Heliyon 10 (2024) e33107

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Assessment of AURKA expression and prognosis prediction in lung adenocarcinoma using machine learning-based pathomics signature

Cuiqing Bai, Yan Sun, Xiuqin Zhang, Zhitong Zuo [*]

Department of Respiratory Disease, Affiliated Hospital of Jiangnan University, Wuxi, China

ARTICLE INFO

Keywords:
Machine learning
Pathomics
AURKA
Lung adenocarcinoma
Prognosis

ABSTRACT

Objective: This study aimed to develop quantitative feature-based models from histopathological images to assess aurora kinase A (AURKA) expression and predict the prognosis of patients with lung adenocarcinoma (LUAD).
Methods: A dataset of patients with LUAD was derived from the cancer genome atlas (TCGA) with information on clinical characteristics, RNA sequencing and histopathological images. The TCGA-LUAD cohort was randomly divided into training (n = 229) and testing (n = 98) sets. We extracted quantitative image features from histopathological slides of patients with LUAD using computational approaches, constructed a predictive model for AURKA expression in the training set, and estimated their predictive performance in the test set. A Cox proportional hazards model was used to assess whether the pathomic scores (PS) generated by the model independently predicted LUAD survival.
Results: High AURKA expression was an independent risk factor for overall survival (OS) in patients with LUAD (hazard ratio = 1.816, 95 % confidence intervals = 1.257–2.623, P = 0.001). The model based on histopathological image features had significant predictive value for AURKA expression: the area under the curve of the receiver operating characteristic curve in the training set and validation set was 0.809 and 0.739, respectively. Decision curve analysis showed that the model had clinical utility. Patients with high PS and low PS had different survival rates (P = 0.019). Multivariate analysis suggested that PS was an independent prognostic factor for LUAD (hazard ratio = 1.615, 95 % confidence intervals = 1.071–2.438, P = 0.022).
Conclusion: Pathomics models based on machine learning can accurately predict AURKA expression and the PS generated by the model can predict LUAD prognosis.

1. Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide, accounting for 21 % of all cancer deaths [1]. Lung adenocarcinoma (LUAD) is the most common pathological subtype of lung cancer among them [2]. In recent years, significant advances have been made in the treatment of patients with lung cancer through surgery, chemotherapy, radiotherapy, targeted therapy and immunotherapy; however, the prognosis of patients with advanced lung cancer remains poor. Traditional prognostic indicators of LUAD, such as carcinoembryonic antigen, and computed tomography (CT) imaging, are no longer able to meet the clinical needs of

[*] Corresponding author. Department of Respiratory Disease, Affiliated Hospital of Jiangnan University, 200 Huihe Road, Wuxi, 214062, China.
E-mail address: zzt555@163.com (Z. Zuo).

https://doi.org/10.1016/j.heliyon.2024.e33107
Received 21 January 2024; Received in revised form 7 June 2024; Accepted 14 June 2024
Available online 14 June 2024
2405-8440/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

precision medicine. Thus, new prognostic markers need to be identified to stratify the prognosis of patients and provide new indicators for the individualization of precision treatment.

Aurora kinase A (AURKA) is an oncogene that is typically expressed in most cell types. AURKA overexpression is associated with several malignant tumor characteristics, including chromosomal instability, mammalian aneuploidy, and abnormal centromere replication [3,4]. Additionally, it is overexpressed in a wide variety of tumors, including breast, ovarian, prostate and head and neck cancers, and is associated with poor prognosis [5–8]. In colorectal cancer models, when ARID1A mutant cells are treated with AURKA inhibitors, these cells are blocked in the G2/M phase and apoptosis is induced; thus targeted inhibition of AURKA may be a new therapeutic strategy for ARID1A mutant colorectal cancer [9,10].

Detection of AURKA expression relies primarily on paraffin-embedded tissue samples, fresh tissue mRNA, and peripheral blood cytokines. However, there are significant flaws in these methods owing to difficulties in real-time detection, specimen collection and detection and high cost. Hematoxylin and eosin (HE)-stained tissue slides are the most readily available imaging materials for clinical diagnosis. Computer-aided image analysis systems have recently been used to evaluate digital pathological images, with the advantages of high accuracy, speed, and consistency [11,12]. The retrieved histopathological imaging features include a variety of morphological and histological data, including cell shape, size, and patterns of the nuclei and cytoplasm texture [13], which reflected information on tumor cells and their surrounding microenvironments [14]. Previous studies have demonstrated that the characteristics of histopathological images have significant potential for outcome prediction, tumor grading, and classification [15,16]. Additionally, digital pathology can act as a link between morphological characteristics and omics profiles (genomics, transcriptomics, and proteomics) to improve tumor characterization and understanding underlying biological processes [17]. In glioblastoma [18] and liver cancer [19], there is a clear link among gene expression, mutations, and histopathological characteristics. Therefore, histopathological image analysis can be used to assess gene expression and evaluate the disease prognosis.

However, no study has focused on the role of AURKA expression in LUAD prognosis using pathomics. Therefore, in this study, we first determined AURKA expression in LUAD samples and then used pathohistological technology to build a pathohistological prediction model to assess AURKA expression in LUAD. We predicted the prognoses of patients with LUAD based on the pathomic scores (PS) generated by the model. Finally, we combined biosynthesis analysis to investigate the underlying biological mechanisms of the histopathological features.

2. Materials and methods

2.1. *Data source*

We collected a dataset for 522 patients with LUAD from the cancer genome atlas (TCGA, https://portal.gdc.cancer.gov/). The inclusion criteria were patients with LUAD, and the exclusion criteria were non-primary non-first diagnosis, missing clinical data, missing follow-up data, a survival time of less than 30 days, and missing RNA-seq samples. Finally, 443 eligible patients were included in this study. We also downloaded the pathology images of 478 patients with LUAD from TCGA, deleted those with poor image quality, and finally screened the H&E-stained histopathology images of 327 patients with corresponding clinical data and RNA-seq, as described above (Fig. 1). Ethical approval was not required, as TCGA database was open to the study. The TCGA database included informed consent for all patients.

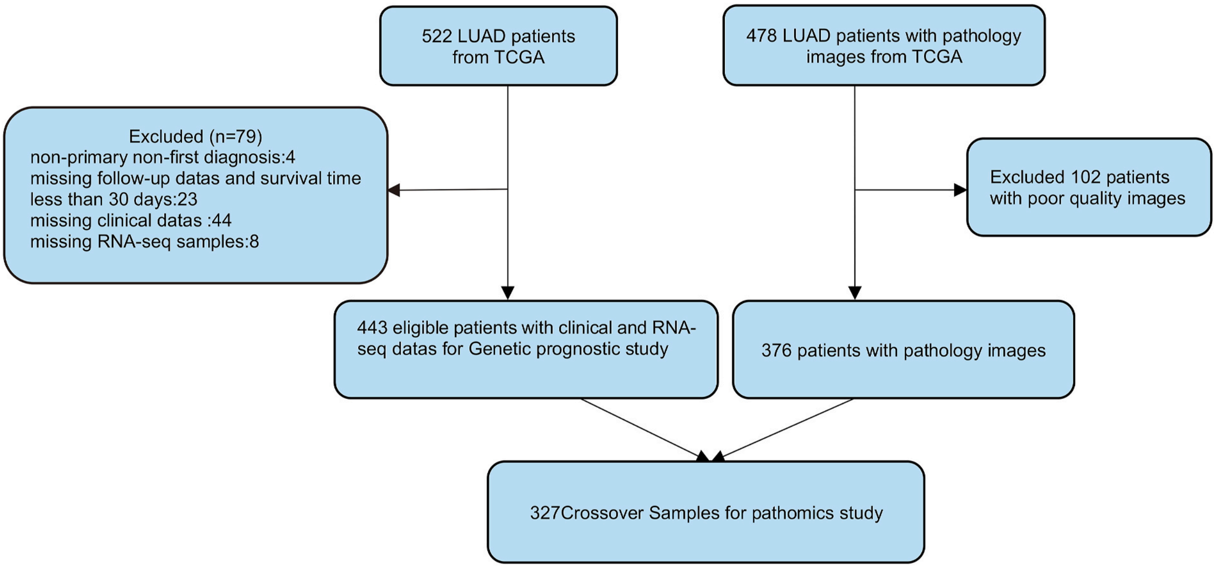

522 LUAD patients from TCGA
478 LUAD patients with pathology images from TCGA
Excluded (n=79)
non-primary non-first diagnosis:4
missing follow-up datas and survival time less than 30 days:23
missing clinical datas :44
missing RNA-seq samples:8
Excluded 102 patients with poor quality images
443 eligible patients with clinical and RNA-seq datas for Genetic prognostic study
376 patients with pathology images
327Crossover Samples for pathomics study

Fig. 1. Inclusion and exclusion flowchart: finally, 443 eligible patients for Genetic prognostic study; 327 eligible patients for pathomics study.

2.2. Genetic prognostic analysis

2.2.1. Groups of patients in the genetic prognosis analysis

The data of 443 patients with LUAD extracted from TCGA were divided into groups with high-expression and low-expression using the cutoff value of AURKA expression (cutoff = 2.749), which was obtained using the R program "survminer" [20,21]. The high-expression group (n = 266) had an expression level >2.749. The low-expression group (n = 177) had an expression level ≤2.749. Wilcox test was used to analyze the relationship between AURKA expression and clinicopathological characteristics (age, sex, and patient age).

2.2.2. Expression of AURKA in tumor tissues and normal tissues

We obtained the LUAD data from TCGA and relevant normal tissue data from Genotype-Tissue Expression (GTEx). The toil technique was used to consistently handle RNA-seq data from TCGA and GTEx in the FPKM format [22]. The Wilcoxon rank-sum test was used to analyze the differences in AURKA expression between tumor and normal tissues.

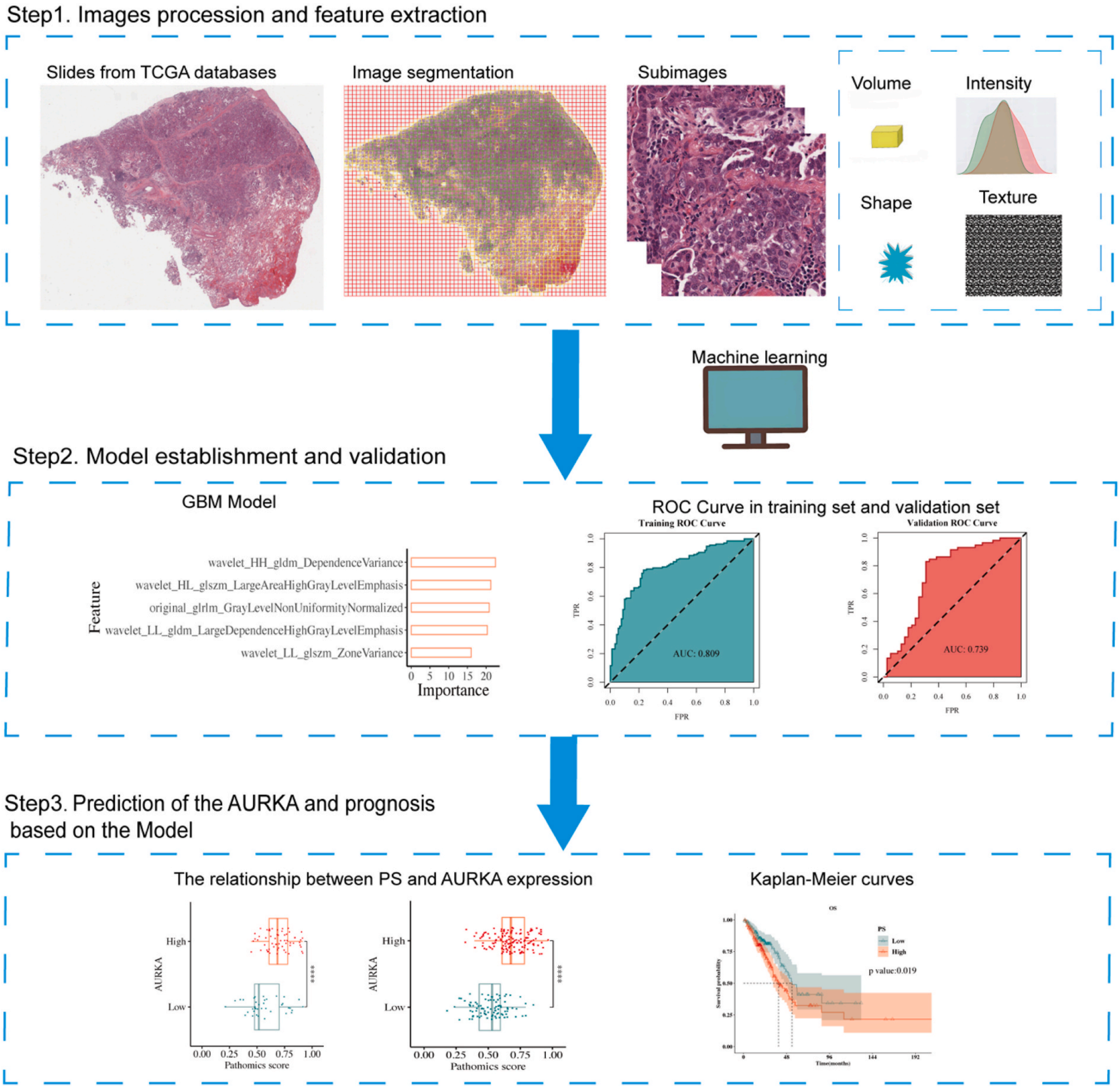

Step1. Images procession and feature extraction
Slides from TCGA databases
Image segmentation
Subimages
Volume
Intensity
Shape
Texture
Machine learning
Step2. Model establishment and validation
GBM Model
Feature
wavelet_HH_gldm_DependenceVariance
wavelet_HL_glszm_LargeAreaHighGrayLevelEmphasis
original_glrlm_GrayLevelNonUniformityNormalized
wavelet_LL_gldm_LargeDependenceHighGrayLevelEmphasis
wavelet_LL_glszm_ZoneVariance
0 5 10 15 20
Importance
ROC Curve in training set and validation set
Training ROC Curve
Validation ROC Curve
TPR
FPR
0.0 0.2 0.4 0.6 0.8 1.0
AUC: 0.809
AUC: 0.739
Step3. Prediction of the AURKA and prognosis based on the Model
The relationship between PS and AURKA expression
AURKA
High
Low
****
0.00 0.25 0.50 0.75 1.00
Pathomics score
Kaplan-Meier curves
OS
PS
Low
High
p value:0.019
Survival probability
1.00 0.75 0.50 0.25 0.00
0 48 96 144 192
Time(months)

Fig. 2. The workflow of data analysis and integration. First, we performed histopathological image processing and feature extraction. Secondly, we integrated features of histopathological images to generate a model using machine learning and evaluated the model's predictive performance in the validation set. Subsequently, the intergroup variability analysis of the gradient boosting algorithm (GBM) model showed that PS could be used to predict AURKA expression. The relationship between PS and OS in patients with LUAD shows that PS could be used to predict LUAD prognosis.

2.2.3. Prognostic analysis of AURKA expression in patients with LUAD

The difference in overall survival (OS) between the AURKA high-expression and low-expression groups was analyzed using Kaplan–Meier survival curves. OS was defined as the time interval from the date of diagnosis to the date of death from any cause, or to the date on which the patient was last known to be alive. The log-rank test was used to test the significance of survival rates between groups. Univariate and multivariate Cox proportional hazards models were used to assess whether AURKA was an independent predictor of survival.

2.2.4. Subgroup and interaction analyses

One-way Cox regression was used for exploratory subgroup analysis to examine the impact of AURKA expression on patient prognosis (high-vs low-expression groups) in various subgroups for each covariate. A likelihood ratio test was used to examine the interaction between AURKA expression and other variables.

2.3. Pathomics method

The overall framework was summarized in Fig. 2. The details of each section were described in following parts.

2.3.1. Histopathological image features

We obtained 478H&E-stained whole-slide histopathological images from TCGA [23,24]. A considerable percentage of white backgrounds (between 40 and 80 %) may be observed in whole-slide histopathological images. However, this is often irrelevant for cancer analyses. Removing such non-informative regions can markedly reduce the computational cost while ensuring the validity of the training samples. Therefore, we used the OTSU algorithm (https://opencv.org/), which is currently the most commonly used pathological image processing algorithm in pathomics research, to obtain the tissue regions of pathology sections [25–28]. To be able to incorporate samples with different magnifications, increase the number of samples for model training, and thus increase the accuracy of the model's predictions, we used upsampling [29], which maintains a consistent input resolution while preserving the details and texture information in the image. Increased sample size improves prediction accuracy, and 20 × and 40 × are common whole slide imaging (WSI) magnifications [30,31]. We divided the 40 × images into multiple sub-images of 1024 × 1024 pixels, and divided the 20 × images into multiple sub-images of 512 × 512 pixels to obtain the same perspective. We then resized the 512 × 512 pixels sub-images to 1024 × 1024 pixels for further analysis. Sub-images with poor image quality (contamination, blurred images, and more than 50 % white background) were excluded [19,32], and 327 eligible images of the pathology sections were obtained. To decrease the computational cost, we randomly selected 10 sub-images from each pathology image for subsequent analysis. As in Fig. 2, we showed representative H&E tiles. Using the open-source package PyRadiomics, 465 pathomic features were extracted from each image tile for analysis. The mean value of each extracted feature from the selected 10 tiles was summarized to represent the value of the corresponding slide for further statistical analyses [33–36].

2.3.2. TCGA pathology crossover samples

Patients (n = 327) with available pathological slides and clinical data were split into high- and low-expression groups using the R program "survminer" and a cutoff value of 2.7486 was set for AURKA expression. We randomly divided the data from these 327 samples into a training set (n = 229) and a validation set (n = 98) in a 7:3 ratio. For the training set, the pathohistological feature values (465 features retrieved by the PyRadiomics software) were z-score-normalized, and the validation set was standardized using the mean and standard deviation of the training set. The between-group variability of the clinical variables across the datasets was analyzed.

2.3.3. Construction of a predictive model

The training set data were used for feature screening and modeling. We first ranked feature importance using maximum relevance minimum redundancy (mRMR) and then selected features using recursive feature elimination(RFE). Finally, the top-ranking features were reserved to construct the model using the GBM algorithm.

2.3.4. GBM model evaluation methodology

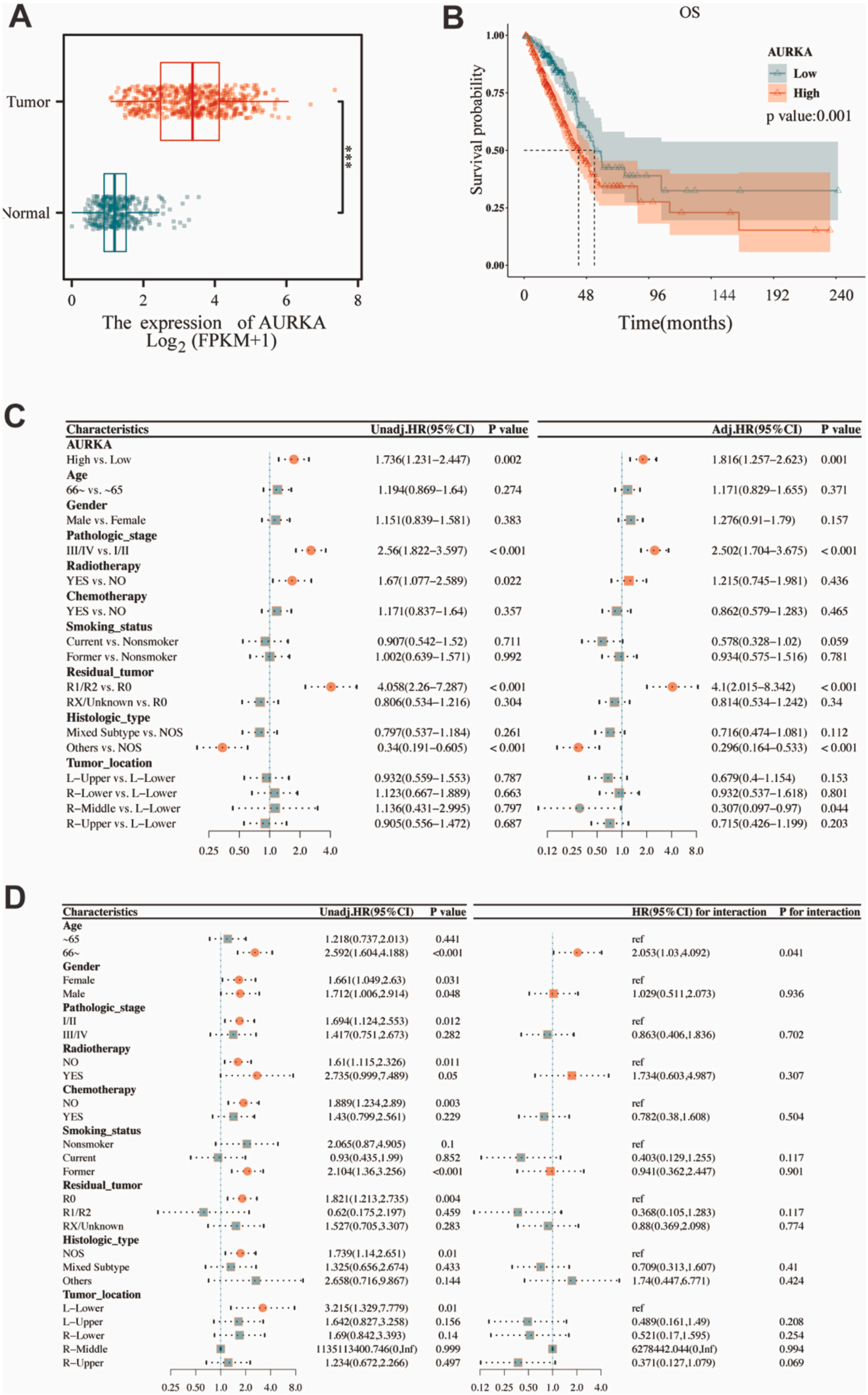
Receiver operating characteristic (ROC) analysis was used to evaluate the efficacy of the GBM model. The evaluation indicators are accuracy (ACC), specificity (SPE), sensitivity (SEN), predictability (PPV), and negativity (NPV). The pROC package in R was used to determine the area under the ROC curve (AUC) and assess the overall performance of the model. Precision-recall (PR) curves were used to evaluate the efficacy of the model by emphasizing positive samples. The PR-AUC is the average of the accuracy values determined for each coverage threshold. Calibration curves were applied to evaluate the goodness-of-fit between the model-predicted value and the true value of AURKA expression. The Brier score was used to quantify the combined performance of the prediction model, and the decision curves were plotted to determine the clinical utility of the model.

2.3.5. Analysis of intergroup variability in the GBM model

We compared the difference in PS between the high and low AURKA subgroups using the Wilcoxon test.

2.3.6. Relationship between PS and LUAD prognosis

Based on the pathomics model, PS was calculated for each of the 327 samples. We merged the PS with clinical data and calculated



Characteristics	Unadj.HR(95%CI)	P value	Adj.HR(95%CI)	P value
AURKA				
High vs. Low	1.736(1.231–2.447)	0.002	1.816(1.257–2.623)	0.001
Age				
66~ vs. ~65	1.194(0.869–1.64)	0.274	1.171(0.829–1.655)	0.371
Gender				
Male vs. Female	1.151(0.839–1.581)	0.383	1.276(0.91–1.79)	0.157
Pathologic_stage				
III/IV vs. I/II	2.56(1.822–3.597)	< 0.001	2.502(1.704–3.675)	< 0.001
Radiotherapy				
YES vs. NO	1.67(1.077–2.589)	0.022	1.215(0.745–1.981)	0.436
Chemotherapy				
YES vs. NO	1.171(0.837–1.64)	0.357	0.862(0.579–1.283)	0.465
Smoking_status				
Current vs. Nonsmoker	0.907(0.542–1.52)	0.711	0.578(0.328–1.02)	0.059
Former vs. Nonsmoker	1.002(0.639–1.571)	0.992	0.934(0.575–1.516)	0.781
Residual_tumor				
R1/R2 vs. R0	4.058(2.26–7.287)	< 0.001	4.1(2.015–8.342)	< 0.001
RX/Unknown vs. R0	0.806(0.534–1.216)	0.304	0.814(0.534–1.242)	0.34
Histologic_type				
Mixed Subtype vs. NOS	0.797(0.537–1.184)	0.261	0.716(0.474–1.081)	0.112
Others vs. NOS	0.34(0.191–0.605)	< 0.001	0.296(0.164–0.533)	< 0.001
Tumor_location				
L–Upper vs. L–Lower	0.932(0.559–1.553)	0.787	0.679(0.4–1.154)	0.153
R–Lower vs. L–Lower	1.123(0.667–1.889)	0.663	0.932(0.537–1.618)	0.801
R–Middle vs. L–Lower	1.136(0.431–2.995)	0.797	0.307(0.097–0.97)	0.044
R–Upper vs. L–Lower	0.905(0.556–1.472)	0.687	0.715(0.426–1.199)	0.203

Characteristics	Unadj.HR(95%CI)	P value	HR(95%CI) for interaction	P for interaction
Age				
~65	1.218(0.737,2.013)	0.441	ref	
66~	2.592(1.604,4.188)	<0.001	2.053(1.03,4.092)	0.041
Gender				
Female	1.661(1.049,2.63)	0.031	ref	
Male	1.712(1.006,2.914)	0.048	1.029(0.511,2.073)	0.936
Pathologic_stage				
I/II	1.694(1.124,2.553)	0.012	ref	
III/IV	1.417(0.751,2.673)	0.282	0.863(0.406,1.836)	0.702
Radiotherapy				
NO	1.61(1.115,2.326)	0.011	ref	
YES	2.735(0.999,7.489)	0.05	1.734(0.603,4.987)	0.307
Chemotherapy				
NO	1.889(1.234,2.89)	0.003	ref	
YES	1.43(0.799,2.561)	0.229	0.782(0.38,1.608)	0.504
Smoking_status				
Nonsmoker	2.065(0.87,4.905)	0.1	ref	
Current	0.93(0.435,1.99)	0.852	0.403(0.129,1.255)	0.117
Former	2.104(1.36,3.256)	<0.001	0.941(0.362,2.447)	0.901
Residual_tumor				
R0	1.821(1.213,2.735)	0.004	ref	
R1/R2	0.62(0.175,2.197)	0.459	0.368(0.105,1.283)	0.117
RX/Unknown	1.527(0.705,3.307)	0.283	0.88(0.369,2.098)	0.774
Histologic_type				
NOS	1.739(1.14,2.651)	0.01	ref	
Mixed Subtype	1.325(0.656,2.674)	0.433	0.709(0.313,1.607)	0.41
Others	2.658(0.716,9.867)	0.144	1.74(0.447,6.771)	0.424
Tumor_location				
L–Lower	3.215(1.329,7.779)	0.01	ref	
L–Upper	1.642(0.827,3.258)	0.156	0.489(0.161,1.49)	0.208
R–Lower	1.69(0.842,3.393)	0.14	0.521(0.17,1.595)	0.254
R–Middle	1135113400.746(0,Inf)	0.999	6278442.044(0,Inf)	0.994
R–Upper	1.234(0.672,2.266)	0.497	0.371(0.127,1.079)	0.069

(caption on next page)

Fig. 3. (A) Expression of AURKA in tumor tissues and normal tissues. (B) Kaplan–Meier survival curves of high- and low-AURKA expression groups in patients with LUAD. (C) Univariate and multivariate Cox regression analyses the effect of AURKA expression and clinicopathologic features on patient prognosis. (D) Subgroup and interaction analyses the effect of AURKA expression on patient prognosis in different subgroups of each covariate.

the cutoff value of the PS using the Survminer package to classify it as a low/high dichotomous variable. We investigated the prognostic role of PS in patients with LUAD, similar to a previous statistical method for molecular prognosis.

2.4. Pathohistologic mechanism analysis

2.4.1. Enrichment analysis of differentially expressed genes between high- and low-PS subgroups

R software "cluster Profiler" was used to perform gene set enrichment analysis (GSEA) of the KEGG (c2.cp.kegg.v7.5.1. symbols. gmt) and Hallmark (h.all.v7.5.1.symbols.gmt) gene sets to explore the underlying molecular mechanisms of gene expression differences between the high/low PS subgroups.

2.4.2. Analysis of differentially expressed genes associated with the cell cycle pathway

The Wilcoxon test was used to assess the differential expression of genes related to the cell cycle pathway between the high- and low-PS groups. Box plots were generated to display the results.

2.4.3. Analysis of differential immune cell abundance

The gene expression matrix of patients with LUAD was uploaded to the CIBERSORTx database (https://cibersortx.stanford.edu/) and the extent of immune cell infiltration in each sample was computed. The R package "limma" was used to examine the extent of immune cell infiltration between high- and low-PS groups. Immune infiltration involves multiple genes for simultaneous analysis, and limma, which can improve the statistical efficacy and reliability of high-throughput gene expression data, is more appropriate for such analyses.

2.4.4. Analysis of differential drug sensitivity

We downloaded the IC50 values of 198 drugs from the GDSC database (http://www.cancerrxgene.org/), used the R package "oncoPredict" to predict the IC50 values of each sample based on RNA-seq, and analyzed the differences in IC50 values between the high- and low-PS groups using the Wilcoxon test.

2.4.5. Mutation and TMB analyses of LUAD

We performed mutation and tumor mutation burden (TMB) analyses of LUAD. Mutation data for TCGA-LUAD patients was downloaded from TCGA Data Portal (https://portal.gdc.cancer.gov/). The sample size at the intersection of the pathomics data was 322 patients. Data for somatic cell variants were stored in the Mutation Annotation Format, and mutation data were analyzed using the R package maftools. The high- and low-PS groups were the same as those described in Section 2.3.6.

2.5. Statistical analysis

Statistical analyses were performed using R version 3.6.3. The Wilcoxon signed-rank test was used to analyze numerical variables. The specificity and sensitivity of the machine-learning-based model were assessed using an ROC curve with an AUC value. Survival outcomes were shown as Kaplan–Meier survival curves and compared using the log-rank test. Cox regression analysis was used to compute the hazard ratio (HR) and 95 % confidence intervals (CIs). Results with $P < 0.05$ were considered statistically significant. The R package "limma" was used to examine the level of immune cell infiltration between high- and low-PS groups.

3. Results

3.1. Genetic prognostic analysis

3.1.1. Expression of AURKA in tumor tissues and normal tissues

AURKA expression in tumor tissues was higher than that in normal tissues. The median difference between the two groups was 2.1231 (1.9951–2.2486), which was statistically significant ($P < 0.001$) (Fig. 3A).

3.1.2. Baseline information of the patients in the genetic prognosis analysis

A total of 266 patients with LUAD were in the high AURKA expression group and 177 patients were in the low AURKA expression group, with a cutoff value of 2.749. No differences were observed in pathological stage, radiotherapy administered, chemotherapy administered, residual tumor after surgery, or tumor location between patients in the high and low AURKA expression groups ($P > 0.05$), but there were significant differences in age, sex, smoking status, and pathological subtype of patients, suggesting that AURKA expression was related to age, sex, smoking status, and pathological subtype of patients (Table 1).

3.1.3. Prognostic value of AURKA in patients with LUAD

Fig. 3B shows significant differences in OS between the high- and low-AURKA expression groups (P = 0.001). The median OS intervals were 41.93 and 54.07 months, respectively. This suggests that increased AURKA expression is associated with a poor prognosis in patients with LUAD.

3.1.4. Cox proportional hazards analysis

Univariate Cox regression analysis (Fig. 3C) revealed that high AURKA expression was a significant risk factor for OS (HR = 1.736, 95 % CI = 1.231–2.447, P = 0.002). Radiotherapy (HR = 1.67, 95 % CI = 1.077–2.589, P = 0.022) and R1/R2 (HR = 4.058, 95 % CI = 2.26–7.287, P < 0.001) versus R0 were significant risk factors for OS in patients with LUAD. Other pathologic subtypes versus not otherwise specified (NOS) pathologic type were protective factors for OS (HR = 0.34, 95 % CI = 0.191–0.605, P < 0.001). Other variables (age, sex, chemotherapy, smoking status, and tumor location) did not have statistically significant effects on OS (P > 0.05).

In multivariate Cox regression analysis (Fig. 3C), we observed that after multifactorial adjustment, high AURKA expression (HR = 1.816, 95 % CI = 1.257–2.623, P = 0.001) remained a statistically significant risk factor for OS. High pathological stage (HR = 2.502, 95 % CI = 1.704–3.675, P < 0.001) and residual tumor R1/R2 (HR = 4.1, 95%CI = 2.015–8.342, P < 0.001) were statistically significant risk factors for OS. The other pathological subtypes (HR = 0.296, 95 % CI = 0.164–0.533, P < 0.001) and tumor location in the middle lobe of the right lung (HR = 0.307, 95 % CI = 0.097–0.97, P = 0.044) were statistically significant protective factors for OS. This indicated that AURKA expression status, pathological stage, residual tumor, pathological subtype, and tumor location in the middle lobe of the right lung were independent prognostic factors for LUAD, and high AURKA expression was associated with poor prognosis in patients with LUAD. Other variables, including age, sex, chemotherapy, radiotherapy, and smoking status, were not significantly associated with the prognosis of patients with LUAD (P > 0.05).

3.1.5. Subgroup and interaction analyses

One-way Cox regression was used for exploratory subgroup analysis to determine the effect of AURKA expression (high vs low expression groups) on patient prognosis in various subgroups for each covariate. We observed that elevated AURKA expression was a risk factor for OS in the subgroup aged ≤65 years (HR = 1.218, 95 % CI = 0.737–2.013, P = 0.441), although not statistically significant; in the subgroup aged >65 years, elevated AURKA was a risk factor for OS (HR = 2.592, 95 % CI = 1.604–4.188, P < 0.001),

Table 1
Descriptive statistics of patient characteristics.

Variables	Total(n = 443)	Low(n = 177)	High(n = 266)	P
Age, n (%)				0.027
≤65	215(49)	74(42)	141(53)	
>65	228(51)	103(58)	125(47)	
Sex, n (%)				<0.001
Female	244(55)	118(67)	126(47)	
Male	199(45)	59(33)	140(53)	
Pathologic-stage, n (%)				0.072
I/II	353(80)	149(84)	204(77)	
III/IV	90(20)	28(16)	62(23)	
Radiotherapy, n (%)				0.144
No	395(89)	163(92)	232(87)	
Yes	48(11)	14(8)	34(13)	
Chemtherapy				0.219
No	289(65)	122(69)	167(63)	
Yes	154(35)	55(31)	99(37)	
Smoking status, n (%)				<0.001
Current	105(24)	20(11)	85(32)	
Former	271(61)	120(68)	151(57)	
Nonsmoker	67(15)	37(21)	30(11)	
Residual tumor, n (%)				0.656
R0	297(67)	117(66)	180(68)	
R1/R2	16(4)	5(3)	11(4)	
RX/Unknown	130(29)	55(31)	75(28)	
Histologic type, n (%)				0.006
Mixed Subtype	97(22)	49(28)	48(18)	
NOS	270(61)	92(52)	178(67)	
Others	76(17)	36(20)	40(15)	
Tumor location, n (%)				0.203
L- Lower	69(16)	34(19)	35(13)	
L- Upper	111(25)	38(21)	73(27)	
R-Lower	85(19)	39(22)	46(17)	
R-Middle	20(5)	8(5)	12(5)	
R-Upper	158(36)	58(33)	100(38)	

Histologic type: NOS: Not Otherwise Specified; Others: Acinar, Bronchioloalveolar, Clear Cell, Micropapillary, Papillary, Signet Ring, Solid, Mucinous.

and this result was statistically significant. The P value of the interaction test was 0.041, which indicated statistical significance and that there was a significant interaction between AURKA expression and the different age subgroups. Thus, the effect of AURKA on OS differed between the two age subgroups. The interaction of AURKA in different subgroups, including sex, pathological stage, radiotherapy, chemotherapy, smoking status, residual tumor, pathological subtype, and tumor location were statistically insignificant, (all $P > 0.05$), suggesting that the effect of AURKA expression on OS was not significantly different between the different subgroups (Fig. 3D).

3.2. Histopathology results

3.2.1. Baseline information on TCGA pathology crossover samples

We discovered that variables, including AURKA expression, age, sex, pathologic stage, radiation therapy, chemotherapy, smoking status, residual tumor, pathologic subtype, and tumor location, did not differ significantly ($P > 0.05$) between the training ($n = 229$) and validation ($n = 98$) sets. This indicated that the baseline conditions of the patients in the training and validation sets were comparable (Table 2).

3.2.2. GBM model building

Using the mRMR approach, the first 30 features were selected, and RFE was used to continue feature screening. As shown in Fig. 4A, a schematic of the RFE feature screening process is shown, and five features were obtained from the final screening process. Fig. 4B illustrates the significance of the filtered characteristics of the GBM algorithm.

3.2.3. GBM model evaluation

The AUC value of the model in the training set was 0.809, with a 95 % CI of 0.752–0.866 (Fig. 4C), and the AUC value in the validation set was 0.739, with a 95 % CI of 0.629–0.849 (Fig. 4D). In the PR curve, the AUC_{PR} values of the model were 0.865 and 0.749

Table 2
Descriptive statistics of patient characteristics in the training and validation sets.

Variables	Total(n = 327)	Train(n = 229)	Validation(n = 98)	P
AURKA, n (%)				1
Low	130(40)	91(40)	39(40)	
High	197(60)	138(60)	59(60)	
Age n (%)				0.744
≤65	164(50)	113(49)	51(52)	
>65	163(50)	116(51)	47(48)	
Sex, n (%)				1
Female	183(56)	128(56)	55(56)	
Male	144(44)	101(44)	43(44)	
Pathologic-stage, n (%)				1
I/II	265(81)	186(81)	79(81)	
III/IV	62(19)	43(19)	19(19)	
Radiotherapy, n (%)				0.785
NO	293(90)	204(89)	89(91)	
YES	34(10)	25(11)	9(9)	
Chemotherapy, n (%)				0.771
NO	219(67)	155(68)	64(65)	
YES	108(33)	74(32)	34(35)	
Smoking-status, n (%)				0.201
Nonsmoker	41(13)	24(10)	17(17)	
Current	82(25)	57(25)	25(26)	
Former	204(62)	148(65)	56(57)	
Residual-tumor, n (%)				0.12
R0	220(67)	159(69)	61(62)	
R1/R2	13(4)	6(3)	7(7)	
RX/Unknown	94(29)	64(28)	30(31)	
Histologic-type, n (%)				0.31
NOS	204(62)	148(65)	56(57)	
Mixed Subtype	67(20)	42(18)	25(26)	
Others	56(17)	39(17)	17(17)	
Tumor-location, n (%)				0.916
L-Lower	56(17)	39(17)	17(17)	
L-Upper	76(23)	50(22)	26(27)	
R-Lower	63(19)	45(20)	18(18)	
R-middle	14(4)	10(4)	4(4)	
R-Upper	118(36)	85(37)	33(34)	
OS, n (%)				0.272
Alive	213(65)	154(67)	59(60)	
Dead	114(35)	75(33)	39(40)	

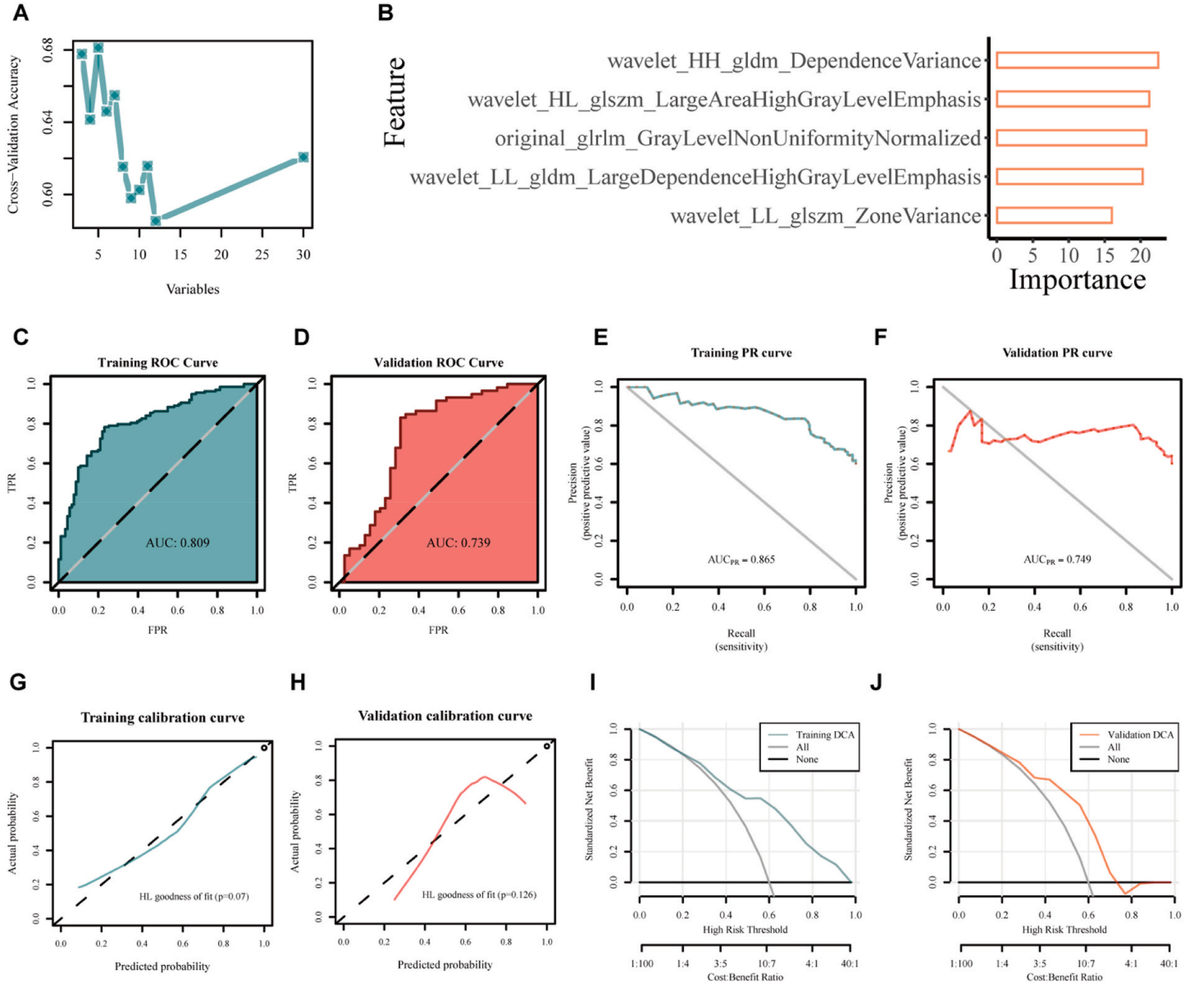


Fig. 4. (A)Schematic of the RFE feature screening. (B)The significance of the filtered characteristics of the GBM algorithm. (C–J) Validation of GBM model for AURKA prediction: C and D were the Receiver operating characteristic (ROC) curves of training set and validation set respectively. E and F were the precision-recall (PR) curves of training set and validation set respectively. G and H were the calibration curves of training set and validation set respectively. I and J were the decision curves (DCA) curves of training set and validation set respectively.

for the training and validation sets, respectively (Fig. 4E and F). These results indicated that the pathohistological model had good predictive ability. In both the training and validation sets, the calibration curves showed good agreement ($P > 0.05$) between the predicted probability and the true value of the AURKA expression (Fig. 4G and H). The decision curves suggested that the model has clinical utility (Fig. 4I and J).

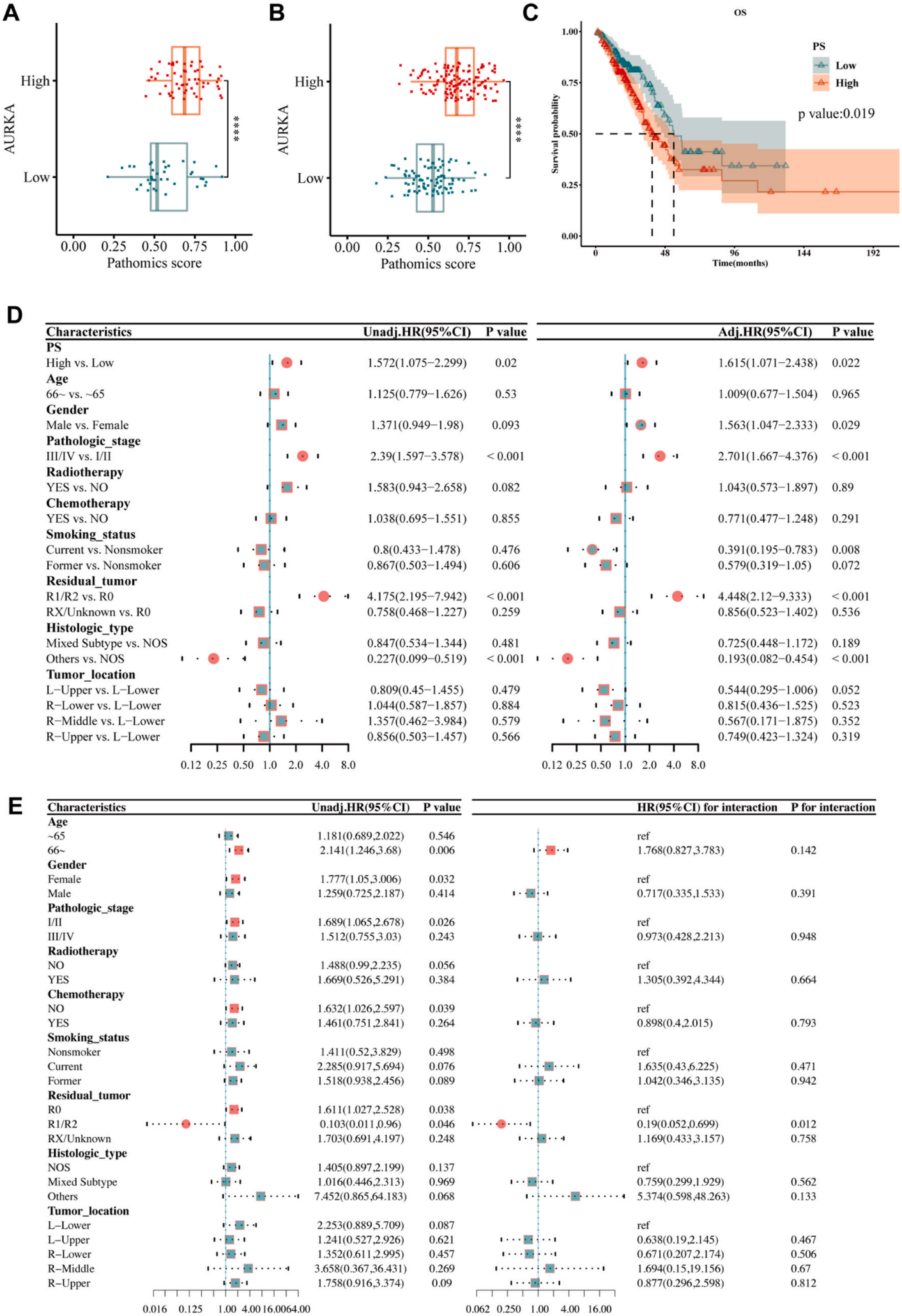
The training set had a threshold of 0.599, an accuracy of 0.777, a sensitivity of 0.783, a specificity of 0.769, and a Brier score of 0.186, while the accuracy of the validation set was 0.745, a sensitivity of 0.78, a specificity of 0.692, and a Brier score of 0.203. This suggests that our established pathomics model has a better accuracy, sensitivity, and specificity.

3.2.4. Intergroup variability analysis of the GBM model

In the training set (Fig. 5A), PS was significantly different between the high AURKA expression group and the low AURKA expression group, $P < 0.001$. In the high AURKA expression group, the PS value was high. Similarly, in the test set (Fig. 5B), PS was significantly different between the high AURKA expression group and the low AURKA expression group, $P < 0.001$. In the high AURKA expression group, PS was high. Moreover, it was suggested that the higher the PS, the higher the AURKA expression. This suggests that the PS generated by the model can be used to predict AURKA expression.

3.2.5. Clinical baseline data for high and low PS groups

We calculated a cutoff value of 0.6098 for PS using the Survminer package and categorized the patients into high PS ($n = 177$) and low PS ($n = 150$) groups. No significant difference was observed in the age distribution between the high- and low-PS groups (P =



Characteristics	Unadj.HR(95%CI)	P value	Adj.HR(95%CI)	P value
PS				
High vs. Low	1.572(1.075−2.299)	0.02	1.615(1.071−2.438)	0.022
Age				
66~ vs. ~65	1.125(0.779−1.626)	0.53	1.009(0.677−1.504)	0.965
Gender				
Male vs. Female	1.371(0.949−1.98)	0.093	1.563(1.047−2.333)	0.029
Pathologic_stage				
III/IV vs. I/II	2.39(1.597−3.578)	< 0.001	2.701(1.667−4.376)	< 0.001
Radiotherapy				
YES vs. NO	1.583(0.943−2.658)	0.082	1.043(0.573−1.897)	0.89
Chemotherapy				
YES vs. NO	1.038(0.695−1.551)	0.855	0.771(0.477−1.248)	0.291
Smoking_status				
Current vs. Nonsmoker	0.8(0.433−1.478)	0.476	0.391(0.195−0.783)	0.008
Former vs. Nonsmoker	0.867(0.503−1.494)	0.606	0.579(0.319−1.05)	0.072
Residual_tumor				
R1/R2 vs. R0	4.175(2.195−7.942)	< 0.001	4.448(2.12−9.333)	< 0.001
RX/Unknown vs. R0	0.758(0.468−1.227)	0.259	0.856(0.523−1.402)	0.536
Histologic_type				
Mixed Subtype vs. NOS	0.847(0.534−1.344)	0.481	0.725(0.448−1.172)	0.189
Others vs. NOS	0.227(0.099−0.519)	< 0.001	0.193(0.082−0.454)	< 0.001
Tumor_location				
L−Upper vs. L−Lower	0.809(0.45−1.455)	0.479	0.544(0.295−1.006)	0.052
R−Lower vs. L−Lower	1.044(0.587−1.857)	0.884	0.815(0.436−1.525)	0.523
R−Middle vs. L−Lower	1.357(0.462−3.984)	0.579	0.567(0.171−1.875)	0.352
R−Upper vs. L−Lower	0.856(0.503−1.457)	0.566	0.749(0.423−1.324)	0.319

Characteristics	Unadj.HR(95%CI)	P value	HR(95%CI) for interaction	P for interaction
Age				
~65	1.181(0.689,2.022)	0.546	ref	
66~	2.141(1.246,3.68)	0.006	1.768(0.827,3.783)	0.142
Gender				
Female	1.777(1.05,3.006)	0.032	ref	
Male	1.259(0.725,2.187)	0.414	0.717(0.335,1.533)	0.391
Pathologic_stage				
I/II	1.689(1.065,2.678)	0.026	ref	
III/IV	1.512(0.755,3.03)	0.243	0.973(0.428,2.213)	0.948
Radiotherapy				
NO	1.488(0.99,2.235)	0.056	ref	
YES	1.669(0.526,5.291)	0.384	1.305(0.392,4.344)	0.664
Chemotherapy				
NO	1.632(1.026,2.597)	0.039	ref	
YES	1.461(0.751,2.841)	0.264	0.898(0.4,2.015)	0.793
Smoking_status				
Nonsmoker	1.411(0.52,3.829)	0.498	ref	
Current	2.285(0.917,5.694)	0.076	1.635(0.43,6.225)	0.471
Former	1.518(0.938,2.456)	0.089	1.042(0.346,3.135)	0.942
Residual_tumor				
R0	1.611(1.027,2.528)	0.038	ref	
R1/R2	0.103(0.011,0.96)	0.046	0.19(0.052,0.699)	0.012
RX/Unknown	1.703(0.691,4.197)	0.248	1.169(0.433,3.157)	0.758
Histologic_type				
NOS	1.405(0.897,2.199)	0.137	ref	
Mixed Subtype	1.016(0.446,2.313)	0.969	0.759(0.299,1.929)	0.562
Others	7.452(0.865,64.183)	0.068	5.374(0.598,48.263)	0.133
Tumor_location				
L−Lower	2.253(0.889,5.709)	0.087	ref	
L−Upper	1.241(0.527,2.926)	0.621	0.638(0.19,2.145)	0.467
R−Lower	1.352(0.611,2.995)	0.457	0.671(0.207,2.174)	0.506
R−Middle	3.658(0.367,36.431)	0.269	1.694(0.15,19.156)	0.67
R−Upper	1.758(0.916,3.374)	0.09	0.877(0.296,2.598)	0.812

(caption on next page)

Fig. 5. (A, B) Intergroup variability analysis of the GBM model. (C) Kaplan–Meier survival curves of high-PS and low-PS groups in patients with LUAD. (D) Univariate and multivariate Cox regression analyses the effect of PS and clinicopathologic features on OS in patients with LUAD. (E) Subgroup and interaction analyses the effect of PS on OS in patients with LUAD in different subgroups of each covariate.

0.408). Significant differences were observed in the distributions of sex, smoking status, and pathologic subtypes between the high- and low-PS groups ($P < 0.05$). No significant differences were observed in the distribution of pathologic stage, radiotherapy, chemotherapy, residual tumor, or tumor location between the high- and low-PS groups ($P > 0.05$) (Table 3).

3.2.6. Relationship between PS and OS in patients with LUAD

The median survival time was 39.03 months in the high PS group and 54.07 months in the low PS group, which was statistically significant and suggested that high PS was significantly associated with poor prognosis in LUAD patients ($P = 0.019$) (Fig. 5C).

3.2.7. Cox proportional hazards analysis in pathomics

High PS was a risk factor for OS in the univariate analysis (HR = 1.572, 95%CI = 1.075–2.299, $P = 0.02$). Pathologic stage III/IV (HR = 2.39, 95 % CI = 1.597–3.578, $P < 0.001$) and Residual tumor R1/R2 type (HR = 4.175, 95 % CI = 2.195–7.942, $P < 0.001$) were significant risk factors for OS. Others pathologic subtypes were protective factors for OS (HR = 0.227, 95 % CI = 0.099–0.519, $P < 0.001$). Other variables (age, sex, radiotherapy, chemotherapy, smoking status, and tumor location) had no significant effect on OS ($P > 0.05$) (Fig. 5D).

In the multifactorial analysis, after adjustment, high PS (HR = 1.615, 95 % CI = 1.071–2.438, $P = 0.022$) was a statistically significant risk factor for OS. Male sex (HR = 1.563, 95 % CI = 1.047–2.333, $P = 0.029$) and pathologic stage III/IV (HR = 2.701, 95 % CI = 1.667–4.376, $P < 0.001$) were significant risk factors for OS. Other pathologic subtypes (HR = 0.193, 95 % CI = 0.082–0.454, $P < 0.001$) were significant protective factors for OS. Other variables (including age, radiotherapy, chemotherapy, smoking status, and tumor location) had no significant effect on OS ($P > 0.05$) (Fig. 5D).

3.2.8. Subgroup analyses and interaction tests

In the subgroup of age less than or equal to 65 years, elevated PS was a risk factor for OS (HR = 1.181, 95 % CI = 0.689–2.022, $P = 0.546$), although not statistically significant; in the subgroup of age greater than or equal to 66 years, elevated PS was a risk factor for OS (HR = 2.141, 95 % CI = 1.246–3.68, $P = 0.006$), which was statistically significant. The p-value of interaction test was 0.142, which

Table 3
Clinical baseline data for the high- and low-PS groups.

Variables	Total(n = 327)	Low(n = 150)	High(n = 177)	p
Age, n (%)				0.408
≤65	164(50)	71(47)	93(53)	
>65	163(50)	79(53)	84(47)	
Sex, n (%)				0.01
Female	183(56)	96(64)	87(49)	
Male	144(44)	54(36)	90(51)	
Pathologic_ stage, n (%)				0.764
I/II	265(81)	120(80)	145(82)	
III/IV	62(19)	30(20)	32(18)	
Radiotherapy, n (%)				0.136
NO	293(90)	139(93)	154(87)	
YES	34(10)	11(7)	23(13)	
Chemotherapy, n (%)				0.821
NO	219(67)	99(66)	120(68)	
YES	108(33)	51(34)	57(32)	
Smoking_ status, n (%)				<0.001
Nonsmokers	41(13)	28(19)	13(7)	
Current	82(25)	27(18)	55(31)	
Former	204(62)	95(63)	109(62)	
Residual_ tumor, n (%)				0.896
R0	220(67)	99(66)	121(68)	
R1/R2	13(4)	6(4)	7(4)	
RX/Unknown	94(29)	45(30)	49(28)	
Histologic_ type, n (%)				0.896
NOS	204(62)	82(55)	122(69)	
Mixed Subtype	67(20)	37(25)	30(17)	
Others	56(17)	31(20)	25(14)	
Tumor_ location, n (%)				0.634
L-Lower	56(17)	27(18)	29(16)	
L-Upper	76(23)	29(19)	47(27)	
R-Lower	63(19)	31(21)	32(18)	
R-Middle	14(4)	6(4)	8(5)	
R-Upper	118(36)	57(38)	61(34)	

indicated lack of significant interaction between PS and different age subgroups. Thus, the effect of PS on OS was similar between the two age subgroups.

The PS interaction between different subgroups for sex, pathological stage, radiotherapy, chemotherapy, smoking status, pathological subtype, and tumor location were statistically insignificant (all $p > 0.05$), suggesting that the effect of PS on OS was the same between these different subgroups. This suggests that PS generated by the model is an independent prognostic factor for LUAD

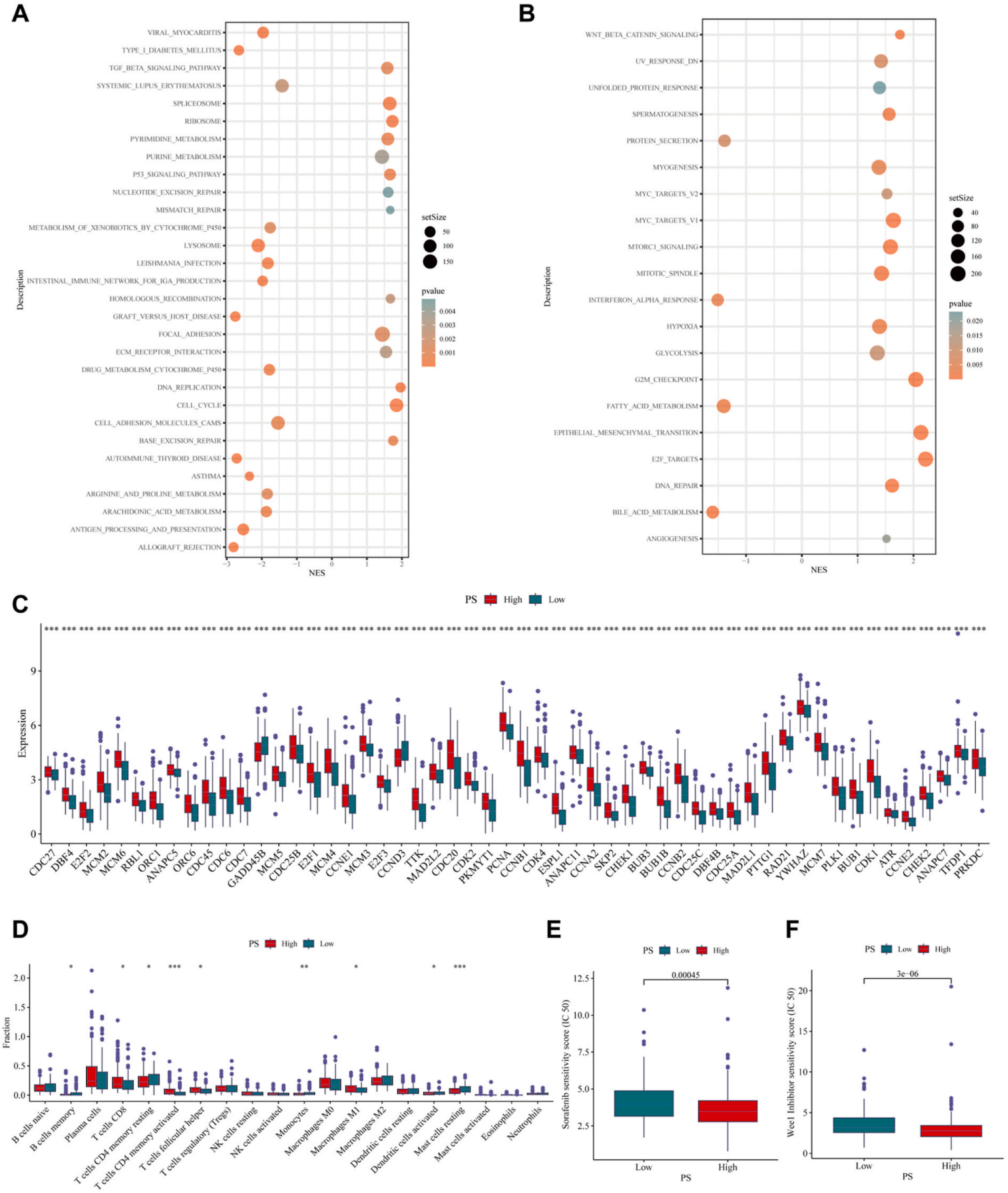


Fig. 6. (A) Enrichment analysis of the KEGG gene set. (B) Enrichment analysis of the Hallmark gene set. (C) Analysis of differentially expressed genes associated with the cell cycle pathway. (D) Analysis of differential immune cell abundance. (E, F) Analysis of differential drug sensitivity.

(Fig. 5E).

3.3. *Pathohistologic mechanism analysis*

3.3.1. *Enrichment analysis of differentially expressed genes between high- and low-PSsubgroups*

The top 30 pathways were visualized using GSEA of the KEGG gene collection. We observed that the differentially expressed genes in the high PS group were significantly enriched in signaling pathways, such as CELL_CYCLE and p53 (P53_SIGNALING_PATHWAY) signaling pathways (Fig. 6A). GSEA revealed the complete spectrum of pathways in the hallmark gene collection. We observed that differentially expressed genes in the PS-high group were significantly enriched in signalling pathways, such as G2M_CHECKPOINT and EMT (EPITHELIAL)_MESENCHYMAL_TRANSITION (Fig. 6B).

3.3.2. *Analysis of differentially expressed genes associated with the cell cycle pathway*

We discovered that the PS high-expression group had significantly higher levels of CDC27 and DBF4 expression ($P < 0.001$) (Fig. 6C).

3.3.3. *Analysis of differential immune cell abundance*

We analyzed immune cell infiltration in LUAD and observed that $CD8^+$ and activated memory $CD4^+$ T cell infiltration was significantly higher in the PS high-expression group ($P < 0.001$), whereas eosinophil cell infiltration did not show a statistically significant difference between the two groups ($P > 0.05$) (Fig. 6D).

3.3.4. *Analysis of differential drug sensitivity*

We observed that patients were more sensitive to sorafenib and Wee1 inhibitors in the PS high-expression group ($P < 0.001$) (Fig. 6E and F). This suggests a positive correlation between PS expression and the sensitivity to sorafenib and Wee1 inhibitors. The higher the PS expression, the stronger the sensitivity to sorafenib and Wee1 inhibitor. This made it possible to predict the sensitivity to Wee1 inhibitors and sorafenib based on the PS generated by the pathomics model.

3.3.5. *Mutation analysis and TMB analysis of LUAD*

As shown in Fig. 7A, B, missense_ mutations are the most frequent in both the high- and low-PS groups. Both TP53 and TTN showed higher mutation rates in the high PS group than in the low PS group. No significant differences were observed in the mutation rate of ARID1A gene(5 %) between the two groups. We speculated that AURKA expression was associated with a poor prognosis in LUAD, possibly due to common mutations in genes such as TP53 and TTN.

4. Discussion

In this study, we first investigated the prognostic value of AURKA expression in LUAD and observed that increased expression of AURKA is an independent risk factor for OS in patients with LUAD. We retrieved histopathological image features and built a

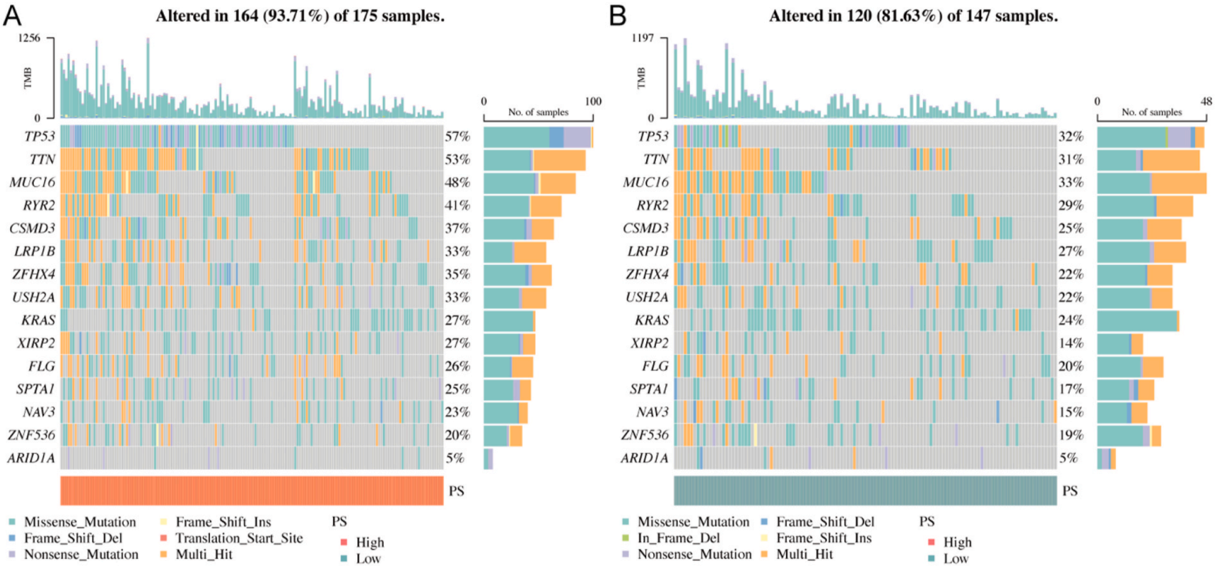


Fig. 7. Mutation analysis and TMB analysis for LUAD. (A). PS high-expression group (B) PS low-expression group; TMB: Tumor Mutation Burden; Multi_ Hit: Genes mutated multiple times in the same sample.

pathohistological prediction model for AURKA expression using a machine learning technique. The results showed that the model achieved outstanding performance in predicting AURKA expression. PS generated by this model was an independent prognostic factor for OS in patients with LUAD. Finally, KEGG and GO enrichment analyses were carried out for functional annotation to identify the underlying biological processes. We discovered that they were mostly concentrated in these key pathways, including the CELL_CYCLE, P53, and G2M_CHECKPOINT signaling pathways. In the mutation and TMB analyses, we observed that TP53 and TTN had significantly higher mutation rates in the high PS group than in the low PS group.

AURKA is a potential low-penetrance gene associated with tumor susceptibility [5,37]. AURKA is overexpressed and associated with a poor prognosis in various malignant tumors [38,39]. NING ZHONG et al. [2] investigated the connection between AURKA expression and lung cancer for the first time and discovered that both human LUAD cells and tissues had significantly high AURKA expression. However, their results were based on only 101 clinical cases. Mengyu Zhang et al. [40] investigated the relationship between AURKA expression status and LUAD in smokers and discovered that such patients had high AURKA expression. Moreover, these patients had a shorter median OS than those with lower AURKA expression. These results were consistent with our findings. The KM curve in our study suggested that patients with high AURKA expression had a worse prognosis than those with low AURKA expression, and multifactorial regression analysis suggested that AURKA overexpression was an independent risk factor for OS in patients with LUAD. However, our study population was not restricted to patients with LUAD who smoked; instead, we included a broader study population.

The quantitative morphological properties of H&E-stained pathology sections provide objective, quantitative measurement of the morphology and texture of the nuclei and cytoplasm. For instance, the Zernike form features label the nucleic pixels as 1 and the cytoplasmic region as 0, and then create Zernike polynomials from binary images [41]. Using expanded structural components to match the texture, granularity calculates the size of an image's texture [42]. Manual inspection typically has difficulty identifying these quantitative picture elements; however, computational approaches provide quick and accurate results. Recently, artificial intelligence-based computer techniques have been developed to convert pathological images into highly accurate, high-volume data containing quantitative features such as textural, morphological, edge-gradient, and biological features, which can be used to quantify pathology, molecular expression, and disease prognosis [14,43,44]. Machine learning from histopathological images could predict the STK11, EGFR, FAT1, SETBP1, KRAS, and TP53 mutations of LUAD (AUCs varied from 0.733 to 0.856) [32]. Yu et al. [15] recognized the predictive significance of the Zernike characteristics in lung cancer. Meanwhile, several studies revealed that histopathological image analysis could identify the presence of gene mutations in ovarian, colorectal, and liver malignancies [24,45,46]. This significantly improves the efficiency of molecular testing and reduces the cost of human resources.

To the best of our knowledge, this is the first study to predict AURKA expression in tissue sections using a machine learning-based pathomics signature. In this study, we created and tested machine learning classifiers to predict AURKA expression in LUAD using an automated procedure that extracts objective elements from histopathological images. As HE-stained images are frequently created in clinical practice, our classifier can be successfully used in everyday practice. Moreover, by upsampling, the samples for model training contained samples at 20 × and 40 × magnifications, which enabled the model to accurately predict samples at different magnifications, thereby improving the generalizability of the model. This may be a practical and economical method for predicting AURKA expression in LUAD. Based on this pathophysiological model, we could predict the prognosis of LUAD. This suggests that our pathomics model has superior accuracy, sensitivity, and specificity. We observed that the higher the PS, the higher the expression of AURKA in the intergroup variability analysis. Using univariate and multivariate Cox regression analyses, we observed that the higher PS, the worse the prognosis. These results demonstrate that the pathomics model created using machine learning is capable of accurately predicting the expression of AURKA and LUAD prognosis. Finally, AURKA is an emerging therapeutic target for cancer. Many preclinical and clinical studies have evaluated AURKA inhibitors (NCT01380756, NCT01431664, NCT01719744, and NCT00766324).The model constructed in this study can help us make objective, quick, batch, and accurate predictions of AURKA expression. Therefore, this study is anticipated to serve as a foundation for directing the use of AURKA inhibitors in clinical settings.

In the analysis of the pathohistologic mechanism, we observed that $CD8^+$ and activated memory $CD4^+$ T cell infiltration was significantly higher in the PS high group ($P < 0.001$). Our model suggested that a high PS was significantly associated with a poor prognosis in patients with LUAD. Immune cells have complex regulatory mechanisms for tumor progression. $CD8^+$ T cells, as the main effector immune cells, play an important role in anti-tumor effects [47]. Infiltrating $CD8^+$ T cells are associated with longer OS in patients with various malignancies, including melanoma, squamous cell carcinoma (SCCs), large-cell lung cancer, and several types of adenocarcinoma [47–49]. However, notable exceptions have been observed for clear cell RCC (ccRCC) [50,51] and prostate cancer [52], where a high density of $CD8^+$ T cell was correlated with a shorter OS. Firstly, some studies have shown that various T cell subtypes may be associated with a poor prognosis in cancer, possibly because of the heterogeneity of T cells. TIGIT + CD8+T, CXCL13+CD8+T, $CD39^+$CD8+T cell infiltration is correlated with poor prognosis in a variety of tumors [53–55]. T helper 1 (TH1) cells are a subtype of $CD4^+$ T cells, and recent studies have observed that a high frequency of TH1 cells are associated with reduced 2-year survival after surgery for non-small cell lung cancer. Second, AURKA inhibitor therapy exerts antitumor effects while inducing PD-L1 upregulation, resulting in a decrease in $CD8^+$ T cells. Moreover, a combination of AURKA inhibitor and PD-L1 antibody therapy can restore $CD8^+$ T cell infiltration, suggesting that the increase in $CD8^+$ T cells may only be a concomitant phenomenon of high AURKA expression [56]. Tumor immunotherapy is the most promising therapeutic modality discovered in recent years, and T cell-rich infiltrated tumors are believed to respond better to immunotherapy with immune checkpoint inhibitors in hot tumors [57]. This implies that the AURKA-based prediction model has the potential to predict responses to immunotherapy. In this study, we focused on analyzing T-cell infiltration in patients in the poor-prognosis group. We plan to further explore this phenomenon in the future. We performed mutation and TMB analyses and observed that TP53 and TTN had higher mutation rates in the high-PS group than in the low-PS group. No significant differences were observed in the mutation rates of ARID1A between the two groups. Previous studies have observed that targeted inhibition of AURKA may be a new therapeutic strategy for ARID1A mutant colorectal cancer [9,10]. However, we observed a low rate of ARID1A mutations (only 5 %) in patients with LUAD. We speculate that AURKA may have different mechanisms in LUAD and colon cancer. AURKA is associated with a poor prognosis in LUAD, possibly due to common mutations in genes such as TP53 and TTN.

This study has several limitations. First, we obtained information from TCGA database and constructed and verified a model for the training and validation groups. As it was challenging to discover other datasets with comprehensive information on histology and omics, this study had a small sample size, was limited to one cohort, and lacked external validation. Therefore, the generalizability of our findings should be considered within these constraints. Second, because this study was retrospective, confounding variables inevitably had an impact. In addition, representative tumor slices were more likely to be uploaded, which may have resulted in a selection bias in TCGA dataset. The usual histological patterns of these tumor slices may aid in their classification [15]. However, as clinicians often use depth data from several slides and microscopic images, further research is required to determine the performance of prediction models in clinical settings.

5. Conclusion

We discovered that pathohistological image features can predict AURKA expression in patients with LUAD through machine learning and that PS based on this model could predict the prognosis of patients with LUAD. This can help in risk stratification and individualized care for these patients in the future.

Ethical approval statement

Approval by an ethics committee was not needed for this study because TCGA databases were publicly available for research. The TCGA database has obtained the informed consent of the subjects.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: the TCGA repository (https://portal.gdc.cancer.gov). Accession number: TCGA-LUAD.

CRediT authorship contribution statement

Cuiqing Bai: Writing – original draft. **Yan Sun:** Conceptualization. **Xiuqin Zhang:** Formal analysis. **Zhitong Zuo:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The authors would like to thank TCGA working group for offering the slide images and the corresponding cancer information.

References

[1] R.L. Siegel, K.D. Miller, N.S. Wagle, A. Jemal, Cancer statistics, CA A Cancer J. Clin. 73 (1) (2023) 17–48, 2023.
[2] N. Zhong, S. Shi, H. Wang, G. Wu, Y. Wang, Q. Ma, et al., Silencing Aurora-A with siRNA inhibits cell proliferation in human lung adenocarcinoma cells, Int. J. Oncol. 49 (3) (2016) 1028–1038.
[3] S. Kanamori, I. Kajihara, S. Kanazawa-Yamada, S. Otsuka-Maeda, H. Ihn, Expression of aurora kinase A expression in dermatofibrosarcoma protuberans, J. Dermatol. 45 (4) (2018) 507–508.
[4] A.E. Lykkesfeldt, B.R. Iversen, M.B. Jensen, B. Ejlertsen, A. Giobbie-Hurder, B.E. Reiter, et al., Aurora kinase A as a possible marker for endocrine resistance in early estrogen receptor positive breast cancer, Acta Oncol. 57 (1) (2018) 67–73.
[5] B. Goldenson, J.D. Crispino, The aurora kinases in cell cycle and leukemia, Oncogene 34 (5) (2015) 537–545.
[6] M. Yan, C. Wang, B. He, M. Yang, M. Tong, Z. Long, et al., Aurora-A kinase: a potent oncogene and target for cancer therapy, Med. Res. Rev. 36 (6) (2016) 1036–1079.
[7] A. Hoque, J. Carter, W. Xia, M.C. Hung, A.A. Sahin, S. Sen, et al., Loss of aurora A/STK15/BTAK overexpression correlates with transition of in situ to invasive ductal carcinoma of the breast, Cancer Epidemiol. Biomarkers Prev. 12 (12) (2003) 1518–1522.
[8] H. Zhou, J. Kuang, L. Zhong, W.L. Kuo, J.W. Gray, A. Sahin, et al., Tumour amplified kinase STK15/BTAK induces centrosome amplification, aneuploidy and transformation, Nat. Genet. 20 (2) (1998) 189–193.
[9] C. Sun, Z. Qu, W. Liu, Z. Qiu, Y. Lü, Z. Sun, The synergistic anti-colon cancer effect of aurora A inhibitors and AKT inhibitors through PI3K/AKT pathway, Anti Cancer Agents Med. Chem. 23 (1) (2023) 87–93.
[10] C. Wu, J. Lyu, E.J. Yang, Y. Liu, B. Zhang, J.S. Shim, Targeting AURKA-CDC25C axis to induce synthetic lethality in ARID1A-deficient colorectal cancer cells, Nat. Commun. 9 (1) (2018) 3212.

[11] J. Hipp, T. Flotte, J. Monaco, J. Cheng, A. Madabhushi, Y. Yagi, et al., Computer aided diagnostic tools aim to empower rather than replace pathologists: lessons learned from computational chess, J. Pathol. Inf. 2 (2011) 25.
[12] X. Zhang, F. Xing, H. Su, L. Yang, S. Zhang, High-throughput histopathological image analysis via robust cell segmentation and hashing, Med. Image Anal. 26 (1) (2015) 306–315.
[13] K. Soliman, CellProfiler: novel automated image segmentation procedure for super-resolution microscopy, Biol. Proced. Online 17 (2015) 11.
[14] A.H. Beck, A.R. Sangoi, S. Leung, R.J. Marinelli, T.O. Nielsen, M.J. van de Vijver, et al., Systematic analysis of breast cancer morphology uncovers stromal features associated with survival, Sci. Transl. Med. 3 (108) (2011) 108ra113.
[15] K.H. Yu, C. Zhang, G.J. Berry, R.B. Altman, C. Re, D.L. Rubin, et al., Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features, Nat. Commun. 7 (2016) 12474.
[16] X. Luo, X. Zang, L. Yang, J. Huang, F. Liang, J. Rodriguez-Canales, et al., Comprehensive computational pathological image analysis predicts lung cancer prognosis, J. Thorac. Oncol. 12 (3) (2017) 501–509.
[17] A. Madabhushi, G. Lee, Image analysis and machine learning in digital pathology: challenges and opportunities, Med. Image Anal. 33 (2016) 170–175.
[18] J. Kong, L.A. Cooper, F. Wang, J. Gao, G. Teodoro, L. Scarpace, et al., Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates, PLoS One 8 (11) (2013) e81049.
[19] T. Zhong, M. Wu, S. Ma, Examination of independent prognostic power of gene expressions and histopathological imaging features in cancer, Cancers 11 (3) (2019).
[20] Y. Yu, Y. Tan, C. Xie, Q. Hu, J. Ouyang, Y. Chen, et al., Development and validation of a preoperative magnetic resonance imaging radiomics-based signature to predict axillary lymph node metastasis and disease-free survival in patients with early-stage breast cancer, JAMA Netw. Open 3 (12) (2020) e2028086.
[21] Q. Fang, H. Chen, The significance of m6A RNA methylation regulators in predicting the prognosis and clinical course of HBV-related hepatocellular carcinoma, Mol. Med. 26 (1) (2020) 60.
[22] J. Vivian, A.A. Rao, F.A. Nothaft, C. Ketchum, J. Armstrong, A. Novak, et al., Toil enables reproducible, open source, big biomedical data analyses, Nat. Biotechnol. 35 (4) (2017) 314–316.
[23] L. Chen, H. Zeng, M. Zhang, Y. Luo, X. Ma, Histopathological image and gene expression pattern analysis for predicting molecular features and prognosis of head and neck squamous cell carcinoma, Cancer Med. 10 (13) (2021) 4615–4628.
[24] H. Zeng, L. Chen, M. Zhang, Y. Luo, X. Ma, Integration of histopathological images and multi-dimensional omics analyses predicts molecular features and prognosis in high-grade serous ovarian cancer, Gynecol. Oncol. 163 (1) (2021) 171–180.
[25] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, et al., Weakly supervised deep learning for whole slide lung cancer image analysis, IEEE Trans. Cybern. 50 (9) (2020) 3950–3962.
[26] N. Otsu, A tlreshold selection method from gray-level histograms, 2EEE TRANSACTIONS ON SYSTREMS, MAN, AND CYBERNETICS SMC-9 (1, JANUARY) (1979).
[27] D. Chen, M. Fu, L. Chi, L. Lin, J. Cheng, W. Xue, et al., Prognostic and predictive value of a pathomics signature in gastric cancer, Nat. Commun. 13 (1) (2022) 6903.
[28] W.F. Qu, M.X. Tian, H.W. Lu, Y.F. Zhou, W.R. Liu, Z. Tang, et al., Development of a deep pathomics score for predicting hepatocellular carcinoma recurrence after liver transplantation, Hepatol Int 17 (4) (2023) 927–941.
[29] S. Kobayashi, J.H. Saltz, V.W. Yang, State of machine and deep learning in histopathological applications in digestive diseases, World J. Gastroenterol. 27 (20) (2021) 2545–2575.
[30] D. Rajput, W.J. Wang, C.C. Chen, Evaluation of a decided sample size in machine learning applications, BMC Bioinf. 24 (1) (2023) 48.
[31] N. Kumar, R. Gupta, S. Gupta, Whole slide imaging (WSI) in pathology: current perspectives and future directions, J. Digit. Imag. 33 (4) (2020) 1034–1040.
[32] N. Coudray, P.S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyo, et al., Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning, Nat. Med. 24 (10) (2018) 1559–1567.
[33] K. Saednia, A. Lagree, M.A. Alera, L. Fleshner, A. Shiner, E. Law, et al., Quantitative digital histopathology and machine learning to predict pathological complete response to chemotherapy in breast cancer patients using pre-treatment tumor biopsies, Sci. Rep. 12 (1) (2022) 9690.
[34] H. Li, L. Chen, H. Zeng, Q. Liao, J. Ji, X. Ma, Integrative analysis of histopathological images and genomic data in colon adenocarcinoma, Front. Oncol. 11 (2021) 636451.
[35] M. Nishio, M. Nishio, N. Jimbo, K. Nakane, Homology-based image processing for automatic classification of histopathological images of lung tissue, Cancers 13 (6) (2021).
[36] J. Cheng, J. Zhang, Y. Han, X. Wang, X. Ye, Y. Meng, et al., Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis, Cancer Res. 77 (21) (2017) e91–e100.
[37] N.J. Taylor, J.T. Bensen, C. Poole, M.A. Troester, M.D. Gammon, J. Luo, et al., Genetic variation in cell cycle regulatory gene AURKA and association with intrinsic breast cancer subtype, Mol. Carcinog. 54 (12) (2015) 1668–1677.
[38] S. Staff, J. Isola, M. Jumppanen, M. Tanner, Aurora-A gene is frequently amplified in basal-like breast cancer, Oncol. Rep. 23 (2) (2010) 307–312.
[39] K.B. Lukasiewicz, W.L. Lingle, A. Aurora, Centrosome structure, and the centrosome cycle, Environ. Mol. Mutagen. 50 (8) (2009) 602–619.
[40] M.Y. Zhang, X.X. Liu, H. Li, R. Li, X. Liu, Y.Q. Qu, Elevated mRNA Levels of AURKA, CDC20 and TPX2 are associated with poor prognosis of smoking related lung adenocarcinoma using bioinformatics analysis, Int. J. Med. Sci. 15 (14) (2018) 1676–1685.
[41] S. Li, M.C. Lee, C.M. Pun, Complex Zernike moments features for shape-based image retrieval, IEEE Trans Syst Man Cyber 39 (2009) 227–237.
[42] L. Vincent, Granulometries and opening trees, Fundam. Inf. 41 (2000) 57–90.
[43] D. Romo-Bucheli, A. Janowczyk, H. Gilmore, E. Romero, A. Madabhushi, Automated tubule nuclei quantification and correlation with oncotype DX risk categories in ER+ breast cancer whole slide images, Sci. Rep. 6 (2016) 32706.
[44] W.K. Moon, Y.W. Lee, Y.S. Huang, S.H. Lee, M.S. Bae, A. Yi, et al., Computer-aided prediction of axillary lymph node status in breast cancer using tumor surrounding tissue features in ultrasound images, Comput. Methods Progr. Biomed. 146 (2017) 143–150.
[45] M. Chen, B. Zhang, W. Topatana, J. Cao, H. Zhu, S. Juengpanich, et al., Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning, npj Precis. Oncol. 4 (2020) 14.
[46] H.J. Jang, A. Lee, J. Kang, I.H. Song, S.H. Lee, Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning, World J. Gastroenterol. 26 (40) (2020) 6207–6223.
[47] D. Bruni, H.K. Angell, J. Galon, The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy, Nat. Rev. Cancer 20 (11) (2020) 662–680.
[48] W.H. Fridman, F. Pages, C. Sautes-Fridman, J. Galon, The immune contexture in human tumours: impact on clinical outcome, Nat. Rev. Cancer 12 (4) (2012) 298–306.
[49] W.H. Fridman, L. Zitvogel, C. Sautes-Fridman, G. Kroemer, The immune contexture in cancer prognosis and treatment, Nat. Rev. Clin. Oncol. 14 (12) (2017) 717–734.
[50] N.A. Giraldo, E. Becht, F. Pages, G. Skliris, V. Verkarre, Y. Vano, et al., Orchestration and prognostic significance of immune checkpoints in the microenvironment of primary and metastatic renal cell cancer, Clin. Cancer Res. 21 (13) (2015) 3031–3040.
[51] R. Remark, M. Alifano, I. Cremer, A. Lupo, M.C. Dieu-Nosjean, M. Riquet, et al., Characteristics and clinical impacts of the immune environments in colorectal and renal cell carcinoma lung metastases: influence of tumor origin, Clin. Cancer Res. 19 (15) (2013) 4079–4091.
[52] F. Petitprez, N. Fossati, Y. Vano, M. Freschi, E. Becht, R. Luciano, et al., PD-L1 expression and CD8(+) T-cell infiltrate are associated with clinical progression in patients with node-positive prostate cancer, Eur Urol Focus 5 (2) (2019) 192–196.
[53] Z. Liu, Q. Zhou, Z. Wang, H. Zhang, H. Zeng, Q. Huang, et al., Intratumoral TIGIT(+) CD8(+) T-cell infiltration determines poor prognosis and immune evasion in patients with muscle-invasive bladder cancer, J Immunother Cancer 8 (2) (2020).

[54] S. Dai, H. Zeng, Z. Liu, K. Jin, W. Jiang, Z. Wang, et al., Intratumoral CXCL13(+)CD8(+)T cell infiltration determines poor clinical outcomes and immunoevasive contexture in patients with clear cell renal cell carcinoma, J Immunother Cancer 9 (2) (2021).

[55] Y. Qi, Y. Xia, Z. Lin, Y. Qu, Y. Qi, Y. Chen, et al., Tumor-infiltrating CD39(+)CD8(+) T cells determine poor prognosis and immune evasion in clear cell renal cell carcinoma patients, Cancer Immunol. Immunother. 69 (8) (2020) 1565–1576.

[56] B. Meng, X. Zhao, S. Jiang, Z. Xu, S. Li, X. Wang, et al., AURKA inhibitor-induced PD-L1 upregulation impairs antitumor immune responses, Front. Immunol. 14 (2023) 1182601.

[57] Y.T. Liu, Z.J. Sun, Turning cold tumors into hot tumors by improving T-cell infiltration, Theranostics 11 (11) (2021) 5365–5386.