Proceedings of Machine Learning Research vol 278:1–13, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

Explainable Deep Neural Network for Lung Squamous Cell Carcinoma Survival Analysis by Integrating Genomic and Clinical Data

Xudan Zhou 416422837@QQ.COM
School of Intelligent Medicine and Biotechnology, Guilin Medical University, Guilin, China

Qinglin Yang[*] YANGQLIN12@QQ.COM
Department of math and art, Guangxi international business vocational college, Nianning, China

Yuxin Zhang 1723287165@QQ.COM
School of Intelligent Medicine and Biotechnology, Guilin Medical University, Guilin, China

Yanyan Hou 1946171377@QQ.COM
School of Intelligent Medicine and Biotechnology, Guilin Medical University, Guilin, China

Changlong Chen 3031476838@QQ.COM
School of Intelligent Medicine and Biotechnology, Guilin Medical University, Guilin, China

Guohui Ma 3401296153@QQ.COM
School of Intelligent Medicine and Biotechnology, Guilin Medical University, Guilin, China

Jin Luo 1544538603@QQ.COM
School of Intelligent Medicine and Biotechnology, Guilin Medical University, Guilin, China

Wei Shu SHUWEI7866@126.COM
School of Intelligent Medicine and Biotechnology, Guilin Medical University, Guilin, China
[*]*Corresponding author*

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

we utilized explainable deep learning methodologies to elucidate critical genes and prospective biomarkers correlated with the prognosis of Lung Squamous Cell Carcinoma (LUSC). Transcriptomic data were systematically acquired from the TCGA repository and underwent comprehensive differential expression profiling to identify candidate genes warranting in-depth exploration. We developed Cox-PASNet, a pathway-aware deep learning model designed to predict survival outcomes in lung squamous cell carcinoma (LUSC) by integrating multi-modal data, including clinical variables, transcriptomic profiles, and curated biological pathways. The model demonstrated robust performance, achieving an AUC of 0.73 in stratifying patients into long- and short-term survival groups. Beyond predictive accuracy, Cox-PASNet offers interpretable insights into key molecular pathways, facilitating the discovery of novel prognostic biomarkers (CCDC181, B2M, BTD, C1orf112, ANAPC7) and their related biological pathways (regulation of cell cycle, DNA repair, cytoskeletal dynamics, tumor microenvironment, and metastasis) associated with LUSC survival. The significance of these genes was validated using external datasets and clinical indicators. Notably, members of the CCDC family were particularly important, with many found to enhance tumor cell proliferation. Elevated expression levels of CCDC proteins demonstrated a significant correlation with adverse clinical outcomes, including diminished overall survival rates and unfavorable prognosis. In summary, through interpretable deep learning and bioinformatics approaches, we identified several relevant genes, with CCDC genes being closely linked to LUSC survival.

Keywords: Lung Squamous Cell Carcinoma (LUSC), Deep Learning, Biomarkers, CCDC Gene Family

1. Introduction

Lung cancer maintains its status as the most lethal malignancy worldwide, with 1.8 million incident cases reported yearly. Notably, its annual mortality rate surpasses the combined death toll of colorectal, breast, prostate, and pancreatic malignancies (Li et al., 2018). Among its histological subtypes, Lung squamous cell carcinoma (LUSC), a predominant subtype of non-small cell lung cancer (NSCLC), accounts for approximately 40% of all lung cancer diagnoses. Compared to Lung Adenocarcinoma (LUAD), LUSC exhibits poorer responses to targeted therapies and still lacks effective molecular targets for such treatments. The pathogenesis of LUSC involves multiple biological pathways and molecular processes. Clinically, LUSC often manifests as chest pain due to tumor invasion of the chest wall or pleura, as well as dyspnea caused by airway obstruction or pleural effusion. Current therapeutic interventions, including chemotherapy, immunotherapy, and targeted therapies, primarily aim to alleviate these symptomatic burdens. However, the fundamental biomarkers and precise molecular targets underlying LUSC development and progression remain unclear. A deeper understanding of the molecular mechanisms driving the discovery of LUSC-associated survival biomarkers is urgently needed, as it could revolutionize early diagnosis, therapeutic strategies, prognostic evaluation, and breakthrough treatments.

Effective processing of medical data and extraction of informative features are crucial for gaining deeper insights into diseases and developing effective treatments. The development of advanced computational learning methods has been leveraged to analyze disease-related data, ushering in new avenues for diagnosis and treatment. LUSC presents significant molecular complexity that exceeds the analytical capabilities of conventional statistical methods, particularly in capturing polygenic risk factors. This limitation underscores the critical need for developing next-generation analytical tools. Recent advances in computational power have enabled the implementation of deep learning-based artificial intelligence models, which show promising potential for both tumor risk prediction and elucidation of multifactorial tumorigenesis. Consequently, the application of explainable deep neural network models to identify and validate potential biomarkers carries substantial clinical relevance (Bade and Dela Cruz, 2020). In the present study, we implemented an interpretable deep learning framework to systematically elucidate genomic determinants associated with survival outcomes in LUSC.

2. Materials and Method

Our computational workflow comprises three key stages (Figure 1):

Preprocessing phase: Systematic screening of differentially expressed genes (DEGs) through differential expression analysis.

Model construction phase: Implementation of an interpretable model, Cox-PASNet (Hao et al., 2019), to build a complex yet sparse deep learning network that integrates genes, pathways, and clinical data relevant to lung squamous cell carcinoma (LUSC) survival.

Validation phase: Application of machine learning techniques to identify hub genes, followed by validation of their prognostic utility using an external dataset and clinical biomarkers (epidermal growth factor receptor, EGFR).

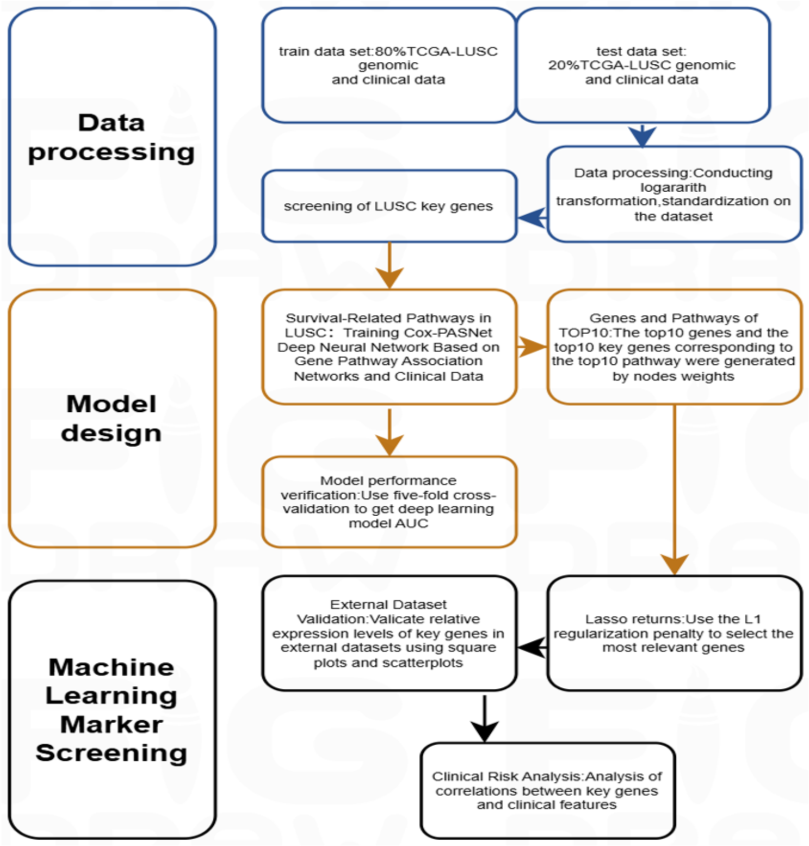

Data processing
train data set:80%TCGA-LUSC genomic and clinical data
test data set: 20%TCGA-LUSC genomic and clinical data
Data processing:Conducting logararith transformation,standardization on the dataset
screening of LUSC key genes
Model design
Survival-Related Pathways in LUSC： Training Cox-PASNet Deep Neural Network Based on Gene Pathway Association Networks and Clinical Data
Genes and Pathways of TOP10:The top10 genes and the top10 key genes corresponding to the top10 pathway were generated by nodes weights
Model performance verification:Use five-fold cross-validation to get deep learning model AUC
Machine Learning Marker Screening
External Dataset Validation:Valicate relative expression levels of key genes in external datasets using square plots and scatterplots
Lasso returns:Use the L1 regularization penalty to select the most relevant genes
Clinical Risk Analysis:Analysis of correlations between key genes and clinical features

Figure 1: Overview of the workflow

2.1. Dataset Download and Preprocessing

Transcriptomic data were obtained from the TCGA database (https://portal.gdc.cancer.gov/) using "LUSC" as the search criterion. The R package sva was implemented for batch effect correction and dataset integration, with detailed sample characteristics presented in Table 1. Eliminating variations between different datasets for subsequent analysis. External datasets GSE30219 and GSE74777 were used for further validation.

Table 1: Dataset Description

Dataset ID	Sequencing Platform	Alive	Dead	Total
TCGA-LUSC	Illumina	291	159	450
GSE30219	GPL570	93	200	293
GSE74777	GPL17586	65	42	107

2.2. Screening of Disease-Related Differentially Expressed Genes in LUSC

Gene expression profiling was performed using Limma (R package), revealing statistically significant differentially expressed genes (DEGs) in LUSC versus normal controls ($P < 0.05$; $|\text{log2FC}| >$

0.25). Volcano plots and heatmaps were generated using the R packages ggplot2 and pheatmap, respectively. The identified DEGs were then input into the Reactome database (https://reactome.org/) to retrieve all pathways associated with these genes. Pathways with fewer than 10 participating genes were excluded, resulting in the final set of Survival-associated pathways in LUSC.

Construction of an Interpretable Pathway-Related Deep Learning Neural Network Model The training protocol comprised: (1) Adam optimizer with learning rates (0.03,0.01,0.001,0.00075) determined via 100-epoch grid search; (2) 500-epoch training for parameter optimization; (3) multi-level regularization (L2 penalty, [λ=0.1,0.01,0.05,0.001]; standard dropout, [p=0.5,0.7]; novel pathway dropout). Input features were z-score normalized, with weights initialized via He-normal distribution.

The framework was implemented using PyTorch 2.5.1 (CUDA 12.0 acceleration) and Keras 3.3, with fixed random seeds ensuring reproducibility. Model selection employed AUC optimization on a held-out validation set (20%), followed by final evaluation on an independent test cohort.

Based on the Cox-PASNet model, a deep learning network model closely related to LUSC survival was constructed. The model consists of five distinct layers: the input layer (gene layer), hidden layer, second hidden layer and clinical layer(H2), output layer. Following data preprocessing, the filtered DEGs dataset served as the training set. Pathways for model construction (input and pathway layers) were sourced from the Reactome database. A mask table was generated based on validated gene-pathway interactions to link these layers. The model employed the Adam optimizer with dropout and L2 regularization to enhance performance and mitigate overfitting.

2.3. Discovery and Validation of LUSC Survival-Related Genes

We constructed a sophisticated deep learning network model, Cox-PASNet, which is intricately intertwined with LUSC-related pathways. The processed training dataset consisted of 7,058 genes, with 860 pathway-related genes from Reactome being allocated to the input and pathway layers respectively. A mask table was constructed using established gene-pathway relationships to intelligently constrain the associations between genes and pathways.

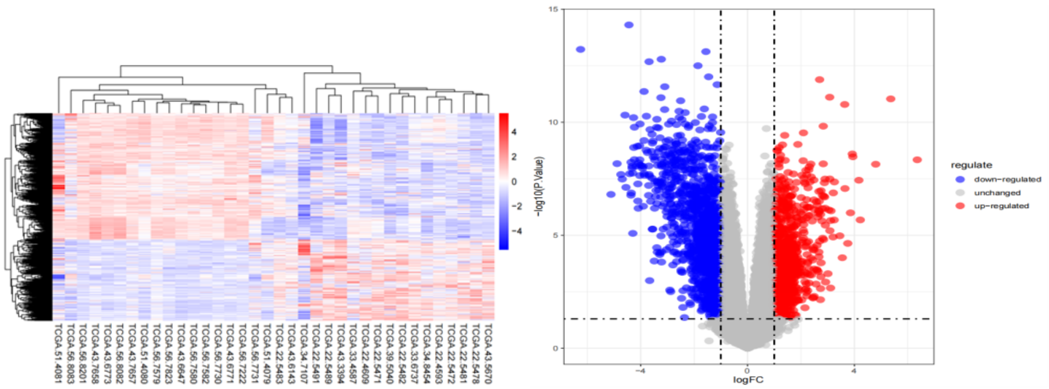
3. Result and Discussion

3.1. Identification and Selection of LUSC-Associated Differentially Expressed Genes for Training Set Construction

The TCGA-LUSC dataset was examined for expression profiling, comprising 501 samples, including 450 LUSC samples and 51 normal control samples. Differential gene expression analysis, as illustrated by the volcano plot and heatmap (Figure 2), highlighted genes exhibiting significant differences between the groups. A cohort of 450 patients was analyzed, encompassing 7,058 differentially expressed genes (DEGs) alongside comprehensive clinical data, including age, survival status, and survival time. These data were utilized as the training and testing samples for the model.

3.2. Explainable AI modeling of LUSC through multimodal integration of ranscriptomic profiles, pathogenic pathways and clinical parameters

To investigate molecular pathways associated with LUSC survival, we developed a neural network model based on Cox-PASNet for LUSC survival pathways. The model integrates both transcriptomic profiles and clinical parameters as input features, which are subsequently linked to a layer



(*A*) (*B*)

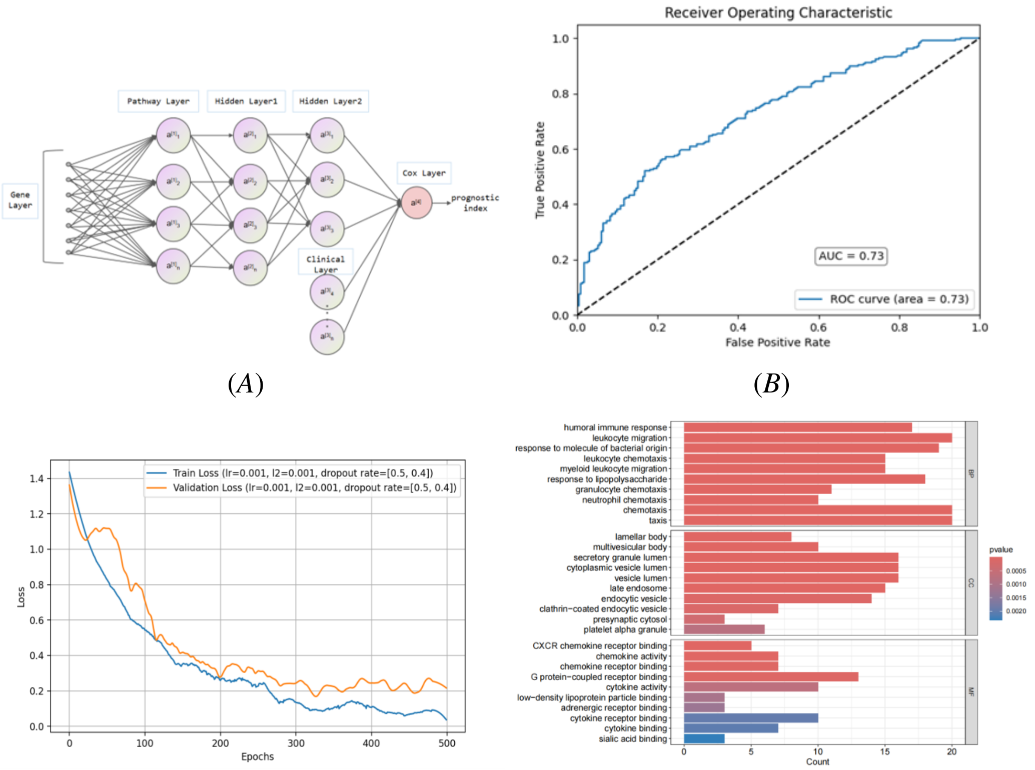
Figure 2: DEG analysis in LUSC. (A-B) Heatmap and volcano plot representation of transcriptomic differences between LUSC ($n = 450$) and normal samples ($n = 51$)

of biological pathways. The hidden layers are designed to extract high-level abstract features from pathway-level representations, thereby achieving dual optimization of nonlinear representation power and model interpretability (Figure 3A).

For model optimization, we set the learning rate to 8E-05 and L2 regularization coefficient to 1E-06 to enhance predictive performance, and the number of epochs to 500. Using a five-fold cross-validation method with Survival-related data in LUSC and control samples, our findings demonstrated robust performance. Model optimization was tracked via loss function evaluation during both training and validation phases. As illustrated in Figure 3C, the training loss exhibited a progressive decrease, whereas the validation loss stabilized after 400 training epochs. The model attained an AUC of 0.73 on the test set (Figure 3B),the Cox-PASNet model demonstrated exceptional capability in handling gene expression data, particularly in capturing its sparsity and extracting hidden signals within this complex domain. Furthermore, the Cox-PASNet method adapts to increases in sample size and dimensionality, facilitating broader research applications.

Addressing the interpretability limitations of conventional deep learning approaches, we implemented a biologically constrained framework that explicitly incorporates curated gene-pathway associations from canonical databases.Consequently,this analysis enabled the identification of key genes and pathways significantly associated with the survival outcomes in LUSC. The analysis of each node's relevance involved a detailed examination of the trained model's weights, leading to the identification of major pathways and their associated genes. These pathways include processes such as cell cycle regulation, DNA repair,cytoskeletal dynamics (Reber et al., 2024), tumor microenvironment (Xiao and Yu, 2021) and metastasis (Steeg, 2016). Understanding the abnormal mechanisms of these pathways aids in the development of new therapeutic strategies and improves the prognosis of LUSC patients (Figure 3D).

Visualization of Node Activation Patterns To enhance Cox-PASNet's biological interpretability, we retrained the model on LUSC samples across 500 hyperparameter-optimized experiments. Using the median prognostic index (PI) from model outputs, we stratified samples into high and low-risk groups. Figures 4 and 5 visualize node activation patterns in integrated layers (H2 layers) and pathway layers across risk groups.



(*A*) (*B*)

(*C*) (*D*)

Figure 3: Model architecture and performance evaluation. (A) Schematic of Cox-PASNet framework, comprising four core components: gene input layer, pathway layer, hidden layers and clinical layer (H2 layer), and Cox output layer for LUSC survival prediction. Gene and clinical feature significance was determined through weight-based validation. (B) Performance metrics demonstrating Cox-PASNet's optimal predictive capability (AUC = 0.73). (C) Loss function analysis reveals training convergence with potential underfitting, as evidenced by the validation loss. (D) GO enrichment analysis of genes associated with the top 10 identified pathways.

The median prognostic index (PI)-stratified subgroups showed significant covariate concordance in Figure 4A, evaluated through log-rank tests (-log(p-value)). Red triangles and blue markers denote significant (-log(p-value) > 1.3) and non-significant covariates, respectively. Top-weighted covariates strongly correlated with survival prediction. Figure 4(B-E) display Kaplan-Meier curves for the top four covariates, demonstrating significant intergroup survival differences. These findings establish the top-ranking covariates as potential prognostic biomarkers.

Pathway layer node values were visualized similarly (Figure 5). Figure 5A presents a heatmap of the top 100 pathway node activations, ranked by mean absolute derivatives, across high and low-risk patient groups. Among 967 pathway nodes, 601 demonstrated statistical significance in survival analysis through log-rank tests. The top four pathways were further investigated using Kaplan-Meier analysis (Figure 5(B-E)), revealing their potential as prognostic biomarkers.

t-SNE visualization in Figure 6 demonstrates the nonlinear distribution of Four top-ranked H2 layer and pathway layer nodes associated with PI. Notably, the hierarchical combination of pathway layer and nonlinear clinical layer exhibited stronger survival associations compared to the pathway layer alone.

- Activation patterns of nodes in the integrated second hidden layer and clinical feature layer (H2).
- Visualization of pathway layer node activation patterns
- t-SNE Visualization of Top Four Cox-PASNet Nodes

3.3. Identification of Biomarkers Associated with LUSC Survival

To identify potential biomarkers for evaluating survival-related genes in Lung Squamous Cell Carcinoma (LUSC), we focused on the top ten genes with the highest weights from core pathways and their associated genes. Multimodal integration revealed 102 key driver genes, for which the model constructed a biologically constrained expression matrix encoding essential pathway relationships.

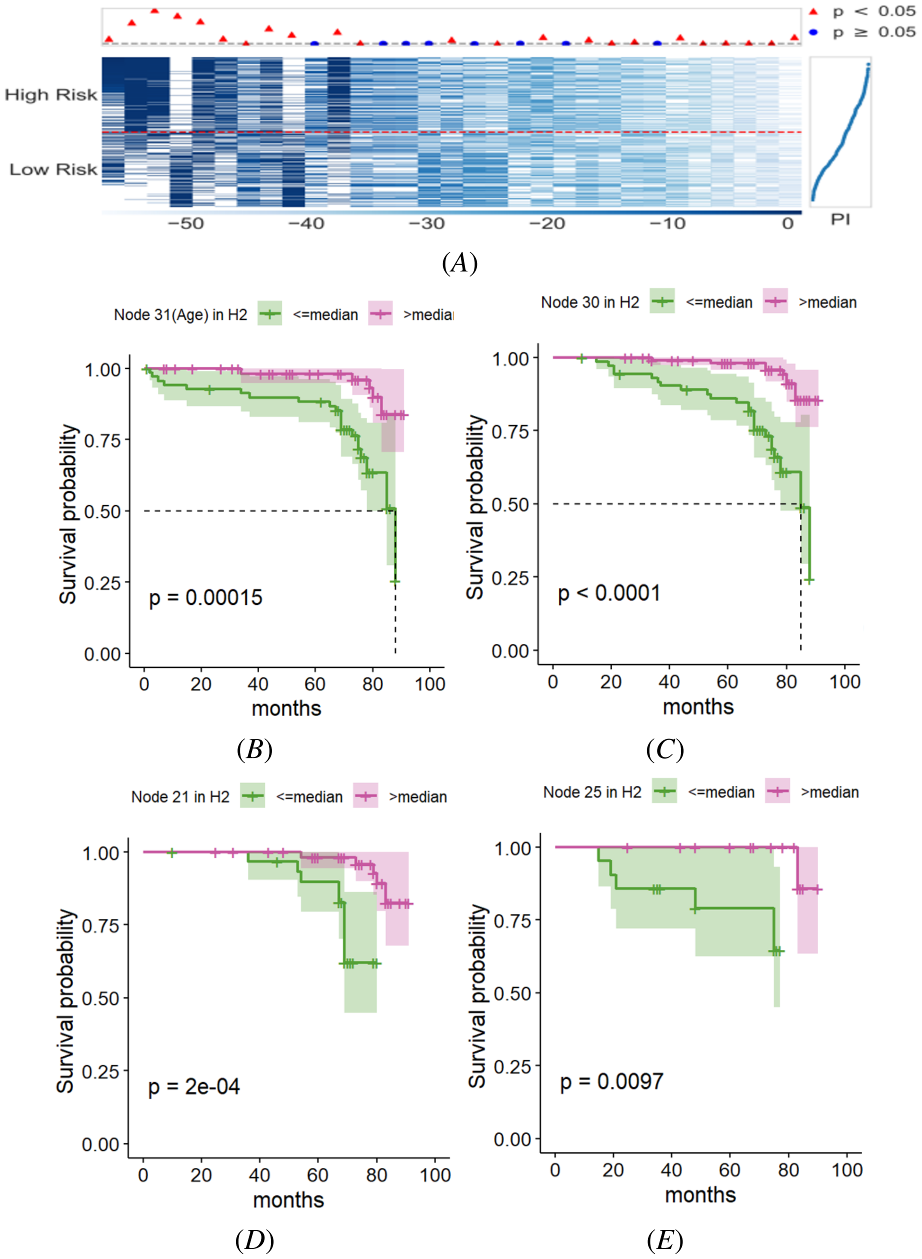
LASSO regression with L1 regularization was employed to identify the most informative genes, effectively eliminating noise variables by setting their coefficients to zero while preserving biologically relevant features. The optimal lambda value was meticulously determined as lambda.1se (Figure 7A and 7B), leading to the identification of genes with non-zero regression coefficients associated with LUSC survival. Validation in the training set confirmed significant differential expression for most candidate genes (Figure 7C), with particularly strong effects observed for CCDC181, B2M, BTD, C1orf112, and ANAPC7 (all $p < 0.01$).

Subsequent correlation analysis of EGFR levels in the GSE74777 and GSE30219 datasets (as shown in Figures 8(B-F) highlighted significant survival correlations. Among these candidates, CCDC181, a member of the CCDC protein family, encodes a coiled-coil domain-containing protein that has been demonstrated to promote tumor cell proliferation, facilitate cancer cell invasion, and drive LUSC progression (Liu et al., 2023). B2M is involved in antigen presentation within the immune system, and its low expression may lead to immune escape, thereby affecting patient response to immunotherapy and survival. BTD expression levels may be linked to metabolic reprogramming in tumor cells, with metabolic abnormalities potentially promoting tumor growth and drug resistance, thus impacting patient survival. ANAPC7 expression levels may be associated with cell cycle regulation and genomic stability in LUSC, and its aberrant expression could lead to uncontrolled tumor proliferation and drug resistance, ultimately affecting patient survival.

Multi-cohort validation (GSE30219/GSE74777) established the generalizability of these findings across diverse populations (Figure 8B-E). The conserved prognostic value of these molecular features suggests immediate translational potential as: (i) stratification biomarkers for clinical trials, and (ii) targets for personalized therapeutic approaches in LUSC.

4. Conclusion

CoxPASNet is a deep learning framework that models the nonlinear hierarchy of biological pathways to identify key survival prognostic factors, achieving a 0.73 AUC in LUSC survival prediction. Furthermore, through pathway-level interpretation of the model's outputs, we elucidated the correlations between risk-associated genes and survival-relevant pathways in LUSC (Table 2). Although neural network models can be applied to polygenic risk analysis in LUSC, there remains room for



(A)

(B)

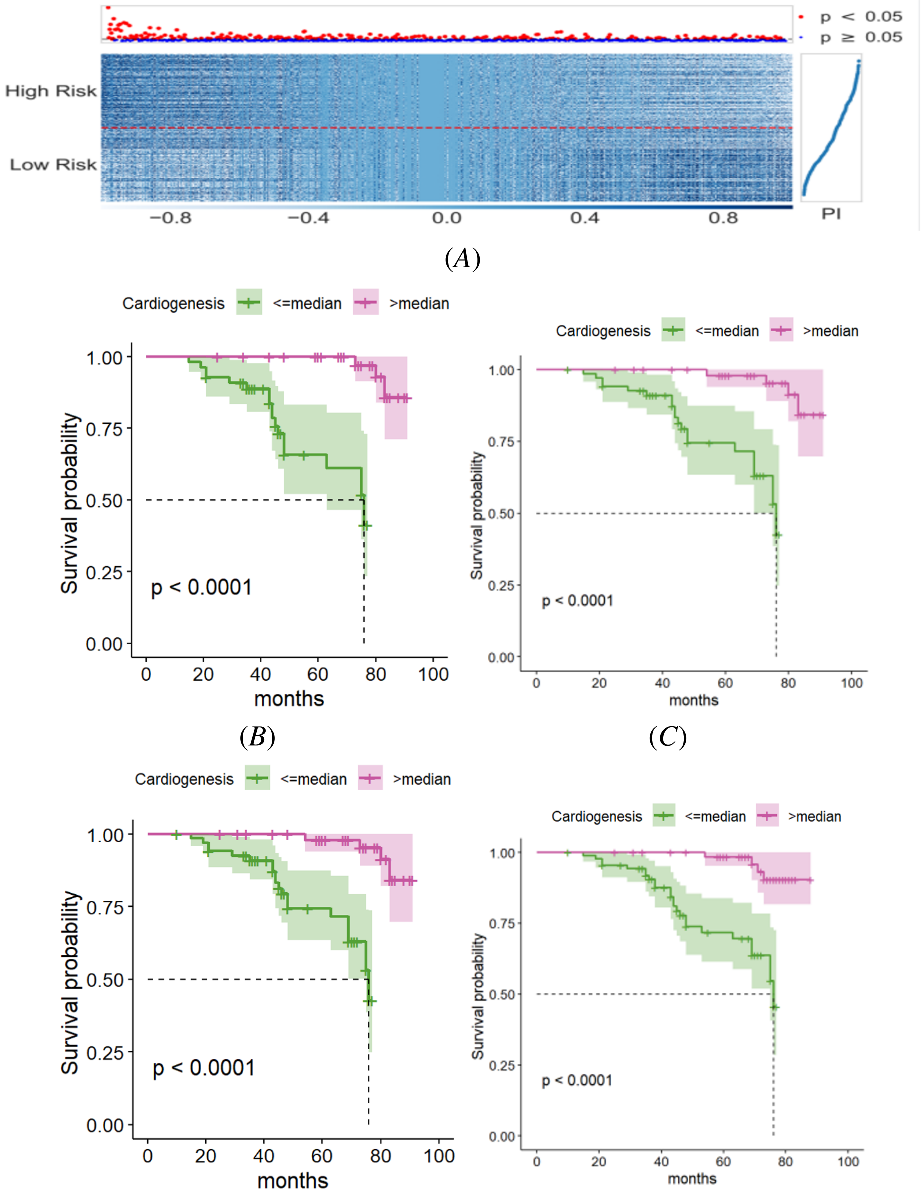
(C)

(D)

(E)

Figure 4: Node value visualization in the second hidden and clinical layers (H2). (A) Heatmap of 31 nodes, stratified by risk groups (red dashed line: upper/lower = high/low-risk patients). Top dot plot shows node importance. Log-rank test results (-log(p-value)) are color-coded (red: significant, blue: non-significant). Right panel displays prognostic indices (PI). (B-E) Kaplan-Meier curves for top four nodes.

performance improvement. First, incorporating additional clinical samples into the classification model would enable more precise capture of disease genetic characteristics. Furthermore, transcriptomic data and clinical parameters obtained from large LUSC patient cohorts allow deeper under-



(*A*)

(*B*)

(*C*)

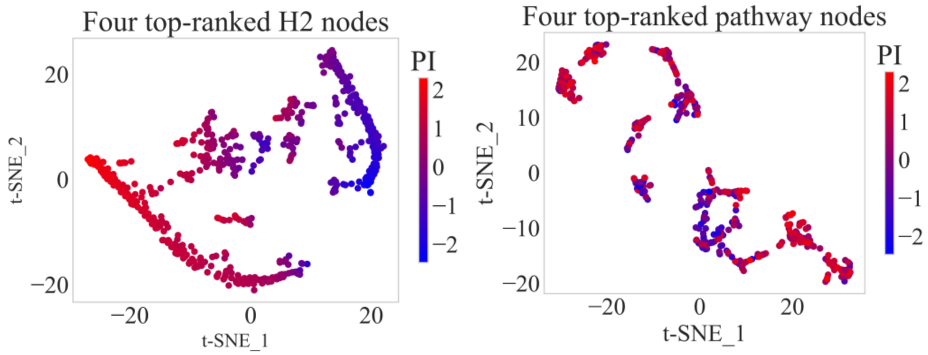
(*D*)

(*E*)

Figure 5: Pathway layer node activation patterns. (A) Heatmap of top 100 nodes, stratified by risk (red dashed line: upper/lower = high/low-risk). Dot plot shows node importance. Log-rank test results (-log(p-value)) are color-coded (red: significant, blue: non-significant). Right panel displays prognostic indices (PI). (B-E) Kaplan-Meier curves for top four pathway nodes.

standing of disease-associated genes, as gene expression data from different batches and platforms may impact model accuracy.

“In conclusion, we have identified and validated five LUSC-associated genes: Coiled-Coil Domain Containing 181 (CCDC181), β-2-microglobulin (B2M), biotinidase (BTD), Chromosome 1



(*A*)

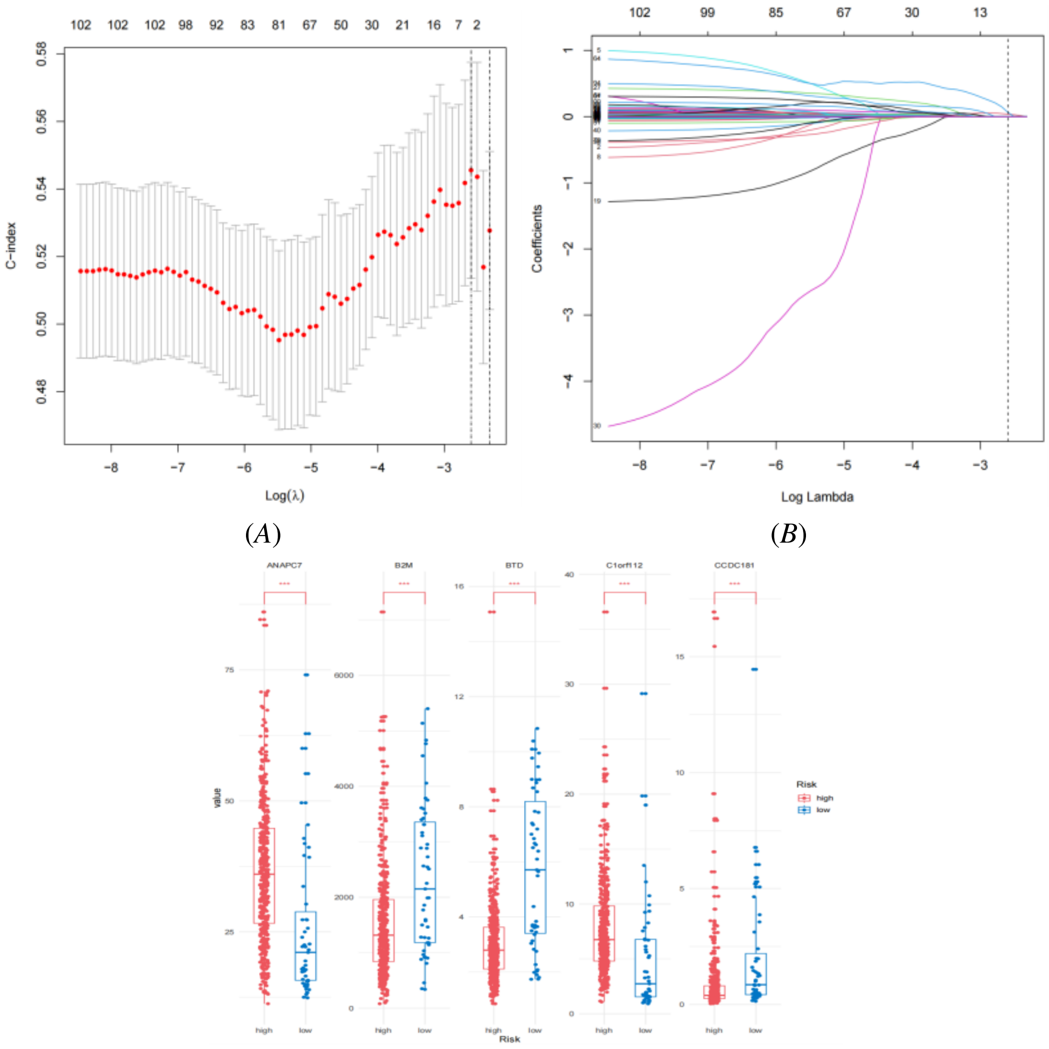
(*B*)

Figure 6: Top-ranked node visualization in Cox-PASNet. (A) t-SNE plot of top four significant pathway nodes. (B) t-SNE plot of top four significant H2 layer nodes.

Table 2: Clinically relevant pathway signatures predictive of LUSC survival outcomes.

Top Pathway	Genes
Transcription of E2F targets (K, 2022)	ABHD3; AFF3; BCL7A; CA9; ANKRD34B
Microtubule-dependent trafficking (N and GC, 2017)	ARHGEF5; C4orf17; ARG2; ADAM23; BOLA2B
Cardiogenesis (Reber et al., 2024)	BTD; C5orf46; B2M; AMY2B; CBLN1
Developmental Cell Lineages (Xiao and Yu, 2021)	BTD; C5orf46; CBLN1; ARTN; ABCA6
Collagen chain trimerization (Steeg, 2016)	ANAPC7; C1orf112; ARHGEF17; CCDC178; AC011498.4
Cooperation of Prefoldin and TriC/CCT (Kim et al., 2024)	BTD; CBLN1; C5orf46; CCDC77; ARTN
Defective homologous recombination repair (HRR) (Liu et al., 2023)	C1orf112; ANAPC7; CARD18; AC011498.4; CCDC178
Polo-like kinase mediated events (Zhang et al., 2023)	BTD; CBLN1; C5orf46; AMY2B; ABCA6
Laminin interactions (Durbeej, 2010)	C1orf112; ANAPC7; CARD18; AC011498.4; CCDC178
ECM Macromolecular Assembly: Collagen & Multimeric Structures (Onursal et al., 2021)	BTD; C5orf46; CBLN1; AATK; AMY2B

Open Reading Frame 112 (C1orf112), and Anaphase Promoting Complex Subunit 7 (ANAPC7). These genes may participate in the survival processes of LUSC patients and provide a foundation for understanding potential novel biomarkers for monitoring LUSC prognosis.



(A)

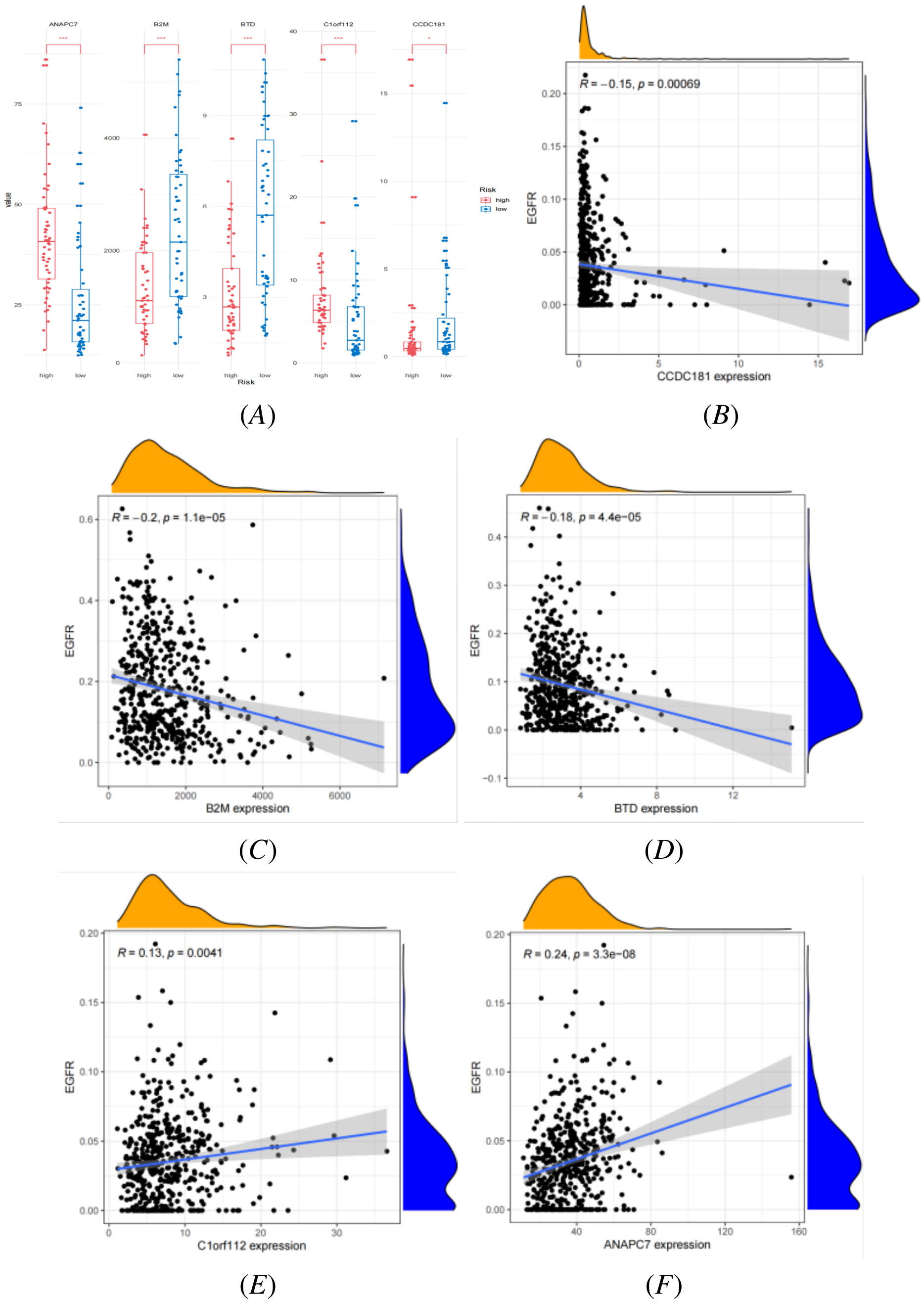
(B)

(C)

Figure 7: LASSO Regression Analysis and Gene Expression Validation. (A) Through 5-fold cross-validation, we determined the optimal regularization parameter (λ) that minimized the mean squared error (MSE). (B) The LASSO coefficient shrinkage profile demonstrates the feature selection process. As λ increased, non-informative variables were progressively eliminated through L1 regularization, with their coefficients compressed to zero. (C) Differential expression analysis of five candidate genes in the TCGA-LUSC cohort. All genes showed statistically significant expression differences (***$p < 0.001$, two-sided t-test) between high- and low-risk patient groups, confirming their prognostic relevance.

Acknowledgments

Financial support was provided by: Two NSFC grants (32060157, 82360098), Central Guiding Local Development Fund (ZY20230103) and GMU's Graduate Research Award (GYYK2022020).



(*A*) (*B*)

(*C*) (*D*)

(*E*) (*F*)

Figure 8: Multi-Cohort Verification of Clinical Correlations in External Patient Populations in GSE30219 and GSE74777. (A) Experimental validation of identified key genes was performed using independent datasets GSE30219 and GSE74777,where Similar expression was observed (***$p < 0.001$). (B-F) Correlation of Core Genes(CCDC181, B2M, BTD, C1orf112, ANAPC7) with Clinical Features(EGFR).

References

Brett C. Bade and Charles S. Dela Cruz. Lung cancer 2020: Epidemiology, etiology, and prevention. *Clin Chest Med*, 41(1):1–24, Mar 2020. doi: 10.1016/j.ccm.2019.10.001.

M Durbeej. Laminins. *Cell Tissue Res*, 339(1):259–68, Jan 2010. doi: 10.1007/s00441-009-0838-2.

J. Hao, Y. Kim, T. Mallavarapu, J. H. Oh, and M. Kang. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Med Genomics*, 12(Suppl 10):189, Dec 23 2019. doi: 10.1186/s12920-019-0624-2.

Engeland K. Cell cycle regulation: p53-p21-rb signaling. *Cell Death Differ*, 29(5):946–960, May 2022. doi: 10.1038/s41418-022-00988-z. Epub 2022 Mar 31.

Hyunmin Kim, Junsun Park, and Soung-Hun Roh. The structural basis of eukaryotic chaperonin tric/cct: Action and folding. *Mol Cells*, 47(3):100012, Mar 2024. doi: 10.1016/j.mocell.2024.100012.

Yin Li, Jie Gu, Fengkai Xu, Qiaoliang Zhu, Di Ge, and Chunlai Lu. Transcriptomic and functional network features of lung squamous cell carcinoma through integrative analysis of geo and tcga data. *Sci Rep*, 8(1):15834, Oct 26 2018. doi: 10.1038/s41598-018-34160-w.

Zhen Liu, Weiwei Yan, Shaohua Liu, Zhan Liu, Ping Xu, and Weiyi Fang. Regulatory network and targeted interventions for ccdc family in tumor pathogenesis. *Cancer Lett*, 565:216225, Jul 1 2023. doi: 10.1016/j.canlet.2023.216225,.

Chatterjee N and Walker GC. Mechanisms of dna damage, repair, and mutagenesis. *Environ Mol Mutagen*, 58(5):235–263, Jun 2017. doi: 10.1002/em.22087. Epub 2017 May 9.

C. Onursal, E. Dick, I. Angelidis, H. B. Schiller, and C. A. Staab-Weijnitz. Collagen biosynthesis, processing, and maturation in lung ageing. *Front Med (Lausanne)*, 8:593874, May 20 2021. doi: 10.3389/fmed.2021.593874.

S. Reber, M. Singer, and F. Frischknecht. Cytoskeletal dynamics in parasites. *Curr Opin Cell Biol*, 86:102277, Feb 2024. doi: 10.1016/j.ceb.2023.102277.

Patricia S. Steeg. Targeting metastasis. *Nat Rev Cancer*, 16(4):201–18, Apr 2016. doi: 10.1038/nrc.2016.25.

Yi Xiao and Dihua Yu. Tumor microenvironment as a therapeutic target in cancer. *Pharmacol Ther*, 221:107753, May 2021. doi: 10.1016/j.pharmthera.2020.107753.

Pengcheng Zhang, Xinglong Zhang, Yongfu Zhu, Yiyi Cui, Jing Xu, and Weiping Zhang. Polo-like kinase 1 suppresses lung adenocarcinoma immunity through necroptosis. *Oncol Res*, 31(6): 937–953, Sep 15 2023. doi: 10.32604/or.2023.030933.