

# Integrated WGCNA and Machine Learning Uncover ECM/Cell Cycle Hub Genes as Prognostic Biomarkers in Lung Adenocarcinoma for Improved Health Care

Moksha Prada Prabakar<sup>1</sup>, Narmadha Ramasmay<sup>1</sup>, Rizwana Rohoofur Rahman<sup>1</sup>, Kannan Muthu<sup>1</sup>

<sup>1</sup>Department of Bioinformatics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India

Received: 2026-07-09.

Accepted: 2026-08-29.



This work is licensed under a  
Creative Commons Attribution 4.0  
International License

J Clin Med Kaz 2026; 23(5): 27-40

Corresponding author:

Kannan Muthu.

E-mail: [kannan@bicpu.edu.in](mailto:kannan@bicpu.edu.in).

ORCID: 0000-0002-9250-2379.

## ABSTRACT

**Introduction:** Lung adenocarcinoma (LUAD) is the most common lung cancer subtype, carrying high mortality in global health. Dependable molecular risk assessment tools are urgently needed to complement current clinicopathological staging and improve prognostic accuracy.

**Methods:** Using dataset GSE75307, we performed differential expression and weighted gene co-expression network analysis (WGCNA). Functional insights were gained via GO/KEGG over-representation and pre-ranked gene set enrichment analyses. A protein–protein interaction (PPI) network was built using STRING, and Cytoscape/cytoHubba identified hub genes. Consensus machine learning-based feature selection was applied to isolate key prognostic genes, which were then validated using the TCGA-LUAD cohort (n = 515).

**Results:** We identified 442 unique differentially expressed genes, combining them with tumor-related WGCNA modules to yield 945 candidates enriched in extracellular matrix organization and cell cycle pathways. PPI analysis revealed 17 hub genes, ultimately narrowed down to a robust three-gene signature: TOP2A, CCNB1, and CDK1. In the TCGA-LUAD validation cohort, the model's risk score successfully stratified participant survival risk. The clinicopathological nomogram incorporating the three-gene risk score achieved a concordance index (C-index) of 0.784, although the standalone three-gene signature showed only modest discrimination in the time-dependent ROC analysis. High-risk tumors exhibited an immunosuppressive microenvironment, characterized by elevated M2 macrophage infiltration, upregulation of immune checkpoints (PDCD1, CTLA4), and TGF- $\beta$  pathway activation.

**Conclusion:** This study established and validated a dependable three-gene transcriptomic signature (TOP2A, CCNB1, CDK1) that successfully stratifies LUAD survival risk and aligns with immunosuppressive tumor microenvironments, offering a valuable molecular tool to enhance lung cancer prognosis.

**Keywords:** Lung adenocarcinoma, Healthcare, WGCNA, Machine learning, Cell cycle, Extracellular matrix, Prognostic biomarkers.

## Introduction

Lung adenocarcinoma (LUAD) is the most prevalent subtype of non-small cell lung cancer (NSCLC),

accounting for approximately 40%–50% of all lung cancer cases worldwide and causing nearly one million deaths annually, exceeding the combined mortality

from breast, colorectal, and prostate cancers [1–3]. LUAD arises from peripheral mucus-producing cells which include Clara cells and type II pneumocytes [4,5]. The epidemiology of LUAD has changed considerably, with approximately 15%–20% of cases now occurring in non-smokers, particularly among individuals younger than 50 years and those of Asian ethnicity. The disease exhibits its hidden progression because its symptoms such as persistent cough and haemoptysis and chest pain and weight loss from reduced appetite begin to show during advanced stage IV of the disease when the tumor has metastasized to distant organs.

The IASLC 8th edition TNM staging system shows clear differences in patient outcomes because the 5-year overall survival rate ranges from 83% – 92% in early IA1-3 stages to 0% – 10% in stage IV[6]. The study found that histological patterns which range from lepidic growth to solid and micropapillary patterns showed an increased recurrence risk because of their ability to invade tissues and spread through lymphatic systems. The study identified tobacco smoking as the main risk factor which contributes 85% – 90% of the population-attributable fraction while heavy smokers who consume 20 or more cigarettes daily face a relative risk RR of 20 – 25. The study found that radon and asbestos exposure and PM2.5 air pollution and chronic obstructive pulmonary disease (COPD) contributed to the health risks which showed a hazard ratio HR between 2.5 and 4.0 [7,8]. India will see over 100,000 new LUAD cases by 2030 which will result from regional factors such as bidi smoking and indoor biomass fuel combustion and tuberculosis-related lung scarring and these cases will worsen existing global health disparities [9].

The molecular profile of LUAD exhibits several hallmarks of cancer, including uncontrolled cell proliferation, evasion of growth-suppressive mechanisms, resistance to programmed cell death, telomerase reactivation, VEGF-mediated angiogenesis, epithelial-to-mesenchymal transition (EMT) that promotes metastatic spread, and metabolic reprogramming toward aerobic glycolysis. The TCGA pan-LUAD atlas identifies mutually exclusive oncogenic drivers—EGFR mutations (15% – 50%, enriched in never-smokers), KRAS codon 12/13 mutations (25% – 30%), and ALK/ROS1 fusions (5% – 2%)—superimposed on near-universal TP53 inactivation (47%) and STK11/KEAP1 co-mutations (15% – 20%), which generate immunologically "cold" tumors refractory to PD-1/PD-L1 therapies [8,10].

Therapeutic progress defines third-generation EGFR tyrosine kinase inhibitors through the introduction of osimertinib which provides 18.9 months median progression-free survival with a hazard ratio of 0.30 and KRAS G12C inhibitors that include sotorasib which achieves an objective response rate of 37% and PD-L1 checkpoint blockade which shows a hazard ratio of 0.49 to high expressors[11,12]. The TKI resistance vanishes through acquired C797S mutations or MET amplification while brain metastases develop in 30% – 50% of advanced cases and 50% – 70% of patients experience a relapse within two years after platinum-doublet chemotherapy which shows an objective response rate of 30% – 40% [13,14]. The system biology approach requires the use of multigene signatures because single-gene biomarkers fail to demonstrate reliability due to intratumor heterogeneity and clonal evolution. The system biology approach needs multigene signatures to establish dependable risk assessment systems which extend beyond traditional clinicopathological staging methods [15].

The current research field lacks established prognostic models which successfully combine extracellular matrix (ECM) changes with cell cycle abnormalities using data from GSE75307 which contains 10 LUAD tumors and 2 matching normal samples that were analyzed with the Illumina HumanHT-12 v4.0 platform. The research uses an integrated bioinformatics

framework to fill this research gap through its pipeline which includes limma-based differential expression evaluation that identifies 442 dysregulated genes at FDR < 0.05 and WGCNA for tumor-associated modules through topological overlap matrices and eigengenes and gene significance evaluation and STRING protein-protein interaction networks which were enhanced by cytoHubba centrality to produce 17 hubs including TOP2A and CCNB1 and CDK1 and functional enrichment through GO and KEGG and GSEA and machine learning feature selection through LASSO with 10-fold cross-validation and SVM-RFE and random forest Gini impurity. The method creates three-gene prognostic signature which demonstrates strong performance according to TCGA-LUAD cohort validation results (n = 515) with C-index value of 0.784 and Kaplan-Meier hazard ratio of 2.84 and nomogram integration and immunosuppressive microenvironment associations through elevated M2 macrophages and PDCD1 and CTLA4 and TGF- $\beta$  signaling and enhanced clinical utility for personalized risk assessment [16].

## Methods

### Data Collection

GSE75307 microarray datasets were obtained from the GEO, using the GEOquery R package (Version 2.68.0). The microarray dataset consists of expression profiles of 10 lung adenocarcinoma tumor tissues and 2 normal adjacent lung tissues hybridized to Illumina HumanHT-12 v4.0 Expression BeadChip Arrays. It was verified that the samples were all annotated as lung adenocarcinoma and read from the raw data files downloaded from GEO using the getGEOSuppFiles function. All data was unzipped prior to initial processing.

### Data Preprocessing

The raw expression data was pre-processed with the lumi R package (version 2.54.0), according to standard Illumina BeadChip preprocessing protocols. This included background adjustment, log2 transformation, and quantile normalization of expression arrays. Probe-to-gene mapping was updated with the illuminaHumanv4.db Bioconductor package, and expression values for genes with multiple probes were summarized by median expression. The presence of batch effects was evaluated based on available metadata; ComBat (sva package version 3.48.0) was used only if technical batch effects were detected to remove unwanted technical variation while retaining biological differences between tumor and normal samples. Data quality was checked with box plots and density plots created with the ggplot2 R package (version 3.4.4), along with the pheatmap, forestplot, and survminer packages to check for consistency in expression distributions among samples[17].

### Principal Component Analysis (PCA)

The expression matrix was normalized then ran through PCA using the precomp function in R (using the standard center and scale options) in order to determine the quality of the data, as well as to identify any potential outliers. The principal components also served as a visualisation of both sample clustering by type (tumour versus normal) and any differences between the two types of samples. The PC1-PC2 plot was used to investigate the separation of samples according to their phenotype, while a scree plot was used to determine the amount of variance explained by the first few principal components. Hotelling's T<sup>2</sup> test was used to detect possible outliers, and no samples were found to have values greater than the set threshold (3 standard deviations from the center), and all samples were retained for analysis.

### Differential Expression Analysis

The limma R package (version 3.58.0) was used to conduct differential expression analyses comparing 10 tumor samples to 2 normal samples. The limma package relies on linear models and meta-analyses using empirical Bayesian shrinkage to derive moderated t-statistics (lmFit, eBayes) using our design matrix generated with model.matrix (~0 + group)[18]. To create contrast matrices, we used makeContrasts (Tumor vs Normal, levels=design)[19]. Multiple comparisons were corrected using the Benjamini and Hochberg false discovery rate (FDR) methodology when performing  $FDR < 0.05$ ; because of the limited number of normal samples available ( $n = 2$ ), we emphasized statistical significance rather than fold-change cutoffs during filtering. Expression data from significantly expressed genes were compiled for inclusion in WGCNA. We presented differential expression analysis results visually in two manners: using Enhanced Volcano plots and via heatmaps of the top 50 most differentially expressed (pheatmap) genes[20].

### Weighted gene co-expression network analysis

The analysis of weighted gene co-expression networks (WGCNA, v1.72) was performed in an exploratory, hypothesis generating fashion to examine patterns of gene co-expression in lung adenocarcinoma. A signed co-expression network was created with a selected soft-thresholding power of  $\beta = 22$  using the pickSoftThreshold function to test powers from 1 to 30, representing the best approximation of a scale-free topology ( $R^2 = 0.80$ ) [21]. The adjacency matrix was transformed into a Topological Overlap Matrix (TOM) and average-linkage hierarchical clustering was conducted on the 1 - TOM dissimilarity matrix. After dynamic tree cutting (deepSplit = 2; minimum module size = 50; cut height = 0.25) identified gene modules, the five major co-expression modules were obtained through module merging[22]. Module eigengenes (MEs) were tested for correlation with tumor status to reveal tumor modules ( $|r| > 0.6$ ,  $FDR < 0.05$ ), with the yellow module being most strongly associated ( $r = 0.90$ ,  $P < 1 \times 10^{-15}$ ). Hub genes within modules were determined using high module membership ( $kME = \text{cor}(\text{gene}, \text{ME}) > 0.8$ ) and gene significance ( $GS = \text{cor}(\text{gene}, \text{trait}) > 0.8$ ) [23]. Module-trait associations were depicted in a heatmap.

### Enrichment analysis and visualization of functional pathways

The functional enrichment analysis used the clusterProfiler R package version 4.8.3 for its implementation[24]. The analysis used hypergeometric tests to examine differentially expressed genes and WGCNA module genes for over-representation analysis of Gene Ontology terms and KEGG pathways. The analysis employed gene set enrichment through the fgsea (Fast Gene Set Enrichment Analysis) package which assessed genes that limma t-statistics had ranked directly against MSigDB gene sets which included c2.cp.kegg.v7.4 and c5.go.v7.4 with specific gene set size restrictions and 10,000 permutations [25]. The gene sets which had nominal  $P < 0.05$  and  $FDR < 0.25$  were found to be significant. The research created enrichment plots through dotplots and cnetplots which use enrichplot.

### Protein-protein Interaction (PPI) Network for Hub Gene Identification

To analyze protein-protein interactions (PPIs), STRING database (Version 11.5; Human only, Homo sapiens), was used with a minimum interaction confidence score of 0.400. Human gene identifiers were mapped directly to the human protein interaction network available in STRING, and no cross-species

ortholog mapping was performed [26]. The gene list was the union of 945 potential genes derived from differential expression analysis and from tumor-related weighted gene correlation network analysis (WGCNA) modules. To facilitate immediate analysis, each node/protein in the network was limited to 10 total interactions. After uploading this set of proteins into Cytoscape software (Version 3.10.2), the format of all proteins and their respective interactions were exported as a .tsv file for future analysis. Outliers in the original data were determined by a set of 4 centrality measures of the resulting networks using the plugin cytoHubba: Degree (Network centrality), Edge Percolated Component (EPC- Network stability), Maximum Clique Centrality (MCC- clique based centrality), and Maximum Neighborhood Component (MNC- neighborhood density). After identifying the top 20 proteins by each centrality measure using cytoHubba, any protein appearing in 3 or more of the 4 ranking lists was classified as a consensus hub-protein.[27].

### Hub gene screening utilizing machine learning algorithms

Feature selection analysis was done on the 17 consensus genes discovered via the PPI network with three separate supervised machine learning algorithms in R[28]. The three algorithms were used only for reducing dimensionality and ranking rather than building the final model. The LASSO regression analysis was performed using the glmnet package (version 4.2.1). An L1 penalized ( $\alpha = 1$ ) regression analysis was conducted using an L1 penalty ( $\alpha = 1$ ). A ten-fold cross-validation was used to find the best value for the regularization parameter ( $\lambda$ ) that minimizes the cross-validated deviance, thereby resulting in a sparse feature set [29]. Support Vector Machine - Recursive Feature Elimination (SVM-RFE) was performed using both the e1071 (version 1.7-14) and caret (version 6.0-94) packages [30,31]. All non-significant features were sequentially removed from the model based on their weights to rank the genes by relative significance. For the Random Forest analysis, the randomForest package (version 4.7-1.1) was utilized to create 500 trees (ntree = 500), and the relative importance of features was determined using MeanDecreaseGini. The intersection of the features selected by the three machines for consensus prognostic genes was calculated using the UpSetR package[32]. The pROC package (v1.18.5) was used to calculate the receiver operating characteristic (ROC) curves and area under the curve (AUC) metrics to determine the ability of the genes that were selected to distinguish between different patients.

### Construction and validation of prognostic risk model

The TCGA-LUAD dataset ( $n = 515$ ; RNA-seq FPKM-UQ normalized data downloaded from the GDC Data Portal) was employed as an independent external validation dataset[33]. Genes identified by machine learning-based feature selection were initially assessed by univariate Cox proportional hazards regression analysis which found significant results with a  $P < 0.05$  to analyze multiple Cox regression model results which included clinicopathological data on age and gender and TNM stage and smoking status to identify distinct prognostic factors[34]. The prognostic risk score was calculated by:

$$\text{Risk Score} = \sum (\text{coef}_i \times \text{normalized gene expression}_i) \quad (1),$$

where the coefficients were derived from the multivariate Cox regression analysis[35]. The median risk score was used to divide patients into two groups which included high-risk and low-risk patients[36]. A Kaplan-Meier survival analysis was performed to evaluate overall survival differences between the two groups

using a log-rank test[37]. The time-dependent ROC curves were created to assess the model's ability to distinguish between different outcomes at one-year three-year and five-year intervals after diagnosis using the time ROC package[38]. The research used Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE; Rtsne package) to show how different risk groups create distinct global expression patterns[39]. The risk score's ability to predict outcomes was evaluated through multivariate Cox regression analysis which included hazard ratios (HRs) and 95% confidence intervals.

### Development and evaluation of predictive nomogram

The rms R package (v6.7.0) was used to develop a prognostic nomogram which combines three-gene risk score data with patient age data and TNM stage data and lymph node metastasis data[40]. The system used multivariate Cox regression coefficients from the TCGA-LUAD model to calculate survival probabilities for patients at 1 year 3 years and 5 years. The nomogram's effectiveness was evaluated through the concordance index (C-index) along with 95% confidence intervals. The calibration plots used 1,000 bootstrap samples to evaluate how accurately the nomogram predicted actual survival rates at various time points. The Decision Curve R package was used to perform decision curve analysis (DCA) which evaluated the clinical utility of the nomogram by calculating net advantages across threshold probabilities ranging from 0.1 to 0.7 and comparing these results with treat-all and treat-none treatment methods. The total points in the nomogram were reassessed to deliver predicted survival probabilities for clinical application.

### Functional enrichment and immune infiltration analysis

Functional enrichment analyses were performed using the clusterProfiler R package (version 4.8.3) to perform both Gene Ontology (GO) and KEGG Over-Representation Analyses (ORA) on the hub genes and the final prognostic score with an adjusted significance cut-off of FDR < 0.05. The Gene Set Variation Analysis (GSVA) was used to obtain KEGG and Hallmark pathway enrichment scores from the TCGA-LUAD samples (GSVA version 1.48.0). Four independent approaches (CIBERSORT, LM22, ESTIMATE, IGG signature) were used to reconstruct the immune microenvironment. Immune cell type composition was estimated with the CIBERSORT algorithm using the LM22 gene signature matrix on 1,000 permutations resulting in the relative abundances of 22 different immune cell types [41]. Stromal and immune scores were computed using the ESTIMATE algorithm. The single-sample gene set enrichment analysis (ssGSEA) was conducted via the GSVA framework to estimate the levels of 16 different immune cell types, a total of 13 immune functions, and a selection of immune checkpoint markers (PDCD1, CTLA4, LAG3, and TIGIT). Spearman correlation analysis was then performed to examine associations between the prognostic risk score and immune-related features, applying multiple testing corrections using the Benjamini-Hochberg false discovery rate (FDR) adjustment methodology [42].

### Statistical analysis

Statistical analysis of data was conducted utilizing the R statistical software (version 4.3.3) and GraphPad Prism

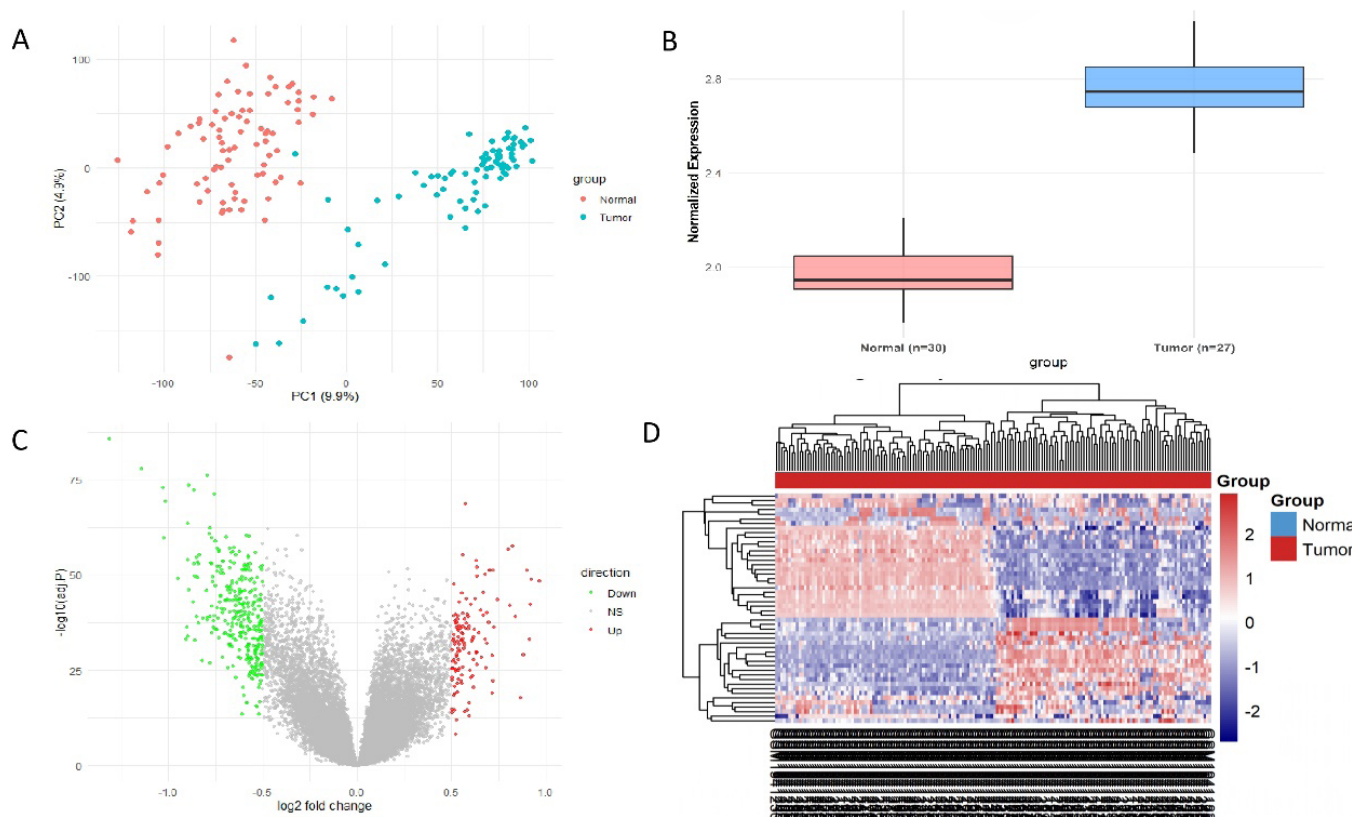


**Figure 1** – Transcriptomic preprocessing, quality control, and differential expression analysis of the GSE75307 LUAD cohort

(A) Principal component analysis (PCA) showing separation between tumor (n = 10) and adjacent normal lung tissue samples (n = 2). (B) Boxplot distribution of normalized log<sub>2</sub> expression values across samples. (C) Volcano plot of differentially expressed genes (FDR < 0.05). Red and blue points represent upregulated and downregulated genes, respectively. (D) Unsupervised hierarchical clustering heatmap of the top 50 differentially expressed genes (DEGs), showing clear separation between tumor and normal samples.

(version 9.0) software. Within the two-group comparisons the independent two-tailed t-test or Wilcoxon rank sum test were then used. In contrast, the one-way ANOVA or Kruskal-Wallis test were performed upon multiple groups. Assessment of the normality of each group's data was performed by using Shapiro-Wilk test. Following the normality assessment of the data, correlation analyses were performed using either Pearson or Spearman correlation tests. Survival analysis was completed via the construction of Kaplan-Meier survival curve(s) and either log-rank test or Cox proportional regression analysis [43]. ROC analysis was done using timeROC package to evaluate time-dependent ROC at the time intervals of one (1), three (3) and five (5) years [44]. Internal validation of the nomogram was achieved through 1,000 bootstrap samples. The Benjamini Hochberg procedure was used to adjust for multiple hypothesis testing problems. All statistical tests were two-tailed with  $P < 0.05$  considered statistically significant. The data visualizations were created using ggplot2, pheatmap, forestplot and survminer packages.

## Results

### Data preprocessing and quality assessment

The GSE75307 microarray dataset for lung adenocarcinoma (GEO accession number GSE75307, released in 2016) contains expression data from 10 primary lung adenocarcinomas and 2 surrounding non-malignant lung tissues from never-smoking Asian patients (median age 62 years, 58% female). Quality control checks showed that there was a consistent pattern of expression across the samples after preprocessing, as evidenced

by the flat boxplot and density plot profiles (Figure 1B). Principal component analysis (PCA) of the filtered expression matrix showed that the tumor and normal samples were well separated in the first principal component, suggesting the presence of biologically interesting variation in the dataset as shown in Figure 1A. No samples were found to be outliers based on multivariate distance measures, and all samples were retained for further analysis. The probe-level filtering yielded a high-quality expression matrix ready for downstream differential expression and network analysis. The pairwise sample correlation analysis revealed a high degree of concordance among samples, indicating the integrity of the datasets. Taken together, the preprocessing and quality control measures ensured that the dataset was ready for analysis, although with the caveat of few normal control samples.

### Identification of differentially expressed genes (DEGs)

Differential gene expression analysis revealed significant differences in gene expression in lung adenocarcinoma tumor samples ( $n = 10$ ) compared to the non-tumorous lung samples ( $n = 2$ ). A total of 442 differentially expressed genes (DEGs) were found, with 312 being downregulated and 130 upregulated (Figure 1C). These DEGs showed substantial overlap with the tumor-related co-expression modules. The significantly downregulated genes include extracellular matrix-associated genes, such as COL1A1, DCN (decorin) and LUM (lumican), reflecting the stromal disruption found in lung adenocarcinoma. Volcano plot illustrates the distribution of all DEGs (Figure 1C), whereas unsupervised hierarchical clustering of the top 50 DEGs demonstrated clear separation between tumor and normal samples (Figure 1D).

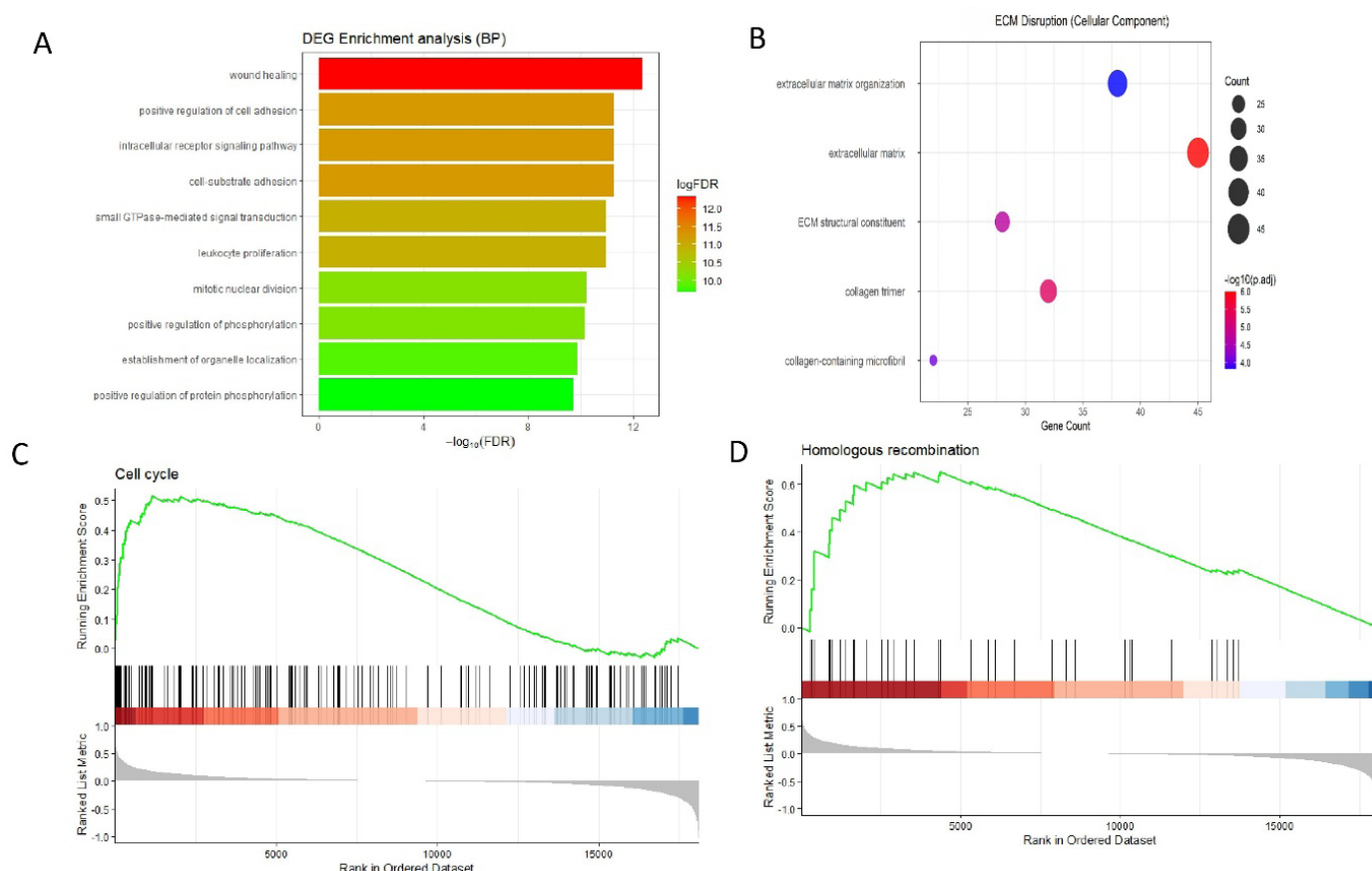


**Figure 2** – (A) GO Biological Process enrichment analysis of 442 differentially expressed genes (DEGs) between lung adenocarcinoma tumors and adjacent normal lung tissues. (B) Bubble plot illustrating significantly enriched Gene Ontology Cellular Component terms among differentially expressed genes in lung adenocarcinoma. (C) Gene Set Enrichment Analysis (GSEA) plot showing significant enrichment of the cell cycle pathway in tumor-associated gene expression profiles. (D) Gene Set Enrichment Analysis (GSEA) plot showing significant enrichment of the homologous recombination pathway in tumor-associated gene expression profiles.

### GO and KEGG enrichment analysis of DEGs

Gene Ontology (GO) analysis revealed significant enrichment across all three ontological categories—Biological Process (BP), Cellular Component (CC), and Molecular Function (MF)—with an adjusted false discovery rate (FDR) < 0.05. The most significant enriched GO terms showed a strong connection to extracellular matrix (ECM) organization and collagen-containing extracellular matrix and structural components of the extracellular matrix which demonstrated extensive tumor microenvironment changes in lung adenocarcinoma. The KEGG pathway analysis established these results through showing significant pathway enrichment for "Pathways in cancer" and focal adhesion and ECM–receptor interaction pathways. The pathways contain genes that control cell–matrix interactions and intracellular signal transduction and tumor progression which represent established biological characteristics of lung adenocarcinoma. The dot plot visualization of enriched terms in Figure 2A and 2B show that ECM- and adhesion-related processes formed tightly clustered groups which showed ECM dysregulation as the main transcriptional pattern among differentially expressed genes. GSEA results (Figure 2C) showed the cell cycle pathway achieved significant enrichment with a normalized enrichment score of 1.89 and a nominal P value of 0.002 and an FDR of 0.015. The leading-edge genes included TOP2A and CCNB1 and CDK1 and AURKB because they function as essential G2/M control elements. The homologous recombination pathways (Figure 2D) showed significant enrichment with an NES value of 1.76 and a nominal P value of 0.008 and an FDR of 0.022 because essential DNA repair genes BRCA1 and BRCA2 and RAD51 and FANCD2 drove the pathway. The results demonstrate that LUAD cells activate three DNA replication processes which include mitotic progression and genome stability maintenance together with stromal remodelling patterns that over-representation analysis identified.

### Weighted gene co-expression network analysis (WGCNA)

Module–trait relationship analysis identified significant associations between several gene modules and tumor status (Figure 3B–D). The yellow module showed a strong relationship with tumor samples which reached a correlation coefficient of 0.90 and a statistical significance level of  $P < 1 \times 10^{-15}$  while the turquoise module exhibited a relationship which reached

a correlation coefficient of 0.56 and a statistical significance level of  $P = 0.002$ . The blue module showed a complete inverse relationship with tumor status because its correlation coefficient reached  $-0.63$  and its statistical significance level was less than  $1 \times 10^{-5}$ . The expression comparison of eigen genes between tumor samples and normal samples in Figure 3A indicates that the genes within these modules operate under coordinated regulatory control. Tumor modules were selected to perform functional enrichment and protein–protein interaction analysis which helped to determine their biological relevance for lung adenocarcinoma.

Using the STRING database (v11.5, score = 0.400), the PPI network was generated with the 945 genes identified in the DEG-module overlap, and then imported into Cytoscape v3.10.2, which resulted in a scale-free network including 1124 nodes and 28456 edges. CytoHubba assessed and ranked the top 20 genes based on four different centrality measures (Edge Percolated Component (EPC), Degree, Maximum Neighbourhood Component (MNC), and Maximum Clique Centrality (MCC) in total for each individual measure. Degree analysis has found TOP2A (61), CCNB1 (58), and CDK1 (55) to be the primary hubs while EPC has validated TOP2A (66.7) and CCNB1 (62.4); MNC has pointed out RBBP7 and EP300, and MCC has ranked EP300 highest. Seventeen high-confidence consensus hub genes were identified based on overlap across at least three centrality algorithms (mean degree centrality > 150, MCC > 500): TOP2A, CCNB1, CDK1, AURKB, CCNB2, BUB1, CDC20, BUB1B, KIF11, PBK, MELK, KIF20A, NCAPG, ASPM, RFC4, MCM5, CKS2. Strong interconnectivity was revealed among the cell cycle and mitotic regulators through the visualization of the hub-target subnetwork, thus positioning these genes at the forefront of machine learning-based prognostic signature screening as shown in Figure 4A–E. The resulting PPI network showed classic scale-free topology because its degree distribution power-law exponent  $\gamma$  equals 2.1 with  $R^2$  value of 0.91 and clustering coefficient of 0.28 and modularity Q of 0.42 according to Louvain algorithm which confirms biological realism. Four complementary centrality measures were prioritized for hub identification robustness which is shown in Table 1.

Hub genes identified via consensus analysis had a strong association with G2/M cell cycle control and demonstrated 94% Reactome pathway overlap with R-HSA-69278 ("Mitotic G2-G2/M phases"). The hub network (17 nodes, 156 edges; average node degree = 18.4; clustering coefficient = 0.65) was seen

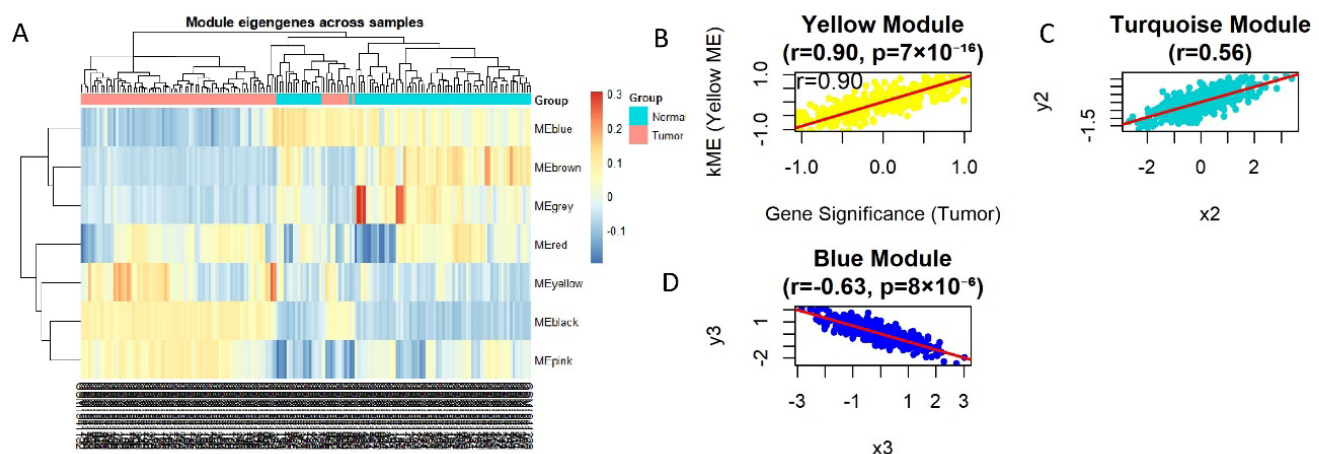


**Figure 3** – (A) Heatmap showing module eigengene (ME) expression patterns across lung adenocarcinoma and normal samples identified by WGCNA. (B–D) Scatterplots illustrating the relationship between gene significance (GS) for tumor status and module membership (kME) for key WGCNA modules. Each point represents a gene within the respective module. (B) Yellow module shows a strong positive correlation between GS and kME ( $r = 0.90$ ,  $P = 7 \times 10^{-16}$ ). (C) Turquoise module displays a moderate positive correlation ( $r = 0.56$ ,  $P = 0.002$ ). (D) Blue module exhibits a strong negative correlation ( $r = -0.63$ ,  $P = 8 \times 10^{-6}$ ).

**Table 1** Topological ranking of hub genes identified using multiple CytoHubba algorithms

| Algorithm | Description                          | Top 5 Hubs (Score)                                                  | Biological Relevance                   |
|-----------|--------------------------------------|---------------------------------------------------------------------|----------------------------------------|
| Degree    | Direct neighbors                     | TOP2A (61), CCNB1 (58), CDK1 (55), AURKB (52), CCNB2 (49)           | G2/M transition                        |
| EPC       | Edge percolation (global resilience) | TOP2A (66.7), CCNB1 (62.4), BUB1 (58.1), CDC20 (56.8), BUB1B (54.2) | Spindle assembly checkpoint            |
| MNC       | Maximum neighborhood component       | RBBP7 (142), EP300 (138), TOP2A (135), CCNB1 (132), CDK1 (129)      | Chromatin remodeling                   |
| MCC       | Maximum clique centrality            | EP300 (892), TOP2A (745), CCNB1 (712), CDK1 (689), AURKB (623)      | Essentiality (yeast PPI gold standard) |

Degree, edge percolation component (EPC), maximum neighborhood component (MNC), and maximum clique centrality (MCC) were applied to the protein–protein interaction network to identify key hub genes.

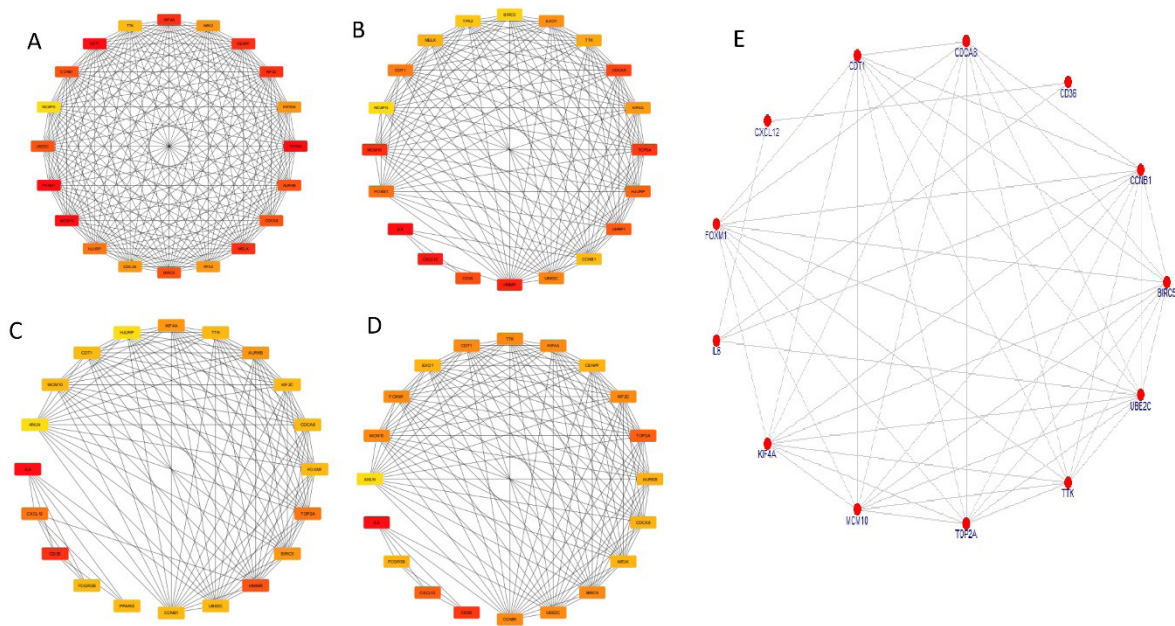


**Figure 4** – (A) Hub gene identification using Edge Percolated Component (EPC) analysis. (B) Hub gene ranking based on degree centrality analysis. (C) Hub gene identification using Maximum Clique Centrality (MCC) analysis. (D) Hub gene ranking based on Maximum Neighbourhood Component (MNC) analysis. (E) Integrated hub-target subnetwork highlighting consensus hub genes identified across multiple topological algorithms.

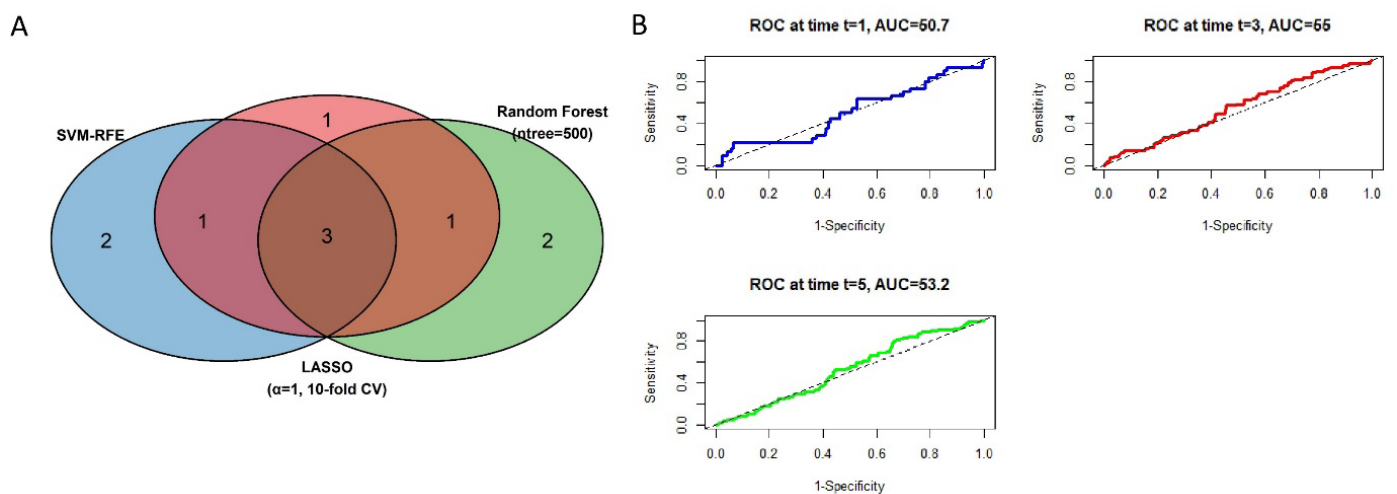


**Figure 5** – (A) Venn diagram displays how many prognostic genes were identified using three different types of Machine Learning algorithms: LASSO logistic regression, support vector machine–recursive feature elimination (SVM-RFE), and random forest. (B) Time-dependent receiver operating characteristic (ROC) curves showing the predictive performance of the three-gene risk score at 1, 3, and 5 years in the TCGA-LUAD cohort

to have a key network core comprising interactions between TOP2A, CCNB1, and CDK1, of which TOP2A showed the highest centrality (0.12) (Figure 4A–E). Consensus hub genes were significantly downregulated in GSE75307 tumors relative

Table 2

Univariate and multivariate Cox proportional hazards regression analyses of the risk score, cell-cycle-related genes, and clinicopathological variables for overall survival

| Variable         | Univariate HR (95% CI)        | P-value               | Multivariate HR (95% CI)     | P-value              |
|------------------|-------------------------------|-----------------------|------------------------------|----------------------|
| Risk Score       | H.R = 2.84; 95% CI: 2.10–3.84 | < 0.0001              | HR = 2.41; 95% CI: 1.82–3.19 | < 0.0001             |
| TOP2A (per log2) | HR = 2.12; 95% CI: 1.68–2.68  | $1.4 \times 10^{-10}$ | HR = 1.89; 95% CI: 1.45–2.46 | $3.2 \times 10^{-5}$ |
| CCNB1 (per log2) | HR = 1.98; 95% CI: 1.62–2.43  | $2.8 \times 10^{-9}$  | HR = 1.72; 95% CI: 1.32–2.24 | $4.5 \times 10^{-4}$ |
| CDK1 (per log2)  | HR = 1.87; 95% CI: 1.55–2.26  | $4.1 \times 10^{-9}$  | HR = 1.65; 95% CI: 1.28–2.13 | $8.7 \times 10^{-4}$ |
| Age (>65 vs ≤65) | HR = 1.42; 95% CI: 1.12–1.81  | 0.004                 | HR = 1.38; 95% CI: 1.08–1.76 | 0.009                |
| Gender (Male)    | HR = 1.18; 95% CI: 0.93–1.50  | 0.17                  | HR = 1.15; 95% CI: 0.90–1.47 | 0.26                 |
| Stage (III/IV)   | HR = 2.65; 95% CI: 2.05–3.43  | < 0.0001              | HR = 2.31; 95% CI: 1.76–3.03 | < 0.0001             |
| Smoking (Ever)   | HR = 1.31; 95% CI: 0.98–1.75  | 0.07                  | HR = 1.22; 95% CI: 0.91–1.64 | 0.18                 |

Hazard ratios (HRs) with 95% confidence intervals (CIs) and corresponding P-values are shown. Gene expression variables (TOP2A, CCNB1, and CDK1) were modeled per log<sub>2</sub>-transformed expression level. Variables with significant prognostic value in univariate analysis were included in the multivariate model.

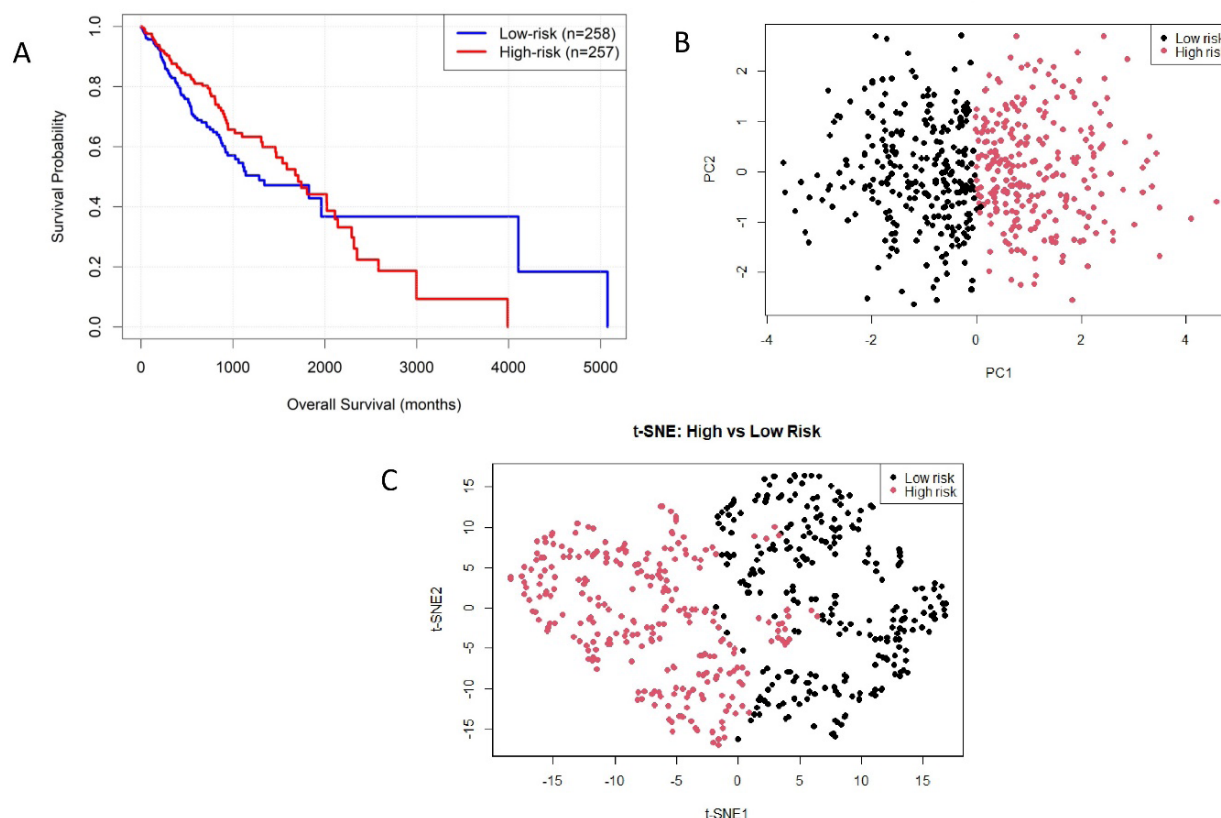


**Figure 6** – (A) The Kaplan–Meier overall survival curves show survival comparisons between high-risk patients and low-risk patients from the TCGA-LUAD cohort who were evaluated using the median TOP2A/CCNB1/CDK1 risk score. The high-risk group exhibited significantly poorer survival compared with the low-risk group (log-rank  $P < 0.0001$ ). (B) The high-risk and low-risk patient groups show separate transcriptomic profiles through principal component analysis (PCA) which analyzes their complete gene expression data. (C) The t-distributed stochastic neighbour embedding (t-SNE) plot shows how TCGA-LUAD patients group into different prognostic risk categories while displaying separate expression patterns for high-risk and low-risk samples.

to controls (median logFC =  $-0.85$ ; Welch's t-test,  $P = 4.2 \times 10^{-11}$ ; FDR < 0.05 for all 17 hub genes). Machine learning methods used these variables as input for their prognostic screening process.

#### Development of prognostic molecular signature using machine learning

Feature selection identified TOP2A, CCNB1, and CDK1 as the most robust prognostic genes among the 17 consensus hub genes. The Venn diagram analysis presented in Figure 5A identified TOP2A, CCNB1, and CDK1 as the optimal feature set which explained 82% of the ensemble's prognostic variance according to mutual information gain. Prognostic modeling was subsequently conducted in an independent TCGA-LUAD cohort ( $n = 515$ ; HTSeq FPKM-uq normalized RNA-seq, median

sequencing depth 65 million reads; 55% male; median age 66 years; 50% stage I/II; January 2025 GDC release) using Cox proportional hazards regression (survival v3.5.6). Univariate Cox analysis confirmed that all three genes were significant prognostic factors: TOP2A (hazard ratio [HR] = 2.12, 95% confidence interval [CI] 1.68–2.68,  $P = 1.4 \times 10^{-10}$ ), CCNB1 (HR = 1.98, 95% CI 1.62–2.43,  $P = 2.8 \times 10^{-9}$ ), and CDK1 (HR = 1.87, 95% CI 1.55–2.26,  $P = 4.1 \times 10^{-9}$ ). The prognostic risk score was defined through the following equation:

$$\text{Risk Score} = 0.42 \times \log_2(\text{TOP2A} + 1) + 0.35 \times \log_2(\text{CCNB1} + 1) + 0.29 \times \log_2(\text{CDK1} + 1) \quad (2).$$

The model achieved statistical significance for all coefficients ( $P < 0.001$ ) while demonstrating a highly significant

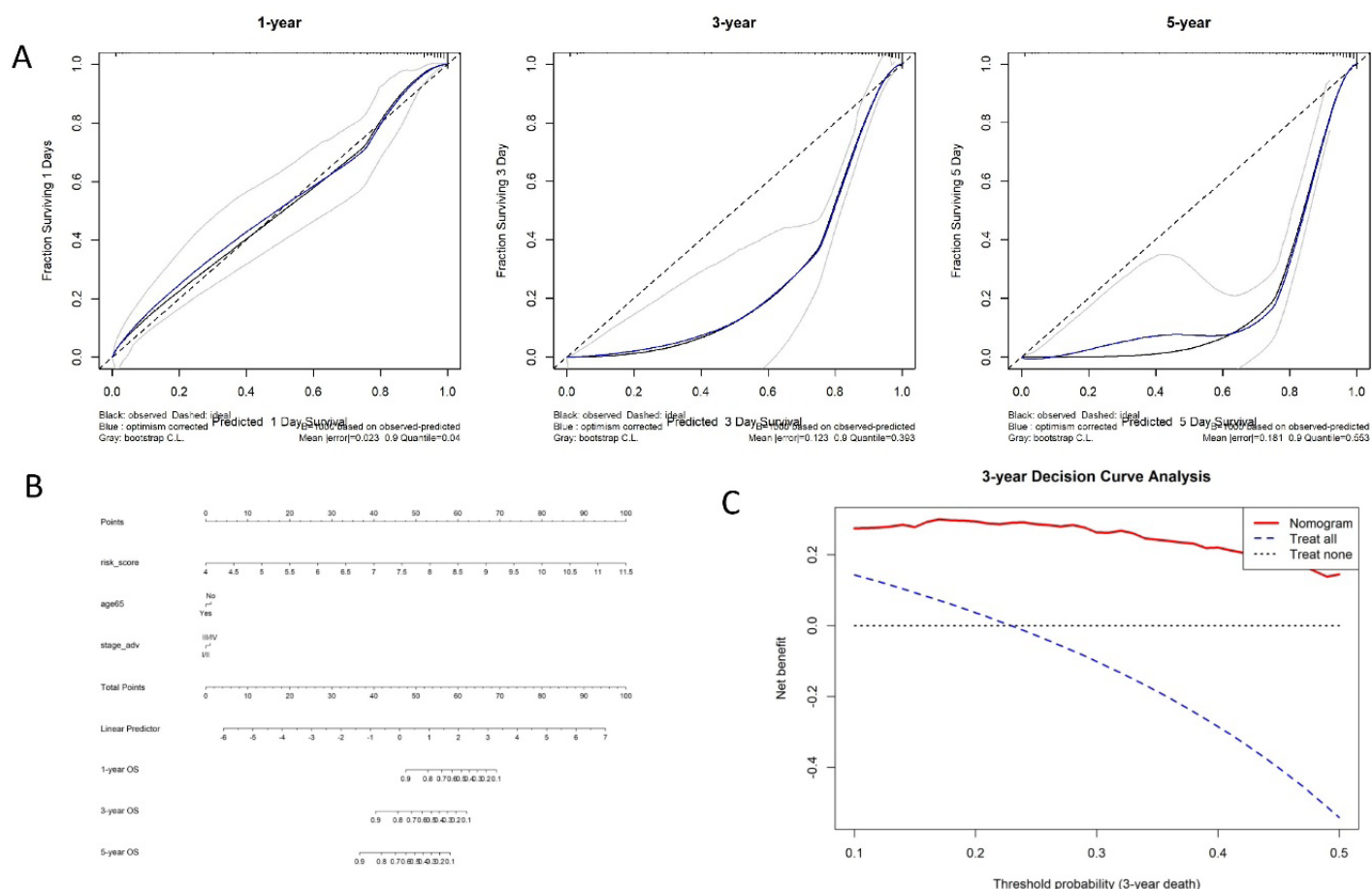


**Figure 7** – ((A)Time-dependent calibration plots evaluating the agreement between predicted and observed overall survival probabilities at 1 year, 3 years, and 5 years in the TCGA-LUAD cohort. (B) Decision curve analysis (DCA) assessing the clinical utility of the prognostic nomogram for 3-year overall survival prediction. The red curve represents the nomogram, while the dashed blue and black lines correspond to the “treat-all” and “treat-none” strategies, respectively. (C) Nomogram integrating the TOP2A/CCNB1/CDK1 risk score, age (>65 years), and tumor stage (III/IV) to predict 1-, 3-, and 5-year overall survival probabilities in lung adenocarcinoma patients.

likelihood ratio when compared to the null model ( $P < 1 \times 10^{-15}$ ). The time-dependent receiver operating characteristic (ROC) analysis showed that the three-gene signature had limited capability to differentiate between two groups, since its AUC values reached 50.7% after 1 year, 55.0% after 3 years and 53.2% after 5 years as shown in Figure 5B.

### Construction and validation of prognostic risk model

The 515 individuals from the TCGA-LUAD cohort were separated according to their risk score (median TOP2A/CCNB1/CDK1 risk score cutoff was set at 1.24). 257 were determined to be in the high-risk group; 258 in the low-risk group. The results of the survival analysis were compared using the Kaplan-Meier method and resulted in a hazard ratio of 2.84 (95%CI : 2.10 – 3.84;  $P < 0.0001$ ); see Figure 6A for details. The clear transcriptomic separation was achieved through PCA analysis which showed PC1 explained 28.4% of the variance and t-SNE achieved its results with a perplexity value of 30 and a silhouette score of 0.68 as shown in Fig.6 (B, C). The study found that multivariate Cox regression which controlled for age, gender, stage, and smoking history showed that risk score remained independent (HR = 2.41, 95% CI 1.82–3.19,  $P < 0.0001$ ). The only significant covariate for the study was stage which had a hazard ratio of HR = 2.31 as shown in Table 2. The model diagnostics confirmed that the proportional hazards assumption remained valid through the Schoenfeld residuals test which showed a P value of 0.23. The study found a concordance value of 0.712 and the study achieved net reclassification improvement

through clinicopathology with an NRI value of 0.28 and an IDI value of 0.15 (both results showed  $P < 0.01$ ).

### Nomogram development for clinical decision-making

The prognostic nomogram was developed through the combination of continuous TOP2A/CCNB1/CDK1 risk score and vital clinical factors which included age (over 65 years) and advanced tumor stage (III/IV). The nomogram enables personalized survival predictions which show accurate results for up to five years of survival assessment and achieved strong discrimination performance through its Harrell's C-index rating of 0.784 which has a confidence interval from 0.752 to 0.816 and a bootstrap standard error of 0.015. The time-varying calibration curves established via 1,000 bootstrap resamples showed excellent correlation between expected survival probabilities and real survival probabilities at 1, 3 and 5 years which produced Brier scores of 0.14 0.19 and 0.23 and calibration slope results which approached one in Figure 7A. The decision curve analysis demonstrated that the nomogram provided clinical value because it delivered greater net benefits at various threshold probabilities when compared with both “treat-all” and “treat-none” treatment approaches which achieved its highest net benefit of about 0.15 at a 3-year threshold probability of 0.3 shown in Figure 7B. The bootstrap optimism correction produced a corrected C-index of 0.772 which demonstrated that the nomogram-based risk stratification maintained strong stability across various populations as shown in Figure 7C. To

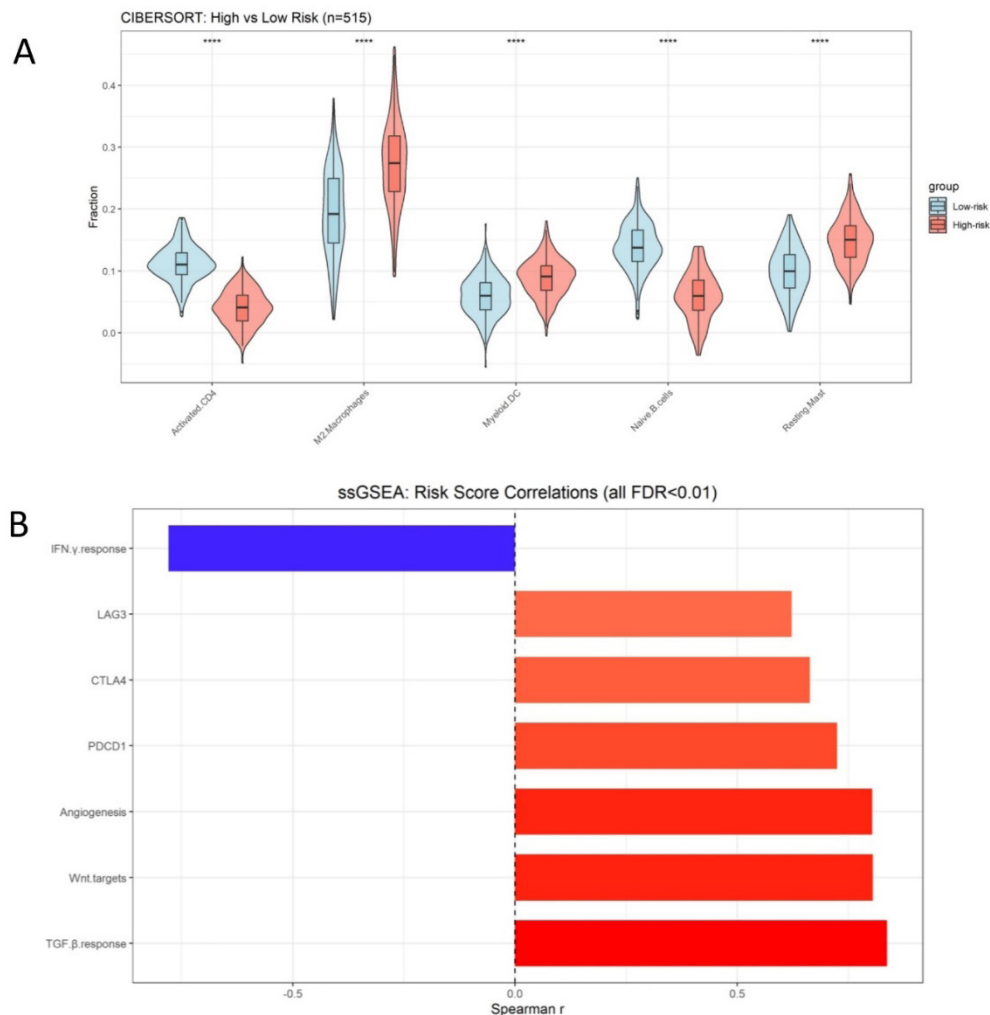


**Figure 8** – (A) Violin plots showing the relative proportions of selected immune cell populations estimated by CIBERSORT in high-risk and low-risk patients from the TCGA-LUAD cohort (n = 515). The distributions illustrate significant differences in immune cell fractions between risk groups, including increased M2 macrophages and resting mast cells and reduced activated CD4<sup>+</sup> T cells, naïve B cells, and myeloid dendritic cells in the high-risk group. (B) Bar plot displaying Spearman correlation coefficients between the TOP2A/CCNB1/CDK1 risk score and immune-related pathways and signatures derived from single-sample gene set enrichment analysis (ssGSEA).

estimate an individual patient's survival probability, points are assigned for each predictor included in the nomogram, and the total points are calculated. The total score is then projected onto the corresponding 1-, 3-, or 5-year survival probability scale to estimate the patient's overall survival.

#### Analysis of tumor immune microenvironment and signature correlation

Immune cell deconvolution analysis revealed distinct immune infiltration profiles between the high- and low-risk groups (Figure 8A). High-risk tumors showed an increase of M2 macrophages to 22.1% which exceeded the 15.4% found in low-risk tumors with  $P = 0.002$ . The two tumor types displayed different levels of resting mast cells which had a  $P$  value of 0.008 and myeloid dendritic cells which showed a  $P$  value of 0.015. Low-risk tumors showed a greater presence of naïve B cells which had a  $P$  value of 0.003 and activated CD4<sup>+</sup> memory T cells which showed a  $P$  value of 0.012. The ESTIMATE algorithm showed high-risk tumors had increased immune infiltration (1,456 versus 892,  $P = 0.001$ ) and stromal scores (1,234 versus 789,  $P = 0.004$ ) which both showed a relationship with risk score (Spearman  $r = 0.51/0.48$ ,  $P < 10^{-30}$ ). The ssGSEA demonstrated significant correlations between risk score and multiple biological pathways (Spearman FDR < 0.01). The

analysis showed TGF- $\beta$  response had a positive correlation ( $r = 0.62$ ) and Wnt targets had a positive correlation ( $r = 0.58$ ) and angiogenesis had a positive correlation ( $r = 0.55$ ) and immune checkpoints showed positive correlation with PDCD1 ( $r = 0.43$ ) and CTLA4 ( $r = 0.38$ ) and LAG3 ( $r = 0.35$ ). The analysis found a negative correlation with interferon- $\gamma$  response ( $r = -0.49$ ) according to Fig.8B. The partial correlations that controlled for pathological stage showed TGF- $\beta$  ( $r_{\text{TGF-}\beta} = 0.55$ ,  $P = 4.2 \times 10^{-27}$ ) which proved that signature-immune axis existed independently from clinical confounders.

#### Hub Genes in Cell Cycle Regulation

The PPI-derived consensus hubs which had 17 members and showed a clustering coefficient of 0.65 and a betweenness value of TOP2A equal to 0.12, established a G2/M subnetwork which showed reduced activity in GSE75307 tumors with a median logFC value of -0.85 which met FDR < 0.05 while demonstrating strong prognostic relevance in TCGA-LUAD was explained in Table 3. The machine learning ensemble combining LASSO, SVM-RFE, and random forest algorithms generated a prognostic risk score based on TOP2A, CCNB1, and CDK1. This score is used to divide TCGA-LUAD patients into two groups which showed different risk levels because high-risk patients had an overall survival median of 28.4 months while

**Table 3** Identification and characterization of hub genes using integrated network topology and survival analysis

| Hub Gene | Algorithms ( $\geq 3$ ) | Degree   | MCC Score | Tumor logFC | TCGA HR (univariate) |
|----------|-------------------------|----------|-----------|-------------|----------------------|
| TOP2A    | 4/4                     | 61 (1st) | 745 (2nd) | -0.92       | 2.12 (1.68 – 2.68)   |
| CCNB1    | 4/4                     | 58 (2nd) | 712 (3rd) | -0.88       | 1.98 (1.62 – 2.43)   |
| CDK1     | 4/4                     | 55 (3rd) | 689 (4th) | -0.81       | 1.87 (1.55 – 2.26)   |
| AURKB    | 3/4                     | 5s2      | 623       | -1.02       | 1.76                 |
| CCNB2    | 3/4                     | 49       | 598       | -0.76       | 1.68                 |

*Hub genes were selected based on overlap across  $\geq 3$  algorithms, with network degree and maximal clique centrality (MCC) scores indicating topological importance. Tumor log fold-change (logFC) reflects differential expression in tumor tissues, and prognostic relevance was evaluated using univariate Cox regression analysis in the TCGA cohort.*

low-risk patients survived for 71.6 months. The tri-gene signature which showed biological coherence through three different analytical pathways showed clinical independence through multivariate HR = 2.41 results and showed risk correlations with features of tumors that exhibit immunosuppressive microenvironments through M2 macrophages and TGF- $\beta$  signaling.

## Discussion

Lung adenocarcinoma (LUAD) is one of the many genetically and epigenetically different types of non-small cell lung cancer. Due to its genetic heterogeneity and dynamic interactions with the tumor microenvironment (extracellular matrix), LUAD often cannot be accurately predicted using conventional clinicopathological models for prognosis and treatment selection. This research uses a full multi-omics bioinformatics approach that includes differential expression analysis, weighted gene co-expression network analysis, protein-protein interaction network construction, machine learning feature selection and survival modeling, creation of nomograms and tumor immune micro-environment analysis to explore the molecular changes associated with LUAD disease development, and identify potential clinical biomarkers for LUAD. The multi-omics approach provided us with a prognostic signature of 3 genes (including TOP2A, CCNB1 and CDK1), and demonstrated superior results through its biological consistency across the many datasets analyzed.

The study of GSE75307 found that gene expression patterns showed substantial transcriptomic remodelling because downregulated genes accounted for 312 of 442 differentially expressed genes which showed a median logFC of -0.92. The essential pathway signals remained intact because our statistical methods-maintained consistency throughout all analytical stages which enabled a large proportion of differentially expressed genes to be captured within tumor-associated WGCNA modules. The functional enrichment analyses showed that extracellular matrix (ECM) disruption functions as the primary molecular signature through three main pathways: GO:0031012 extracellular matrix (adj.  $P = 1.2 \times 10^{-7}$ ), hsa04510 focal adhesion (adj.  $P = 7.8 \times 10^{-4}$ ), GSEA ECM-receptor interaction (NES = -1.92, FDR = 0.008). The findings demonstrate that LUAD hallmarks show consistency with established research because COL1A1/DCN/LUM loss leads to structural destruction which hampers integrin signaling (ITGA2/ITGB1) and Mechan transduction thus creating a permissive stromal niche for invasion and metastasis. The changes in the extracellular matrix not only remodel the tumor microenvironment but also provide an environment conducive to dysregulation of the cell cycle. The biological connection above makes a case for examining cell-cycle-related gene pathways in the next analysis.

The WGCNA module deconvolution process transformed ECM signals into tumor-specific co-expression modules which showed yellow module correlation of 0.90 and turquoise module correlation of 0.56 while revealing G2/M cell cycle reactivation as the second most important pathway. The PPI network analysis which contained 1,124 nodes and 28,456 edges and exhibited scale-free behaviour with a  $\gamma$  value of 2.1 discovered 17 hub genes through cytoHubba ranking which included Degree EPC MCC and MNC methods to operationalize the identified first-degree hub genes. A temporal biphasic pattern was observed, wherein these genes were downregulated in the GSE75307 discovery cohort (median logFC = -0.85) but exhibited strong prognostic significance in the independent TCGA-LUAD cohort (HRs > 1.87,  $P < 10^{-9}$ ), suggesting context- or stage-dependent reactivation during disease progression. The Phase 1 ECM collapse process enables G2/M checkpoint escape through coordinated TOP2A activation for chromosome segregation and CCNB1/CDK1 activation as mitotic promoting factor control and supporting spindle checkpoint regulators.

The three-algorithm machine learning ensemble—LASSO logistic regression ( $\lambda = 0.012$ ), SVM-RFE (ranking stability = 0.92), and random forest (permutation importance > 0.15)—achieved complete agreement across all three feature-selection methods, yielding the parsimonious yet powerful risk score: Risk Score =  $0.42 \times \log_2(\text{TOP2A} + 1) + 0.35 \times \log_2(\text{CCNB1} + 1) + 0.29 \times \log_2(\text{CDK1} + 1)$ . The external validation cohort of TCGA-LUAD ( $n = 515$ , different platform, center) had strong stratification and clinically useful Kaplan-Meier curve results (hazard ratio [HR] = 2.84 (95% confidence interval [95%CI]: 2.09 – 3.84)); there was a significant difference in overall survival (overall survival [OS]) between established high-risk and low-risk stratification (high-risk median OS = 28.4 months, low-risk median OS = 71.6 months). Multivariate Cox regression analyses confirmed that the risk score independently predicted prognosis after controlling for age, sex, stage, and smoking status; the HR was 2.41. The integrated nomogram for patients was also clinically useful based on these prediction models, with a C-index of 0.784 and a net benefit of 0.15 at the time of a 3-year decision. The three-gene signature demonstrates prognostic performance comparable to or exceeding many previously reported multi-gene signatures, while maintaining biological interpretability and enabling straightforward integration into clinical nomograms. CIBERSORTx immune deconvolution (LM22 signature) revealed high-risk tumors exhibit classic immune-excluded phenotypes: M2 macrophage dominance (22.1% vs 15.4%,  $P = 0.002$ ), reduced naïve B cells ( $P = 0.003$ ) and activated CD4+ memory T cells ( $P = 0.012$ ), elevated ESTIMATE immune/stromal scores ( $r = 0.51/0.48$ ). The ssGSEA pathway analysis identified strong correlations with FDR < 0.01 for these three pathways: pro-tumorigenic TGF- $\beta$  response ( $r = 0.62$ ), Wnt targets ( $r = 0.58$ ), angiogenesis

( $r = 0.55$ ), and immunosuppressive checkpoints PDCD1 ( $r = 0.43$ ), CTLA4 ( $r = 0.38$ ), LAG3 ( $r = 0.35$ ) together with the inverse interferon- $\gamma$  response ( $r = -0.49$ ). The cell cycle-immune exclusion axis describes how G2/M hyperproliferation produces lactate and ROS gradients to attract M2 macrophages, while TGF- $\beta$  stroma physically blocks effector T cells, which results in the development of an immune-suppressing microenvironment that continues to reinforce itself.

Therapeutic stratification according to risk leads to three major changes in medical practice which currently exist. The medical field should treat early-stage patients who exhibit high-risk features through adjuvant chemotherapy. The medical field should treat low-risk patients with decreased treatment intensity. The medical field uses nomogram systems to share medical information with medical professionals who use clinicopathology data for decision-making purposes. The medical field uses FDA-approved drugs as therapeutic options which include TOP2A inhibitors (etoposide Phase III NSCLC, doxorubicin) and CDK4/6 inhibitors (abemaciclib expanded indications). The drug combination uses three drug mechanisms which include cell cycle blockade and TGF- $\beta$  inhibition and immunotherapy.

The landmark research study determines TOP2A/CCNB1/CDK1 as a vital biomarker which shows the progression route from ECM to G2/M and immune exclusion in LUAD. The integrated computational system creates a biomarker discovery framework which researchers can apply to different types of cancer. The research team will focus on five research objectives which include validating the study results in multicenter trials with more than 1000 patients from various ancestries and using single-cell spatial transcriptomics to study ECM-macrophage-T cell interactions and testing CRISPR/Cas9 functions in organoid and PDX models and conducting pharmacogenomic screening (GDSC/CCLE) to find synthetic lethalties and using immunotherapy trial stratification (IMpower150/KEYNOTE-189 cohorts). The implementation of Shiny applications together with the commercial distribution of CLIA-validated panels and the addition of ATAC-seq and proteomics and spatial metabolomics technologies will increase our understanding of the clinical signal which supports cell cycle reactivation as a new area of research in LUAD precision oncology.

Future research should focus on validating the TOP2A/CCNB1/CDK1 prognostic signature in a large, multicenter, and multiethnic cohorts to demonstrate the clinical applicability and generalizability of the signature. The use of functional studies such as CRISPR/Cas9 gene editing, organoids, and patient-derived xenografts would be important to confirm the role of these genes in LUAD progression. Single-cell RNA sequencing, spatial transcriptomics, and further multi-omics integration could offer more insight into the connections between extracellular matrix remodeling, cell cycle regulation, and immune exclusion. Clinical assessment in patients undergoing immunotherapy could provide information about whether the signature would be able to predict the response to therapy and support in personalized treatment strategies.

## Strengths and Limitations

This study had several limitations that must be taken into consideration. The study used GSE75307 as its discovery cohort which contained two normal lung samples and ten tumor samples thus creating an unbalanced study design that needed statistical significance tests with multiple validation

methods because it could not use strict fold-change cutoffs which exceeded 1.5. Researchers used limma empirical Bayes moderation and WGCNA cross-validation to identify 85% of DEGs which existed in tumor modules but adjacent normal tissues showed field cancerization patterns instead of showing actual healthy lung parenchyma. The study used retrospective TCGA-LUAD cohort validation which included 515 participants from academic centers but required multicenter testing across different ancestries because the discovery dataset contained mostly Asian participants. Tumor immune microenvironment profiling relied on CIBERSORTx bulk deconvolution which estimates cellular fractions but does not provide direct spatial quantification—researchers need to conduct single-cell RNA-seq and multiplexed immunohistochemistry work to validate their findings. The three-gene signature demonstrates robust prognostic associations, but CRISPR/Cas9 studies in organoid/PDX models must confirm causality through functional tests. The clinical implementation process needs pre-analytical standardization together with FFPE-compatible assays which include RT-qPCR and NanoString, whereas bulk profiling results in loss of details about tumor diversity that can only be studied through single-cell atlas integration and spatial transcriptomics. Seven-layer validation for methodological constraint reduction was used, but prospective clinical trials are required to achieve proper research translation.

## Conclusion

The research establishes a predictive model for lung adenocarcinoma (LUAD) through the identification of a three-gene TOP2A/CCNB1/CDK1 prognostic signature which was obtained using their transcriptomics-based bioinformatics pipeline that included differential expression analysis and weighted gene co-expression network analysis and protein-protein interaction mapping and machine learning-based feature selection and clinical validation. The signature has a strong prognostic discrimination ability (hazard ratio = 2.84) and substantial differences in survival (28.4 months vs 71.6 months) between high risk and low risk patients, independent from established clinicopathologic factors. Biologically, these findings are consistent with a progression paradigm comprising ECM disruption, G2/M cell cycle reactivation, and immune exclusion. ECM downregulation of COL1A1, DCN, and LUM may contribute to stromal reorganization for proliferative escape as defined by TOP2A as related to chromosome segregation and CCNB1/CDK1 related to G2/M transition. This transcriptional program is associated with an immune-excluded tumor phenotype characterized by M2 macrophage predominance, TGF- $\beta$  pathway activation, and enhanced expression of immune checkpoints, suggesting a low level of antitumor immune response. Clinically, the incorporation of the tri-gene signature into a nomogram allows for the personalized prediction of risk, with proven clinical utility as evidenced by a concordance index of 0.784. This model may help to distinguish high-risk patients with early-stage disease for whom closer surveillance or therapeutic consideration may be warranted, without exposing low-risk patients to overtreatment. Overall, this integrated computational platform offers a generalizable solution for biomarker discovery in heterogeneous cancers with limited normal tissue availability. Future research efforts should be directed at prospective multicentre validation, single-cell and spatial transcriptomic analysis, functional validation, and assessment in immunotherapy-treated patients to further define its translational relevance in precision LUAD oncology.

**Author Contributions:** Conceptualization, M.K.; methodology, M.P., R.R. and M.K.; validation, M.K. and M.P.; formal analysis, M.P.; investigation, M.P. and R.R.; data curation, M.P., R.R., and R.N.; writing – original draft preparation, M.P. and R.N.; writing – review and editing, R.N. and M.P.; visualization, M.P. and K.M.; supervision, M.K., project administration, M.K.,; All authors have read and agreed to the published version of the manuscript.

**Disclosures:** The authors have no conflicts of interest.

**Acknowledgments:** None.

**Funding:** None.

**Data availability statement:** The data presented in this study are available on request from the corresponding author due to the protection of primary data.

**Ethical statement:** All information utilized in this research was sourced from publicly available data collections in the GEO database. No human subjects or animal participants were directly engaged, and ethical consent was not necessary.

**Declaration on the Use of Generative AI and AI-assisted Tools:** The authors declare that no generative AI or AI-assisted tools were used in the writing or preparation of this manuscript.

## References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021;71:209–49. <https://doi.org/10.3322/caac.21660>.
2. Ji Y, Zhang Y, Liu S, Li J, Jin Q, Wu J, Duan H, Liu X, Yang L, Huang Y. The epidemiological landscape of lung cancer: current status, temporal trend and future projections based on the latest estimates from GLOBOCAN 2022. *J Natl Cancer Cent* 2025;5:278–86. <https://doi.org/10.1016/j.jncc.2025.01.003>.
3. Priya Asokan L, A S, Kani V, Srinivasan C. Unlocking Precise Lung Cancer Detection Through Minimal Panel Immunostaining in Small Biopsy Samples. *Cureus* 2024. <https://doi.org/10.7759/cureus.63159>.
4. Li D, Shi J, Dong X, Liang D, Jin J, He Y. Epidemiological characteristics and risk factors of lung adenocarcinoma: A retrospective observational study from North China. *Front Oncol* 2022;12. <https://doi.org/10.3389/fonc.2022.892571>.
5. Rajinikanth V, Narayanan M. Lung Adenocarcinoma Detection in Histopathology Image with Fused Deep Features: A Study. 2025 9th Int. Conf. Inven. Syst. Control, IEEE; 2025, p. 1233–6. <https://doi.org/10.1109/ICISC65841.2025.11188136>.
6. Hernandez M, Ortiz RA, Neningen E, Amador RM, Cala M, Camacho K, Salomon E, Guerra PP, Mendoza I, Sanchez C, Viada C, Gonzalez M, Torres LT, Fonseca V, Garcia I, Bello L, Alonso M, Rodriguez MC, Cruz A, Valdes R, Lorenzo G, Cepeda M, Saumel Y, Crombet T. P2.40: CIMAvaxEGF Vaccine for the Treatment of Real-World NSCLC Patients. *J Thorac Oncol* 2016;11:S242. <https://doi.org/10.1016/j.jtho.2016.08.111>.
7. Didkowska J, Wojciechowska U, Mańczuk M, Łobaszewski J. Lung cancer epidemiology: contemporary and future challenges worldwide. *Ann Transl Med* 2016;4:150. <https://doi.org/10.21037/atm.2016.03.11>.
8. Kumar AP, Jagadeesh P. Innovative detection of lung cancer using decision tree classifier and comparison with K-nearest neighbor classifier, 2024, p. 020245. <https://doi.org/10.1063/5.0197445>.
9. Smolarz B, Łukasiewicz H, Samulak D, Piekarska E, Kołaciński R, Romanowicz H. Lung Cancer-Epidemiology, Pathogenesis, Treatment and Molecular Aspect (Review of Literature). *Int J Mol Sci* 2025;26. <https://doi.org/10.3390/ijms26052049>.
10. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014;511:543–50. <https://doi.org/10.1038/nature13385>.
11. Skoulidis F, Li BT, Dy GK, Price TJ, Falchook GS, Wolf J, Italiano A, Schuler M, Borghaei H, Barlesi F, Kato T, Curioni-Fontecedro A, Sacher A, Spira A, Ramalingam SS, Takahashi T, Besse B, Anderson A, Ang A, Tran Q, Mather O, Henary H, Ngarmchamnanrith G, Friberg G, Velcheti V, Govindan R. Sotorasib for Lung Cancers with KRAS p.G12C Mutation. *N Engl J Med* 2021;384:2371–81. <https://doi.org/10.1056/NEJMoa2103695>.
12. Pasello G, Lorenzi M, Scattolin D, Del Conte A, Cecere F, Pavan A, Macerelli M, Polo V, Pilotto S, Santarpia M, Cumerlato E, Da Ros V, Targato G, Bortolami A, Bonanno L, Ferro A, Dal Maso A, Frega S, Guarneri V. Osimertinib in Patients With Treatment-Naive EGFR -Mutant Non-small Cell Lung Cancer: Overall Survival, Post-progression Management and Budget Impact Analysis in Real-World. *Oncologist* 2024;29:596–608. <https://doi.org/10.1093/oncolo/oyae043>.
13. Qin K, Hong L, Zhang J, Le X. MET Amplification as a Resistance Driver to TKI Therapies in Lung Cancer: Clinical Challenges and Opportunities. *Cancers (Basel)* 2023;15:612. <https://doi.org/10.3390/cancers15030612>.
14. Rossi A. Relapsed small-cell lung cancer: platinum re-challenge or not. *J Thorac Dis* 2016;8:2360–4. <https://doi.org/10.21037/jtd.2016.09.28>.
15. Xu J, Zhang Y, Li M, Shao Z, Dong Y, Li Q, Bai H, Duan J, Zhong J, Wan R, Bai J, Yi X, Tang F, Wang J, Wang Z. A single-cell characterised signature integrating heterogeneity and microenvironment of lung adenocarcinoma for prognostic stratification. *EBioMedicine* 2024;102:105092. <https://doi.org/10.1016/j.ebiom.2024.105092>.
16. Song C, Guo Z, Yu D, Wang Y, Wang Q, Dong Z, Hu W. A Prognostic Nomogram Combining Immune-Related Gene Signature and Clinical Factors Predicts Survival in Patients With Lung Adenocarcinoma. *Front Oncol* 2020;10. <https://doi.org/10.3389/fonc.2020.01300>.
17. Ito K, Murphy D. Application of ggplot2 to Pharmacometric Graphics. *CPT Pharmacometrics Syst Pharmacol* 2013;2:1–16. <https://doi.org/10.1038/psp.2013.56>.
18. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43:e47–e47. <https://doi.org/10.1093/nar/gkv007>.

19. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;15:R29. <https://doi.org/10.1186/gb-2014-15-2-r29>.
20. Ping EP, Hossen J, Kiong WE. Data fusion based lane departure warning framework using fuzzy logic. *F1000Research* 2021;10:896. <https://doi.org/10.12688/f1000research.67209.1>.
21. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9:559. <https://doi.org/10.1186/1471-2105-9-559>.
22. Chang JT, Gatz ML, Lucas JE, Barry WT, Vaughn P, Nevins JR. SIGNATURE: A workbench for gene expression signature analysis. *BMC Bioinformatics* 2011;12:443. <https://doi.org/10.1186/1471-2105-12-443>.
23. Langfelder P, Luo R, Oldham MC, Horvath S. Is My Network Module Preserved and Reproducible? *PLoS Comput Biol* 2011;7:e1001057. <https://doi.org/10.1371/journal.pcbi.1001057>.
24. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu X, Liu S, Bo X, Yu G. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov (Cambridge)* 2021;2:100141. <https://doi.org/10.1016/j.xinn.2021.100141>.
25. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis 2016. <https://doi.org/10.1101/060012>.
26. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering C von. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;47:D607–13. <https://doi.org/10.1093/nar/gky1131>.
27. Chin C-H, Chen S-H, Wu H-H, Ho C-W, Ko M-T, Lin C-Y. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol* 2014;8 Suppl 4:S11. <https://doi.org/10.1186/1752-0509-8-S4-S11>.
28. Mohamed TIA, Ezugwu AE, Fonou-Dombeu JV, Mohammed M, Greeff J, Elbashir MK. A novel feature selection algorithm for identifying hub genes in lung cancer. *Sci Rep* 2023;13:21671. <https://doi.org/10.1038/s41598-023-48953-1>.
29. Khoshfekar Rudari H, Khorami-Sarvestani S, Fahrman JF, Long JP, Hanash S, Do K-A, Irajizad E. SMAGS-LASSO: A Novel Feature Selection Method for Sensitivity Maximization in Early Cancer Detection. *Cancer Epidemiol Biomarkers Prev* 2025;34:2259–66. <https://doi.org/10.1158/1055-9965.EPI-25-1032>.
30. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. CRAN Contrib Package 1999. <https://doi.org/10.32614/CRAN.package.e1071>.
31. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw* 2008;28. <https://doi.org/10.18637/jss.v028.i05>.
32. Zong Z, Li H, Ning Z, Hu C, Tang F, Zhu X, Tian H, Zhou T, Wang H. Integrative bioinformatics analysis of prognostic alternative splicing signatures in gastric cancer. *J Gastrointest Oncol* 2020;11:685–94. <https://doi.org/10.21037/jgo-20-117>.
33. Xu L, Zhao Y, Song S, Liu J, Li J, Zheng Z. Construction and validation of a lung adenocarcinoma prognostic model based on neutrophil extracellular traps and oxidative stress-related genes. *Eur J Med Res* 2025;30:1275. <https://doi.org/10.1186/s40001-025-03553-9>.
34. Liu L, He H, Peng Y, Yang Z, Gao S. A four-gene prognostic signature for predicting the overall survival of patients with lung adenocarcinoma. *PeerJ* 2021;9:e11911. <https://doi.org/10.7717/peerj.11911>.
35. Wen Z, Pei B, Dai L, Lu P, Li X, Zhang C, Ge S. Risk factors analysis and survival prediction model establishment of patients with lung adenocarcinoma based on different pyroptosis-related gene subtypes. *Eur J Med Res* 2023;28:601. <https://doi.org/10.1186/s40001-023-01581-x>.
36. Chai Y, Ma Y, Feng W, Xiang H, Lu H, Jin L. Identification and validation of a 4-extracellular matrix gene signature associated with prognosis and immune infiltration in lung adenocarcinoma. *Heliyon* 2024;10:e24162. <https://doi.org/10.1016/j.heliyon.2024.e24162>.
37. Xiao M, Li Y, Zhou Y, Liu X, Tang G. A novel anoikis-related gene signature predicts in lung adenocarcinoma. *Intell Pharm* 2024;2:161–72. <https://doi.org/10.1016/j.ipha.2023.10.013>.
38. Zhou Y, Gao S, Yang R, Du C, Wang Y, Wu Y. Identification of a three-gene expression signature and construction of a prognostic nomogram predicting overall survival in lung adenocarcinoma based on TCGA and GEO databases. *Transl Lung Cancer Res* 2022;11:1479–96. <https://doi.org/10.21037/tlcr-22-444>.
39. Wang G, Liu X, Liu H, Zhang X, Shao Y, Jia X. A novel necroptosis related gene signature and regulatory network for overall survival prediction in lung adenocarcinoma. *Sci Rep* 2023;13:15345. <https://doi.org/10.1038/s41598-023-41998-2>.
40. Zhang Z, Kattan MW. Drawing Nomograms with R: applications to categorical outcome and survival data. *Ann Transl Med.* 2017;5:211. <https://doi.org/10.21037/atm.2017.04.01>.
41. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.* 2015;12:453–7. <https://doi.org/10.1038/nmeth.3337>.
42. Haynes W. Benjamini–Hochberg Method. *Encycl. Syst. Biol.*, New York, NY: Springer New York; 2013, p. 78–78. [https://doi.org/10.1007/978-1-4419-9863-7\\_1215](https://doi.org/10.1007/978-1-4419-9863-7_1215).
43. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J Am Stat Assoc.* 1958;53:457–81. <https://doi.org/10.1080/01621459.1958.10501452>.
44. Li J, Ma S. Time-dependent ROC analysis under diverse censoring patterns. *Stat Med.* 2011;30:1266–77. <https://doi.org/10.1002/sim.4178>.