computed from cross-validation. (A) Comparison of performance of six ML models of mRNAs via ROC AUC. (B) Comparison of performance of six ML models of mRNAs via PRC AUC. (C) Boxplot showing the classification performance with the six ML methods of mRNAs. Abscissa represents six ML algorithm. The ordinate axis represents the CE for the model. (D) Comparison of performance of six ML models of miRNAs via ROC AUC. (E) Comparison of performance of six ML models of miRNAs via PRC AUC. (F) Boxplot showing the classification performance with the six ML methods of miRNAs. (G) Comparison of performance of six ML models of lncRNAs via ROC AUC. (H) Comparison of performance of six ML models of lncRNAs via PRC AUC. (I) Boxplot showing the classification performance with the six ML methods of lncRNAs.

Validation the diagnostic efficacy of biomarkers for early-stage LUAD

The diagnostic validity of ADRB2, FAM189A2, CLEC3B, AGER, CAT, and RS1 genes for LUAD was verified through ROC analysis in the TCGA-LUAD dataset. The heatmap depicts the expression patterns of six mRNAs between LUAD and normal tissues in the TCGA database (Fig. 4A). The results of ROC analysis demonstrated that ADRB2 (AUC = 0.994), CLEC3B (AUC = 0.994), AGER (AUC = 0.998), CAT (AUC = 0.994), and RS1 (AUC = 0.990) exhibited AUC values exceeding 0.98, indicating high specificity and sensitivity (Fig. 4B). To further validate the diagnostic value of these genes, the microarray dataset GSE7670 from the GPL570 platform was selected as an external validation set. The heatmap depicts the expression patterns of six mRNAs between LUAD and normal tissues in the GSE 7670 database (Fig. 4C). In the validation set, ADRB2 (AUC = 0.988),

Learner	auc_train	auc_test	prauc_train	prauc_test
classif.kknn	1.0000000	1.0000000	1.0000000	1.0000000
classif.naive_bayes	1.0000000	0.9973684	1.0000000	0.9997300
classif.ranger	1.0000000	1.0000000	1.0000000	1.0000000
classif.rpart	0.9844472	0.9337719	0.9942892	0.9881881
classif.svm	1.0000000	1.0000000	1.0000000	1.0000000
classif.xgboost	0.9979483	0.9364035	0.9996251	0.9884611

Table 2. Evaluation of mRNA machine learning models. PRAUC, the area under the precision-recall curve; AUC, the area under the ROC curve.

Learner	auc_train	auc_test	prauc_train	prauc_test
classif.kknn	0.9991565	0.9803571	0.9998591	0.9869748
classif.naive_bayes	0.9694598	0.9803646	0.9905558	0.9906283
classif.ranger	1.0000000	0.9665007	1.0000000	0.9812384
classif.rpart	0.9469609	0.9704365	0.9824365	0.9806327
classif.svm	0.9905906	0.9859375	0.9982290	0.9952889
classif.xgboost	0.9808584	0.9704365	0.9937417	0.9806327

Table 3. Evaluation of miRNA machine learning models. PRAUC, the area under the precision-recall curve; AUC, the area under the ROC curve. ACC, accuracy.

Learner	auc_train	auc_test	prauc_train	prauc_test
classif.kknn	0.9998633	0.9985294	0.9999711	0.9996638
classif.naive_bayes	0.9995248	0.9985294	0.9999014	0.9996638
classif.ranger	1.0000000	1.0000000	1.0000000	1.0000000
classif.rpart	0.9860368	0.9833333	0.9942586	0.9947368
classif.svm	1.0000000	1.0000000	1.0000000	1.0000000
classif.xgboost	0.9997312	0.9833333	0.9999442	0.9947368

Table 4. Evaluation of lncRNA machine learning models. PRAUC, the area under the precision-recall curve; AUC, the area under the ROC curve.

FAM189A2 (AUC=0.988), AGER (AUC=1.000), and CAT (AUC=0.961) demonstrated AUC values exceeding 0.96, also exhibiting high specificity and sensitivity. However, RS1 (AUC=0.826) exhibited a relatively low predictive accuracy in the validation set (Fig. 4D). Overall, the results indicate that ADRB2, FAM189A2, AGER, CAT, and RS1 genes exhibited high diagnostic validity in both the TCGA-LUAD dataset and the external validation set.

Immune infiltration ssGSEA analysis

The TCGA-LUAD dataset encompassed 294 stage I LUAD samples and 59 normal samples. To accurately distinguish the significant differences in immune cell infiltration between stage I LUAD samples and normal samples, two distinct algorithms, namely LASSO regression and wilcoxon test, were employed. The results of LASSO regression (as shown in the Fig. 4E-G) revealed that seven immune cells were significantly included, exhibiting distinct infiltration patterns in the early stages of LUAD compared to normal samples. Additionally, the wilcoxon test results were presented through a violin plot (Fig. 4H), demonstrating significant changes ($p < 0.05$) in the infiltration levels of six immune cells in early-stage LUAD samples. These cells play crucial roles in the early stages of LUAD. Further analysis of immune infiltration in early adenocarcinoma revealed an increase in the number of Th2 cells, Tcm, and T helper cells, along with a decrease in the number of macrophages, eosinophils, and Tem cells during the early stages of LUAD.

Construction of a prognostic prediction model for LUAD based on the training dataset

With the application of thresholds of $p.\text{adj} < 0.05$ and $|\text{Log}_2 \text{FC}| \geq 1$, a total of 444 mRNAs were identified as differentially expressed between Stage III-IV and Stage I-II. This differential expression was graphically represented using a volcano plot (Fig. 5A). To establish a correlation between mRNAs and the OS of patients with LUAD, univariate cox regression analysis was conducted. Statistical survival analysis was performed using the R packages survival and survminer, revealing an association between 4996 genes and the overall survival of LUAD patients ($P < 0.05$). Among these genes, 64 mRNAs were common to both groups, as depicted in a venn diagram (Fig. 5B). To prevent overfitting, LASSO regression was utilized to screen 13 mRNAs as potential

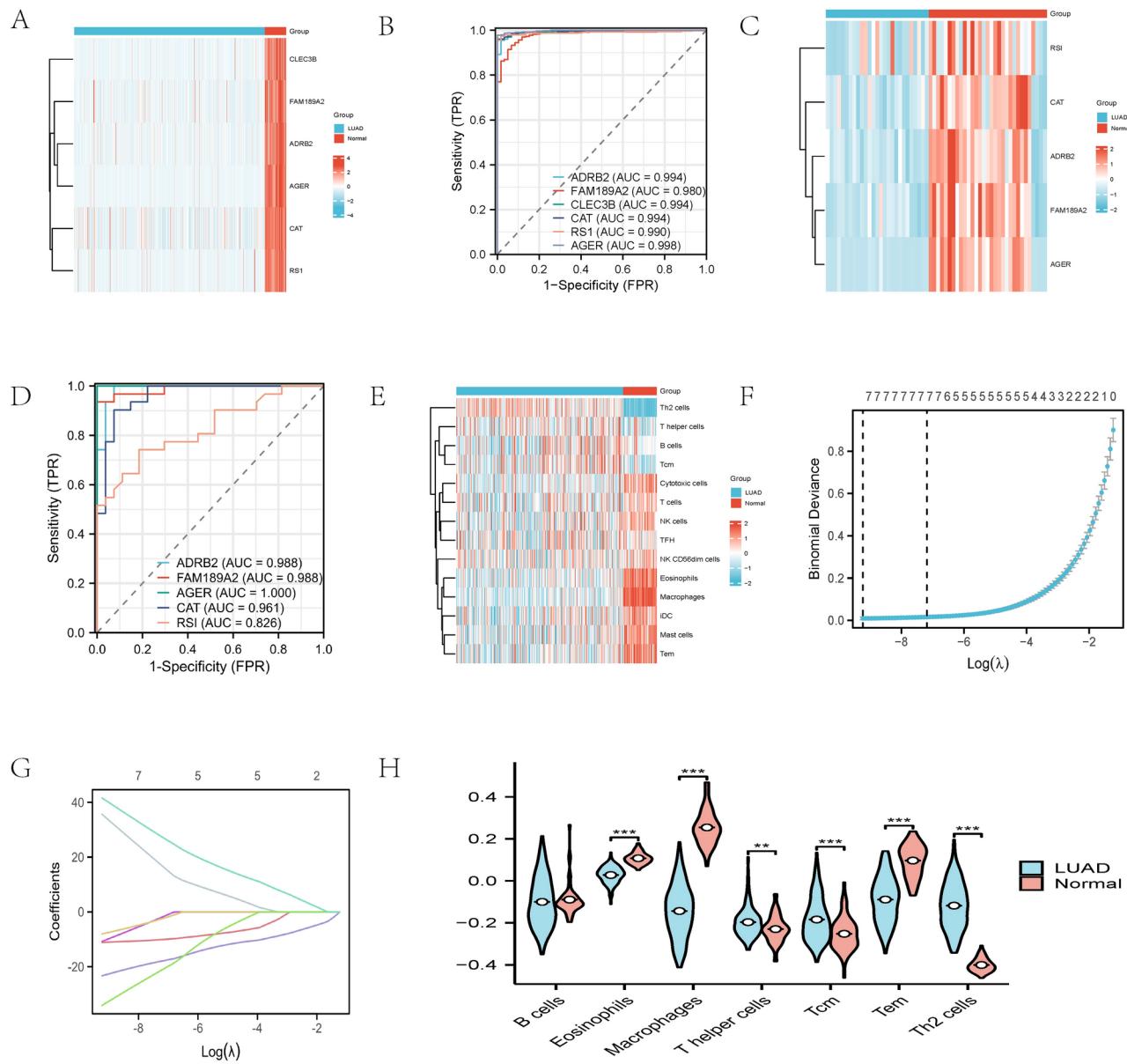


Fig. 4. Validation in TCGA-LUAD and GSE 7670 data sets and immune infiltration analysis. **(A)** The clustering heat map of tumor samples and non-cancer tissues based on key molecular markers in TCGA-LUAD data sets. **(B)** ROC curve analysis of key molecular markers in TCGA-LUAD data sets. **(C)** The clustering heat map of tumor samples and non-cancer tissues based on key molecular markers in GSE 7670 data sets. **(D)** ROC curve analysis of key molecular markers in GSE 7670 data sets. **(E)** The clustering heat map of tumor samples and non-cancer tissues based on infiltrating immune cells in TCGA-LUAD data sets. **(F)** A coefficient profile plot was produced against the log (lambda) sequence in the LASSO model. The optimal parameter (lambda) was selected as the first black dotted line indicated. **(G)** LASSO coefficient profiles of immune infiltration cells. **(H)** The expression of key immune infiltrating cells in normal tissues and stage I LUAD tissues.

variables for subsequent univariate and multivariate cox regression (Fig. 5C,D). Variables exhibiting significance in the univariate cox regression analysis ($P \leq 0.01$) were further analyzed using multivariate cox regression, ultimately leading to the establishment of a prognostic model for LUAD patients comprising 12 key variables: RHOV, PITX3, DKK1, CRYGN, FLNC, FAIM2, RHCG, IGFBP1, FAM83A, MS4A1, KCNF1, and MUC2. The findings of both univariate and multiple cox regression analysis are summarized in Table 5. The predictive model was formulated as a linear combination of the expression levels of the twelve mRNAs, with each mRNA weighted by its corresponding coefficient obtained from the multivariate cox regression analysis. The risk score for each patient was calculated using the following formula: risk score = $-0.55070176 + (0.04452972 \times \text{RHOV}) + (0.67298431 \times \text{PITX3}) + (0.0723084 \times \text{DKK1}) + (-0.16380385 \times \text{CRYGN}) + (0.07934765 \times \text{FLNC}) + (-0.16859844 \times \text{FAIM2}) + (0.01703192 \times \text{RHCG}) + (0.05784092 \times \text{IGFBP1}) + (0.07334462 \times \text{FAM83A}) +$

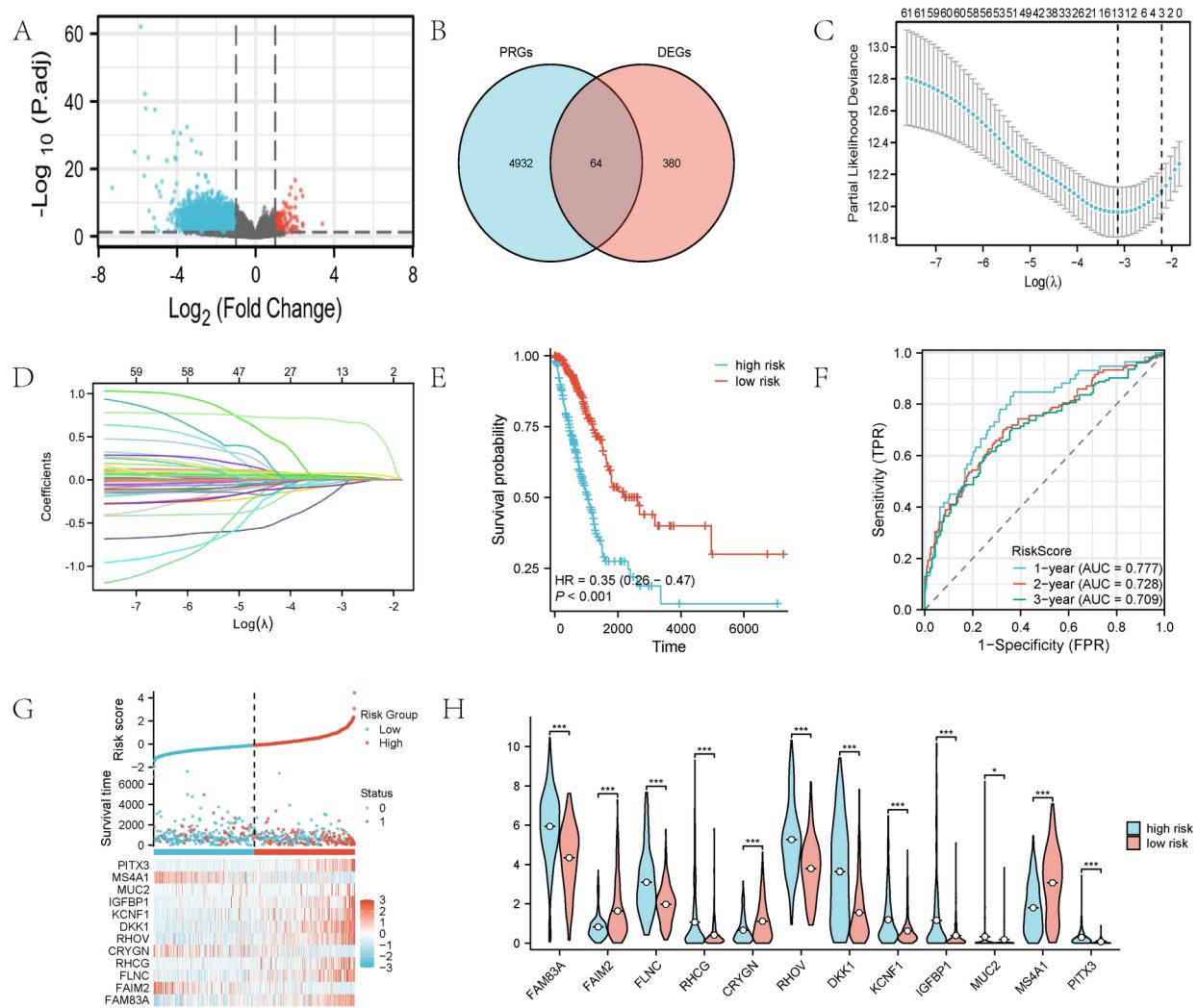


Fig. 5. (A) Volcano plot of differentially expressed genes between Stage III-IV and Stage I-II. (B) Venn diagram of differentially expressed genes and prognosis -related genes. (C) A coefficient profile plot was produced against the $\text{log}(\lambda)$ sequence in the LASSO model. The optimal parameter (λ) was selected as the first black dotted line indicated. (D) LASSO coefficient profiles of the 64 mRNAs. (E) K-M survival curves of the 12-mRNA signature between the high-and low-risk score groups. (F) Time-dependent ROC curve analyses. (G) The distribution of risk scores (upper), survival time (middle) and mRNA expression levels (below). The black dotted lines represent the median risk score cut-off dividing patients into low- and high-risk groups. The red dots and lines represent the patients in high-risk groups. The blue dots and lines represent the patients in low-risk groups. (H) Expression level of 12-mRNA signature between the high- and low-risk groups.

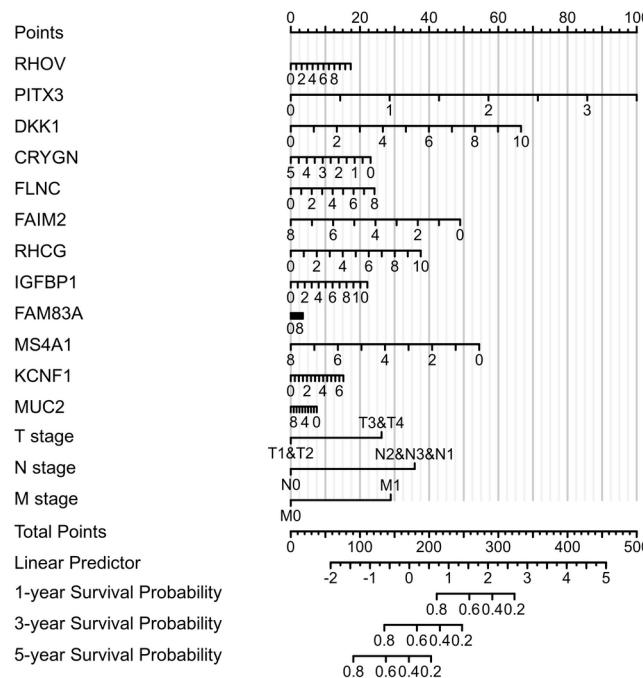
$(-0.12202164 \times \text{MS4A1}) + (0.04530024 \times \text{KCNF1}) + (0.05228372 \times \text{MUC2})$. Using the median risk score as the cutoff point, patients in the training set were stratified into two groups: low-risk ($n = 261$) and high-risk ($n = 261$). K-M analysis demonstrated that the low-risk group exhibited superior OS with a median of 1046 days, compared to the high-risk group with a median of 2620 days (log-rank test, $P < 0.001$), as depicted in Fig. 5E. In time-dependent ROC curve analyses, the prognostic model employing the twelve mRNA signatures exhibited an AUC of 0.777 at one year, 0.728 at two years, and 0.730 at four years of OS (Fig. 5F). This suggests that the model possesses good discriminatory power, with a maximum AUC of 0.777. Figure 5G presents the distribution of risk scores, survival status, and mRNA expression profiles across patients. A comparative gene expression analysis was conducted between the high-risk and low-risk groups for the twelve genes. The results revealed that the expressions of RHOV, PTX3, DKK1, FLNC, RHCG, IGFBP1, FAM83A, KCNF1, and MUC2 were significantly upregulated in the high-risk group, indicating a positive correlation between their high expression and the high-risk score. Conversely, the expressions of CRYGN, FAIM2, and MS4A1 were significantly downregulated in the high-risk group, suggesting a positive correlation between their low expression and the high-risk score (Fig. 5H).

The optimism-corrected c-index values (for OS, C-index = 0.746, 95% CI = 0.724–0.767) demonstrated that the proposed nomograms could precisely predict the 1-, 3-, and 5-year OS of LUAD patients (Fig. 6A). Furthermore, the 1-, 3-, and 5-year calibration curves of OS visually confirmed a good agreement between

Characteristics	Total(N)	Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
FAM83A	526	1.262 (1.158–1.375)	<0.001	1.082 (0.983–1.192)	0.109
FAIM2	526	0.761 (0.650–0.891)	<0.001	0.841 (0.714–0.989)	0.037
FLNC	526	1.248 (1.142–1.365)	<0.001	1.085 (0.978–1.203)	0.124
RHCG	526	1.239 (1.135–1.352)	<0.001	1.012 (0.900–1.139)	0.839
CRYGN	526	0.762 (0.632–0.919)	0.005	0.846 (0.699–1.024)	0.086
RHOV	526	1.244 (1.153–1.342)	<0.001	1.044 (0.944–1.154)	0.405
DKK1	526	1.210 (1.143–1.282)	<0.001	1.079 (1.005–1.158)	0.036
KCNF1	526	1.271 (1.138–1.419)	<0.001	1.049 (0.907–1.213)	0.521
IGFBP1	526	1.235 (1.134–1.345)	<0.001	1.058 (0.952–1.176)	0.295
MUC2	526	1.232 (1.084–1.399)	0.001	1.049 (0.904–1.217)	0.528
MS4A1	526	0.831 (0.755–0.915)	<0.001	0.882 (0.797–0.977)	0.016
PITX3	526	3.503 (2.576–4.764)	<0.001	1.920 (1.287–2.865)	0.001

Table 5. Univariate and multivariate Cox analysis of the twelve-gene prognostic signature in TCGA-LUAD patients. CI: confidence interval. Significant values are in [bold].

A



B

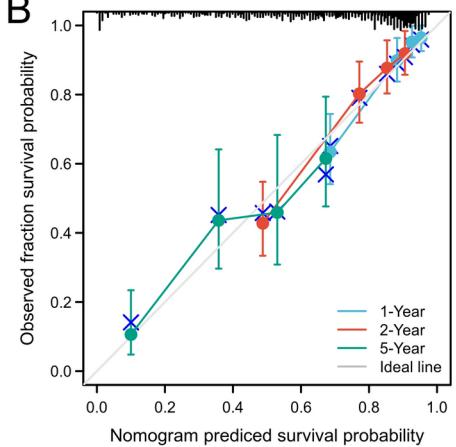


Fig. 6. The nomogram and calibration curve developed for model. (A) Establishment of nomograms for the prediction of OS in patients with LUAD. To use the nomogram, the value of individual patients with LUAD is shown on each variable axis, and a line is depicted upward to determine the number of points received for each variable value. Subsequently, the sum of these numbers is located on the total point axis, and a line is drawn downward to the survival axes to determine the likelihood of 1- 3- and 5-year survival of OS. (B) Calibration curve for predicting the 3- and 5-year survival of OS in LUAD patients in the training cohort. The actual OS rates are plotted on the y-axis and nomogram-predicted OS rates are plotted on the x-axis.

predicted survival and observed survival, which could also validate the predictive accuracy of the prognostic nomograms (Fig. 6B).

Prognostic value of twelve-gene signature in the validation set

To validate our model, we utilized separate datasets: GSE30219, GSE31210, GSE50081, GSE37745, and a merged dataset (combining GSE31210, GSE50081, and GSE37745). Below, we present the outcomes of K-M curves and mRNA expression profiles in the high-risk and low-risk groups within these validation sets. The K-M curves for the validation samples in GSE30219 confirmed that patients in the low-risk group exhibited superior OS

compared to those in the high-risk group (HR 0.63, 95% CI 0.47–0.83, $P=0.001$) (Fig. 7A). Analysis of mRNA expression levels in the high-risk and low-risk groups indicated that the expressions of RHOV, PITX3, DKK1, FLNC, and FAM83A were significantly upregulated in the high-risk group, suggesting a positive correlation between their high expressions and the high-risk score. Conversely, the expression of MS4A1 was significantly downregulated in the high-risk group, indicating a positive correlation between its low expression and the high-risk score ($p<0.05$) (Fig. 7B). The K-M curves for the validation samples in GSE31210 validated that patients with low-risk had superior OS compared to those with high-risk (HR 0.39, 95% CI 0.19–0.80, $P=0.01$) (Fig. 7C). Analysis of the expression levels of twelve mRNAs in the high-risk and low-risk groups indicated that the expressions of RHOV, PITX3, DKK1, FLNC, and FAM83A were significantly upregulated in the high-risk group, while the expression of MS4A1 was significantly downregulated ($p<0.05$) (Fig. 7D). Similarly, the K-M curves for the validation samples in GSE50081 confirmed a better OS in low-risk patients compared to those with high-risk (HR 0.40, 95% CI 0.22–0.72, $P=0.002$) (Fig. 7E). Analysis of the mRNA expression profiles revealed that the expressions of RHOV, PITX3, DKK1, FLNC, and FAM83A were significantly increased in the high-risk group, whereas the expression of MS4A1 was significantly decreased ($p<0.05$) (Fig. 7F). Furthermore, the K-M curves for the validation samples in GSE337745 also supported the finding that low-risk patients exhibited improved OS compared to high-risk patients (HR 0.63, 95% CI 0.40–0.99, $P=0.043$) (Fig. 7G). These findings provide further validation for the prognostic accuracy of our nomograms. Analysis of the expression levels of twelve mRNAs in the high-risk and low-risk groups revealed that the expressions of RHOV, PITX3, DKK1, FLNC, and FAM83A were significantly upregulated in the high-risk group compared to the low-risk group, while the expression of MS4A1 was significantly downregulated in the high-risk group ($p<0.05$) (Fig. 7H). The K-M curves for the merged data set validated that patients in the low-risk group exhibited superior OS compared to those in the high-risk group (HR 0.27, 95% CI 0.19–0.39, $P<0.001$) (Fig. 8A). Notably, the expression patterns of these twelve mRNAs remained consistent with those observed in the training set, despite the distinct populations represented in the training and validation sets. This consistency demonstrates the robustness and reliability of the model. The distribution of risk scores, survival status, and mRNA expression levels among patients are presented in Fig. 8B. The results clearly demonstrate that patients with lower risk scores generally exhibit better survival outcomes compared to those with higher risk scores. To comprehensively evaluate the predictive performance of the models, time-dependent ROC analysis was conducted for OS at one, three, and five years. This analysis discriminated between patients who experienced an event and those who remained event-free, resulting in ROC curves with AUC values of 0.809, 0.749, and 0.719 in the merged data set. These findings suggest that the model demonstrates good discriminative ability, with a maximum AUC of 0.809 (Fig. 8C).

GO, KEGG and GSEA enrichment analyses

A correlation analysis was conducted on the twelve-gene signature in the TCGA-LUAD cohorts, and the genes were ranked based on their spearman's rank correlation coefficient ($p<0.05$). Subsequently, functional enrichment analysis and GSEA analysis were performed on the top 20 positively and negatively correlated genes with the twelve-gene signature in TCGA-LUAD patients. GO enrichment analysis revealed that the positively correlated genes were enriched in functions such as extracellular matrix organization, collagen-containing extracellular matrix, and extracellular matrix structural constituent. KEGG pathway analysis demonstrated that these genes were enriched in ECM-receptor interaction, among others (Fig. 9A) (Table 6). Conversely, the negatively correlated genes were enriched in functions related to organellar ribosome, oxidoreductase activity acting on the CH-OH group of donors, with NAD or NADP as acceptor. KEGG pathway analysis indicated that these genes were enriched in tyrosine metabolism, glycolysis/gluconeogenesis, among others (Fig. 9B) (Table 7). Gene set enrichment analysis demonstrated that reactome extracellular matrix organization, wp pi3kakt signaling pathway, wp focal adhesion pi3kakt mtor signaling pathway, kegg ecm receptor interaction, and pid integrin1 pathway were enriched (Fig. 9C) (Table 8).

Discussion

In this study, we employed ML algorithms to identify diagnostic and prognostic biomarkers. Subsequently, an immune infiltration analysis was conducted, leading to the final identification of ADRB2, FAM189A2, CLEC3B, AGER, CAT, and RS1 as mRNA biomarkers for early diagnosis of LUAD. The reliability of these biomarkers was further validated using an external dataset, GSE7670. The study found that CLEC3B serves as a potential diagnostic and prognostic biomarker in lung cancer, and it is associated with the immune microenvironment⁹. Furthermore, another study revealed that overexpression of AGER reduced the proliferation, invasion, and migration abilities of H1299 cells, while increasing cellular apoptosis. This suggests that AGER may serve as a potential molecular marker for NSCLC¹⁰. CAT demonstrates a significant correlation with PFS in lung cancer patients receiving platinum-based chemotherapy, potentially serving as a biomarker for predicting the prognosis of lung cancer patients treated with platinum-based chemotherapy¹¹. The treatment with RS-1 elicited a significant antitumor response in mouse models¹². The above reports are consistent with the LUAD diagnostic markers we have obtained. Although there are currently no reports linking ADRB2 and FAM189A2 to LUAD, their expression has been associated with breast cancer. For instance, ADRB2 is a potential protective gene for the breast¹³. Additionally, ENTRP/FAM189A2 encodes a new ITCH ubiquitin ligase activator that is downregulated in breast cancer¹⁴. These are potential molecules related to LUAD that require further investigation in the future.

A comprehensive and dynamic understanding of the immune microenvironment is crucial for developing effective treatment strategies. Therefore, this study delved into the infiltration of immune cells in early-stage LUAD and utilized LASSO regularization technique to identify key immune cells. The study found that Th2 cells release anti-inflammatory cytokines and promote tumor progression¹⁵. Furthermore, Th2 cells can also promote angiogenesis and hinder the apoptosis of tumor cells by remodeling the cytokine environment infiltrated by macrophages and eosinophils¹⁶. Consistent with this, our study found that the infiltration of Th2 cells increases

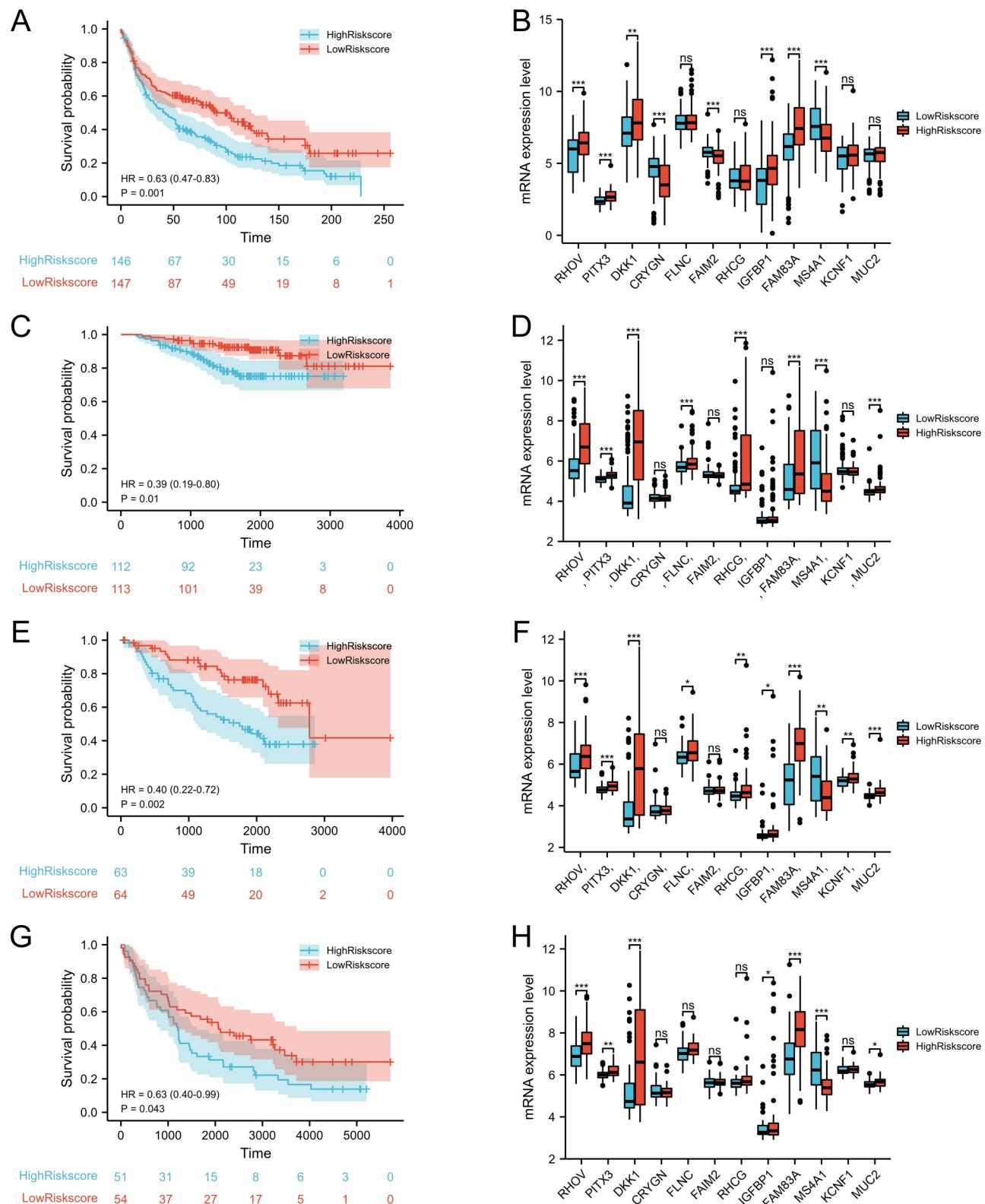


Fig. 7. Validation of the prognostic signature in the GSE30219, GSE1210, GSE50081 and GSE37745 data sets. (A,C,E,G) K-M curves of the twelve-gene prognostic signature between the high- and low-risk score groups in the GSE30219, GSE1210, GSE50081 and GSE37745 data sets, HR indicates hazard ratio. (B,D,F,H) expression level analysis of 12-mRNAs prognostic signature between the high- and low-risk groups in the GSE30219, GSE1210, GSE50081 and GSE37745 data sets. $P < 0.05$ (*), $P < 0.01$ (**), $P < 0.001$ (***) (Wilcoxon rank sum test).

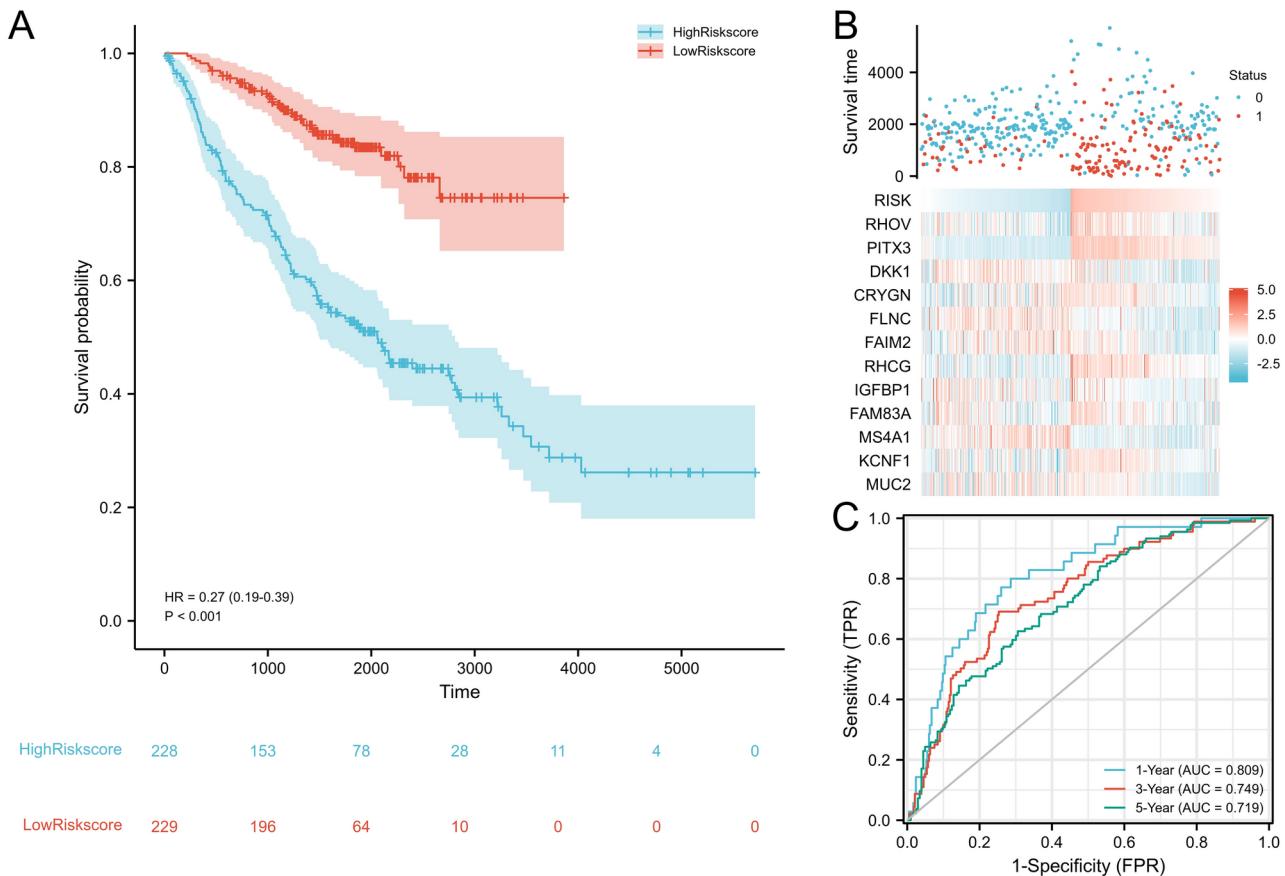


Fig. 8. Validation of the 12-mRNAs prognostic signature in the merged data set. (A) K-M curves of the 12-mRNA signature between the high- and low-risk score groups. (B) The distribution of risk scores (upper), survival time (middle) and miRNA expression levels (below). (C) Time-dependent ROC curve analyses.

in the early stage of LUAD. It is possible that in the early stage of LUAD, Th2 cells may promote the infiltration of immune cells in the tumor environment by releasing cytokines. Although Th2 cells are generally associated with tumor growth and progression, their presence in the early stage may reflect the initial response of the immune system to tumor cells. Tcm cells possess memory functions and can rapidly generate effector T cells under the stimulation of tumor antigens¹⁷. Previous studies have shown that the cytotoxic activity and proliferative capacity of effector memory T cells are superior to those of central memory T cells (TCM) in vitro. However, TCM has the potential to induce immune memory and exert more durable antitumor activity¹⁸. Therefore, the infiltration of Tcm cells in the early stage of LUAD may indicate that the immune system is actively responding to the tumor and attempting to establish long-term memory to cope with possible tumor recurrence. T-helper cells are generally associated with cellular immunity, capable of activating and guiding other immune cells to attack tumors. In the early stage of LUAD, the infiltration of T-helper cells may represent the immune system's attempt to eliminate tumor cells. In the early stage of LUAD, macrophages may more frequently manifest as tumor-associated macrophages (TAMs), which typically have immunosuppressive functions and can promote tumor growth and angiogenesis^{19,20}. Therefore, the negative correlation between macrophages and immune infiltration may suggest that immunosuppression has not yet dominated in the early stages of tumor development. Eosinophils usually play a role in allergic reactions and parasitic infections, while in the tumor environment, they may be associated with immunosuppression or tumor progression²¹. Therefore, the negative correlation between eosinophils and immune infiltration may suggest that these cells have not yet infiltrated into the tumor tissue in large numbers in the early stage of LUAD. Effector memory T cells (Tem cells) usually exert anti-tumor effects in the tumor environment. However, in some cases, Tem cells may be functionally impaired due to the influence of the tumor microenvironment²². Therefore, the negative correlation between Tem cells and immune infiltration may suggest that the anti-tumor function of these cells has not yet been fully exerted in the early stage of LUAD. This is consistent with our findings on immune infiltration.

In this study, our diagnostic and prognostic prediction models have both demonstrated exceptional results. Specifically, the AUC value of the diagnostic model ranges from 0.9979483 to 1, indicating a high degree of accuracy in distinguishing LUAD patients from non-patients, with CE below 10%. This further proves its reliability in early diagnosis. Meanwhile, our prognostic prediction model has also been effectively validated on four independent Gene Expression Omnibus datasets, demonstrating not only high accuracy but also good generalization ability, enabling it to adapt to different datasets and patient populations. Despite these

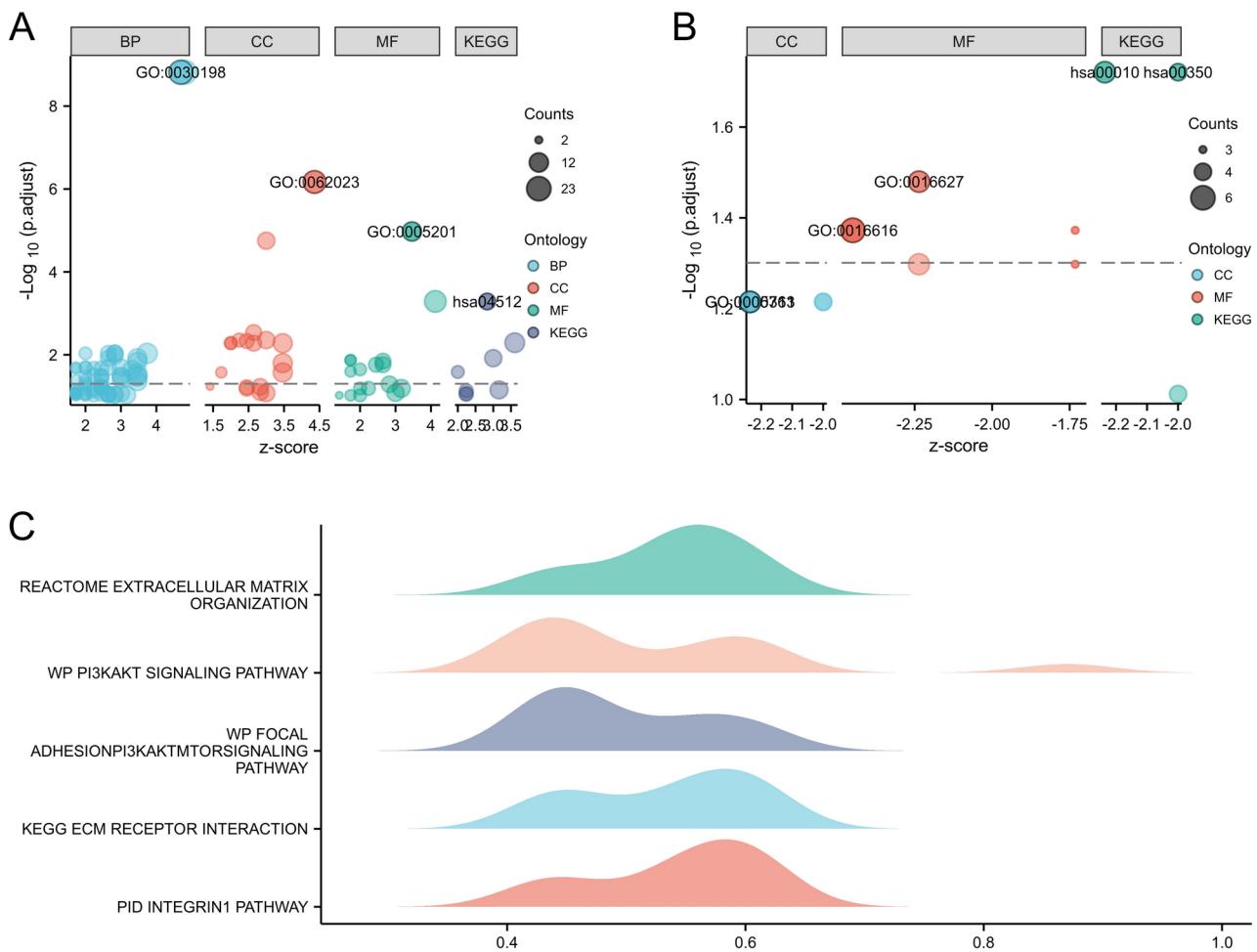


Fig. 9. GO, KEGG and GSEA enrichment analyses. (A) GO and KEGG pathway enrichment analysis of the top 20 positively correlated genes with twelve-gene signature. The ordinate represents a P.value, the higher the y values, the more reliable results of the enrichment analysis, the abscissa represents zscores. (B) GO and KEGG pathway enrichment analysis of the top 20 negatively correlated genes with twelve-gene signature. (C) Gene set enrichment analysis demonstrated that the reactome extracellular matrix organization, wp pi3kakt signaling pathway, wp focal adhesionpi3kaktmtorsignalng pathway, kegg ecm receptor interaction and pid integrin1 pathway were enriched.

achievements, our models still have limitations and shortcomings in LUAD diagnosis and prognostic prediction, which require continuous improvement and optimization in our future work. Firstly, through comparisons with the current state-of-the-art models, we have gained a clearer understanding of the strengths and limitations of our study. Notably, some research not only focuses on improving diagnostic accuracy but also delves into the biological mechanisms of cancer, particularly the role of immune infiltration in cancer progression and patient prognosis. In this regard, the immunotherapy prognostic prediction signature (IPPS) model has achieved significant results in predicting immunotherapy prognosis for LUAD patients, providing valuable references for our work²³. Compared to the IPPS model, our model differs in terms of its objectives and application scenarios. The IPPS model primarily focuses on predicting the prognosis of immunotherapy, while our model aims at early diagnosis of LUAD and overall prognosis prediction, with a particular emphasis on the role of immune infiltration. This difference reflects the varying research directions and priorities of the two models. However, we also recognize that by integrating the advantages of different models, we can more comprehensively assess patients' conditions and prognosis, providing more accurate and comprehensive bases for clinical decision-making. Therefore, in future research, we will consider leveraging the successful experience of the IPPS model to further optimize our model, better serving the diagnosis and prognosis prediction of LUAD patients. Secondly, we have noticed that immune infiltration and tumor-associated lymphoid structures (TLS), as crucial components of the cancer immune microenvironment, both play significant roles in the immune therapeutic response and patient survival²⁴. While our current study primarily focuses on immune infiltration in the early stages, research on TLS has been more concentrated in gastrointestinal cancers. However, we speculate that there may be potential correlations between these two phenomena within the cancer immune microenvironment. To gain a more comprehensive understanding of the cancer immune microenvironment and develop more accurate diagnostic and prognostic prediction models, future research should further explore the interactions

Category	ID	Term	adj_pval
BP	GO:0030198	Extracellular matrix organization	1.52542E-09
BP	GO:0043062	Extracellular structure organization	1.52542E-09
BP	GO:0043588	Skin development	0.009218619
BP	GO:0048565	Digestive tract development	0.009218619
BP	GO:0022600	Digestive system process	0.009218619
BP	GO:0007586	Digestion	0.009218619
BP	GO:0030277	Maintenance of gastrointestinal epithelium	0.009218619
BP	GO:0055123	Digestive system development	0.010558369
BP	GO:0070268	Cornification	0.012095677
CC	GO:0031589	Cell-substrate adhesion	0.012095677
CC	GO:0062023	Collagen-containing extracellular matrix	6.76853E-07
CC	GO:0005604	Basement membrane	1.77046E-05
CC	GO:0030175	Filopodium	0.002895847
CC	GO:0044420	Extracellular matrix component	0.00440781
CC	GO:0098858	Actin-based cell projection	0.00440781
CC	GO:0005581	Collagen trimer	0.004661637
CC	GO:0008305	Integrin complex	0.005193845
MF	GO:0005201	Extracellular matrix structural constituent	1.05889E-05
MF	GO:0050839	Cell adhesion molecule binding	0.000515535
MF	GO:0015250	Water channel activity	0.013488077
MF	GO:0005372	Water transmembrane transporter activity	0.013488077
KEGG	hsa04512	ECM-receptor interaction	0.00051659
KEGG	hsa04151	PI3K-Akt signaling pathway	0.005110996
KEGG	hsa04510	Focal adhesion	0.012081601

Table 6. GO and KEGG pathway enrichment analysis of the top 20 positively correlated genes with 12-gene signature in TCGA-LUAD patients. *GO* Gene Ontology, *KEGG* Kyoto Encyclopedia of Genes and Genomes, *TCGA* The Cancer Genome Atlas, *LUAD* Lung adenocarcinoma, *BP* Biological Processes, *CC* Cellular Component, *MF* Molecular Function.

Category	ID	Term	adj_pval
CC	GO:0000313	Organellar ribosome	0.060966007
CC	GO:0005761	Mitochondrial ribosome	0.060966007
CC	GO:0000315	Organellar large ribosomal subunit	0.060966007
CC	GO:0005762	Mitochondrial large ribosomal subunit	0.060966007
MF	GO:0016627	Oxidoreductase activity, acting on the CH-CH group of donors	0.033132866
MF	GO:0016616	Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	0.042403306
MF	GO:0016614	Oxidoreductase activity, acting on CH-OH group of donors	0.042403306
MF	GO:0004602	Glutathione peroxidase activity	0.042403306
MF	GO:0008106	Alcohol dehydrogenase (Nadp+) activity	0.042403306
MF	GO:0004497	Monooxygenase activity	0.050363647
KEGG	hsa00350	Tyrosine metabolism	0.019019156
KEGG	hsa00010	Glycolysis / Gluconeogenesis	0.019019156
KEGG	hsa00590	Arachidonic acid metabolism	0.097255883

Table 7. GO and KEGG pathway enrichment analysis of the top 20 positively correlated genes with 12-gene signature in TCGA-LUAD patients. *GO* Gene Ontology, *KEGG* Kyoto Encyclopedia of Genes and Genomes, *TCGA* The Cancer Genome Atlas, *LUAD* Lung adenocarcinoma, *BP* Biological Processes, *CC* Cellular Component.

and potential associations between immune infiltration and TLS in cancer. By integrating information from both phenomena, we hope to provide patients with more personalized treatment options, optimize clinical decision-making processes, and ultimately improve the efficacy of cancer treatments and patient survival rates. Finally, in the research of high-performance concrete (HPC) and ultra-high-performance concrete (UHPC), ML models have exhibited remarkable capabilities in predicting material strength^{25,26}. Particularly, the GRU model, which excels in handling time-series data, is not only applicable to the progression analysis of LUAD but also to the

Rank	NAME	SIZE	ES	NES	pvalue
1	REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION	16	0.624925949	2.368176957	0.00135318
2	WP_PI3KAKT_SIGNALING_PATHWAY	15	0.564222412	2.092175863	0.001356852
3	WP_FOCAL_ADHESIONPI3KAKTMTORSIGNALING_PATHWAY	14	0.58933644	2.124350376	0.001381215
4	KEGG_ECM_RECECTOR_INTERACTION	10	0.678478813	2.170877396	0.0013947
5	PID_INTEGRIN1_PATHWAY	10	0.609154269	1.949064891	0.0013947
6	KEGG_FOCAL_ADHESION	11	0.61066653	2.023599428	0.002754821
7	NABA_CORE_MATRISOME	12	0.635506652	2.170637813	0.002777778
8	WP_FOCAL_ADHESION	10	0.602496137	1.927761367	0.0027894
9	NABA_MATRISOME	33	0.431242153	2.067369944	0.003558719
10	REACTOME_NERVOUS_SYSTEM_DEVELOPMENT	10	0.566123066	1.811381199	0.011157601

Table 8. Representative results of GSEA. GSEA Gene Set Enrichment Analysis, ES enrichment score, NES normalized enrichment score.

prediction of concrete strength over time²⁷. This study's achievements in LUAD diagnosis and prognosis share similarities with concrete compressive strength prediction research, both leveraging advanced machine learning to enhance prediction accuracy²⁸. Specifically, the introduction of the SHapley Additive exPlanations (SHAP) method in the field of concrete compressive strength prediction not only provides in-depth explanations for the prediction results but also reveals the influence of different variables on the outcomes^{29,30}. In LUAD research, we can gain insights from analyzing prediction results to comprehend tumor microenvironment, immune infiltration, and their impact on tumor progression. Future work should explore machine learning models to unravel these complexities, offering tailored tumor treatment guidance. Additionally, concrete compressive strength prediction research underscores the potential of machine learning to drive innovation and development across diverse fields^{31,32}. Whether it is achieving precision medicine in the healthcare sector or promoting the development of environmentally friendly materials in the construction industry, ML provides us with powerful tools²⁵. Therefore, cross-disciplinary learning and borrowing will help us better tackle various challenges and promote the development of different fields.

Methods

Data sources

This study acquired Level 3 RNA-seq expression data consisting of 535 LUAD samples (Stage I: n=294, Stage II: n=123, Stage III: n=84, Stage IV: n=26) and normal lung samples (n=59) from TCGA (<https://portal.gdc.cancer.gov/>). The tumor expression datasets and corresponding clinical information in the validation set were downloaded from the Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) with accession numbers GSE7670 (paired samples from 27 patients) (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7670>), GSE30219 (14 normal and 293 cancer) (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30219>), GSE31210 (226 Stage I-II cancers) (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31210>)³³, GSE50081 (127 cancers) (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50081>)³⁴, and GSE37745 (106 cancers) (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37745>)³⁵. Expression profiling was performed using the Affymetrix Human Genome U133 Plus 2.0 Array. This study does not include any experiments involving human participants or animals conducted by the authors. Baseline demographic and clinical information are presented in Table 1.

Data preprocessing

Level 3 RNA-seq expression data were acquired from TCGA (<https://portal.gdc.cancer.gov/>). The expression data were log2-transformed to stabilize the variance and a value of 1 was added to avoid zero values during the log2 transformation. The raw and series matrix files for GSE7670, GSE30219, GSE31210, GSE37745, and GSE50081 were downloaded. For the raw data, the probe expression matrices were extracted and normalized using the R package “affy” (version 4.2.1). The probe expression matrices were then converted into gene expression matrices using platform annotation files. For genes corresponding to multiple probes, the average value was taken as the expression value for that gene. Probes corresponding to multiple molecules were removed. The results were visualized using the R package “ggplot2”.

Feature selection

To select the most relevant features for classification, we performed feature selection on an initial set of important features. In the training set, we first utilized the randomForest package to train a RF classifier (with the ‘ntree’ parameter set to 800) and subsequently employed the recursive feature elimination function from the caret package to identify key features. Following this, we applied the support vector machine recursive feature elimination (SVM-RFE) method, which ranks the features of our classification problem by training an SVM with a radial basis function kernel and removing the features with the smallest ranking criterion. Specifically, the SVM implementation utilized the “e1071” package in R. For SVM, we explored different kernel functions, including linear, polynomial, and radial basis function. The seed value was set to 2023 to ensure reproducibility of the results. For the SVM with a linear kernel, we used RNA expression and clinical information as inputs to the tune.svm function, with the “kernel” parameter set to “linear” and the cost parameter “cost” set to a range

of candidate values such as “c(0.001, 0.01, 0.1, 1, 5, 10)”. For the SVM with a polynomial kernel, we employed mRNA expression and clinical information as inputs, with the “kernel” parameter set to “polynomial”, the degree of the polynomial “degree” set to “c(3, 4, 5)”, and the “coef0” parameter also set to a range of candidate values such as “c (0.1, 0.5, 1, 2, 3, 4)”. For the SVM with a radial basis function kernel, we again used RNA expression and clinical information as inputs, with the “kernel” parameter set to “radial”, and the “gamma” parameter also set to a range of candidate values such as “c(0.1, 0.5, 1, 2, 3, 4)”. To evaluate the performance of the models, we calculated the kappa value, sensitivity, specificity, and accuracy using the confusionMatrix function from the caret package. The optimal parameter values were determined through tenfold cross-validation to minimize the misclassification rate.

Additionally, LASSO regression coupled with cross-validation was employed for feature selection, aiming to identify the pivotal features pertinent to a specific binary classification problem. The feature variables were structured as a matrix, encompassing multiple potential explanatory variables, while the response variable was represented as a vector, indicating the class labels for each sample. To explore the impact of varying regularization strengths on model performance, an arithmetic sequence of 200 values ranging from 0 to 0.5 was generated, serving as candidate values for the regularization parameter λ in lasso regression. The cv.glmnet function from the glmnet package in R was utilized to execute Lasso regression with cross-validation. Specifically, α was set to 1 to conduct Lasso regression, and the previously generated sequence of lambdas was employed as the range for λ . A tenfold cross-validation scheme was adopted to evaluate model performance, and the family = “binomial” argument was specified to accommodate the binary classification nature of the problem. Through this cross-validation process, estimates of the average error for the model across different λ values were obtained. Based on these estimates, two key λ values were identified: λ_{min} , which minimizes the cross-validation error, and λ_{1se} , the leftmost λ value within one standard error of the minimum cross-validation error. Visualizations were generated to gain insights into the trend of model performance with varying regularization strengths. The relationship between cross-validation error and λ values was plotted, allowing for an intuitive observation of how model performance changes with the intensity of regularization. Additionally, LASSO coefficient paths were illustrated to show the evolution of individual coefficients as λ varies.

To identify features that significantly impact the prediction of patients’ disease status from the dataset, the RF algorithm was employed for feature selection. Initially, the target variable in the dataset was factorized to ensure it was correctly treated as a categorical variable by the model. Subsequently, two RF models were constructed using the randomForest function from the randomForest package. The first model was utilized for initial exploration and visualization, while the second model was dedicated to final prediction and feature importance evaluation. During model construction, the number of trees (ntree = 800) was set to ensure model stability and accuracy, and the options important = TRUE and proximity = TRUE were enabled to calculate feature importance and sample proximity, respectively.

Finally, we took the intersection of the results from the three methods (RF, SVM and LASSO) and these intersecting features were utilized for further analysis, considered as the most reliable biomarkers.

Immune cell infiltration analysis

Using the software R, the immune cell infiltration between stage I LUAD and normal tissues was explored through ssGSEA. Significant immune cells ($p < 0.05$) between the two groups were screened using the Mann-Whitney U test and LASSO regression analysis, respectively, to identify the final significantly different immune infiltrating cells. The immune infiltration algorithm was based on the ssGSEA algorithm provided by the R package GSVA [1.44.5]. The Mann-Whitney U test was conducted with the stats package in R, while LASSO regression was performed using the “glmnet” package [4.1.7]. The visualization of the results was carried out using the “ggplot2” package [3.3.6], and heatmap visualization was achieved through the ComplexHeatmap package [2.13.1].

Analyzing the diagnostic validity of biomarkers in TCGA-LUAD and GSE7670

In the GSE7670 dataset, ROC analysis was conducted using the “Proc” package [1.18.0] within the R software, and visualization was performed utilizing the “ggplot2” package. Subsequently, the AUC was employed to predict the diagnostic effectiveness of LUAD biomarkers. Additionally, heatmap visualization was achieved through the utilization of the ComplexHeatmap package.

Construction prognosis model of lung adenocarcinoma

Differential analysis between stage I-II and stage III-IV LUAD patients in TCGA-LUAD was performed using DESeq2, $P\text{value} < 0.05$ and $\log FC > 1$ as threshold. Overall survival (OS)-related mRNAs with $P\text{value} < 0.05$ were determined using univariate cox regression analysis. R packages survminer survival were used for statistical analysis and data visualization in TCGA-LUAD (11). The intersection of mRNAs associated with poor prognosis and differentially expressed mRNAs (DEGs) between stage I/II and stage III/IV was screened out as candidate genes for subsequent analysis. We performed the lasso cox regression using the glmnet R package. Subsequently, multivariate cox regression analysis was performed to build survival model to predict OS, candidate genes were selected to build the risk signature. Finally, we calculated a risk score for each sample by the following formula: “risk score” = \sum (regression coefficient) \times (expression value of each prognostic mRNA). A median risk score was used to divide patients into high- and low-risk groups. Then we performed K-M analysis between the two groups to compare survival outcomes. K-M analysis was performed using R survminer and survival packages, with $p < 0.05$ taken as significant. In addition, the package “timeROC” and “ggplot2” were used to evaluate the sensitivity and specificity of the prediction model through the AUC of the ROC. A nomogram based on the TNM staging system and prognostic signature was created by R software, using the “rms” package. Calibration curves

were assessed graphically by comparing observed and predicted survival probabilities. The discrimination of the nomogram was measured by the concordance index (C-index).

Validation prognosis model of lung adenocarcinoma

GSE30219, GSE31210, GSE50081, and GSE37745 data set were used for validation of the models, with the same cutoff and the risk score calculating formula for training set. Similarly, expression levels of gene signature and K-M survival curves in the low-risk and high-risk groups were plotted for each of the validation sets.

GO, KEGG and GSEA enrichment analyses

A spearman correlation analysis of gene signature was performed in TCGA-LUAD cohorts, and the top 20 positively and negatively correlated genes were included in functional enrichment analysis. The spearman correlation was calculated using the cor.test function in the R stats package. GO, KEGG, and GSEA analyses were conducted using the clusterProfiler R package, the R package org.Hs.eg.db, and the R package Goplot, respectively. GO and KEGG analyses were conducted with $p.\text{adj} < 0.05$ and $\text{qvalue} < 0.2$ as the threshold, while GSEA was performed using default parameters.

Data availability

The data involved in this study are primarily sourced from two public databases: the TCGA Data Portal and the NCBI GEO Database. The following are the access methods for these data sources. For TCGA data, we utilized datasets provided by the Cancer Genome Atlas (TCGA) project. Readers can access the relevant data by visiting the official TCGA Data Portal at the following URL: <https://portal.gdc.cancer.gov/> For gene expression data, we also utilized the NCBI Gene Expression Omnibus (GEO) Database. The complete URL strings for the GEO datasets used in this study are as follows: GSE7670: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7670> GSE30219: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30219> GSE31210: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31210> GSE50081: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50081> GSE37745: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37745> These GEO datasets contain rich gene expression data, which are crucial for validating the conclusions of this study. Readers can directly copy the above URL strings into their browsers to access the corresponding pages on the GEO Database, view detailed descriptions of the datasets, download raw data, and obtain related analysis tools and documentation.

Received: 17 April 2024; Accepted: 18 September 2024

Published online: 27 September 2024

References

- Cheng, T. Y. *et al.* The international epidemiology of lung cancer: Latest trends, disparities, and tumor characteristics. *J. Thorac. Oncol.* **11**, 1653–1671. <https://doi.org/10.1016/j.jtho.2016.05.021> (2016).
- Kim, J. W. *et al.* Antitumor activity of an engineered decoy receptor targeting CLCF1-CNTFR signaling in lung adenocarcinoma. *Nat. Med.* **25**, 1783–1795. <https://doi.org/10.1038/s41591-019-0612-2> (2019).
- Mission, P. P. *et al.* Autoantibody signature enhances the positive predictive power of computed tomography and nodule-based risk models for detection of lung cancer. *J. Thorac. Oncol.* **12**, 578–584. <https://doi.org/10.1016/j.jtho.2016.08.143> (2017).
- Noonan, S. A. *et al.* Identifying the appropriate FISH criteria for defining MET copy number-driven lung adenocarcinoma through oncogene overlap analysis. *J. Thorac. Oncol.* **11**, 1293–1304. <https://doi.org/10.1016/j.jtho.2016.04.033> (2016).
- Hashimoto, D. A., Rosman, G., Rus, D. & Meireles, O. R. Artificial intelligence in surgery: Promises and perils. *Ann. Surg.* **268**, 70–76. <https://doi.org/10.1097/SLA.0000000000002693> (2018).
- Xia, R. & Kais, S. Quantum machine learning for electronic structure calculations. *Nat. Commun.* **9**, 4195. <https://doi.org/10.1038/s41467-018-06598-z> (2018).
- Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. Artificial intelligence in radiology. *Nat. Rev. Cancer.* **18**, 500–510. <https://doi.org/10.1038/s41568-018-0016-5> (2018).
- Khoury, M. J., Armstrong, G. L., Bunnell, R. E., Cyril, J. & Iademarco, M. F. The intersection of genomics and big data with public health: Opportunities for precision public health. *PLoS Med.* **17**, e1003373. <https://doi.org/10.1371/journal.pmed.1003373> (2020).
- Sun, J. *et al.* CLEC3B as a potential diagnostic and prognostic biomarker in lung cancer and association with the immune microenvironment. *Cancer Cell Int.* **20**, 106. <https://doi.org/10.1186/s12935-020-01183-1> (2020).
- Wang, Q. *et al.* Effect of AGER on the biological behavior of non-small cell lung cancer H1299 cells. *Mol. Med. Rep.* **22**, 810–818. <https://doi.org/10.3892/mmr.2020.11176> (2020).
- Liu, J. S. *et al.* Association of variations in the CAT and prognosis in lung cancer patients with platinum-based chemotherapy. *Front Pharmacol.* **14**, 1119837. <https://doi.org/10.3389/fphar.2023.1119837> (2023).
- Mason, J. M. *et al.* The RAD51-stimulatory compound RS-1 can exploit the RAD51 overexpression that exists in cancer cells and tumors. *Cancer Res.* **74**, 3546–3555. <https://doi.org/10.1158/0008-5472.CAN-13-3220> (2014).
- Wei, X. *et al.* ADRB2 is a potential protective gene in breast cancer by regulating tumor immune microenvironment. *Transl. Cancer Res.* **10**, 5280–5294. <https://doi.org/10.21037/tcr-21-1257> (2021).
- Tsunoda, T. *et al.* ENTREP/FAM189A2 encodes a new ITCH ubiquitin ligase activator that is downregulated in breast cancer. *EMBO Rep.* **23**, e51182. <https://doi.org/10.15252/embr.202051182> (2022).
- Yang, Y., Yang, Y., Yang, J., Zhao, X. & Wei, X. Tumor microenvironment in ovarian cancer: Function and therapeutic strategy. *Front Cell Dev. Biol.* **8**, 758. <https://doi.org/10.3389/fcell.2020.00758> (2020).
- Zhang, Y., Guan, X. Y. & Jiang, P. Cytokine and Chemokine signals of t-cell exclusion in tumors. *Front Immunol.* **11**, 594609. <https://doi.org/10.3389/fimmu.2020.594609> (2020).
- Galluzzi, L., Spranger, S., Fuchs, E. & Lopez-Soto, A. WNT signaling in cancer immunosurveillance. *Trends Cell Biol.* **29**, 44–65. <https://doi.org/10.1016/j.tcb.2018.08.005> (2019).
- Cai, Q., Zhang, M. & Li, Z. Potential strategies against resistance to CAR T-cell therapy in haematological malignancies. *Ther. Adv. Med. Oncol.* **12**, 1758835920962963. <https://doi.org/10.1177/1758835920962963> (2020).
- Kondratova, M. *et al.* A multiscale signalling network map of innate immune response in cancer reveals cell heterogeneity signatures. *Nat. Commun.* **10**, 4808. <https://doi.org/10.1038/s41588-019-12270-x> (2019).
- Yu, J. *et al.* Reconfigurable open microfluidics for studying the spatiotemporal dynamics of paracrine signalling. *Nat. Biomed. Eng.* **3**, 830–841. <https://doi.org/10.1038/s41551-019-0421-4> (2019).

21. Grootjans, J., Kaser, A., Kaufman, R. J. & Blumberg, R. S. The unfolded protein response in immunity and inflammation. *Nat. Rev. Immunol.* **16**, 469–484. <https://doi.org/10.1038/nri.2016.62> (2016).
22. Fujita, M. *et al.* Classification of primary liver cancer with immunosuppression mechanisms and correlation with genomic alterations. *EBioMedicine* **53**, 102659. <https://doi.org/10.1016/j.ebiom.2020.102659> (2020).
23. Li, Z. *et al.* Advancing predictive markers in lung adenocarcinoma: A machine learning-based immunotherapy prognostic prediction signature. *Environ. Toxicol.* <https://doi.org/10.1002/tox.24284> (2024).
24. Li, Z. *et al.* Development and validation of a machine learning model for detection and classification of tertiary lymphoid structures in gastrointestinal cancers. *JAMA Netw. Open* **6**, e2252553. <https://doi.org/10.1001/jamanetworkopen.2022.52553> (2023).
25. Das, P. & Kashem, A. Hybrid machine learning approach to prediction of the compressive and flexural strengths of UHPC and parametric analysis with shapley additive explanations. *Case Stud. Constr. Mater.* <https://doi.org/10.1016/j.cscm.2023.e02723> (2024).
26. Islam, N., Kashem, A. & Das, P. Prediction of high-performance concrete compressive strength using deep learning techniques. *Asian J. Civ. Eng.* <https://doi.org/10.1007/s42107-023-00778-z> (2024).
27. Kashem, A., Das, P., Kashem, A. & Das, P. Compressive strength prediction of high-strength concrete using hybrid machine learning approaches by incorporating SHAP analysis. *Asian J. Civ. Eng.* <https://doi.org/10.1007/s42107-023-00707-0> (2023).
28. Haque, M. A., Chen, B., Kashem, A., Qureshi, T. & Ahmed, A. A. M. Hybrid intelligence models for compressive strength prediction of MPC composites and parametric analysis with SHAP algorithm. *Mater. Today Commun.* <https://doi.org/10.1016/j.mtcomm.2023.105547> (2023).
29. Paul, S., Das, P. & Kashem, A. Sustainable of rice husk ash concrete compressive strength prediction utilizing artificial intelligence techniques. *Asian J. Civ. Eng.* <https://doi.org/10.1007/s42107-023-00847-3> (2023).
30. Kashem, A., Karim, R., Das, P., Datta, S. D. & Alharthai, M. Compressive strength prediction of sustainable concrete incorporating rice husk ash (RHA) using hybrid machine learning algorithms and parametric analyses. *Case Stud. Constr. Mater.* <https://doi.org/10.1016/j.cscm.2024.e03030> (2024).
31. Kashem, A. *et al.* Hybrid data-driven approaches to predicting the compressive strength of ultra-high-performance concrete using SHAP and PDP analyses. *Case Stud. Constr. Mater.* <https://doi.org/10.1016/j.cscm.2024.e02991> (2024).
32. Karim, R., Islam, M. H., Datta, S. D. & Kashem, A. Synergistic effects of supplementary cementitious materials and compressive strength prediction of concrete using machine learning algorithms with SHAP and PDP analyses. *Case Stud. Constr. Mater.* <https://doi.org/10.1016/j.cscm.2023.e02828> (2024).
33. Yamauchi, M. *et al.* Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma. *PLoS One* **7**, e43923. <https://doi.org/10.1371/journal.pone.0043923> (2012).
34. Der, S. D. *et al.* Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. *J. Thorac. Oncol.* **9**, 59–64. <https://doi.org/10.1097/JTO.0000000000000042> (2014).
35. Botling, J. *et al.* Biomarker discovery in non-small cell lung cancer: Integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clin. Cancer Res.* **19**, 194–204. <https://doi.org/10.1158/1078-0432.CCR-12-1139> (2013).

Acknowledgements

This work was supported by Grants from the Natural Science Foundation of Heilongjiang Province of China (Grant No. LH2020H060).

Author contributions

The conception and design of this study were jointly accomplished by YB and LL. During the process of data collection and organization, LL played a leading role, while the analysis and interpretation of the data were primarily undertaken by LL. In the manuscript writing, YB and LL collaborated closely, and the final manuscript was reviewed and approved by both YB and LL.

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/s41598-024-73498-2>.

Correspondence and requests for materials should be addressed to Y.B.

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

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

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

© The Author(s) 2024