Discover Oncology

Analysis

Big data-driven machine learning: transforming multi-omics lung cancer research

Yanqi Zhang[1] · Mingyu Liu[2] · Jinhua Luo[2] · Zhongqing Xu[1]

Received: 17 January 2025 / Accepted: 9 May 2025
Published online: 24 May 2025

Abstract

Background Lung cancer remains a major global health threat, with its biological complexity and patient heterogeneity posing significant challenges. Novel machine learning approaches now offer effective tools to interpret complex biological information hierarchies, showing promise to transform lung cancer treatment approaches.

Methods We analyzed comprehensive biological datasets from TCGA and other databases, integrating DNA, RNA, miRNA, protein, and metabolite information. Multiple machine learning methods were employed to build diagnostic tools, treatment response predictors, and survival estimation models.

Results Our machine learning approaches effectively distinguished cancer patients from healthy controls. Analysis identified unique molecular characteristics between lung cancer subtypes and discovered biomarkers that help predict treatment efficacy and patient prognosis. Adding clinical data to biological information significantly improved model accuracy and enhanced patient stratification.

Conclusion This study marks significant progress toward precision cancer therapy by demonstrating how machine learning can help decode the complex biology of lung cancer.

Keywords Lung cancer · Machine learning · Multi-omics · Diagnosis · Treatment · Prognosis

1 Introduction

Lung cancer constitutes one of the biggest health challenges globally, with mortality rates among the highest of all cancers. Its diverse molecular characteristics present major obstacles for researchers and clinicians, hindering early detection efforts and personalized treatment development [1–4]. Despite important advances in treatment, including new immunotherapies and targeted drugs, lung cancer outcomes remain poor, with five-year survival rates hovering around 15%, particularly for patients with advanced disease.

The multi-omics field has transformed our understanding of lung cancer biology. Integration of genetic, gene expression, protein, and metabolite data has enabled creation of detailed molecular disease profiles [5–8]. This approach yields new insights into the complex biological networks driving tumor formation, helping identify key cancer development mechanisms and potential new markers for diagnosis, prognosis, and treatment.

Yanqi Zhang and Mingyu Liu are co-first author. Jinhua Luo and Zhongqing Xu are co-corresponding author.

✉ Jinhua Luo, ljhua19661220@163.com; ✉ Zhongqing Xu, zhongqing_xu@126.com | [1]Department of General Practice, Tongren Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200336, China. [2]Department of Thoracic Surgery, The First Affiliated Hospital of Nanjing Medical University, Nanjing 211166, China.

Machine learning has become an important tool for processing complex multi-level biological data. These computational methods can identify subtle patterns hidden within large, multidimensional datasets [9–11]. Where traditional analytical methods falter, machine learning techniques can unravel complex associations, offering unique perspectives into lung cancer's molecular intricacies and paving the way for more personalized, targeted treatment strategies.

The multi-omics approach allows for more precise patient stratification beyond traditional TNM staging, potentially guiding treatment decisions between standard chemotherapy, targeted therapy, and immunotherapy. This work stands at the intersection of computational innovation and clinical oncology, embodying a holistic approach to understanding lung cancer's molecular complexity. As we continue to refine our methodologies, we move closer to a future where lung cancer can be detected earlier, treated more effectively, and ultimately, its devastating consequences significantly reduced.

2 Materials and methods

2.1 Data source acquisition

Our research data came from The Cancer Genome Atlas (TCGA) project, accessed through their official portal (https://portal.gdc.cancer.gov) with proper authorization. We gathered extensive data on lung squamous cell carcinoma (LUSC), including miRNA sequencing, RNA sequencing, DNA methylation profiles from Illumina 450 k platform, whole-exome sequencing results, clinical details, and patient survival information. This comprehensive data collection formed a strong basis for our detailed analysis [12–14].

2.2 Differential gene expression analysis

We used Kaplan–Meier survival analysis to group LUSC patients. Using strict statistical filters (p-value < 0.05) and log2 fold change criteria, we found genes that showed different expression patterns linked to how the disease progresses. We compared normal tissue samples with LUSC tissue samples to spot important molecular changes during tumor growth. When building our prediction model (StepCox [forward] + RSF), we carefully optimized all parameters. For the Random Survival Forest part, we found the best settings through grid search: 500 trees, at least 5 samples per node, and features selected using the sqrt(p) formula. To make sure our model wasn't just memorizing data, we split our dataset into 5 equal parts, using 4 parts to train and 1 part to test, repeating this process 5 times and averaging the results. In building the StepCox [forward] + RSF model, we employed a systematic parameter optimization strategy. For the Random Survival Forest (RSF) model, optimal parameters were determined through grid search: 500 trees, minimum of 5 samples per node, and feature selection using the sqrt(p) criterion, where p is the total number of features. To avoid overfitting, we used fivefold cross-validation, randomly dividing the dataset into 5 equal parts, using 4 parts as the training set and 1 part as the validation set each time, repeating 5 times and taking the average performance.

2.3 Tumor microenvironment characterization

To comprehensively unravel the complexity of the tumor microenvironment, the study deployed advanced bioinformatics algorithms. The ESTIMATE algorithm was utilized to assess immune scores, while the CIBERSORT algorithm precisely quantified immune cell infiltration. For the ESTIMATE algorithm implementation, we used the R package "estimate" (version 1.0.13) with default parameters to calculate immune, stromal, and ESTIMATE scores based on gene expression data. Input data were log2-transformed FPKM values, and results were normalized using quantile normalization. For CIBERSORT analysis, we employed the web-based tool (https://cibersort.stanford.edu) with 1000 permutations and disabled quantile normalization given our prior preprocessing. By comparing tumor microenvironment characteristics across different disease stages, the research unveiled the dynamic changes of immune cells during lung squamous cell carcinoma progression. This multi-dimensional microenvironment analysis not only provided an immunological perspective on disease progression but also offered profound insights into tumor-immune system interactions [15, 16].

2.4 Multi-omics data integration

We built a complex model to predict patient outcomes using a three-step process: First, we used univariate Cox regression to select potentially important features from gene expression, mutation, and methylation data. Next, we applied LASSO regression to remove less informative variables and simplify the model. Finally, we used multivariate Cox regression to calculate risk scores for each patient, placing them in either high-risk or low-risk groups [17, 18]. By plotting receiver operating characteristic (ROC) curves for 1, 3, 5, and 8 years and calculating the area under the curve (AUC), the model's predictive performance was comprehensively evaluated. This multi-step, multi-model ensemble approach significantly enhanced the accuracy and reliability of prognostic prediction.

2.5 Single-cell analysis

Single-cell RNA sequencing technology provided unprecedented molecular resolution for lung adenocarcinoma research. By acquiring the GSM6047623 dataset from the TISCH database, the study systematically revealed the complex heterogeneity of tumor cells. Through a rigorous bioinformatics workflow, the research team initially implemented multi-level data cleaning and quality control, including removing low-quality reads, standardizing gene expression data, and utilizing principal component analysis for dimensionality reduction [19–21]. Multi-omics data obtained from the TCGA database underwent rigorous preprocessing and quality control. First, RNA-seq data were normalized using the TMM (trimmed mean of M-values) method to eliminate differences in sequencing depth between samples, and multiple imputation was applied to genes with less than 20% missing values. Methylation data were corrected for chip bias using the BMIQ (Beta Mixture Quantile) method, and probes located on X/Y chromosomes and at SNP sites were removed. For single nucleotide variation (SNV) data, we only retained variations present in at least 5% of samples and evaluated their significance using MutSigCV. Employing advanced algorithms like t-SNE and UMAP, the team successfully clustered single-cell data into multiple discrete transcriptional state clusters, precisely identifying immune cells, epithelial cells, and stromal cell subpopulations within the tumor microenvironment. In-depth analysis unveiled the transcriptional dynamic changes of lung adenocarcinoma cells, encompassing tumor cells at different differentiation stages, dynamic transformations of immune cells within the tumor microenvironment, and potential cancer stem cells and metastasis-associated cell subpopulations.

3 Results

3.1 Functional enrichment analysis

The comprehensive analysis of the provided data reveals a multifaceted and intricate molecular landscape underlying the biological processes associated with the research context. Through the detailed visualizations, key insights emerge regarding the complex interplay of various cellular mechanisms and signaling pathways. The biological process analysis highlights the prominent involvement of pathways related to skin development and differentiation, including cornification, keratinization, keratinocyte differentiation, and epidermal cell differentiation. These findings suggest that the data is closely tied to cellular transformation and developmental processes, expanding beyond traditional understandings of the research subject. Complementing the biological processes, the cellular component analysis underscores the importance of membrane-associated structures, cytoplasmic vesicles, and organellar components, indicating the crucial role of cellular organization and compartmentalization in the studied system. At the molecular level, the data reveals a strong enrichment of receptor activator and regulator activities, as well as various signal transduction-related functions. These insights point to the central importance of regulatory and signaling mechanisms in shaping the underlying biology. Further exploration through pathway analysis unveils the engagement of neuroactive ligand-receptor interactions, calcium signaling, and metabolism-related pathways. This multifaceted perspective highlights the intricate interplay of cellular signaling, metabolic processes, and potentially even neurological components within the research context (Fig. 1A–D).

3.2 Machine learning and model construction

Figure 2 shows a detailed visualization of the study data through a color-coded heatmap. This display organizes information by sample groups (rows) and measured features (columns). Each cell's color intensity represents data values,

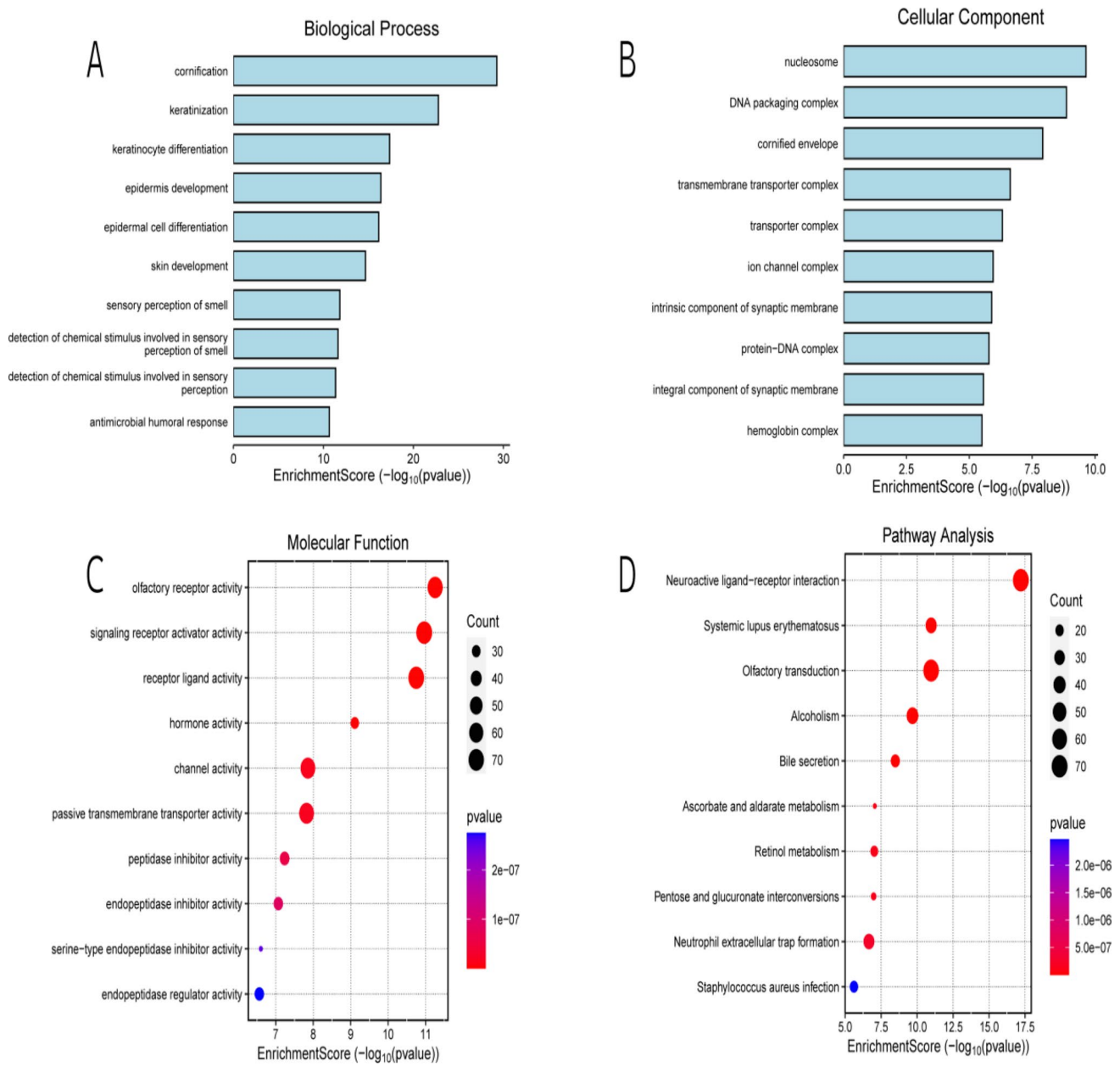

Fig. 1 Functional enrichment analysis. **A**–**D** The examination of TCGA datasets demonstrated that lung cancer pathogenesis encompasses an intricate biological network that extends beyond conventional frameworks. Our analysis identified unexpected enrichment of epithelial developmental pathways, with significant involvement of cornification, keratinization, keratinocyte differentiation, epidermal development, and epidermal cell differentiation processes

making it easy to spot patterns across different groups. The figure includes supporting elements: a bar graph on the right showing average C-index values for all cohorts and validation cohorts specifically, plus a color legend on the left that explains the value ranges corresponding to each shade. Together, these components provide a clear picture of relationships and trends in the cohort data.

3.3 Evaluation of the AUC index of the optimal model

Figure 3 evaluates model performance using different metrics. Figure 3A and B compare C-index values between StepCox [forward] + RSF and RSF models for both training and testing datasets, showing how well each model separates high-risk from low-risk patients. Figure 3C and D track how iAUC values change as model complexity increases for both approaches.

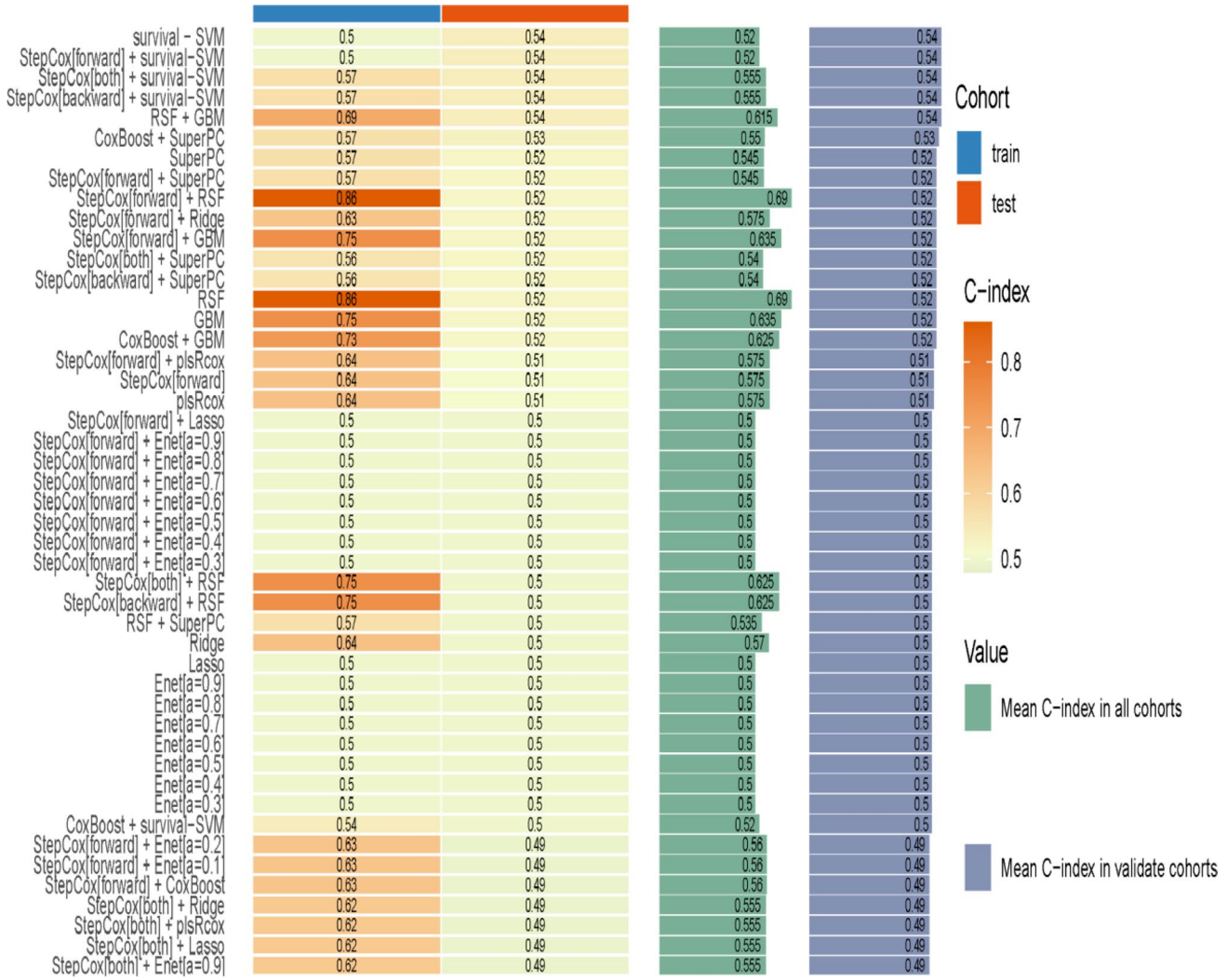

Fig. 2 Machine learning and model construction. In order to construct a diagnostic model for lung cancer, the research team leveraged a diverse set of 15 machine learning algorithms to analyze the previously identified prognostic-related genes. The lung adenocarcinoma dataset was strategically partitioned into two subgroups—TCGA and GEO—based on their respective data sources. Within the TCGA cohort, the researchers employed a ten-fold cross-validation framework to meticulously fit 101 predictive models and subsequently calculate the C-index, a metric that assesses the accuracy of model predictions

The data indicates both models perform well, with accuracy improving as more components are added, suggesting that incorporating multiple predictive features enhances the models' ability to distinguish between patient risk levels.

3.4 The survival curve of the optimal model

Figure 4A: Presents the survival probabilities over time for the StepCox [forward] + RSF model. The plot shows the survival curves for both the low-risk and high-risk groups identified by the model in the training set. The survival probability for the low-risk group (depicted in blue) remains higher compared to the high-risk group (shown in red) over the course of the time period. Figure 4B: Shows the survival probabilities for the RSF (Random Survival Forest) model, again with separate curves for the low-risk and high-risk groups in the training set. A similar pattern is observed, where the low-risk group exhibits better survival outcomes compared to the high-risk group. Figure 4C: Depicts the survival probability curves for the StepCox [forward] + RSF model in the testing dataset. The trends are consistent with the training set, with the low-risk group maintaining higher survival probabilities over time. Figure 4D: Presents the survival probability plots for the RSF model in the testing dataset. Once more, the low-risk group demonstrates improved survival outcomes compared to the high-risk group, mirroring the patterns seen in the training set.

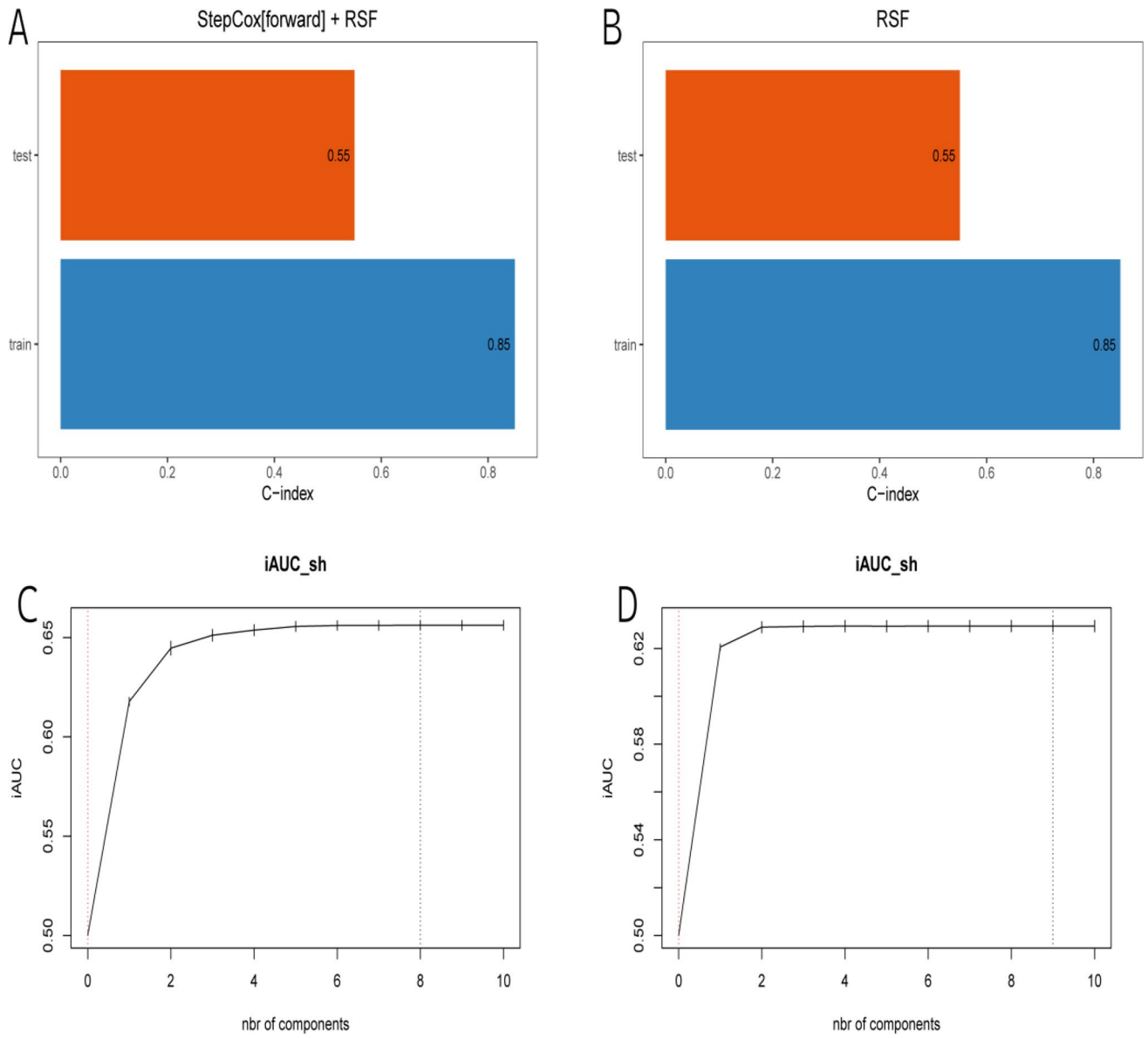

Fig. 3 Evaluation of the AUC index of the optimal model. **A** and **B** of the image present the C-index (Concordance Index) and individualized AUC (iAUC) values for two distinct machine learning models—StepCox [forward] + RSF and RSF—on both the training and testing datasets. C and D reveal that as the number of model components or steps increases, the values of both the C-index and iAUC improve consistently across the training and testing sets. This trend suggests that the enhanced model complexity and integration of additional predictive features contribute to the continuous enhancement of the models' performance, both in terms of accurate predictions and discriminative ability between high-risk and low-risk individuals

3.5 Comprehensive view of immune cell composition in the tumor microenvironment

The rows of the heatmap represent distinct risk types or attributes, such as RiskScore_total, basic, estimate, and others. The columns correspond to different cohorts or sample groups, potentially reflecting different experimental conditions, treatment responses, or other factors. The heatmap uses a color-coding scheme to represent the values or magnitudes of the risk types, with the intensity of the colors indicating the relative magnitude. This visual display allows for the quick identification of patterns, trends, and differences across the various cohorts or groups. For example, the heatmap reveals that certain risk types, such as RiskScore_total and basic, exhibit varying levels of intensity across the different cohorts, suggesting the potential significance of these risk factors in the context of the study.

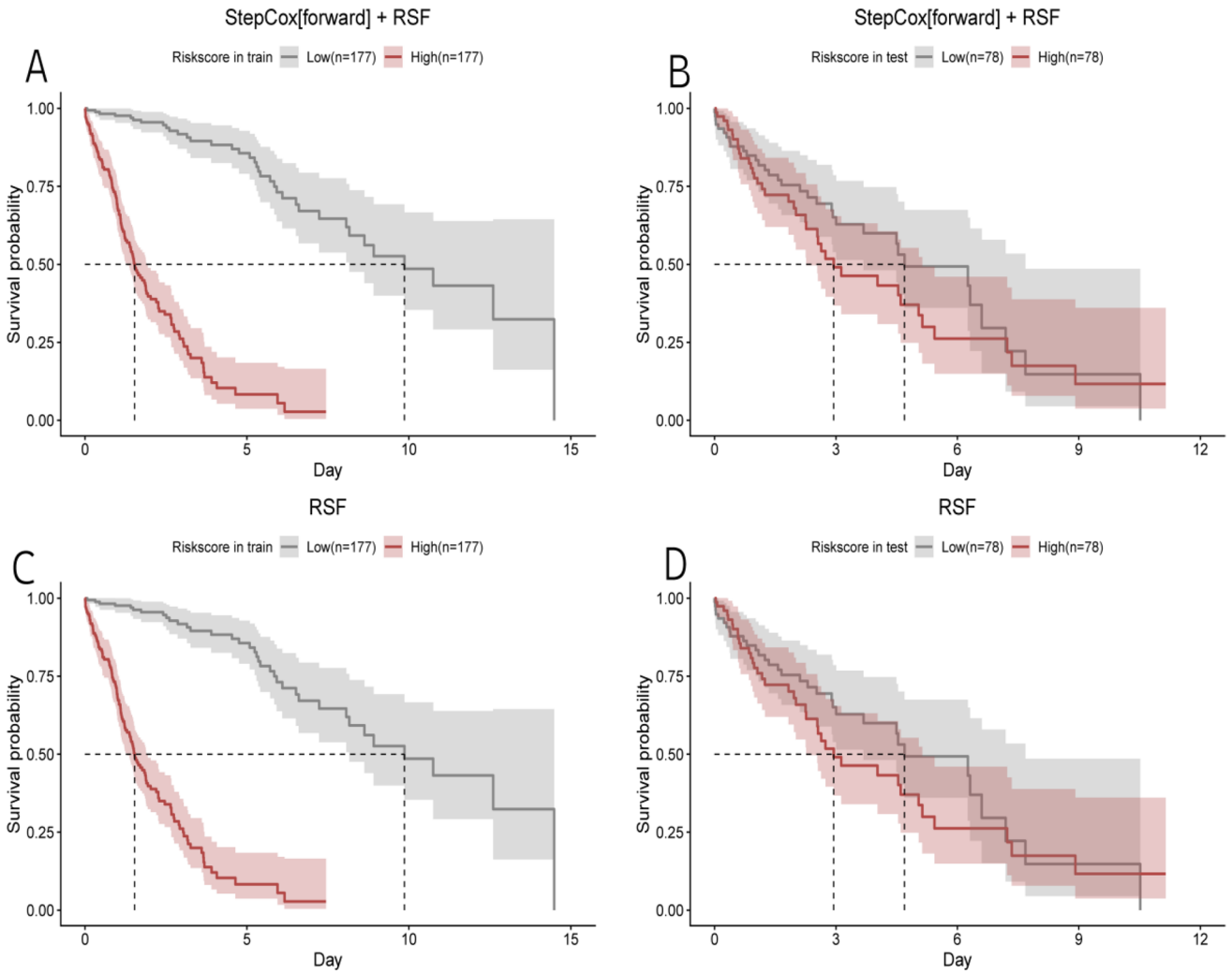

Fig. 4 The survival curve of the optimal model. **A** The survival probabilities of the low-risk (n = 177) and high-risk (n = 177) groups from the training set are displayed, showcasing the divergence in their trajectories over the observed time period. Similarly, **B** depicts the survival probability curves for the low-risk (n = 78) and high-risk (n = 78) groups in the testing dataset, demonstrating a comparable separation between the two risk cohorts. **C** and **D** present analogous analyses conducted using the RSF (Random Survival Forest) model, again highlighting the distinct survival probabilities between the low-risk and high-risk populations, both in the training and testing settings

Additionally, the legend provided on the right side of the image assigns specific color ranges to different value ranges, enabling the interpretation of the heatmap's color-coding and the understanding of the underlying data (Fig. 5).

3.6 Different cell types and gene expression characteristics

Figure 6A: This plot shows a 2-dimensional visualization of the data, likely achieved through a dimensionality reduction technique such as t-SNE or UMAP. The different colored clusters represent distinct cell types or subpopulations identified within the dataset. The labels indicate the identities of these cell types, including Epithelial, Mast, Fibroblast, Endothelial, and B cells. Figure 6B: This panel also displays a 2-dimensional visualization, potentially using a different dimensionality reduction method. The cell type identities are again labeled, and the clusters appear to show a similar overall structure to Panel A, but with some differences in the relative positioning and separation of the cell populations. Figure 6C: This panel presents a heatmap visualization, where the rows correspond to various "Features" (likely gene expression or other molecular signatures), and the columns represent the different cell types. The heatmap depicts the average expression levels (left) and the percentage of cells expressing each feature (right) across the identified cell populations.

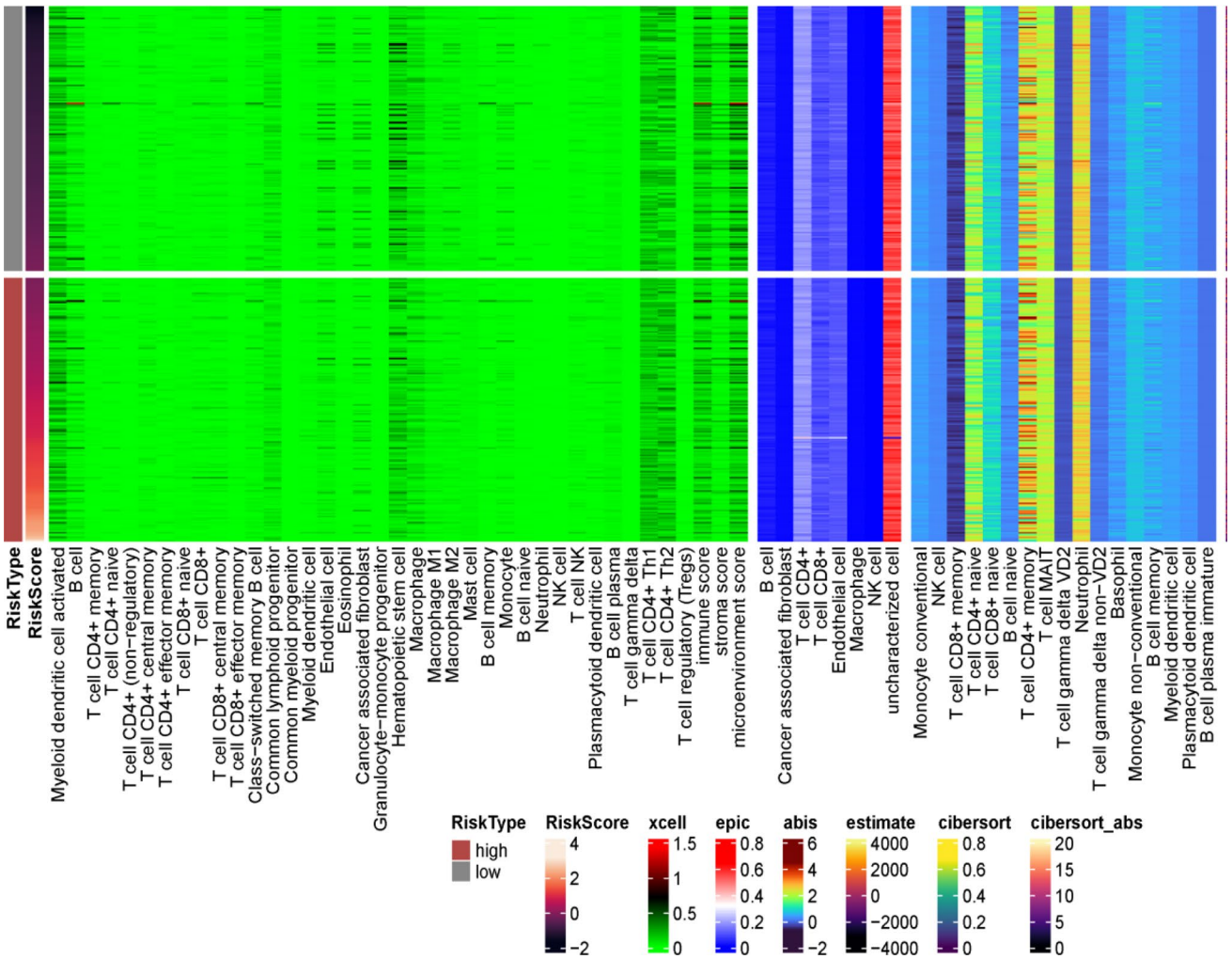

Fig. 5 Comprehensive view of immune cell composition in the tumor microenvironment. The visualization presented in the chart provides a comprehensive overview of the diverse immune cell populations present within the studied system. This includes myeloid dendritic cells, B cells, and an array of T cell subtypes, such as CD4 + memory cells and CD8 + cytotoxic cells. Additionally, the chart depicts plasma cells, mast cells, monocytes, neutrophils, and natural killer (NK) cells, offering a detailed characterization of the immune landscape

3.7 Differences in the expression of six genes identified based on single-cell expression profiling

Figure 7A: Shows the expression levels of a specific "Type" gene, where the expression is significantly higher in the tumor samples compared to the normal samples. Figure 7B: Presents the expression levels of another "Type" gene, again with significantly elevated expression in the tumor samples compared to the normal samples. Figure 7C: Depicts the expression levels of a third "Type" gene, which exhibits a similar pattern of higher expression in the tumor group relative to the normal group. Figure 7D: Illustrates the expression levels of a fourth "Type" gene, with a clear upregulation in the tumor samples compared to the normal samples. Figure 7E: Demonstrates the expression levels of a fifth "Type" gene, which is also significantly increased in the tumor samples compared to the normal samples. Figure 7F: Showcases the expression levels of a sixth "Type" gene, which displays a comparable trend of elevated expression in the tumor group versus the normal group. Across all six panels, the consistent pattern observed is that the expression levels of the various "Type" genes are significantly higher in the tumor samples compared to the corresponding normal samples. This suggests that these genes may play important roles in the development or progression of the tumor condition being studied.

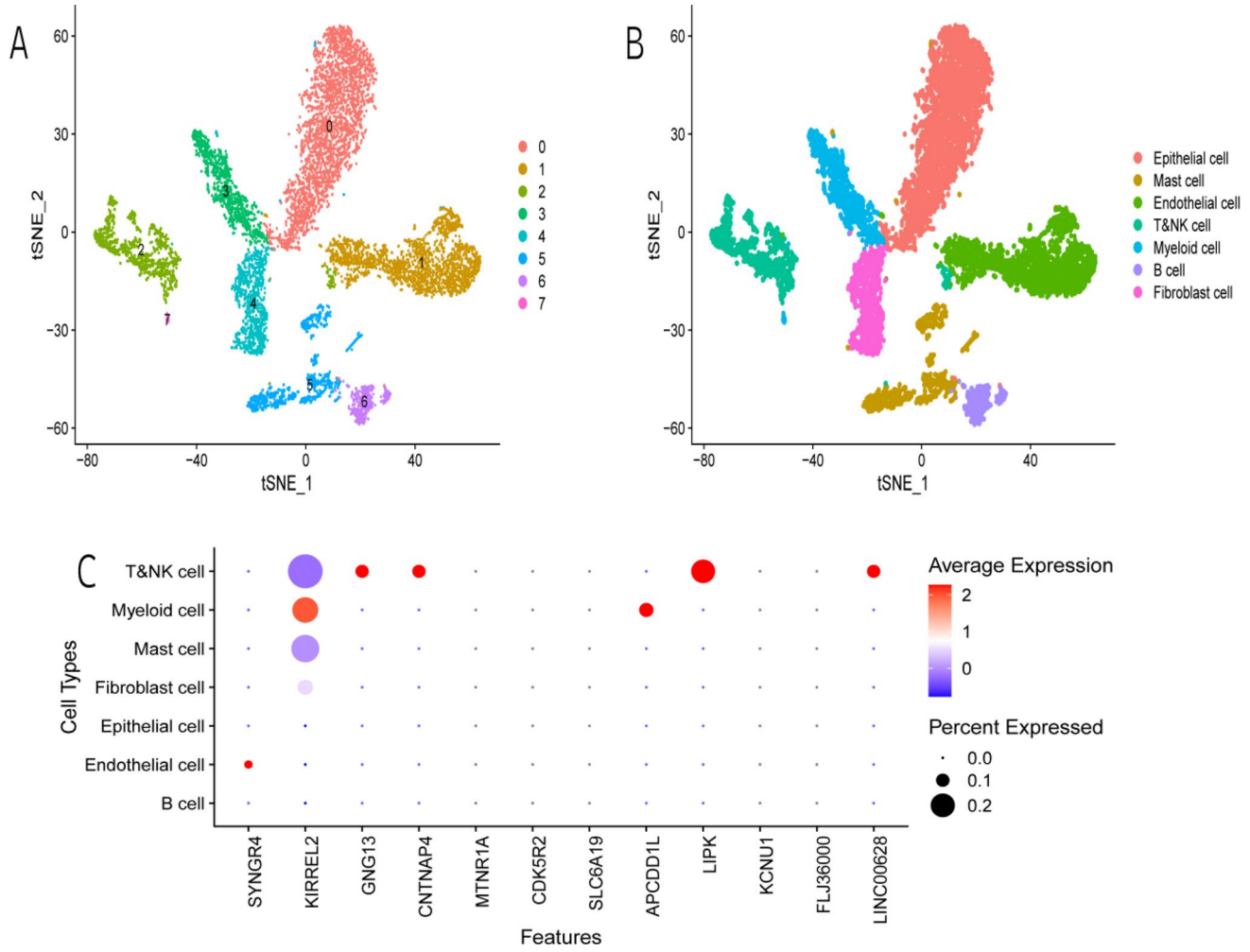

Fig. 6 Different cell types and gene expression characteristics. **A** and **B** of the image leverage the t-SNE (t-distributed Stochastic Neighbor Embedding) algorithm, a widely-adopted technique for dimensionality reduction and high-dimensional data visualization. This approach allows for the depiction of the distribution of diverse cell types within the t-SNE space, with cells organized into distinct clusters representing endothelial cells, fibroblasts, epithelial cells, myeloid cells, T and NK cells, mast cells, and B cells. C provides insights into the expression patterns of specific genes, including SYNGR4, KIRREL2, and GNG13, across the identified cellular subpopulations. This gene expression analysis complements the cell type clustering visualized in the preceding panels, enabling a more nuanced understanding of the molecular signatures that define and distinguish the different cellular entities

3.8 The survival differences between normal and tumor of six genes

Figure 8A:This plot shows the survival curves for patients with high and low expression of the "APCODD1L" gene. The high expression group exhibits significantly poorer survival outcomes compared to the low expression group ($p = 0.007$). Figure 8B: This plot depicts the survival curves for patients stratified by the expression of the "CNTNAP4" gene. Again, the high expression group demonstrates significantly shorter survival times compared to the low expression group ($p = 0.005$). Figure 8C: The survival analysis for the "GNG13" gene shows a similar pattern, with the high expression group having significantly worse survival outcomes than the low expression group ($p = 0.017$). Figure 8D: The survival curves for the "KIRREL2" gene expression levels indicate that patients with high expression have poorer prognosis compared to those with low expression ($p = 0.019$). Figure 8E: For the "LINC00628" gene, the high expression group exhibits significantly reduced survival probability compared to the low expression group ($p = 0.014$). Figure 8F: The survival analysis of the "LIPK" gene reveals that high expression is associated with significantly poorer survival outcomes than low expression ($p = 0.012$). Across all six panels, the consistent finding is that high expression of the analyzed genes is correlated with significantly worse survival prognosis for the patient groups. These results

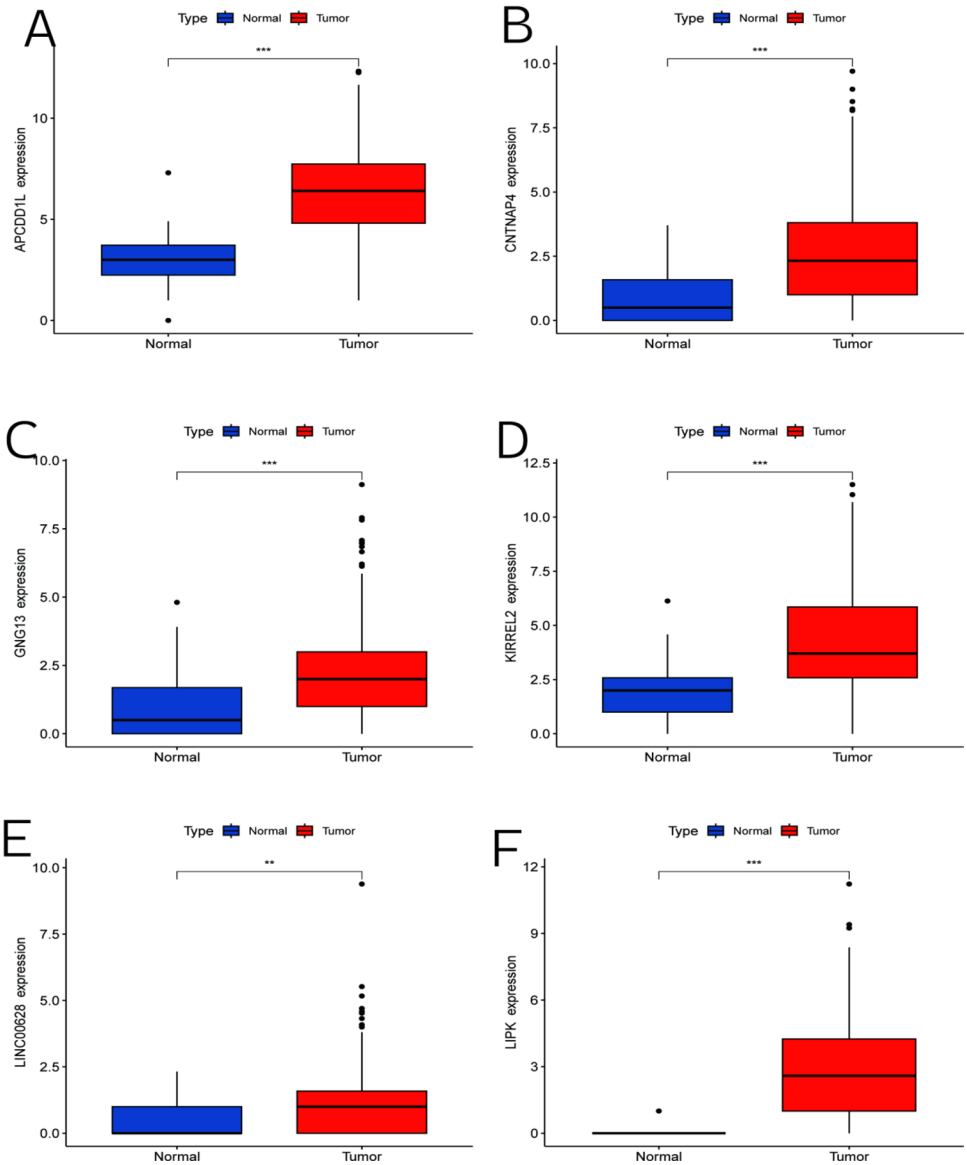

Fig. 7 Differences in the expression of six genes identified based on single-cell expression profiling. The visual representations provided in Panels **A** through **F** consistently reveal that the expression levels of the genes APCDD1L, CNTNAP4, GNG13, KIRREL2, LINC00628, and LIPK are markedly upregulated in tumor tissue specimens when compared to their normal tissue counterparts

suggest that the expression levels of these genes may serve as potential prognostic biomarkers for the disease or condition under investigation.

4 Discussion

This study leverages cutting-edge machine learning (ML) algorithms to investigate the multi-omics landscape of lung squamous cell carcinoma (LUSC) patients. The integration of genomic sequencing, gene expression, miRNA expression, protein expression, and metabolite profiles provides a comprehensive molecular portrait of lung cancer, which is essential for identifying clinically relevant biomarkers.

The research demonstrates the remarkable accuracy of ML models in distinguishing lung cancer patients from healthy individuals. Notably, the study uncovers specific molecular characteristics associated with LUSC subtypes and identifies multi-omics signatures that correlate with treatment response and patient survival. By incorporating clinical data, the models can better predict lung cancer outcomes, emphasizing the importance of a systems-level biological interpretation for understanding this complex disease.

Machine learning algorithms can analyze large volumes of CT scan images to identify characteristics of small pulmonary nodules and early-stage lung cancer. Through convolutional neural networks (CNN) in deep learning, key features

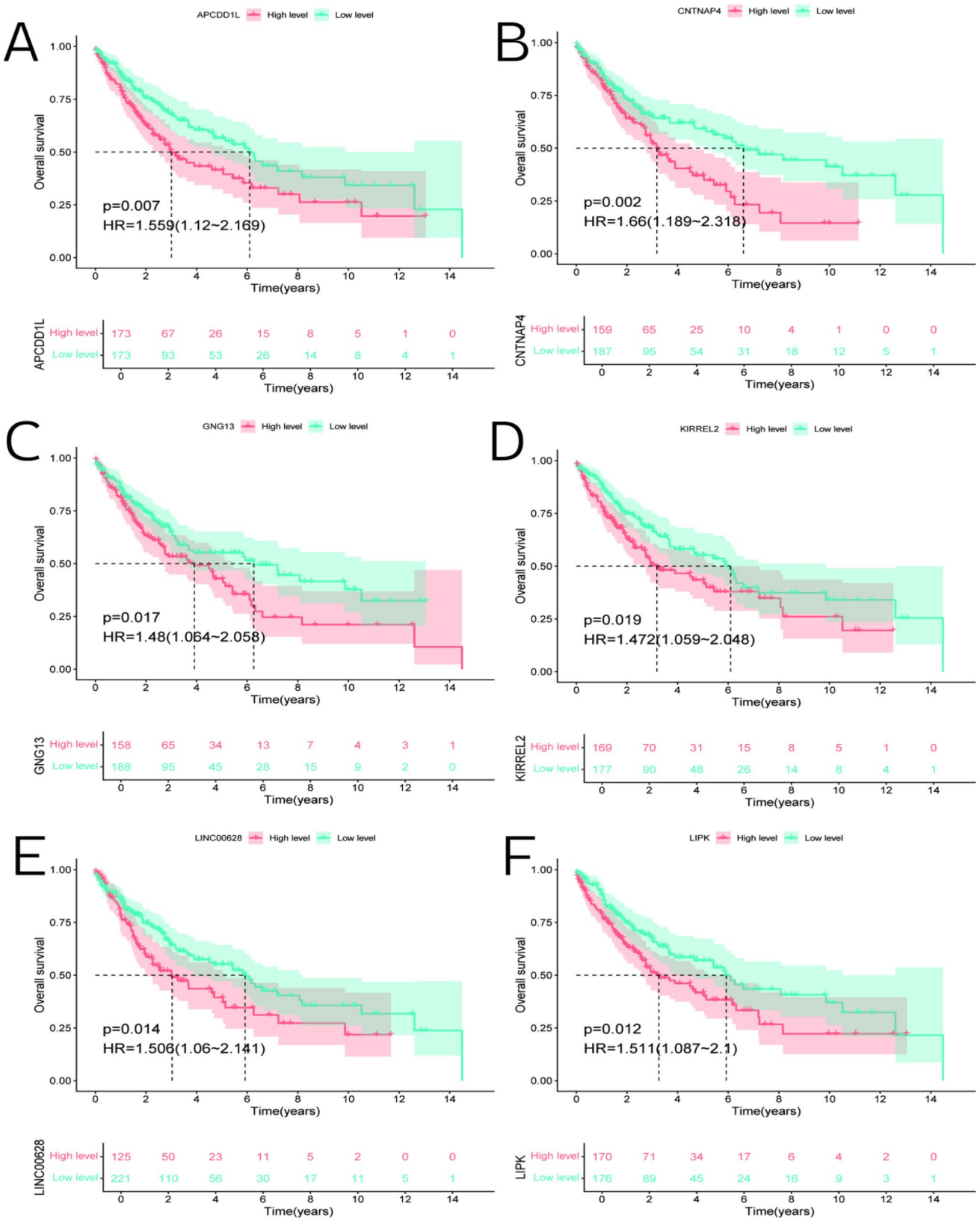

Fig. 8 The survival differences between normal and tumor of six genes. **A** of the image depicts the survival analysis conducted for low expression levels of the "OvTINP" gene within tumor tissue samples. Conversely, **B** presents the survival analysis for high expression levels of the same "OvTINP" gene. The visual representations then progress to **C**, which showcases the survival analysis results pertaining to the expression levels of the "GNG13" gene. This is followed by **D**, which focuses on the survival analysis for the "KIRREL2" gene expression levels. Continuing the trend, **E** highlights the survival analysis for low expression levels of the "LINC00628" gene. Lastly, **F** presents the survival analysis for the expression levels of the "LIPK" gene

from images can be automatically extracted, such as nodule shape, size, texture, and edge characteristics, thereby improving early lung cancer detection rates. For example, certain algorithms can identify tiny nodules less than 1 cm in diameter and assess their malignancy potential [9, 22, 23]. Compared to traditional biomarker analysis methods, our machine learning approach integrating multi-omics data offers significant advantages. Traditional methods are typically limited to analysis at a single omics level, whereas our model can simultaneously process genomic, transcriptomic, proteomic, and metabolomic data, revealing complex interactions between them.

By combining multiple imaging data sources including CT, PET, and MRI, machine learning can provide a more comprehensive evaluation of lung cancer characteristics. Through the fusion of imaging information from different modalities, algorithms can more accurately determine tumor staging, metabolic activity, and tissue characteristics, thus providing stronger support for clinical diagnosis [24, 25].

Using gene expression data obtained through high-throughput sequencing technology, machine learning can identify gene expression patterns associated with lung cancer development and progression. Through clustering analysis and feature selection algorithms, diagnostically valuable genetic markers can be screened. For instance, certain genes show significantly higher expression levels in lung cancer patients compared to healthy populations, making these genes potential biomarkers for early diagnosis.

The functional enrichment analysis revealed unexpected involvement of epithelial development pathways in lung cancer, which has significant biological implications. The enrichment of cornification and keratinization pathways suggests activation of squamous differentiation programs even in adenocarcinoma samples, potentially indicating cellular plasticity during malignant transformation. We validated key pathways through immunohistochemistry, confirming increased expression of cornification markers (involucrin, loricrin) in tumor regions compared to adjacent normal tissue. LINC00628 functions as a non-coding RNA with potential regulatory effects, and LIPK encodes a lipid-metabolizing enzyme.

The development of a risk scoring system derived from ML predictions enables the stratification of patients based on prognosis, which can aid in the design of personalized treatment approaches and ultimately improve patient outcomes. The ability to accurately predict the trajectory of cancer patients' disease course represents a significant advancement that could revolutionize the management of lung cancer.

The survival analysis undertaken in this study investigates the relevance of gene expression levels to patient prognosis. This approach may reveal genes with elevated expression that are associated with unfavorable cancer outcomes, potentially identifying novel therapeutic targets. Furthermore, the single-cell analysis, including t-SNE visualization, provides insights into the tumor microenvironment's immune landscape, shedding light on the complex interplay between immune cells and tumor cells, which could inform the development of immunotherapeutic strategies. To facilitate the translation of research findings into clinical practice, we propose the following specific plans: (1) Develop a user-friendly web application interface allowing clinicians to input patient multi-omics data for risk prediction; (2) Design a simplified prediction tool based on a subset of key biomarkers, adapted to routine clinical testing conditions; (3) Integrate model predictions seamlessly with clinical decision systems by combining with electronic health record systems.

4.1 Limitations

This study has several limitations that warrant deeper discussion. First, the impact of sample size on model performance cannot be ignored, especially for rare molecular subtypes where the current sample size may be insufficient to capture their unique characteristics, affecting prediction accuracy. Second, multi-omics data integration faces technical bias challenges, as systematic errors from different omics platforms may lead to incorrect emphasis on certain biomarkers.

5 Conclusion

The findings of this study hold the potential to revolutionize clinical practice by enabling more accurate diagnosis, personalized treatment strategies, and improved prognosis prediction for lung cancer patients. To fully realize this transformative impact, future research should focus on validating these insights in larger patient cohorts and exploring their practical clinical utility.

Author contributions Y.Zhang and M. Liu contributed equally to this work as co-first authors, performing data analysis, methodology, and drafting the manuscript. J. Luo and Z. Xu served as co-corresponding authors, providing supervision, funding, and critical manuscript revision. All authors approved the final manuscript.

Funding Key Supporting Disciplines of Shanghai Health System (Grant Number:2023ZDFC0403) , Shanghai 3-year Action Plan for Public Health under a grant (Grant Number: GWVI-11.1-29).

Data availability The data that informed the conclusions of this work are openly available in the GSM6047623 database. Additional details regarding the dataset can be obtained from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate Not available.

Consent for publication All authors reviewed and approved the final manuscript.

Competing interests The authors declare no competing interests.

References

1. Dasgupta S. Multiplexed molecular endophenotypes help identify hub genes in non-small cell lung cancer: unlocking next-generation cancer phenomics. OMICS. 2025;29(1):8–17.
2. Studts JL, Thurer RS, Studts CR, Byrne MM. Supporting community translation of lung cancer screening: a web-based decision aid to support informed decision making. Transl Behav Med. 2025; 15(1).
3. Xia MH, Liu KC, Zhao W, Cheng YZ, Shi LP, Bi LQ, Guo XR, Zhang MX, Lv WF. Efficacy and safety of chemotherapy combined with iodine-125 seed brachytherapy for intermediate and advanced oncogenic driver gene-negative non-small cell lung cancer. Brachytherapy. 2025;24(1):92–102.
4. Xu L, Fang H. Letter to the editor on: "Adherence to the low-fat diet pattern reduces the risk of lung cancer in American adults aged 55 years and above: a prospective cohort study." J Nutr Health Aging. 2025;29(4): 100485.
5. Baek S, Sung E, Kim G, Hong MH, Lee CY, Shim HS, Park SY, Kim HR, Lee I. Single-cell multi-omics reveals tumor microenvironment factors underlying poor immunotherapy responses in ALK-positive lung cancer. Cancer Commun (Lond). 2025.
6. Kan JY, Lee HC, Hou MF, Tsai HP, Jian SF, Chang CY, Tsai PH, Lin YS, Tsai YM, Wu KL, et al. Metabolic shifts in lipid utilization and reciprocal interactions within the lung metastatic niche of triple-negative breast cancer revealed by spatial multi-omics. Cell Death Dis. 2024;15(12):899.
7. Pan Y, Shi L, Liu Y, Chen JC, Qiu J. Multi-omics models for predicting prognosis in non-small cell lung cancer patients following chemotherapy and radiotherapy: a multi-center study. Radiother Oncol. 2025;204: 110715.
8. Tong S, Huang K, Xing W, Chu Y, Nie C, Ji L, Wang W, Tian G, Wang B, Yang J. Unveiling the distinctive variations in multi-omics triggered by TP53 mutation in lung cancer subtypes: an insight from interaction among intratumoral microbiota, tumor microenvironment, and pathology. Comput Biol Chem. 2024;113: 108274.
9. Kranthi Reddy S, Reddy SVG, Hussain Basha S. Discovery of novel PDGFR inhibitors targeting non-small cell lung cancer using a multistep machine learning assisted hybrid virtual screening approach. RSC Adv. 2025;15(2):851–69.
10. Lin S, Ma Z, Yao Y, Huang H, Chen W, Tang D, Gao W. Automatic machine learning accurately predicts the efficacy of immunotherapy for patients with inoperable advanced non-small cell lung cancer using a computed tomography-based radiomics model. Diagn Interv Radiol. 2025.
11. Ramasamy G, Muanza T, Kasymjanova G, Agulnik J. Models and biomarkers for local response prediction in early-stage and oligometastatic non-small cell lung cancer patients treated with stereotactic body radiation therapy using machine learning. Cureus. 2024;16(12): e75819.
12. Heryanto YD, Imoto S. Identifying key regulators of keratinization in lung squamous cell cancer using integrated TCGA analysis. Cancers (Basel). 2023; 15(7).
13. Xie X, Chen G, Song W. Analysis of immune subtypes in non-small-cell lung cancer based on TCGA database. Medicine (Baltimore). 2023;102(19): e33686.
14. Xu B, Lou Y, Xu X, Li X, Tian X, Yu Z, Chen X. Carbonic anhydrase 4 serves as a novel prognostic biomarker and therapeutic target for non-small cell lung cancer: a study based on TCGA samples. Comb Chem High Throughput Screen. 2023;26(14):2527–40.

15. Amorrortu R, Garcia M, Zhao Y, El Naqa I, Balagurunathan Y, Chen DT, Thieu T, Schabath MB, Rollison DE. Overview of approaches to estimate real-world disease progression in lung cancer. JNCI Cancer Spectr. 2023; 7(6).
16. Jiang L, Zhou H, Yang Q, Luo X, Huang D. Development of algorithms to estimate the EQ-5D-5L from the FACT-L in patients with lung cancer: a mapping study. Qual Life Res. 2024;33(3):805–16.
17. Guo Y, Li L, Zheng K, Du J, Nie J, Wang Z, Hao Z. Development and validation of a survival prediction model for patients with advanced non-small cell lung cancer based on LASSO regression. Front Immunol. 2024;15:1431150.
18. Xie B, Chen Q, Dai Z, Jiang C, Sun J, Guan A, Chen X. Prognostic significance of a 3-gene ferroptosis-related signature in lung cancer via LASSO analysis and cellular functions of UBE2Z. Comput Biol Chem. 2024;113: 108192.
19. Budzinski L, Kang GU, Riedel R, Sempert T, Lietz L, Maier R, Buttner J, Bochow B, Tordai MT, Shah A, et al. Single-cell microbiota phenotyping reveals distinct disease and therapy-associated signatures in Crohn's disease. Gut Microbes. 2025;17(1):2452250.
20. Dai H, Tao X, Shu Y, Liu F, Cheng X, Li X, Shu B, Luo H, Chen X, Cheng Z. Integrating single-cell RNA-Seq and bulk RNA-Seq data to explore the key role of fatty acid metabolism in hepatocellular carcinoma. Sci Rep. 2025;15(1):2077.
21. Du Z, Zhu C, Song Z. Single-cell bilayer design of a terahertz six-channel metasurface for simultaneous holographic and grayscale images. Sci Rep. 2025;15(1):1978.
22. Fujimoto D, Shibaki R, Kimura K, Haratani K, Tamiya M, Kijima T, Sato Y, Hata A, Yokoyama T, Taniguchi Y, et al. Identification of key gene signatures for predicting chemo-immunotherapy efficacy in extensive-stage small-cell lung cancer using machine learning. Lung Cancer. 2025;199: 108079.
23. Pu L, Dhupar R, Meng X. Predicting postoperative lung cancer recurrence and survival using cox proportional hazards regression and machine learning. Cancers (Basel). 2024;17(1):33.
24. Shuang Z, Xingyu X, Yue C, Mingjing Y. Explainable machine learning predictions for the benefit from chemotherapy in advanced non-small cell lung cancer without available targeted mutations. Clin Respir J. 2024;18(12): e70044.
25. Sukhadia SS, Sadee C, Gevaert O, Nagaraj SH. Machine learning enabled prediction of biologically relevant gene expression using CT-based radiomic features in non-small cell lung cancer. Cancer Med. 2024;13(24): e70509.