Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Characterization of spatial distribution of immune cells in the microenvironment of lung cancer and its computational modeling study

Jianming Peng[1], Youai Dai[2], Huishen Yan[1], Rui Liang[3,]✉

[1] *Shool of Medicine, Yangzhou Polytechnic College, Yangzhou, Jiangsu, 225009, China*
[2] *Laboratory of Organ Transplantation Research Institute, Wuxi People's Hospital Affiliated to Nanjing Medical University, Wuxi, Jiangsu, 214023, China*
[3] *Shool of Basic Medical Science, Suzhou Vocational Health College, Suzhou, Jiangsu, 215009, China*

ABSTRACT

Lung cancer is the most common malignant tumor in humans and the leading cause of cancer-related deaths worldwide. In this study, we focused on the immune cells in the microenvironment of lung cancer at the protein expression level by IHC as well as mIHC techniques to explore the spatial distribution characteristics of immune cells within the tumor. To predict the prognosis of NSCLC patients and their potential response to immunotherapy, a machine learning-based immune-related prognostic model for lung cancer was constructed by combining Cox regression analysis, random survival forest and XGBoost algorithm, and the effect of the prognostic model was verified on the relevant dataset. The results showed that there were some differences in the immune cells between lung adenocarcinoma and lung squamous carcinoma in the lung cancer microenvironment, and the spatial distribution heterogeneity of CD3+ T cells and MHC class II antigen-presenting cells was higher in lung adenocarcinoma ($P<0.05$).The overall survival of high-risk patients was lower than that of the low-risk group in both LUAD and LUSC ($P<0.01$), and the immuno-associated prognostic model of lung cancer had a stable performance in the AUC value in multiple independent cohorts with stable performance, and the IRS model maintained high accuracy and stable performance in the training set and test set, which indicates that IRS has great potential for clinical application.

Keywords: lung cancer, immune cells, spatial distribution, machine learning, prognostic modeling, randomized survival forests

✉ Corresponding author.
E-mail address: pengjianming2025@163.com (R. Liang).
Received 15 September 2024; Revised 05 November 2024; Accepted 20 February 2025; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-211

1. Introduction

With the wide application of computed tomography examination and the enhancement of health awareness, the detection rate of early lung cancer patients has increased significantly, and comprehensive treatment can prolong the survival of patients, but some patients may have drug resistance due to long-term radiotherapy, which ultimately leads to treatment failure due to recurrence and metastasis [1-2]. Therefore, more effective treatment strategies are still urgently needed. Immunotherapy is profoundly changing the treatment of lung cancer because of its high safety, the ability of immune memory to bring sustained immune response to effectively control lung cancer, and its effectiveness in treating patients with advanced lung cancer [3-4]. Immune checkpoint inhibitors are now widely used in the treatment of lung cancer [5]. However, the therapeutic effect of such monoclonal antibodies against specific antigens often depends on the mutational load of the tumor, and a large proportion of patients do not benefit from the treatment [6]. Immune cells are a very important component of the immune system, and recent studies have refined our understanding of the impact of overt immune cell therapy on lung cancer and how it can be enhanced, significantly broadening the possibilities of immunotherapy for lung cancer [7-8].

CAR-T therapy, i.e. chimeric antigen receptor-T cell immunotherapy, is a classic component of over-the-counter immunocytotherapy, in which medical technicians activate T cells through genetic engineering technology and load them with “localization navigation device”-tumor chimeric antigen receptor to form CAR-T cells [9-10]. Specifically recognize tumor cells in the body, and release a large number of multiple effectors through the immune effect, efficiently kill tumor cells, so as to achieve the purpose of treating malignant tumors [11]. In this process, genetic engineering technology plays a very important role, which is a complex technology to manipulate genes at the molecular level and is an important branch of synthetic biology [12-13]. Synthetic biology is an emerging, the use of molecular biology techniques and tools for the design and manipulation of the behavior of cells in the field of research, through synthetic biology methods can be designed and generated with specific functions of living cells, has been applied to vaccine development, regenerative medicine and cell therapy, etc., the cell therapy of cancer is an important part of its cellular therapy, based on the synthetic biology technology of the immune cell modification, is expected to create smarter immune cells to be applied in the treatment of lung cancer [14-16].

The study introduces the immunotherapy of lung cancer, analyzes the immune cell infiltration of the immune microenvironment in NSCLC patients by techniques such as IHC as well as mIHC staining, calculates the Shannon Entropy of the immune cells in different regions, and evaluates the spatial heterogeneity of the distribution of the immune cells between different regions. On this basis, immune-related prognostic genes were screened using one-way Cox regression analysis, and immune-related prognostic models of lung cancer were constructed for LUAD and LUSC based on Random Survival Forest and XGBoost algorithms, respectively. Using the R package TCGAbiolinks download data as the training set and the GEO dataset as the validation set, Kaplan-Meier survival analysis was performed by the survival and survminer packages to evaluate the prognostic value of the IRS model, and the ROC curves were plotted using the timeROC package to evaluate the effectiveness and accuracy of the IRS model.

2. Immunotherapy in lung cancer

2.1. Lung Cancer and Immunotherapy

Globally, the incidence and death rates of lung cancer have been high, of which non-small cell lung cancer (NSCLC) is the main pathological type, accounting for about 85% of the total number of lung cancers, and lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) are the two most common pathological types of NSCLC. With the improvement of health awareness of the population

and the popularization of screening methods, the number of patients with early diagnosis and interventional treatment of lung cancer has gradually risen, but it is still the main cause of tumor-related deaths. Currently, early and intermediate stage NSCLC patients are mostly treated by surgery to remove the lesions to achieve radical treatment, which in turn improves the prognosis. For advanced patients who have lost the chance of surgery, the 5-year survival rate is not satisfactory. Traditional tumor therapies, such as surgery, radiotherapy, chemotherapy and targeted therapies, although able to alleviate the disease to a certain extent, are often unable to meet the growing clinical demand due to their limitations. The emerging immunotherapy technology, due to its unique efficacy and potential, is becoming a hot research focus in the field of tumor treatment. However, the response rate of immune checkpoint inhibitors (ICIs) is only 14-20% in unscreened patients, so it is crucial to explore biomarkers with predictive value for the efficacy and prognosis of NSCLC immunotherapy.

2.2. Basic mechanisms of immunotherapy

Immunotherapy is a therapeutic approach targeting the body's immune system, aiming to activate or enhance the function of the immune system to fight against diseases. Currently, ICIs are widely used in the treatment of a variety of tumors, and can enhance T cell activity and proliferation capacity by blocking interactions between relevant molecules, promoting anti-tumor immune responses, and reconnecting the interactive network between CD8+ T cells, antigen-presenting cells, and tumor cells. Tumor mutational load (TMB) is a measure of the frequency of genetic mutations in tumor cells. TMB is assessed based on the analysis of DNA sequences of tumor tissues to quantify the number of mutated cells therein, and patients with a high TMB usually have a better response to immunotherapy and a relatively better prognosis. However, an increase in the number of non-synonymous mutations may contribute to the formation of new tumor-specific antigens that can be accurately recognized by the immune system, increase the sensitivity of tumor recognition, and be cleared by adaptive immune cells in the immune response triggered by treatment with ICIs.

3. Characterization of the spatial distribution of immune cells in the microenvironment of lung cancer

The tumor immune microenvironment (TIME) includes tumor cells, a variety of T cells, B cells, neutrophils, macrophages, dendritic cells, and natural killer cells (NKs). The interactions between these components together determine the ultimate trend of the body's anti-tumor immunity. The distribution characteristics of various types of immune cells as well as infiltration in TIME were analyzed using hematoxylin-eosin staining (HE staining), immunohistochemistry (IHC), and multiple immunohistochemistry (mIHC).

3.1. Materials and Methods

3.1.1. Sample Selection. One hundred patients diagnosed with NSCLC were included in this study, and a few samples with low tumor content, contamination or damage were excluded after strict quality control. After sequencing analysis using the GeneseeqPrimeTM 425-gene panel protocol, 50 samples were excluded after quality control for reasons such as low tumor purity or DNA contamination, resulting in 250 tissue samples available for analysis.

3.1.2. Experimental methods:

1) Preparation of tissue microarrays. From each tumor tissue, different regions of tissue samples were obtained by the 4-quadrant segmentation method, and paraffin-embedded specimens were fabricated into tissue microarrays for further studies such as subsequent HE staining, IHC, and mIHC staining.

2) Immunohistochemical staining. Immunohistochemical CD3 and CD20 staining was used to label T cells and B cells, respectively. The antibody information was as follows: CD3 (Dako A045229), CD20 (Dako IS60430).

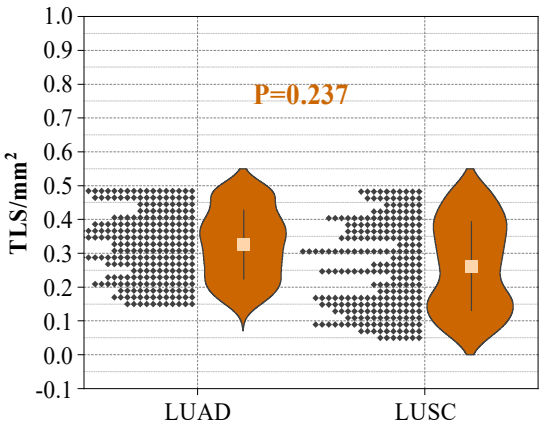
3) Multiple immunohistochemical staining. Paraffin tissues were subjected to 4 μm sections. Multiple rounds of immunohistochemical staining were performed using different antibodies and fluorescent dyes. The antibody information was as follows: CD56: (Dako CD56 Clone 123C3, IS62830), CD68: (Dako CD68, Clone PG-M1, IS61330), CD8: (Dako CD8 clone C8/144B, IS62330), HLADR: (Dako HLADR: Clone TAL.1B5, M074601), PANCK: (Dako CK Clone 34betaE12, M063001).

4) Immunohistochemical score. The immunohistochemical score (H-score) of each tumor tissue sample was calculated by Halo software on a scale of 0-3 from weak to strong (0 = unstained, 1 = weakly stained, 2 = moderately stained, 3 = strongly stained).

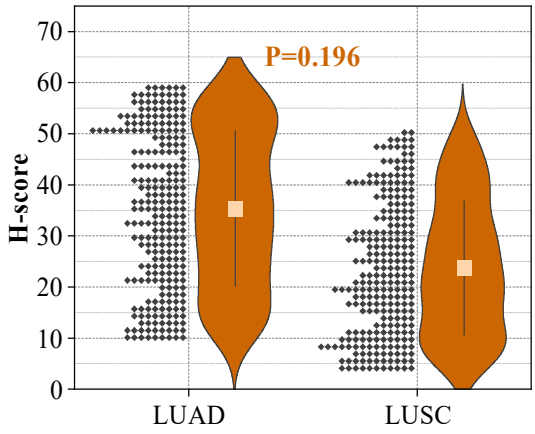
5) Statistical analysis. Statistical analysis in this study was done using R (Version 4.1.3) and GraphPad Prism (Version 8.0.1). For comparisons of continuous variables between two groups and between multiple groups, the Mann-Whitney U test was used. For the evaluation of spatial heterogeneity, Shannon Entropy was calculated using the "Entropy" program package in R. To compare the prognostic differences of the cases, Log-rank test was used. All p-values were two-sided, and $p < 0.05$ was considered statistically significant.

3.2. Analysis of results

3.2.1. Analysis of variances. Various types of immune cell infiltration were obtained by HE staining, IHC staining for CD3, CD20 and mIHC staining for CD56, CD68, CD8, HLADR and PANCK, and immune cell infiltration as well as the distribution of TLSs were quantitatively analyzed after calculating immunohistochemical scores. Differential analysis of immune infiltration between lung adenocarcinoma (LUAD) and lung squamous carcinoma (LUSC) is shown in Figure 1, with (a) to (h) showing the differences in the distribution of various types of immune cells. Compared with lung squamous carcinoma, tumor tissues of lung adenocarcinoma patients were infiltrated with more CD56+ NK cells (P = 0.018), CD68+ macrophages (P < 0.001), and MHC class II antigen-presenting cells (P < 0.001), and additionally tumor tissues of lung adenocarcinoma patients had fewer PANCK-labeled tumor cells (P < 0.001).



(a)TLS



(b)CD3Tcell

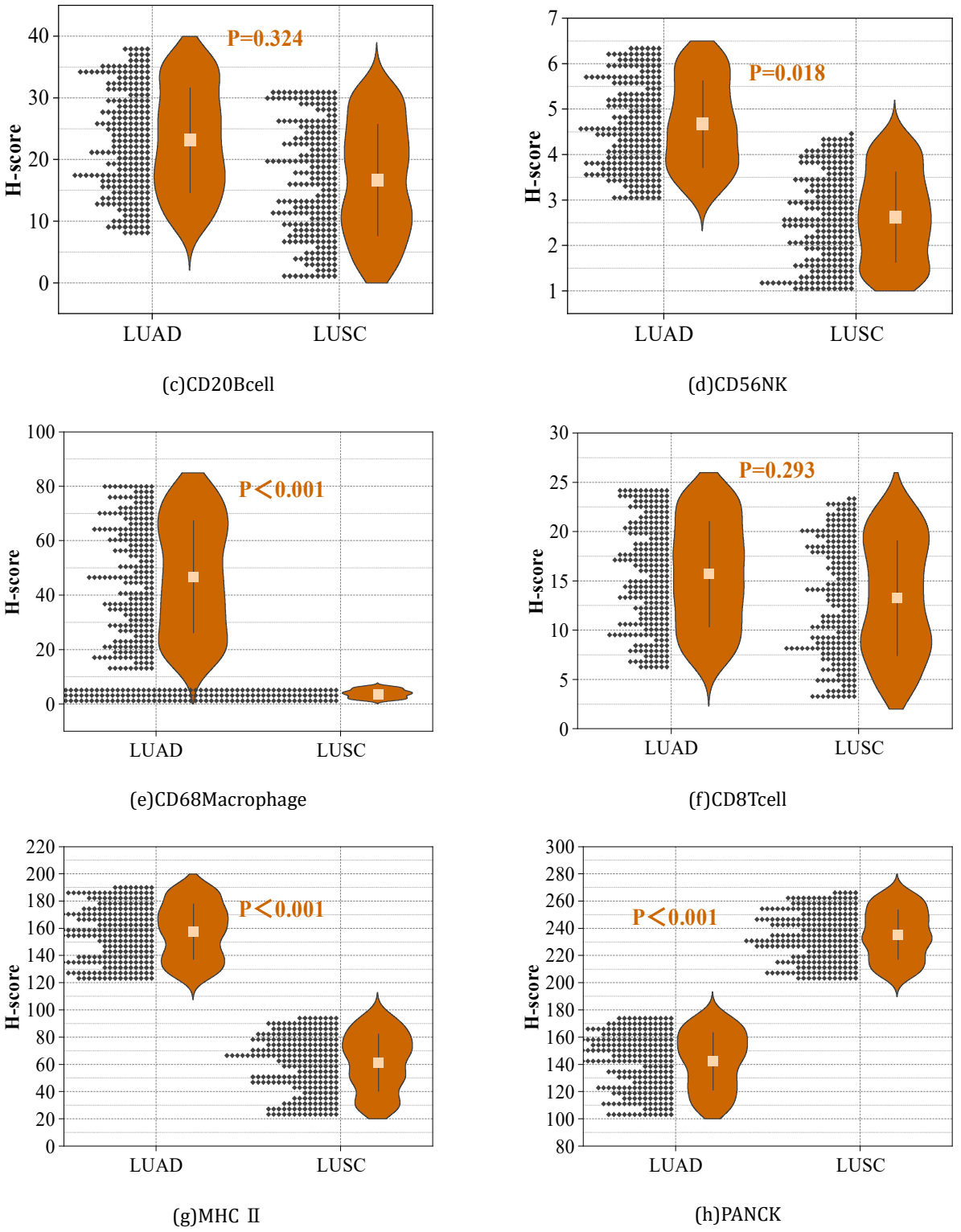
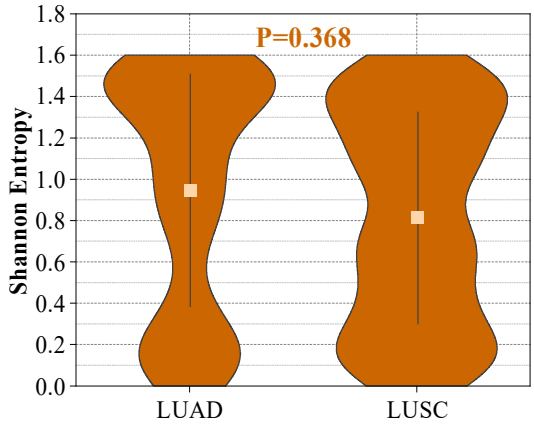


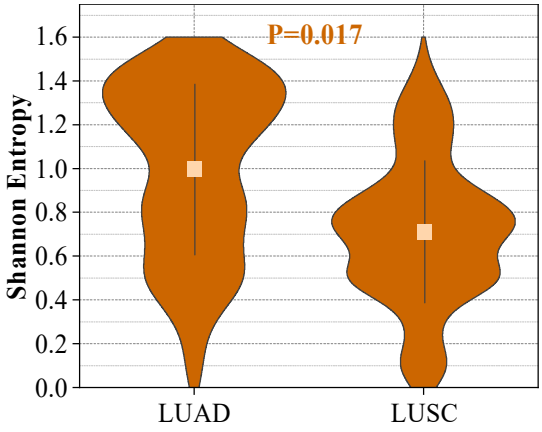
Fig. 1. Differences analysis of the immune infiltration between LUAD and LUSC

3.2.2. Analysis of spatial heterogeneity. In this study, tissue samples from different regions within the tumor tissues of the same patient were obtained by multipoint sampling, and the immune cell infiltration in different regions was analyzed, and the spatial heterogeneity of immune cell distribution between these different regions was evaluated by calculating the Shannon Entropy, with high Shannon Entropy to indicate high heterogeneity. The analysis of spatial heterogeneity of immune cell distribution is shown in Figure 2, and (a) to (h) show the results of spatial heterogeneity of various

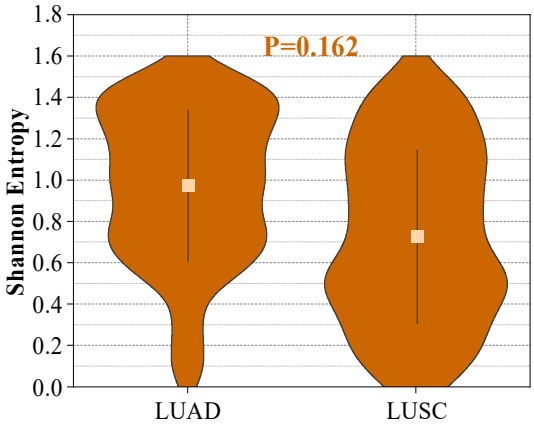
types of immune cells. Compared with squamous lung cancer, the spatial distribution heterogeneity of CD3+ T cells (P=0.017) was higher in the tumor tissues of patients with lung adenocarcinoma, and in addition, there was a trend of increased spatial distribution heterogeneity of MHC class II antigen-presenting cells (P=0.037). In contrast, there was no statistical difference in the spatial distribution of TLSs, CD20+ B cells, CD56+ NK cells, CD68+ macrophages, CD8+ T cells, and PANCK-labeled tumor cells.



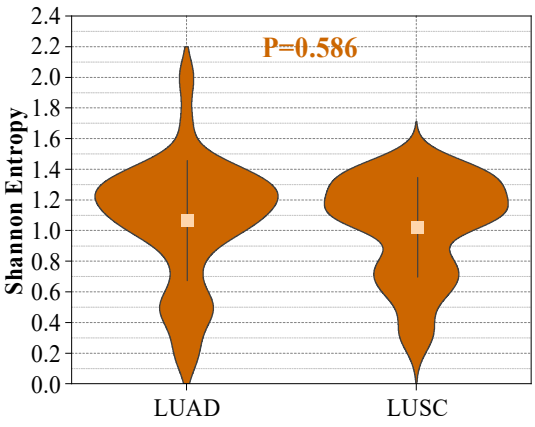
(a)TLS



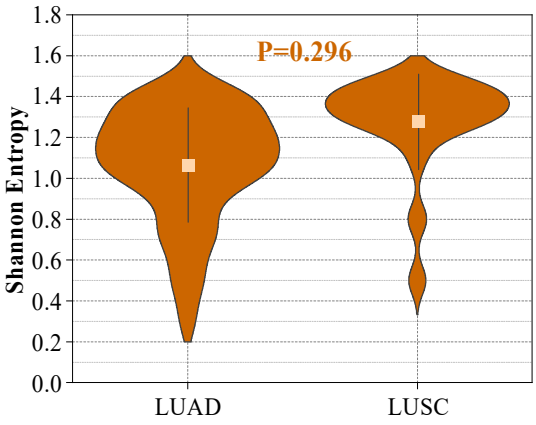
(b)CD3Tcell



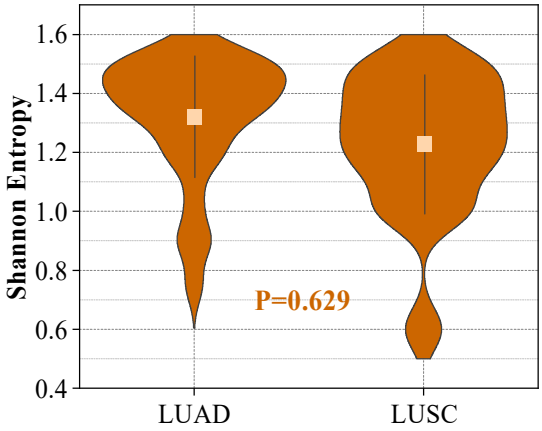
(c)CD20Bcell



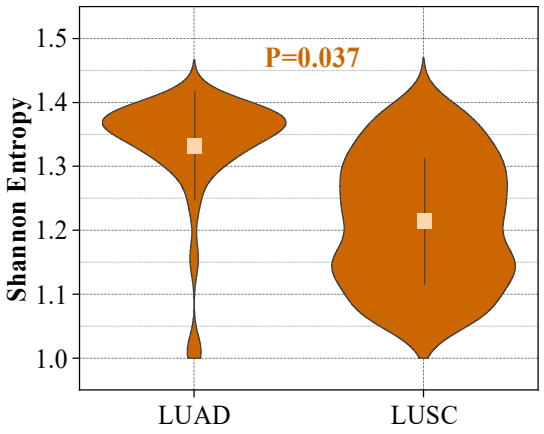
(d)CD56NK



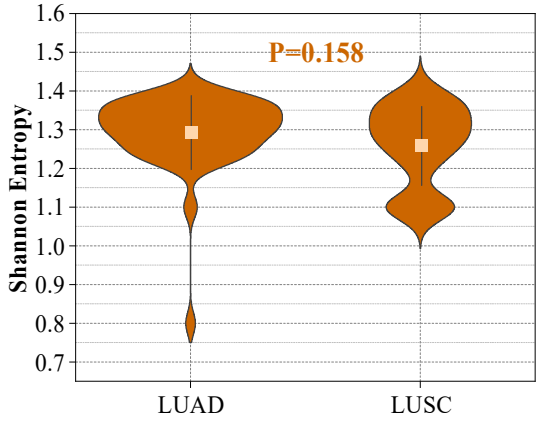
(e)CD68Macrophage



(f)CD8Tcell



(g)MHC Ⅱ



(h)PANCK

Fig. 2. Spatial heterogeneity analysis of immune cell distribution

4. Machine learning-based immune-related prognostic model for lung cancer

Mining safe and effective prognostic biomarkers and prognostic models can more accurately predict immune response and patient prognosis. To explore two lung cancer types (LUAD and LUSC) more comprehensively, this chapter utilizes machine learning algorithms to construct an immune-related prognostic model (IRS) for lung cancer.

4.1. Materials and Methods

4.1.1. Data preparation and pre-processing. Expression profiling data and clinical characterization data (e.g., gender, age, tumor stage, and survival time, etc.) of LUAD and LUSC, specifically including expression profiling data and clinical characterization data of 500 LUAD samples and data of 500 LUSC samples, were downloaded as a training set using the R package TCGAbiolinks. The GEO datasets were downloaded from the GEO database as validation sets, including GSE31210, GSE72094, GSE26939, and GSE30219, and the gene expression data, clinical data, and microarray platform files of these validation sets were downloaded, respectively. According to the corresponding platform annotation information, the probe IDs were converted to gene symbols, the probes of multiple genes were excluded, and the average value of the corresponding genes of multiple probes was calculated to complete the gene annotation. For datasets that were not normalized, log2ization was performed.

4.1.2. Cox regression analysis. The Cox regression model is a semiparametric regression model capable of exploring both single and multiple factors. Before introducing the Cox regression model, a few related concepts are introduced.

1) Survival function. Assuming that the characteristic variable of the observed sample object is $X=(X_1,X_2,\ldots,X_m)$, the survival function represents the probability that the survival time T of the sample is greater than a certain moment t:

$$S(t,X)=P(T>t,X) \tag{1}$$

It is called the survival function. Survival function $S(t,X)$ is also known as cumulative survival.

2) Death function. The death function is literally the opposite of the survival function, which represents the probability that the sample object will survive for a time T no greater than a certain moment t:

$$F(t,X)=P(T\le t,X) \tag{2}$$

is called the mortality function. In practice, the death function $F(t,X)$ can be understood as the actual meaning at a certain definite observation time is when the observation follows up to the moment t, the cumulative mortality rate of the observed subject.

3) Death density function. The death density function represents the instantaneous mortality rate of the sample object at a certain observation time t, with the following formula:

$$f(t,X)=\lim_{\Delta t\to 0}\frac{P(t<T\le t+\Delta t,X)}{\Delta t}=F'(t,X) \tag{3}$$

4) Hazard rate (risk) function. The difference between the risk function and the mortality density function is that the risk function expresses the instantaneous mortality rate at moment t after the survival time of the observed sample has reached t, which can be interpreted as a conditional instantaneous mortality rate:

$$h(t,X)=\lim_{\Delta t\to 0}\frac{P(t<T\le t+\Delta t\mid T\ge t,X)}{\Delta t}=\frac{f(t,X)}{S(t,X)} \tag{4}$$

The hazard rate function $h(t,X)$ is actually a conditional instantaneous mortality rate.

5) Censoring. In the practical application and study of survival analysis, for some samples of observations, there will be no events of interest occurring during the entire observation period, which is called censoring.

The core of survival analysis is the relationship between the characteristic variables X and $S(t,X)$. When $S(t,X)$ is affected by many factors, i.e., X is a vector, due to the existence of deletion, the survival data become a special class of research data, in this type of research problems, time t usually no longer meets the Gaussian normal distribution and does not have variance uniformity, so at this time, the traditional multiple regression method is no longer applicable, and further introduction of the Cox regression model which is more excellent in dealing with survival data.

Rather than examining the relationship between $S(t,X)$ and X directly, the Cox regression model shifts the focus to $h(t,X)$, constructing a model of the basic form:

$$h(t,X)=h_0(t)\exp(\beta_1X_1+\beta_2X_2+\ldots+\beta_mX_m) \tag{5}$$

where $\beta_1,\beta_2,\ldots,\beta_m$ is the partial regression coefficient of the independent variable, the parameter to be determined. $h_0(t)$ is the baseline hazard rate when the X vector is 0. $h(t,X)$ is the quantity to be estimated. The above model is referred to as Cox regression model.

The Cox regression model has a high degree of flexibility and versatility in practical application because it makes no assumptions about the baseline hazard rate $h_0(t)$; moreover, in many cases, the study is interested only in the parameters β rather than $h_0(t)$ in the model, while the estimation of β does not depend on $h_0(t)$. Containing $h_0(t)$ but estimating the parameters without the aid of $h_0(t)$, the Cox regression model is therefore a semiparametric model.

The formula can be transformed into:

$$\ln\left[\frac{h(t,X)}{h_0(t)}\right]=\ln RR=\beta_1X_1+\beta_2X_2+\ldots+\beta_mX_m \tag{6}$$

Univariate and multivariate Cox regression analyses were performed on the clinical factors in both the training and validation sets, and when a factor was significant ($p < 0.05$) in both the univariate and multivariate Cox regression analyses, then that factor was an independent prognostic factor.

4.1.3. Random survival forests. In order to establish an immune-related prognostic model with high robustness and accuracy, this paper uses the random survival forest algorithm as an immune-related prognostic model for constructing LUAD. The random survival forest algorithm is based on the Bagging framework, and the random forest algorithm has excellent features such as low computational complexity, parallel operation, and significant prediction effect, which is a very widely used algorithm in integrated learning. The random survival forest model consists of a binary survival tree, and its basic idea is consistent with the decision tree. When the data passes through the split node of the tree, the split criterion is to maximize the survival difference. The specific process of the random survival forest algorithm is as follows:

1) Firstly, a subsample of data is randomly sampled from the original data via Bootstrap and 63% of the sample is set as experimental data and 37% of the sample is set as out-of-bag (OOB) data. This self-sampling increases the generalization ability of the algorithm, and the OOB data is used to calculate the prediction correctness of the model.

2) Each subsample is grown into a binary survival tree model. At each split node, the split criterion is survival difference maximization, and the log-rank score is usually used as the split criterion. Grow the tree model when the terminal node is not less than the threshold set by the algorithm, otherwise stop growing the tree model.

3) Using Nelson-Aalen estimation the cumulative risk function (CHF) of each tree can be calculated, and then the average cumulative risk function of the model can be calculated using the cumulative risk function of each tree.

For a survival tree, the cumulative risk function for this terminal node h is assumed to be at some time $t_{i,h}$, with $d_{i,h}$ denoting the number of people who died at that moment and $S_{i,h}$ denoting the number of people who are at risk at that moment:

$$\hat{H}_h(t) = \sum_{t_{i,h} \le t} \frac{d_{i,h}}{S_{i,h}} \tag{7}$$

Assume that the covariate X_i for each sample has dimension d, since the tree model is a binary survival tree and X_i will only be in a single terminal node h. Therefore, the cumulative risk function of terminal node h conditional on X_i is:

$$H(t \mid X_i) = \hat{H}_h(t) \tag{8}$$

And the average cumulative risk function is calculated as:

$$H_e^*(t \mid X_i) = \frac{1}{B} \sum_{b=1}^{B} H_b^*(t \mid X_i) \tag{9}$$

where B denotes the number of survival trees and $H_b^*(t \mid X_i)$ denotes the cumulative risk function for each sample b.

4) The predictive correctness of the model is tested by out-of-bag (OOB) data, and the stochastic survival forest uses a consistency index to measure the predictive effectiveness of the model.

4.1.4. XGBoost Algorithm. For LUSC, the extreme gradient boosting algorithm is selected as the immunocorrelation feature model. Extreme Gradient Boosting (XGBoost) is an optimization algorithm based on GBDT, which extends the objective function to the second order and retains more information, in addition to introducing a regular term, which effectively reduces the risk of overfitting and improves the performance of the model.

The objective function of XGBoost is as follows:

$$L(\Phi)=\sum_{i=1}^{n} l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{10}$$

where $l(y_i,\hat{y}_i)$ is the loss function and $\Omega(f_k)$ is the canonical term defined as follows:

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\,|\,\omega|^2 \tag{11}$$

λ is the penalty strength, ω denotes the vector formed by the values of the leaf nodes of the regression tree, and T denotes the number of regression trees.

The loss function is then learned incrementally as follows:

$$\hat{y}_t^{(t)}=\sum_{k=1}^{t} f_k(x_i)=\hat{y}_t^{(t-1)}+f_t(x_i) \tag{12}$$

Above, $f_t(x_i)$ is the prediction function for round t and $\hat{y}_t^{(t)}$ is the prediction result for round t, followed by substituting $\hat{y}_t^{(t)}$ into Eq. (10) to obtain:

$$L(\Phi)^{(t)}=\sum_{i=1}^{n} l(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\sum_{k=1}^{K}\Omega(f_k) \tag{13}$$

$$L(\Phi)^{(t)}=\sum_{i=1}^{n} l(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_k)+C \tag{14}$$

After t-1 rounds of training, the prediction result y_i as well as the loss function $l(y_i,\hat{y}_i)$ are constants, the above equation is subjected to a second-order Taylor expansion, and the Taylor expansion formula is as follows:

$$f(x+\Delta x)\cong f(x)+f'(x)\Delta x+\frac{1}{2}f''(x)\Delta x^2 \tag{15}$$

can be obtained by removing the constant terms y_i and $l(y_i,\hat{y}_i)$:

$$L(\Phi)^{(t)}\cong\sum_{i=1}^{n} l(y_i,\hat{y}_i^{(t-1)})+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)+\Omega(f_t) \tag{16}$$

where g_i is the first order derivative and h_i is the second order derivative with the following expression:

$$g_i=\partial_{\hat{y}_i^{(t-1)}} l(y_i,\hat{y}_i^{(t-1)}) \tag{17}$$

$$h_i=\partial^2_{\hat{y}_i^{(t-1)}} l(y_i,\hat{y}_i^{(t-1)}) \tag{18}$$

The objective function is further calculated to obtain the objective function:

$$L(\Phi)^{(t)}=\sum_{i=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j+\frac{1}{2}\left(\sum_{i\in I_j} h_i+\lambda\right)w_j^2\right]+\gamma T \tag{19}$$

where $I_j=i\,|\,q(x_i)=j$ is the set of all samples in the leaf node, and w_j denotes the weight corresponding to leaf node j, and the expression is:

$$w_j^*=-\frac{G_j}{H_j+\lambda} \tag{20}$$

Seek further:

$$L(\Phi)^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T \tag{21}$$

where G_j and H_j are the optimal weights corresponding to the leaf nodes with the following expression:

$$G_j = \sum_{i \in Ij} g_i \tag{22}$$

$$H_j = \sum_{i \in I_j} h_i \circ w_j^* \tag{23}$$

Iterate through all the potential split points in the features of the sample point and calculate the gain situation before and after the split respectively, the information gain after the split is as follows:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{24}$$

Compare the values of Gain and select the feature with the largest value as the optimal division.

4.2. IRS model assessment analysis

4.2.1. LUAD model analysis. In order to evaluate the prognostic ability of the LUAD model, the risk score of each patient in the training dataset was first calculated by the IRS model of LUAD, and using the surv_cutpoint function in the survminer package, the cutoff value that can significantly distinguish the risk score was selected, and then all the patients were assigned to the high-risk and low-risk groups based on the cutoff value, and then the high- and low-risk groups were subjected to a survival analysis. In addition, this paper further verified the accuracy of IRS in three GEO validation datasets by calculating the risk score of each patient in the three validation datasets and the cutoff value of the risk score in each validation set through the IRS model, and then classified the patients into high- and low-risk groups according to the cutoff value of the risk score, and drew the Kaplan-Meier survival curves for the high- and low-risk groups.

The KM survival curves of the LUAD dataset are shown in Fig. 3, and (a) to (d) represent the KM survival curves of the training set, the validation set GSE31210, GSE72094, and GSE26939, respectively, with the p-values based on the Log-Rank test. The overall survival (OS) of patients in the high-risk group in the training set was significantly lower than that of the low-risk group ($p < 0.0001$), and consistently showed significantly lower OS of patients in the high-risk group than that of the low-risk group in GSE31210, GSE72094, and GSE26939 ($p < 0.0001$, $p < 0.0001$, and $p = 0.0053$). These results further indicate that the IRS model constructed in this paper has good prognostic ability and can effectively differentiate the prognostic risk profiles of LUAD patients.

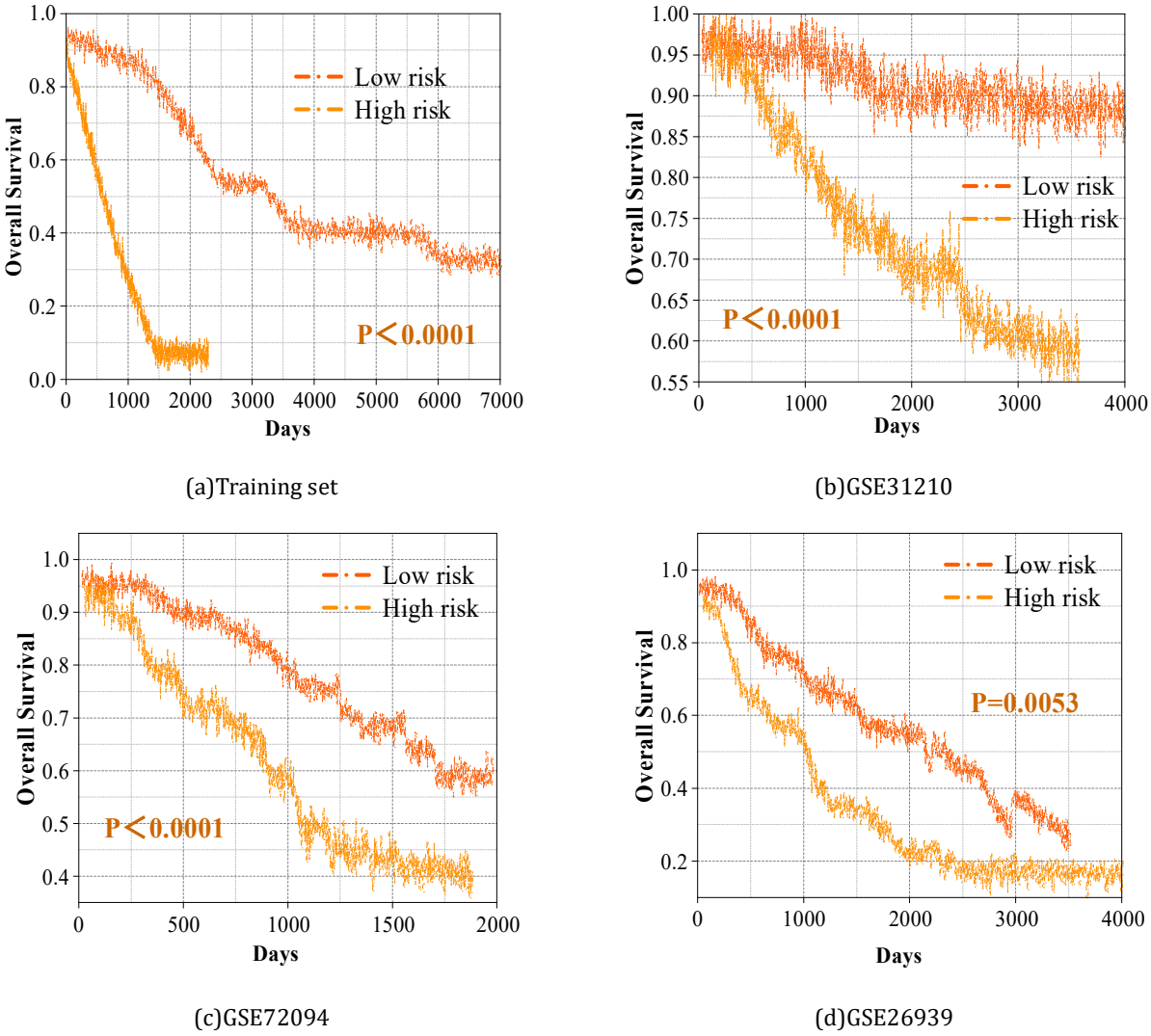
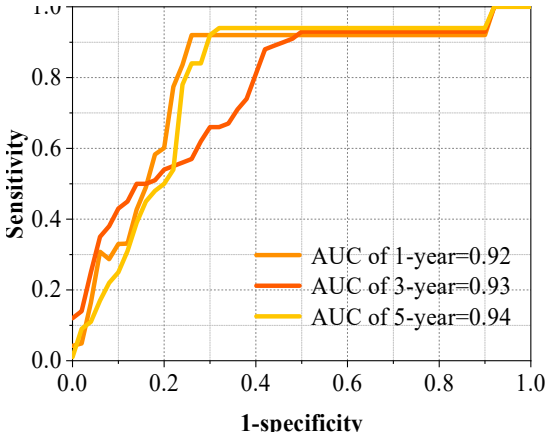
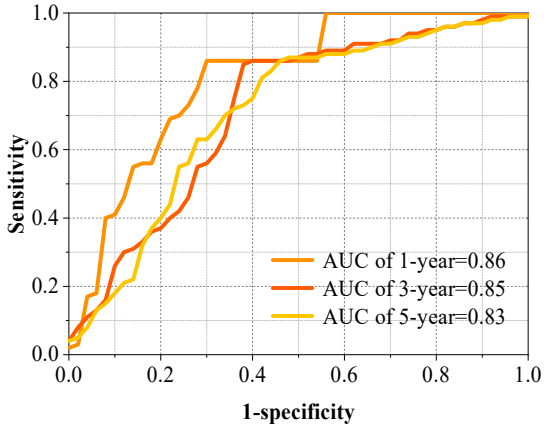


Fig. 3. The KM survival curve of the LUAD data set

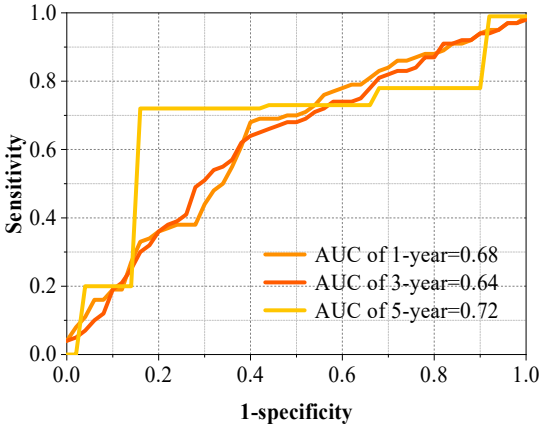
In order to evaluate the accuracy of the IRS model of LUAD. In this paper, we further show the AUC values for 1, 3 and 5 years for the training set, GSE31210, GSE72094 and GSE26939, respectively.The ROC curves of the LUAD dataset are shown in Fig. 4, and (a) to (d) denote the ROC curves of each dataset, respectively. The AUC values for 1, 3 and 5 years in the training set are 0.92, 0.93 and 0.94, respectively, in the GSE31210 dataset are 0.86, 0.85 and 0.83, respectively, in the GSE72094 dataset are 0.68, 0.64 and 0.72, respectively, in the GSE26939 dataset are 0.70, 0.65 and 0.63.All these metrics indicate stable performance of IRS in multiple independent cohorts, confirming the prognostic value of IRS.



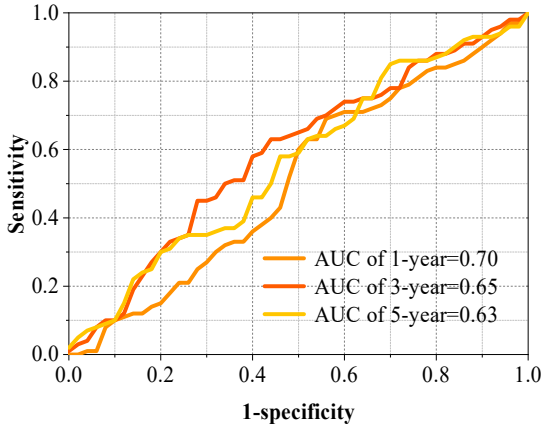
(a)Training set



(b)GSE31210



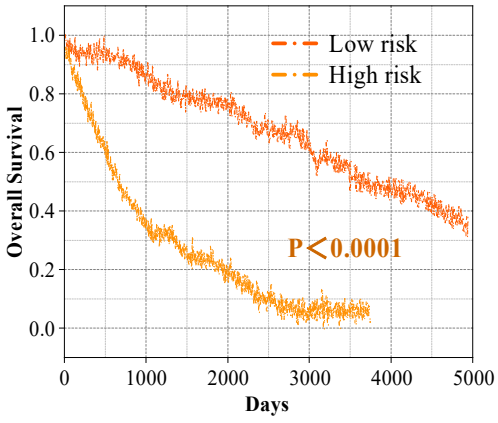
(c)GSE72094



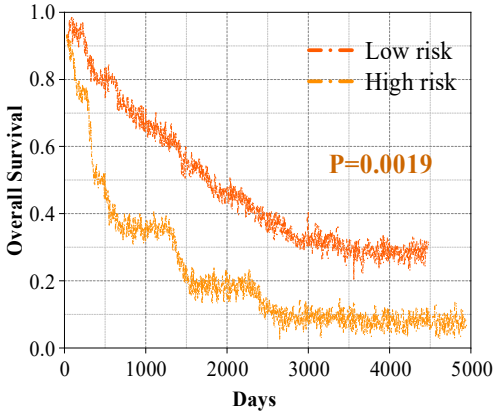
(d)GSE26939

Fig. 4. ROC curve of LUAD data set

4.2.2. LUSC model analysis. The KM survival curves for the LUSC dataset are shown in Figure 5, and (a) and (b) are the KM survival curves for the training set and validation set GSE30219, respectively. Similarly, the OS of patients in the high-risk group was found to be significantly lower than that of the low-risk group in the training set of LUSC and the test set of GSE30219 ($p<0.0001$, $p=0.0019$), indicating that the IRS constructed in this paper has good prognostic ability.



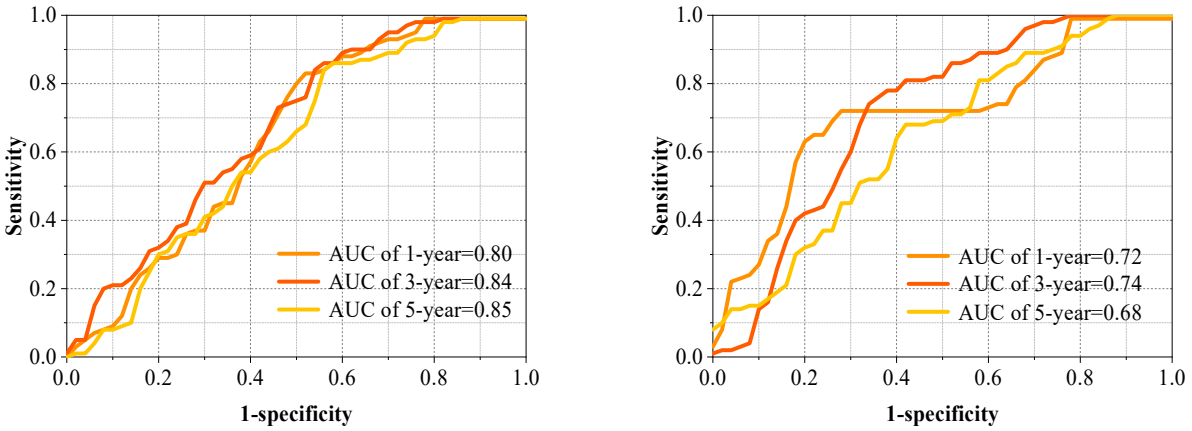
(a)Training set



(b)GSE30219

Fig. 5. The KM survival curve of the LUSC data set

Figure 6 shows the ROC curves for the LUSC dataset, and (a) and (b) show the ROC curves for the training set and validation set GSE30219, respectively. The AUC values were 0.80, 0.84 and 0.85 for 1, 3 and 5 years in the training set and 0.72, 0.74 and 0.68 in the GSE30219 dataset, respectively.The AUC metrics indicate that the IRS model has a stable performance in multiple independent cohorts, reflecting the prognostic accuracy of IRS.



(a)Training set (b)GSE30219

Fig. 6. ROC curve of LUSD data set

5. Conclusion

Currently, the incidence and mortality rates of lung cancer rank among the top of all types of malignant tumors worldwide. In this paper, we mainly use IHC as well as mIHC technology to explore the spatial distribution characteristics of immune cells in the microenvironment of lung cancer. Machine learning algorithms are used to construct mathematical models based on prognosis-related immune genes, which provide more analytical perspectives for cancer research, and IRS models are evaluated in the selected datasets.

1) The immune microenvironment of lung adenocarcinoma and lung squamous carcinoma tumor tissues differed somewhat, with a greater number of CD56+ NK cells, CD68+ macrophages, and MHC class II antigen-presenting cells infiltrating in lung adenocarcinoma tissues ($P<0.05$).The spatial distribution of CD3+ T cells ($P=0.017$) and MHC class II antigen-presenting cells ($P=0.037$) in lung adenocarcinoma showed greater heterogeneity, and the spatial distribution of other immune cells in lung adenocarcinoma and squamous lung cancer tumor tissues did not differ significantly.

2) Survival analysis found that all the data of LUAD and LUSC demonstrated that the overall survival of patients in the high-risk group was significantly lower than that of the low-risk group ($P<0.01$), and the accuracy and robustness of the model was further verified by the GEO dataset, with the AUC values of the model in the range of 0.63 to 0.94 under the data of different years, which reflected the good prognostic ability of the constructed IRS model.

In this paper, the spatial distribution of immune cells in the lung cancer microenvironment and the construction of a prognostic model were investigated to provide a reference for prognostic prediction, clinical decision-making, biomarker discovery and personalized treatment of NSCLC patients. However, there are some limitations in this study, and only two NSCLC subtypes were considered, which is far from enough to systematically explore the molecular characteristics of different lung cancer types, and a more comprehensive study can be conducted on all lung cancer types in the future.

Acknowledgements

1. The Excellent Teaching Team Fund of the Cyan-Blue Project at Yangzhou Polytechnic College;
2. Scientific and Technological Innovation Team Building Program of Suzhou Vocational Health College (No. SZWZYTD202205);
3. The Youth Fund of the Cyan-Blue Project at Jiangsu Province (Su Teacher Letter [2022] No. 51).

References

[1] Stankovic, B., Bjørhovde, H. A. K., Skarshaug, R., Aamodt, H., Frafjord, A., Müller, E., ... & Corthay, A. (2019). Immune cell composition in human non-small cell lung cancer. Frontiers in immunology, 9, 3101.

[2] Muppa, P., Terra, S. B. S. P., Sharma, A., Mansfield, A. S., Aubry, M. C., Bhinge, K., ... & Kosari, F. (2019). Immune cell infiltration may be a key determinant of long-term survival in small cell lung cancer. Journal of Thoracic Oncology, 14(7), 1286-1295.

[3] Liu, X., Wu, S., Yang, Y., Zhao, M., Zhu, G., & Hou, Z. (2017). The prognostic landscape of tumor-infiltrating immune cell and immunomodulators in lung cancer. Biomedicine & Pharmacotherapy, 95, 55-61.

[4] Tamminga, M., Hiltermann, T. J. N., Schuuring, E., Timens, W., Fehrmann, R. S., & Groen, H. J. (2020). Immune microenvironment composition in non-small cell lung cancer and its association with survival. Clinical & translational immunology, 9(6), e1142.

[5] Kargl, J., Busch, S. E., Yang, G. H., Kim, K. H., Hanke, M. L., Metz, H. E., ... & Houghton, A. M. (2017). Neutrophils dominate the immune cell composition in non-small cell lung cancer. Nature communications, 8(1), 14381.

[6] Zhong, R., Chen, D., Cao, S., Li, J., Han, B., & Zhong, H. (2021). Immune cell infiltration features and related marker genes in lung cancer based on single-cell RNA-seq. Clinical and Translational Oncology, 23, 405-417.

[7] Soo, R. A., Chen, Z., Teng, R. S. Y., Tan, H. L., Iacopetta, B., Tai, B. C., & Soong, R. (2018). Prognostic significance of immune cells in non-small cell lung cancer: meta-analysis. Oncotarget, 9(37), 24801.

[8] Pullamsetti, S. S., Kojonazarov, B., Storn, S., Gall, H., Salazar, Y., Wolf, J., ... & Savai, R. (2017). Lung cancer–associated pulmonary hypertension: Role of microenvironmental inflammation based on tumor cell–immune cell cross-talk. Science Translational Medicine, 9(416), eaai9048.

[9] Riemann, D., Cwikowski, M., Turzer, S., Giese, T., Grallert, M., Schütte, W., & Seliger, B. (2019). Blood immune cell biomarkers in lung cancer. Clinical & Experimental Immunology, 195(2), 179-189.

[10] Anichini, A., Perotti, V. E., Sgambelluri, F., & Mortarini, R. (2020). Immune escape mechanisms in non small cell lung cancer. Cancers, 12(12), 3605.

[11] Busch, S. E., Hanke, M. L., Kargl, J., Metz, H. E., MacPherson, D., & Houghton, A. M. (2016). Lung cancer subtypes generate unique immune responses. The Journal of Immunology, 197(11), 4493-4503.

[12] Fucà, G., Galli, G., Poggi, M., Russo, G. L., Proto, C., Imbimbo, M., ... & Signorelli, D. (2019). Modulation of peripheral blood immune cells by early use of steroids and its association with clinical outcomes in patients with metastatic non-small cell lung cancer treated with immune checkpoint inhibitors. ESMO open, 4(1), e000457.

[13] Mark, N. M., Kargl, J., Busch, S. E., Yang, G. H., Metz, H. E., Zhang, H., ... & Houghton, A. M. (2018). Chronic obstructive pulmonary disease alters immune cell composition and immune checkpoint inhibitor efficacy in non–small cell lung cancer. American Journal of Respiratory and Critical Care Medicine, 197(3), 325-336.

[14] Fucikova, J., Becht, E., Iribarren, K., Goc, J., Remark, R., Damotte, D., ... & Cremer, I. (2016). Calreticulin expression in human Non–Small cell lung cancers correlates with increased accumulation of antitumor immune cells and favorable prognosis. Cancer research, 76(7), 1746-1756.

[15] Zheng, X., Hu, Y., & Yao, C. (2017). The paradoxical role of tumor-infiltrating immune cells in lung cancer. Intractable & rare diseases research, 6(4), 234-241.

[16] Tuminello, S., Veluswamy, R., Lieberman-Cribbin, W., Gnjatic, S., Petralia, F., Wang, P., ... & Taioli, E. (2019). Prognostic value of immune cells in the tumor microenvironment of early-stage lung cancer: a meta-analysis. Oncotarget, 10(67), 7142.