scientific reports

Check for updates

OPEN

Identification and verification of immune and oxidative stress-related diagnostic indicators for malignant lung nodules through WGCNA and machine learning

Zhou An[1], Meichun Zeng[2] & Xianhua Wang[3]✉

Early detection of lung nodules (LNs) is critical for prevention and treatment of lung cancer. However, current noninvasive diagnostic methods face significant challenges in reliably distinguishing benign from malignant nodules. Thus, there is an urgent need for novel molecular biomarkers or pathways to facilitate accurate identification of truly malignant LNs. Using the Gene Expression Omnibus (GEO) database and the "limma" package, we identified differentially expressed genes (DEGs) in lung nodules (LNs) by comparing benign and malignant samples. The oxidative stress-related genes were downloaded from the GenCards database. Subsequently, genes associated with immunity and oxidative stress were analyzed using weighted gene co-expression network analysis (WGCNA). A protein–protein interaction (PPI) network was constructed and hub genes were extracted using 12 centrality-based algorithms in the CytoHubba plugin. Shared DEGs from these analyses were subjected to functional enrichment analysis. To develop a diagnostic model for LNs, we investigated 113 combinations of 12 machine-learning algorithms, employing 10-fold cross-validation on the training set, followed by external validation of the test set. A total of 31 shared differentially expressed genes associated with immunity and oxidative stress were identified, including two hub genes, CDK2 and MCL1. Immune infiltration analysis revealed distinct patterns of immune cell infiltration in malignant LNs compared to those in benign controls. A promising 11-gene diagnostic signature was developed, which exhibited superior performance to existing LNs diagnostic models in both training and testing cohorts. This study developed a diagnostic model for malignant LNs, focusing on the shared genes associated with immunity and oxidative stress pathways. Furthermore, the identified hub genes facilitate a deeper understanding of the pathobiological mechanisms underlying the different types of LNs.

Keywords Lung nodules (LNs), Lung cancer, WGCNA, Machine learning, Immune cell infiltration, Diagnostic indicators, Oxidative stress, Immune

Lung cancer is one of the most prevalent and lethal malignancies worldwide, imposing a significant burden on global public health[1]. The early detection of lung cancer is critical for improving patient survival rates and treatment outcomes[2]. Lung nodules (LNs), which are often the initial indicators of possible lung cancer, require early and accurate identification and classification as essential steps in disease management[3].

Early detection of lung nodules (LNs) is critical for prevention and treatment of lung cancer. However, the current detection and diagnostic methods pose significant challenges[4]. Conventional imaging techniques such as computed tomography (CT) can identify nodules, but distinguishing between benign and malignant lesions remains challenging[5]. Experienced radiologists may offer subjective interpretations, potentially leading to misdiagnosis. For example, some benign nodules may mimic malignant lesions on CT scans, while malignant nodules can resemble benign ones[6].

[1]Department of Thoracic Surgery, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, Zhejiang, China. [2]Department of Pathology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, Zhejiang, China. [3]Department of Endoscope, Hangzhou Hospital of Traditional Chinese Medicine, 453 Stadium Road, Hangzhou 310007, Zhejiang, China. ✉email: xhwang1212@163.com

To address these challenges, there is an urgent need to explore novel molecular pathways that can aid in identifying malignant nodules. Recent studies have sought to develop and validate machine-learning models for analyzing pathological data in lung cancer patients[7]. Furthermore, understanding the molecular mechanisms underlying LNs formation and progression may offer critical insights for creating more accurate diagnostic tools. Recently, the roles of immune and oxidative stress-related pathways in various diseases, including cancer[8], have been increasingly recognized.

Oxidative stress arises from an imbalance between reactive oxygen species (ROS) production and antioxidant defenses[9]. This process contributes to the pathogenesis of lung cancer[10] and other lung diseases[11]. Excessive ROS induce damage to cellular components, including DNA, proteins, and lipids, leading to genomic instability and aberrant cell growth, hallmarks of cancer[12,13]. A recent study by Barsan et al. demonstrated that oxidative stress is closely associated with the development of nonfibrotic cellular LNs. While curcumin mitigates the effects of oxidative stress on alveolar epithelia, it fails to prevent nodule-related vascular and airway remodeling[14].

The immune system is central to cancer cell surveillance and elimination. Numerous studies have reported aberrant immune cell infiltration and cytokine production within the LNs microenvironment[15,16]. For example, Qu et al. demonstrated that the presence of a specific subset of immune cells in LNs is associated with a more aggressive tumor phenotype[17].

Bioinformatic approaches, including weighted gene co-expression network analysis (WGCNA) and machine learning algorithms, have become powerful tools for analyzing large-scale gene expression datasets and identifying key genes and pathways implicated in diseases. WGCNA identifies co-expressed gene modules under specific conditions and offers insight into functional gene relationships[18]. Conversely, machine learning algorithms construct predictive models from complex data patterns, enabling high-accuracy classification of lung nodules (LNs) as benign or malignant[19].

In this study, we integrated WGCNA and machine learning techniques to identify genes associated with immunity and oxidative stress that could serve as diagnostic biomarkers for malignant LNs. By analyzing gene expression data from publicly available repositories, we sought to develop a promising diagnostic model for distinguishing between benign and malignant LNs. This approach has the potential to enhance the early detection of malignant LNs, enabling timely and effective treatment strategies for patients at risk of developing lung cancer.

Materials and methods

Source of data

Two LNs datasets were selected for analysis. The GSE108375 dataset (Platform: GPL10558) comprised 164 malignant nodule cases and 151 benign nodule control samples[20]. The GSE20189 dataset (Platform: GPL571), used as a validation set, included 81 malignant nodule cases and 18 benign nodule controls[21]. Additionally, 817 oxidative stress (OS)-related genes were identified from the GeneCards database (https://www.genecards.org) using the keyword "oxidative stress" and a screening threshold with a relevance score of ≥ 7, following previously described methods[22].

Batch effect removal

We merged the two LNs datasets (GSE108375 and GSE20189) and employed the "ComBat" function from the "sva" package (v. 3.52.0, https://code.bioconductor.org/browse/sva/) to correct for batch effects. To assess the efficacy of this correction, we compared data quality before and after batch removal using principal component analysis (PCA).

Identification of differentially expressed genes (DEGs)

We established the screening criteria of $|\log_2$ fold-change (FC)$| > 0.1$ and false discovery rate (FDR) < 0.05. Using the "limma" R package (v.3.60.6, https://code.bioconductor.org/browse/limma/), differentially expressed genes (DEGs) were identified in the GSE108375 dataset. The "pheatmap" (v.1.0.12, https://CRAN.R-project.org/package=pheatmap) and "ggplot2" (v.3.5.2, https://ggplot2.tidyverse.org) R packages were used to generate a heatmap and volcano plot, respectively, to visualize the identified DEGs. We then intersected these DEGs with 817 oxidative stress-related genes, enabling the identification of oxidative stress-associated DEGs.

Immune cell infiltration analysis of these DEGs

Single-sample gene set enrichment analysis (ssGSEA) was used to quantify the abundance of distinct types of infiltrating immune cells within each sample[23]. These cells were classified into three primary categories: (1) lymphoid lineage cells, including activated lymphocytes (e.g., activated B cells, activated $CD4^+$ T cells, and activated $CD8^+$ T cells), regulatory T cells (Tregs), follicular helper T cells, T helper 1 (Th1), Th17, and Th2 cells, gamma delta T cells, natural killer T (NKT) cells, and natural killer (NK) cells (further subdivided into $CD56^{bright}$ and $CD56^{dim}$ subsets). Immature B cells also belong to this lymphoid-derived group; (2) myeloid lineage cells, including macrophages, monocytes, neutrophils, eosinophils, and dendritic cells (both activated and immature forms, including plasmacytoid dendritic cells); and (3) other immune-related cells, such as myeloid-derived suppressor cells (MDSCs) and mast cells, which complete the immune cell infiltration profile.

Construction of weighted gene co-expression networks

CIBERSORT employs a deconvolution algorithm to infer the composition and abundance of immune cell subtypes within cellular mixtures, using transcriptomic data. In this study, we first used the CIBERSORT algorithm to assess the fractional distribution of immune cell types in benign and malignant LNs samples from the GSE108375 dataset[24].

Weighted Gene Co-expression Network Analysis (WGCNA) was performed using the R package "WGCNA" (v.1.73, https://CRAN.R-project.org/package=WGCNA) to identify the gene modules most strongly associated with clinical phenotypes and immune cell profiles in patients[25]. The sample data were pre-processed to remove outliers. The "WGCNA" package generates a correlation matrix, which is transformed into an adjacency matrix using an optimally determined soft threshold. From this adjacency matrix, a topological overlap matrix (TOM) was constructed and a TOM-based dissimilarity metric was used to cluster genes with similar expression patterns into modules via average linkage hierarchical clustering. The two modules most strongly correlated with clinical phenotypes and immune cell signatures were identified as key modules for subsequent analyses.

Finally, genes from the differentially expressed oxidative stress-related genes and key modules were intersected, and the overlapping genes were designated as differentially expressed immune and oxidative stress-related genes (DEIOGs) for in-depth investigation.

Functional enrichment characterization of shared differentially expressed immune and oxidative stress-related genes (DEIOGs)

To elucidate the biological roles and underlying mechanisms of the shared genes, we performed Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, Gene Set Enrichment (GSEA), and Disease Ontology Semantic and Enrichment (DOSE) enrichment analyses. These analyses were conducted using the "org.Hs.eg.db" (v.3.19.1, https://bioconductor.org/packages/org.Hs.eg.db/), "ggplot2", "clusterProfiler" (v.4.12.6, https://code.bioconductor.org/browse/clusterProfiler/), and "enrichplot" (v.1.24.4, https://code.bioconductor.org/browse/enrichplot/) libraries in the R computing environment (v. 4.4.2, https://cran.r-project.org/bin/macosx/). Statistical significance was determined by applying a threshold of $P < 0.05$ to identify enriched functional categories.

Protein–protein interaction (PPI) networks construction and hub genes identification of DEIOGs

The STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins, https://cn.string-db.org/) was used to construct the protein–protein interaction (PPI) network[26]. The PPI network was built from the differentially expressed immune and oxidative stress-related genes (DEIOGs) identified using the STRING interactome. Twelve algorithms in the cytoHubba plugin (v.3.10.1, https://apps.cytoscape.org/apps/cytohubba) were used to assess the network structure of the DEIOGs. For each method, top-ranked genes (10 per method) were selected and overlapping candidates were identified to establish a set of consensus genes. Genes identified through both intersection analysis and cytoHubba's ranking system were designated hub genes[27].

Machine learning algorithms

To develop the diagnostic model, we employed 12 machine-learning algorithms: naive Bayes, support vector machine (SVM), generalized linear model boosting (glmBoost), linear discriminant analysis (LDA), ridge, lasso, stepglm, partial least squares regression for generalized linear models (plsRglm), generalized boosted regression modeling (GBM), elastic net (Enet), random forest (RF), and XGBoost. We systematically explored 113 combinations of the algorithms.

Machine learning techniques were used to conduct comprehensive dataset analysis. Preprocessing steps included removing incomplete entries and extreme outliers to ensure data quality, followed by Z-score normalization to standardize features (mean = 0, standard deviation = 1), and addressing dimensional disparities among variables. The dataset was randomly partitioned into a 70% training set for model development and a 30% test set for independent validation, enabling assessment of model generalizability.

During the algorithm development, predictive models were tested for diagnostic performance, including elastic net regression ($\lambda = 0.1$), LASSO regression ($\lambda = 0.05$), ridge regression ($\lambda = 1.0$), SVM ($\gamma = 0.01$), LDA, GBM (learning rate = 0.1, 100 trees), RF (200 decision trees), and XGBoost (learning rate = 0.01, 150 estimators). Each model was iteratively trained on the training subset with hyperparameters optimized via k-fold cross-validation to enhance robustness and mitigate overfitting.

In the final evaluation, classification metrics were quantified using the test cohort. The area under the curve (AUC) scores (cutoff ≥ 0.7) were computed for each algorithm, providing a standardized discriminative measure. A heatmap generated via the "pheatmap" package visualized the model rankings, with the optimal model identified by the highest AUC. Calibration curves were constructed to assess the agreement between predicted probabilities and observed outcomes, ensuring diagnostic reliability.

The GSE20189 cohort served as the external validation dataset. Following the established criteria, the model with the highest AUC across the training and test cohorts was designated optimal[28].

Calibration curves were used to evaluate the model's predictive probability accuracy, and the DeLong test was used to compare its diagnostic performance with previously published LNs diagnostic models[29,30]. Statistical significance was defined as $P < 0.05$.

Statistical analysis

To evaluate baseline differences, categorical variables were analyzed using chi-squared (χ^2) tests, while continuous variables were analyzed using one-way ANOVA. All statistical computations were performed using R (v.4.4.2; R Foundation for Statistical Computing, https://www.r-project.org). A two-tailed significance threshold of $\alpha = 0.05$ (two-tailed) was applied across all tests. Kaplan–Meier survival analysis was used to evaluate differences in pack-years between groups. Calibration curves were used to assess the agreement between the model-predicted and actual outcomes.

Results

Identification of differentially expressed genes (DEGs) in benign and malignant lung nodules (LNs)

A flowchart of the study design is shown in Fig. 1. To identify pathogenic genes associated with benign and malignant LNs, "limma" package was performed. In the benign vs. malignant LNs cohort, 1,432 DEGs were identified, including 857 upregulated and 575 downregulated genes (Fig. 2A, B) (Supplementary Table S1 and S2).

Data preprocessing of the two LNs datasets

Raw malignant LNs and control transcriptomic data were obtained from Gene Expression Omnibus (GEO), integrated after batch effect removal, and normalized to generate 245 malignant nodule cases and 232 benign nodule control-processed cohorts. Fig. S confirmed that batch effects were significantly mitigated after correction (Supplementary Fig. S1).

Immune infiltration analysis of DEGs

The boxplot distributions in Fig. 2C indicate elevated proportions of mast cells, natural killer T cells, neutrophils, and plasmacytoid dendritic cells in the malignant LNs cohort. Conversely, activated B cells, activated CD4 + T cells, eosinophils, and natural killer cells showed reduced abundance compared to benign controls (Fig. 2C).

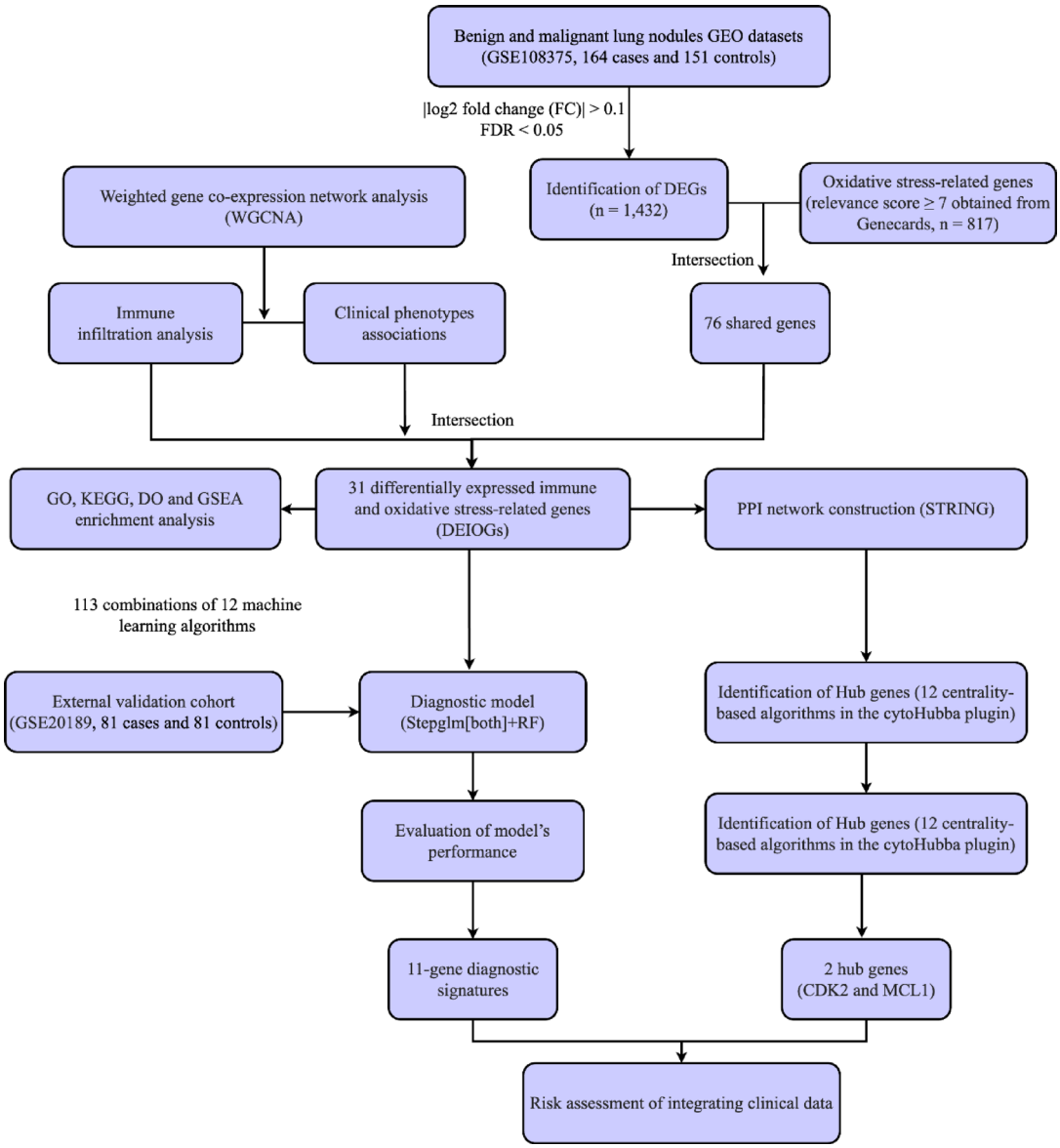

Fig. 1. Study design flowchart.

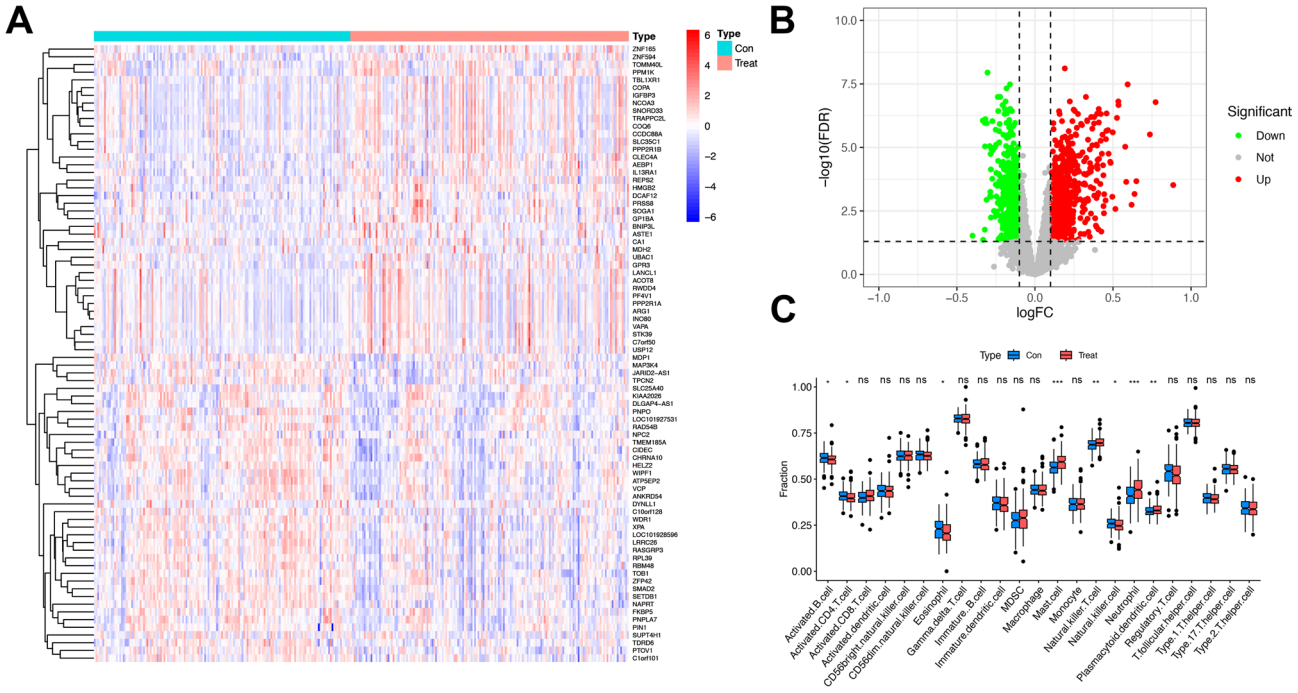

Fig. 2. Identification of differentially expressed genes (DEGs) and immune cell profiles in lung nodules (LNs). (A) Heatmap of DEGs in the GSE108375 cohort. (B) Volcano plot of DEGs ($|\log_2 FC| > 0.1$, FDR < 0.05). (C) Boxplots comparing immune cell abundances between malignant and benign LNs. Statistical significance: $^{***}P < 0.001$, $^{**}P < 0.01$, $P < 0.05$, ns = non-significant. Con, benign lung nodules; Treat, malignant lung nodules; DEGs, differentially expressed genes; LNs, lung nodules.

Construction of weighted gene co-expression networks (WGCNA)

WGCNA was used to explore the associations between clinical phenotypes and gene expression using the GSE108375 dataset. After batch effect correction, cluster analysis grouped patients based on their gene expression profiles, with each cluster representing an independent sample (Fig. 3A). Soft-thresholding power ($\beta = 5$) was selected via scale-free topology analysis (SFT), achieving an R^2 fit index of 0.85 (Fig. 3B) and identifying 36 distinct modules (Fig. 3C).

Three modules (light yellow, dark green, and brown) exhibited significant positive correlations with malignant LNs, whereas the sky-blue module showed a strong negative correlation. These four modules were prioritized for further analysis (Supplementary Table S3).

WGCNA was used to assess the relationship between immune cell infiltration profiles and transcriptomic patterns. Using soft-threshold power ($\beta = 5$) (Fig. 4A), nine gene modules were identified (Fig. 4B, C). The green and pink modules are associated with $CD56^{bright}$ natural killer cells, macrophages, plasmacytoid dendritic cells, regulatory T cells, and Th17/Th2 helper cells. The yellow module was correlated with immature B cells, macrophages, monocytes, natural killer cells, neutrophils, regulatory T cells, and Th1 helper cells. The turquoise module was linked to activated B cells, CD8 + T cells, eosinophils, gamma-delta T cells, myeloid-derived suppressor cells (MDSCs), mast cells, natural killer T cells, and neutrophils (Fig. 4C). Green, pink, yellow, and turquoise modules were selected for downstream analysis (Supplementary Tables S2 and S4).

Functional enrichment of differentially expressed immune and oxidative stress-related genes (DEIOGs)

As shown in Fig. 5A, DEIOGs were derived from benign/malignant LNs DEGs, WGCNA modules associated with clinical phenotypes and immune infiltration, and 817 oxidative stress markers were obtained from GeneCards (Supplementary Table S5). The integration of these approaches yielded a final set of 31 DEIOGs.

GO enrichment analysis revealed significant biological processes, including positive regulation of mitochondrial organization, autophagy, and neuronal apoptotic processes. Enriched cellular components include the mitochondrial matrix, Bcl-2 family protein complex, and the external side of the plasma membrane. The molecular functions included tau protein binding and NAD(P) + nucleosidase activity (Fig. 5B). The KEGG pathways included colorectal cancer, endometrial cancer, HIV-1 infection, and lipid metabolism/atherosclerosis (Fig. 5C). Disease Ontology analysis showed that DEIOGs were associated with mitochondrial metabolism disorders, oxidative phosphorylation deficiency, colon cancer, and nephritis (Fig. 5D). GSEA highlighted the enrichment for the CDC42 GTPase cycle, chromatin organization, Fcγ receptor-dependent phagocytosis, and carbohydrate metabolism (Fig. 5E).

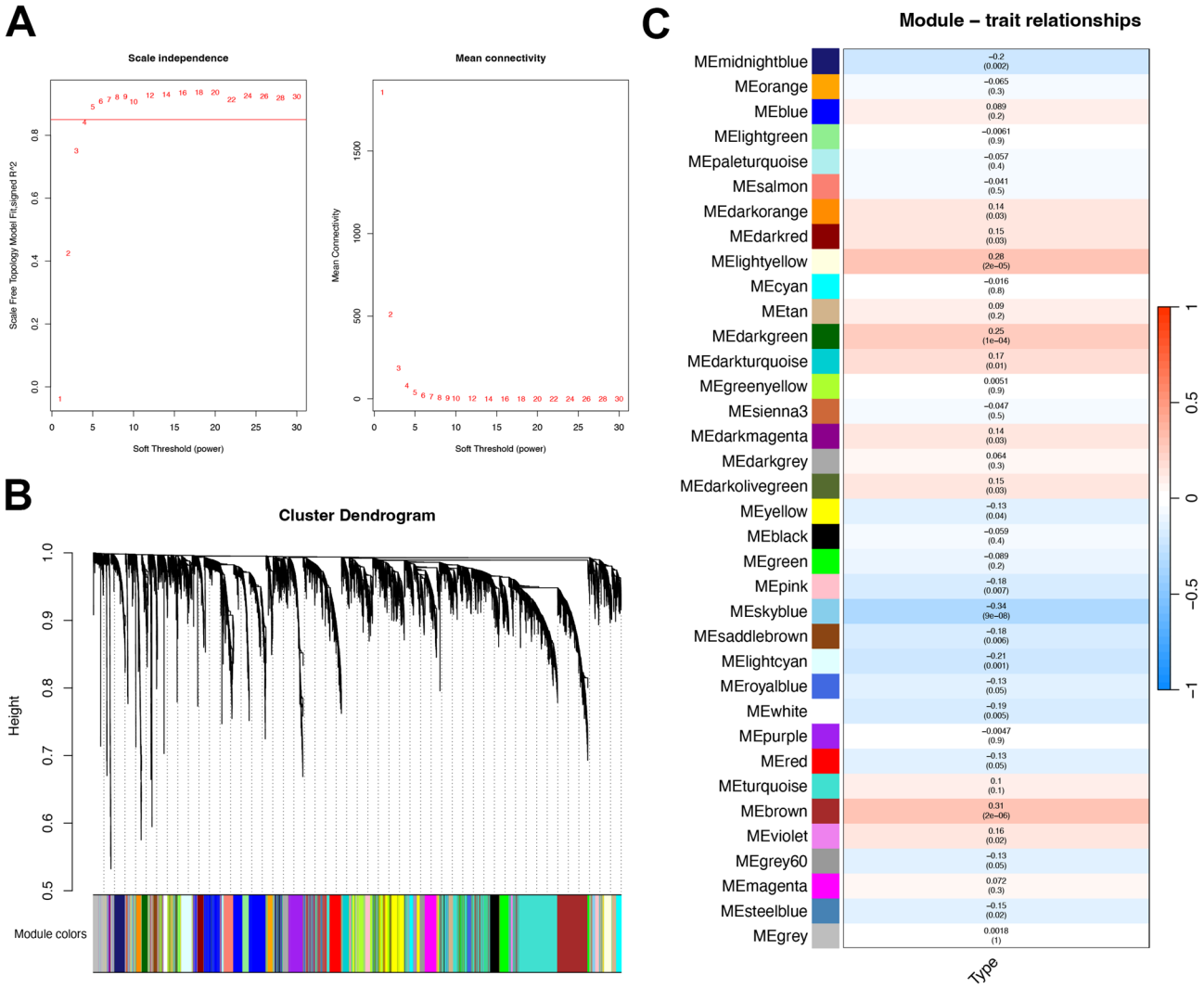

Fig. 3. Weighted gene co-expression network analysis (WGCNA) of lung nodules (LNs). (A) Determination of the soft-thresholding power (β = 5) for scale-free network construction. (B) Dendrogram of gene clustering and module merging. (C) Heatmap of correlations between module eigengenes and LNs malignancy. WGCNA, weighted gene co-expression network analysis.

Development of a diagnostic model for malignant LNs via machine learning

Using 31 DEIOGs, a 10-fold cross-validation framework was used to evaluate 12 machine learning algorithms across the training dataset (GSE108375) and external validation dataset (GSE20189) (Fig. 6A-D). The top-performing model combined Stepglm[both] and Random Forest (RF) to identify 11 key genes: IL1R1, GADD45 A, SMAD2, AHSP, ENDOG, CLEC4 A, SHC1, SDHAF1, PXN, SUOX, and BAD. Calibration curves (Fig. 6E) showed strong agreement between the predicted probabilities and observed outcomes in both cohorts, indicating robust calibration.

Compared to the existing models (Guan's and Nishio's), our approach demonstrated superior diagnostic performance, as measured by AUC, in both the training and testing datasets (Fig. 6F). Transcriptomic profiling revealed that seven genes (SDHAF1, GADD45 A, AHSP, CLEC4 A, SHC1, SUOX, and IL1R1) were upregulated in malignant LNs (MN), while ENDOG, BAD, and SMAD2 were downregulated (all $P < 0.05$) (Fig. 6G). BAD expression varied significantly across MN pathological stages ($P < 0.05$) (Fig. 6H), underscoring its role in pulmonary malignancy progression.

Hub gene screening and clinical correlation analysis

The STRING database constructed a protein-protein interaction (PPI) network for the 31 DEIOGs. GSK3B exhibited the highest node and edge connectivity, followed by TLR4 and CASP9 (Fig. 7A). Using the cytoHubba plugin, two hub genes (CDK2 and MCL1) were identified using multiple algorithms (Fig. 7B) (Supplementary Table S6).

Kaplan-Meier survival analysis showed that CDK2 overexpression correlated with longer packyears (10×) (survival curve above the low-expression group, $P = 0.012$; Fig. 7C), while MCL1 overexpression was associated

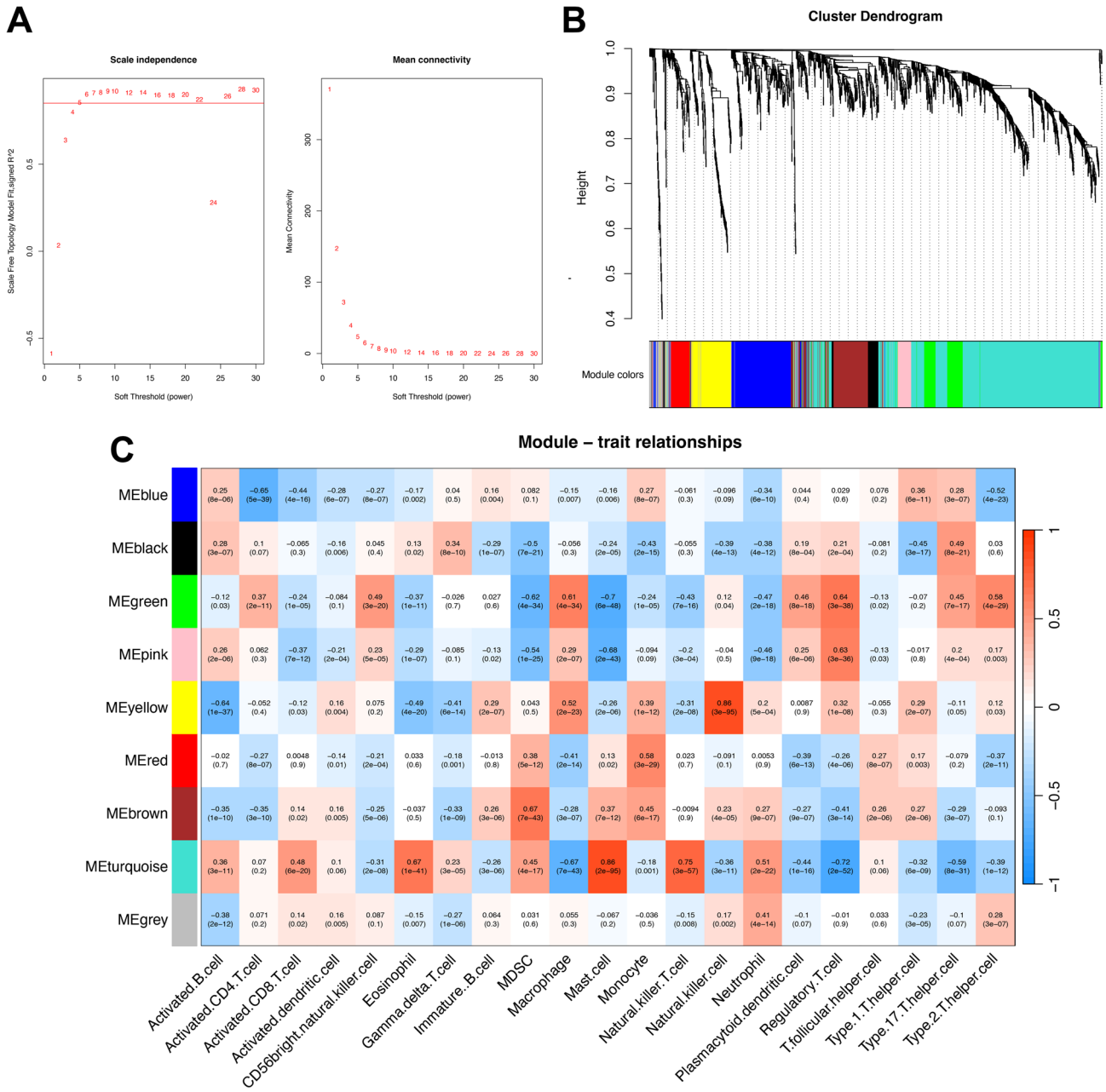

Fig. 4. Construction of immune infiltration-related weighted gene co-expression networks (WGCNA). (A) Selection of the optimal soft-thresholding power (β = 5) for immune cell profile correlation. (B) Dendrogram of gene clustering and module merging for immune-related modules. (C) Heatmap displaying nine gene modules identified by WGCNA. WGCNA, weighted gene co-expression network analysis.

with shorter packyears (survival curve below the low-expression group, $P = 0.018$; Fig. 7D). A clinical relevance heatmap (Fig. 7E), stratifying patients by LNs size (≥ 10 mm as high risk), clustered 13 genes (11 DEIOGs + 2 hub genes) into two groups: one with three downregulated genes and CDK2 (low-risk enrichment) and another with seven upregulated genes and MCL1 (high-risk enrichment). Expression differences were consistent across stage and type groups, linking these genes to LNs size and biological behavior.

Discussion

The early detection of malignant lung nodules (LNs) remains a critical challenge in lung cancer management, as current imaging techniques often struggle to differentiate benign from malignant lesions[31]. This study integrated weighted gene co-expression network analysis (WGCNA) and machine learning to identify genes associated with immunity and oxidative stress as diagnostic biomarkers for malignant LNs. The 11-gene signature derived from this approach exhibited promising performance across both the training and validation cohorts, outperforming

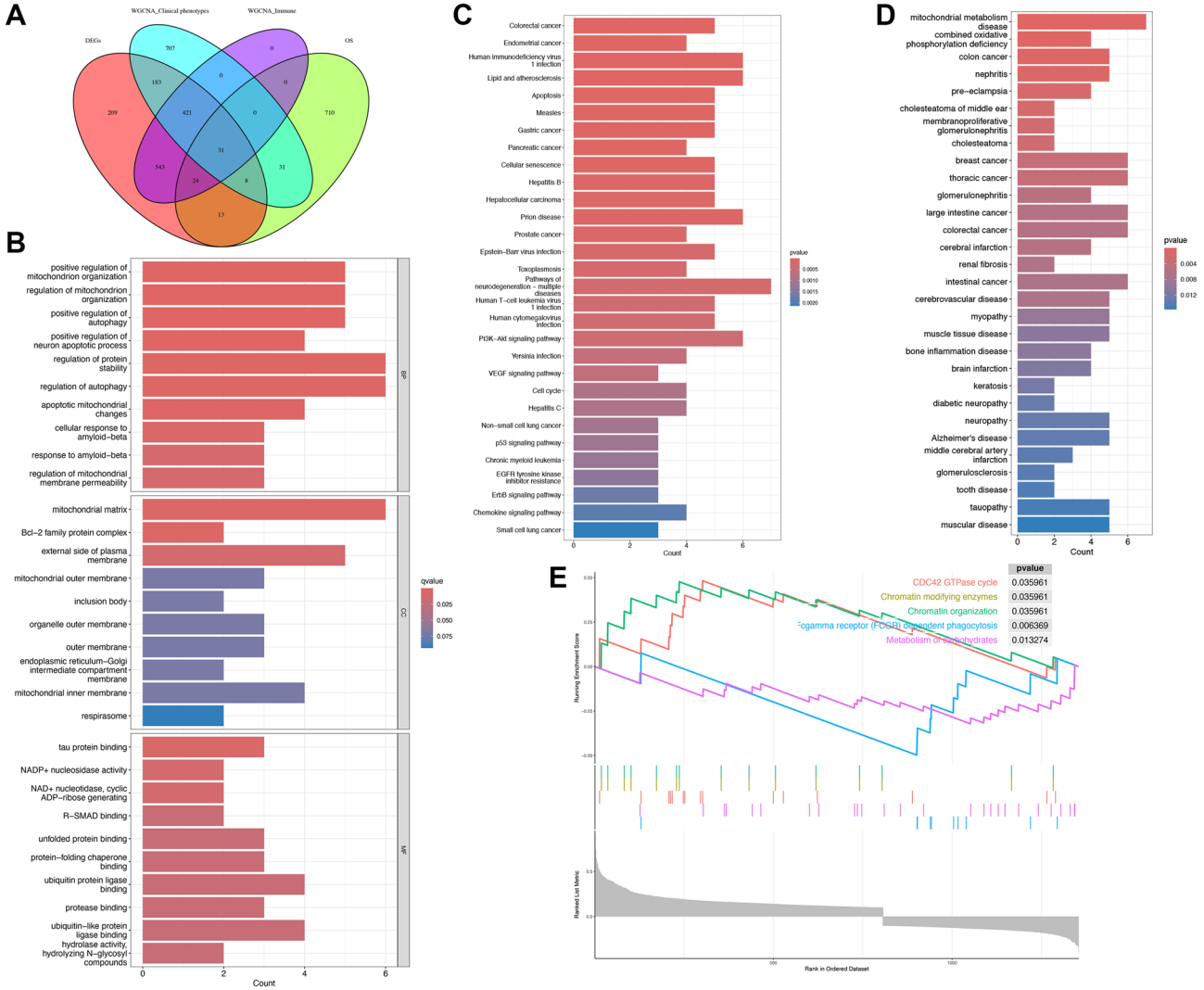

Fig. 5. Identification and functional enrichment of differentially expressed immune and oxidative stress-related genes (DEIOGs). (A) Venn diagram depicting the integration of DEGs, WGCNA modules, and oxidative stress genes to identify DEIOGs. (B) Bar plots of Gene Ontology (GO) enrichment for biological processes, cellular components, and molecular functions. (C) Bar plots of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment. (D) Bar plots of Disease Ontology (DO) enrichment. (E) Gene Set Enrichment Analysis (GSEA) showing significantly activated pathways in malignant vs. benign LNs. DEIOGs, differentially expressed immune and oxidative stress-related genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DO, Disease Ontology; GSEA, Gene Set Enrichment Analysis; LNs, lung nodules.

existing diagnostic models. These findings underscore the potential of integrating bioinformatics and machine learning to enhance LNs diagnosis and classification.

The 31 identified immune and oxidative stress-related genes were enriched in pathways pivotal for cancer progression, including cytokine signaling, lipid metabolism, and mitochondrial metabolic disorders. This finding aligns with emerging evidence linking oxidative stress to genomic instability and immune evasion in lung cancer[12,32]. Recent investigations have shown that PI3 K/Akt and p53 signaling pathways are associated with mitochondrial oxidative stress in lung cancer cells[33,34]. Enrichment of interleukin signaling[35] and mitochondrial dysfunction[36] among these genes further highlights their role in tumor microenvironment (TME) remodeling. Immune infiltration analysis revealed distinct profiles of T cell and macrophage subsets in malignant lung nodules (LNs), supporting the hypothesis that immune dysregulation contributes to nodule malignancy[37].

Weighted gene co-expression network analysis (WGCNA) identified modules that strongly correlated with clinical phenotypes and immune cell infiltration. These modules overlapped with oxidative stress-related differentially expressed genes (DEGs), suggesting crosstalk between immune activation and redox imbalance[38,39]. Previous studies have highlighted the role of mast cells in promoting angiogenesis[40] and immunosuppression[41] during cancer progression. Moreover, eosinophils may serve as prognostic indicators in lung cancer patients[42].

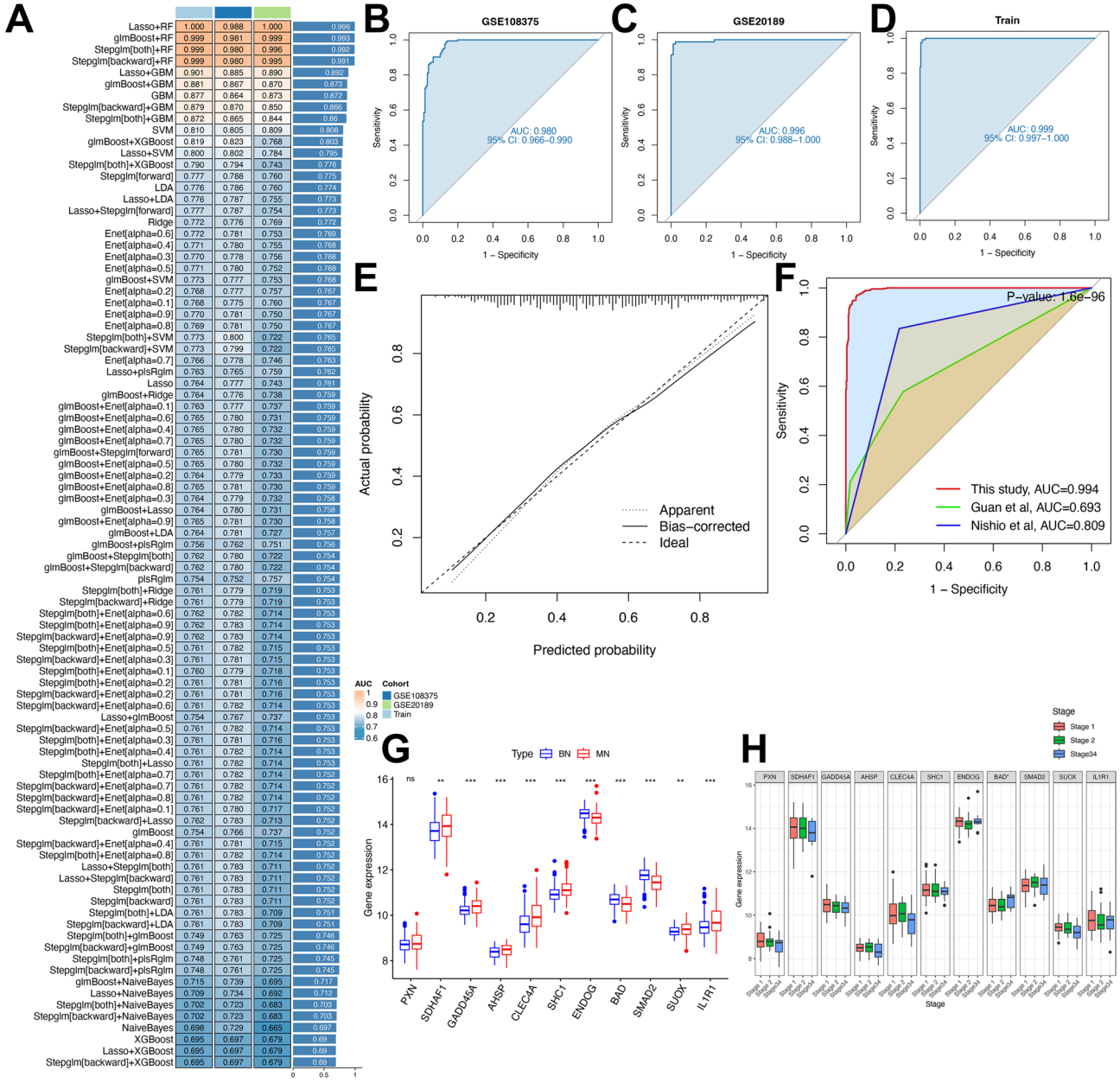

Fig. 6. Diagnostic performance evaluation of the 11-gene signature model. (A) Schematic of 113 machine learning algorithm combinations evaluated via 10-fold cross-validation. The three cohorts were presented from left to right as follows: the training cohort (derived from 70% GSE108375), GSE108375, and an external validation cohort GSE20189. The average AUC scores of these three cohorts were sorted in ascending order. (B–D) Receiver operating characteristic (ROC) curves for the GSE108375 cohort, external validation GSE20189 cohort, and training cohort (derived from 70% GSE108375). (E) Calibration curve demonstrating agreement between predicted probabilities and observed outcomes. (F) ROC curves comparing the proposed model with existing LNs diagnostic models. (G) Expression levels of the 11-gene signature in blood samples from malignant (MN) vs. benign (BN) lung nodules. (H) Expression levels of the 11-gene signature across pathological stages in malignant lung nodules. Statistical significance: ***$P < 0.001$, **$P < 0.01$, $P < 0.05$, ns = non-significant. BN, benign lung nodules; MN, malignant lung nodules.

Thus, the WGCNA approach offers a systems-level view of gene co-expression networks[43], enabling the prioritization of genes with dual roles in immunity and oxidative stress.

The stepglm[both] + RF model exhibited superior diagnostic accuracy compared to existing methods, including Guan's[30] and Nishio's models[29]. This performance can be ascribed to the ability of the algorithm to capture the nonlinear relationships between genes and clinical outcomes. The calibration curves showed a strong agreement between the predicted probabilities and observed outcomes, which is crucial for validating its clinical utility.

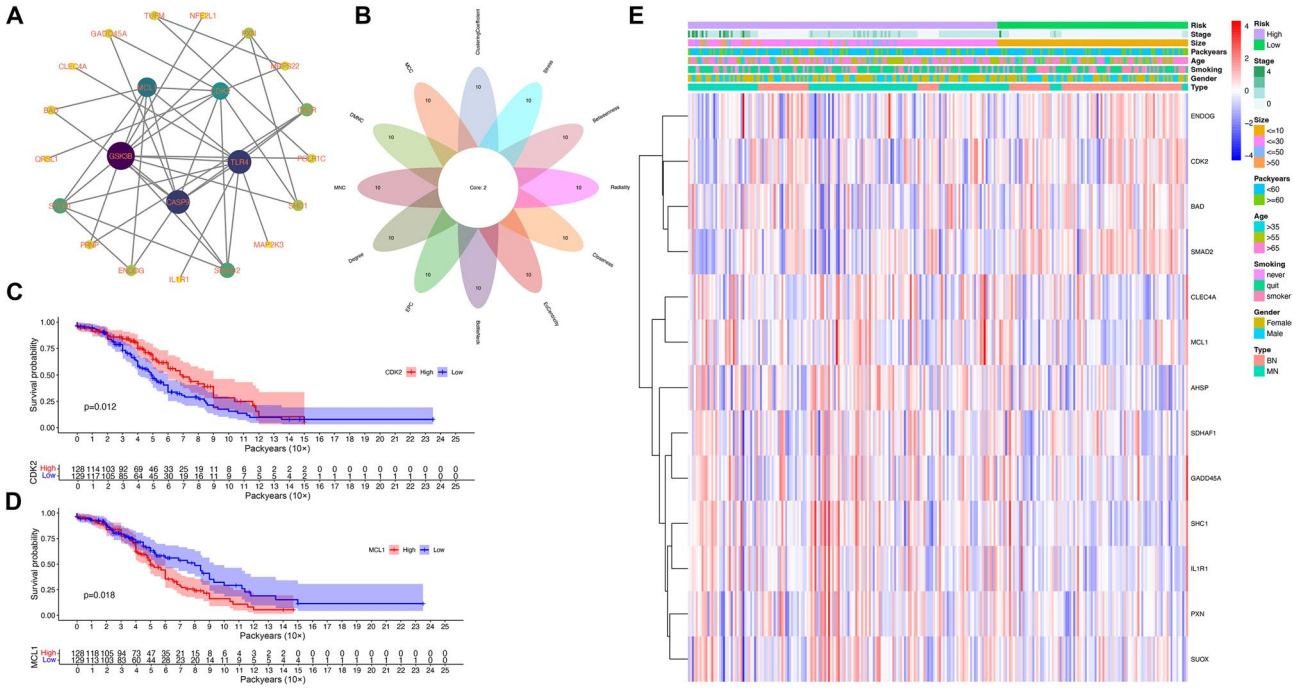

Fig. 7. Hub gene identification via protein-protein interaction (PPI) networks and clinical correlation analysis. (A) PPI network of 31 DEIOGs. Each node represents a gene, and different sizes and color indicate the degree of centrality of the gene in the PPI network, while larger sizes and darker color indicate greater importance in the network. (B) Venn diagram of hub genes identified by 12 cytoHubba algorithms. (C–D) Kaplan-Meier curves for pack-years (10×) stratified by CDK2 (C) and MCL1 (D) expression levels. (E) Heatmap of the correlations between 13-gene expression profiles and clinical parameters (nodule size and stage). Note: High-risk group was defined as nodule size > 10 mm. PPI, protein-protein interaction; DEIOGs, differentially expressed immune and oxidative stress-related genes.

The model incorporated 11 genes, such as CLEC4 A, which show differential expression in the peripheral blood mononuclear cells of early-stage lung cancer patients[44]. In lung cancer cells, S4 significantly upregulates GADD45 A, a gene associated with cell cycle arrest and apoptosis. This upregulation underscores GADD45 A's potential as a molecular mediator for inhibiting the growth and proliferation of lung cancer cells[45]. Moreover, SHC1 is a potential target in combination therapy to overcome crizotinib resistance in lung cancer[46]. Additionally, downregulation of BAD may enhance the therapeutic effect in lung cancer patients receiving EGFR-TKI therapy[47]. Recent findings have also emphasized the potential of the SMAD2 pathway as a therapeutic option for metastatic lung cancer[48]. ENDOG plays a crucial role in the DNA damage response (DDR) pathway. Its involvement in the mechanisms influencing tumor repopulation after radiotherapy in lung cancer has been highlighted[49]. Thus, downregulation of the ENDOG gene is likely to disrupt the pro-tumorigenic DDR and the Cox-2/PGE_2 axis, potentially slowing tumor progression.

Single-cell RNA sequencing revealed that the expression of IL1R1 in T cells was higher in lung cancer biopsies than in pleural effusions[50]. Baothman et al. proposed SUOX as a promising molecular target for lung cancer, suggesting that further experimental research is needed to clarify its exact role and therapeutic implications in the disease[51]. Sun et al. found that paxillin (PXN) is a critical downstream mediator of the CXCL5-driven signaling cascade. Its phosphorylation activates the AKT pathway to upregulate PD-L1 expression in lung cancer cells, promoting immune evasion by increasing CD8 + T cell exhaustion. This highlights PXN as a potential molecular link between chemokine signaling and immune checkpoint dysfunction in lung cancer[52].

Hub genes identified via protein–protein interaction (PPI) networks and Cytoscape software, including CDK2 and MCL1, were significantly associated with lung cancer treatment outcomes[53,54] and disease progression[55]. In lung cancer, differential expression of CDK2 and MCL1 correlated with packyears, a critical metric of cumulative smoking exposure, with distinct patterns: higher CDK2 expression was associated with greater packy values, and elevated MCL1 expression was linked to lower packy values. Elevated CDK2 expression associated with prolonged cumulative smoking exposure may reflect its role in cell cycle regulation and DNA repair. This may mediate cellular adaptation to smoking-induced oxidative stress or genotoxic damage, thereby influencing disease progression in smokers[56]. Conversely, increased MCL1 expression observed in individuals with lower packages aligns with its well-characterized anti-apoptotic function, potentially promoting the survival of smoking-damaged cells and facilitating oncogenic transformation[57].

The incorporation of these genes into a 13-gene signature, stratified by clinical parameters such as nodule size and stage, highlighted their role in distinguishing low- versus high-risk biological behaviors of LNs. Specifically, CDK2's association with the low-risk cluster and MCL1's presence in the high-risk cluster suggests opposing

effects on cell cycle control and apoptosis, contributing to LNs malignancy. These genes may serve as biomarkers for risk stratification in smoking-related lung cancer. Collectively, these findings emphasize the importance of integrating gene expression profiles with clinical metrics, such as packyears, to refine prognostic models and develop precision therapies for smokers, a population with elevated lung cancer risk.

Limitations

This study had several limitations that merit consideration. Reliance on public datasets introduces inherent biases, and the external validation dataset (GSE20189), comprising 162 samples, raises concerns regarding the relatively small sample size. Of the 13 identified genes, 11 selected by machine learning models and two core proteins from the protein–protein interaction (PPI) network—two (AHSP, SDHAF1) lacked reported associations with lung cancer in the existing literature. Therefore, these genes require experimental validation in clinical samples using techniques such as quantitative polymerase chain reaction (qPCR) or western blotting.

The absence of functional validation experiments limits our mechanistic understanding of the identified genes. Prospective validation across larger multicenter cohorts is essential to confirm the clinical utility of this model. Functional studies, including CRISPR-Cas9-mediated knockout in lung cancer cell lines, could elucidate their role in malignant transformation. The integration of multi-omics datasets may reveal additional biomarkers or therapeutic targets. Finally, investigating the potential of these genes to predict patient responses to immunotherapy or oxidative stress-modulating agents could advance precision medicine for LNs treatment.

Conclusion

This study established a promising framework for identifying immune and oxidative stress-related diagnostic biomarkers in malignant lung nodules (LNs). Integration of weighted gene co-expression network analysis (WGCNA) and machine learning identified an 11-gene signature with superior diagnostic performance, offering novel insights into the molecular mechanisms driving LNs malignancy. While prospective validation in independent cohorts is warranted, these findings advance the development of noninvasive tools for early LNs diagnosis, ultimately enhancing patient outcomes in lung cancer management.

Data availability

All data generated or analyzed in this study are available in the Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/geo/).

Received: 16 April 2025; Accepted: 28 May 2025
Published online: 01 July 2025

References

1. Hammad, M. et al. Explainable AI for lung cancer detection via a custom CNN on CT images. *Sci Rep*. **13** (1), 12707. https://doi.org/10.1038/s41598-025-97645-5 (2025).
2. Torok, Z. et al. Serum and exosome WNT5A levels as biomarkers in non-small cell lung cancer. *Respir Res*. **13** (1), 141. https://doi.org/10.1186/s12931-025-03216-7 (2025).
3. Taje, R. et al. Molecular characterization of early-stage lung adenocarcinoma presenting as subsolid nodules in a real-life European cohort. *BMC Cancer*. **9** (1), 647. https://doi.org/10.1186/s12885-025-13998-0 (2025).
4. Zhu, X. et al. Early screening and subtype identification of High-Risk lung nodules via breathprint by graphene eNose platform: A large cohort study. *ACS Sens*. **7** https://doi.org/10.1021/acssensors.5c00314 (2025).
5. Singh, A., Hammer, M. M. & Byrne, S. C. Incidentally detected adrenal nodules on lung Cancer screening CT. *J. Am. Coll. Radiol.* **22** (3), 291–296. https://doi.org/10.1016/j.jacr.2024.12.003 (2025).
6. Mao, Y. et al. Assessments of lung nodules by an artificial intelligence chatbot using longitudinal CT images. *Cell Rep. Med* **18** (3), 101988. https://doi.org/10.1016/j.xcrm.2025.101988 (2025).
7. Zhang, Z. et al. Prediction of STAS in lung adenocarcinoma with nodules = 2 cm using machine learning: a multicenter retrospective study</at. *BMC Cancer* **7** (1), 417. https://doi.org/10.1186/s12885-025-13783-z (2025).
8. Selvaraj, N. R. et al. Oxidative stress and redox imbalance: common mechanisms in Cancer stem cells and neurodegenerative diseases. *Cells* **29** (7). https://doi.org/10.3390/cells14070511 (2025).
9. Li, Z. et al. Oxidative stress, inflammation, and apoptosis contribute to the pathogenesis of Amyloodinium ocellatum infection in Yellowfin seabream (Acanthopagrus latus). *Fish Shellfish Immunol*. **10**, 110312. https://doi.org/10.1016/j.fsi.2025.110312 (2025).
10. Jiang, Z. et al. Effects of MIR-451 on LKB1-AMPK signal pathway and oxidative stress index in lung Cancer. *Ann Clin. Lab. Sci*. **54** (6), 774–781 (2024).
11. Jamali, T. et al. Mustard lung with a unique oxidative stress profile as an independent pulmonary disease. *Int Immunopharmacol.* **6**, 149:114210. https://doi.org/10.1016/j.intimp.2025.114210 (2025).
12. Cheng, X. et al. The mediating role of oxidative stress on the association between oxidative balance score and Cancer-Related cognitive impairment in lung Cancer patients: A Cross-Sectional study. *Nutrients*. **27** (23). https://doi.org/10.3390/nu16234090 (2024).
13. Que, Z. et al. Jingfukang induces anti-cancer activity through oxidative stress-mediated DNA damage in Circulating human lung cancer cells. *BMC Complement. Altern. Med.*. **7** (1), 204. https://doi.org/10.1186/s12906-019-2601-x (2019).
14. Barsan, M. et al. A study of oxidative stress and pulmonary damage after silica instillation in rats and the effect of Curcumin administration. *Med Pr.*. **30** (3), 239–247. https://doi.org/10.13075/mp.5893.01050 (2021).
15. Ren, Y. F. et al. Single-cell RNA sequencing reveals immune microenvironment niche transitions during the invasive and metastatic processes of ground-glass nodules and part-solid nodules in lung adenocarcinoma. *Mol Cancer*. **23** (1), 263. https://doi.org/10.1186/s12943-024-02177-7 (2024).
16. Zhao, C., Xiao, R., Jin, H. & Li, X. The immune microenvironment of lung adenocarcinoma featured with ground-glass nodules. *Thorac Cancer*. **15** (19), 1459–1470. https://doi.org/10.1111/1759-7714.15380 (2024).
17. Qu, R. et al. Distinct cellular immune profiles in lung adenocarcinoma manifesting as pure ground glass opacity versus solid nodules. *J Cancer Res. Clin. Oncol* . **149** (7), 3775–3788. https://doi.org/10.1007/s00432-022-04289-3 (2023).
18. Zhang, J. & Lin, L. Immunoinfiltration analysis of mitochondrial Damage-Related genes in lung adenocarcinoma and construction of a classification and prognostic model integrated with WGCNA and machine learning algorithms. *Cancer Med*. **14** (2), e70590. https://doi.org/10.1002/cam4.70590 (2025).

19. Wei, W. et al. Machine learning for early discrimination between lung Cancer and benign nodules using routine clinical and laboratory data. *Ann Surg. Oncol.* **31** (12), 7738–7749. https://doi.org/10.1245/s10434-024-15762-3 (2024).
20. Kossenkov, A. V. et al. A gene expression classifier from whole blood distinguishes benign from malignant lung nodules detected by Low-Dose CT. *Cancer Res.* **1** (1), 263–273. https://doi.org/10.1158/0008-5472.CAN-18-2032 (2019).
21. Rotunno, M. et al. A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma. *Cancer Prev. Res. (Phila)* **4** (10), 1599–1608. https://doi.org/10.1158/1940-6207.CAPR-10-0170 (2011).
22. Xu, S. et al. Oxidative stress gene expression, DNA methylation, and gut microbiota interaction trigger Crohn's disease: a multi-omics Mendelian randomization study. *BMC Med.* **21** (1), 179. https://doi.org/10.1186/s12916-023-02878-8 (2023).
23. Cai, J. et al. Development and validation of a novel Endoplasmic reticulum stress-related LncRNA prognostic signature and candidate drugs in breast cancer. *Front. Genet.* **13**, 949314. https://doi.org/10.3389/fgene.2022.949314 (2022).
24. Sui, S. et al. An immune cell infiltration-based immune score model predicts prognosis and chemotherapy effects in breast cancer. *Theranostics* **10** (26), 11938–11949. https://doi.org/10.7150/thno.49451 (2020).
25. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 17. https://doi.org/10.2202/1544-6115.1128 (2005).
26. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **4** (D1), D362–D368. https://doi.org/10.1093/nar/gkw937 (2017).
27. Chin, C. H. et al. CytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **8** (Suppl 4), S11. https://doi.org/10.1186/1752-0509-8-S4-S11 (2014).
28. Qin, H. et al. Integrated machine learning survival framework develops a prognostic model based on inter-crosstalk definition of mitochondrial function and cell death patterns in a large multicenter cohort for lower-grade glioma. *J Transl Med* **2** (1), 588. https://doi.org/10.1186/s12967-023-04468-x (2023).
29. Nishio, M. et al. Computer-aided diagnosis of lung nodule using gradient tree boosting and bayesian optimization. *PLoS One.* **13** (4), e0195875. https://doi.org/10.1371/journal.pone.0195875 (2018).
30. Guan, X. et al. Construction of the XGBoost model for early lung cancer prediction based on metabolic indices. *BMC Med. Inf. Decis. Mak.* **13** (1), 107. https://doi.org/10.1186/s12911-023-02171-x (2023).
31. Zaharudin, N. et al. Prevalence and clinical characteristics of malignant lung nodules in tuberculosis endemic area in a single tertiary centre. *BMC Pulm Med.* **29** (1), 328. https://doi.org/10.1186/s12890-022-02125-5 (2022).
32. Shanmugam, G., Subramaniyam, K., George, M. & Sarkar, K. HDAC Inhibition regulates oxidative stress in CD4(+)Thelper cells of chronic obstructive pulmonary disease and non-small cell lung cancer patients via mitochondrial transcription factor a (mtTFA) modulating NF-kappaB/HIF1alpha axis. *Int Immunopharmacol* **122**, 110661. https://doi.org/10.1016/j.intimp.2023.110661 (2023).
33. Wang, X., Zhang, T., Qu, L., Zhang, Y. & Gao, G. Auriculasin induces mitochondrial oxidative stress and drives ferroptosis by inhibiting PI3K/Akt pathway in non-small cell lung cancer. *Naunyn Schmiedebergs Arch. Pharmacol.* **398** (1), 967–977. https://doi.org/10.1007/s00210-024-03328-9 (2025).
34. Krishnamoorthy, K., Natarajan, S. R., Veeraraghavan, V. P. & Jayaraman, S. Blueberry extract and its bioactive compounds mitigate oxidative stress and suppress human lung cancer cell (A549) growth by modulating the expression of p53/EGFR/STAT3/IL6-mediated signaling molecules. *Cell Biochem. Funct.* **42** (4), e4027. https://doi.org/10.1002/cbf.4027 (2024).
35. El-Deeb, M. M. K., El-Sheredy, H. G. & Mohammed, A. F. The possible role of Interleukin (IL)-18 and nitrous oxide and their relation to oxidative stress in the development and progression of breast Cancer. *Asian Pac. J. Cancer Prev* **1** (9), 2659–2665. https://doi.org/10.31557/APJCP.2019.20.9.2659 (2019).
36. Zhang, H. et al. Effective amplification of oxidative stress and calcium manipulation mediated mitochondrial dysfunction based on engineered nanozyme for primary and metastatic breast Cancer therapy. *Small.* **21** (14), e2411299. https://doi.org/10.1002/smll.202411299 (2025).
37. Simon, L., Arazo-Rusindo, M., Quest, A. F. G., Mariotti-Celis, M. S. & Phlorotannins Novel orally administrated bioactive compounds that induce mitochondrial dysfunction and oxidative stress in Cancer. *Antioxidants (Basel)* **7** (9). https://doi.org/10.3390/antiox12091734 (2023).
38. Kimono, D. et al. Dysbiosis-Associated enteric glial cell Immune-Activation and redox imbalance modulate tight junction protein expression in Gulf war illness pathology. *Front. Physiol.* **10**, 1229. https://doi.org/10.3389/fphys.2019.01229 (2019).
39. Cao, P. et al. BRMS1L confers anticancer activity in non-small cell lung cancer by transcriptionally inducing a redox imbalance in the GPX2-ROS pathway. *Transl Oncol* **41**, 101870. https://doi.org/10.1016/j.tranon.2023.101870 (2024).
40. Niczyporuk, M. et al. A lack of correlation between mast cells, angiogenesis, and outcome in non-small cell lung cancer. *Exp Lung Res.* **38** (6), 281–285. https://doi.org/10.3109/01902148.2012.686559 (2012).
41. Lv, Y. et al. Increased intratumoral mast cells foster immune suppression and gastric cancer progression through TNF-alpha-PD-L1 pathway. *J Immunother Cancer.* **26** (1), 54. https://doi.org/10.1186/s40425-019-0530-3 (2019).
42. Sultana, H. et al. Prognostic value of blood eosinophils for predicting survival and treatment outcomes in people with non-small cell lung cancer. *Cochrane Database Syst. Rev.* **7** (2(2), CD015783. https://doi.org/10.1002/14651858.CD015783 (2025).
43. Zhang, T. & Wong, G. Gene expression data analysis using Hellinger correlation in weighted gene co-expression networks (WGCNA). *Comput. Struct. Biotechnol. J.* **20**, 3851–3863. https://doi.org/10.1016/j.csbj.2022.07.018 (2022).
44. Pakvisal, N. et al. Differential expression of immune-regulatory proteins C5AR1, CLEC4A and NLRP3 on peripheral blood mononuclear cells in early-stage non-small cell lung cancer patients. *Sci Rep.* **2** (1), 18439. https://doi.org/10.1038/s41598-022-21891-0 (2022).
45. Demircan, T., Yavuz, M. & Boluk, A. Unveiling the potential of S4 on Non-small cell lung Cancer cells: impact on proliferation, apoptosis, senescence, and metabolome profile. *Anticancer Agents Med. Chem.* **22** https://doi.org/10.2174/0118715206350735241224073200 (2025).
46. Hu, Y. et al. AXL and SHC1 confer Crizotinib resistance in patient-derived xenograft model of ALK-driven lung cancer. *iScience* **20** (9), 110846. https://doi.org/10.1016/j.isci.2024.110846 (2024).
47. Liu, J. et al. Pitavastatin sensitizes the EGFR-TKI associated resistance in lung cancer by inhibiting YAP/AKT/BAD-BCL-2 pathway. *Cancer Cell. Int.* **28** (1), 224. https://doi.org/10.1186/s12935-024-03416-z (2024).
48. Hu, C. et al. The hidden potential of solasonine: targeting non-small cell lung cancer (NSCLC) metastasis through GATM and Smad2 pathways. *Phytomedicine.* **139**, 156455. https://doi.org/10.1016/j.phymed.2025.156455 (2025).
49. Zhao, M. et al. Caspase-3 knockout attenuates radiation-induced tumor repopulation via impairing the ATM/p53/Cox-2/PGE(2) pathway in non-small cell lung cancer. *Aging (Albany NY).* **7** (21), 21758–21776. https://doi.org/10.18632/aging.103984 (2020).
50. Mahmood, K. et al. Differences in microenvironment of lung cancer and pleural effusions by single-cell RNA sequencing. *Lung Cancer.* **193**, 107847. https://doi.org/10.1016/j.lungcan.2024.107847 (2024).
51. Baothman, O. et al. Multi-targeted therapeutic potential of stigmasterol from the Euphorbia ammak plant in treating lung and breast cancer. *Comput Biol. Chem.* **110**, 108037. https://doi.org/10.1016/j.compbiolchem.2024.108037 (2024).
52. Sun, D. et al. CXCL5 impedes CD8(+) T cell immunity by upregulating PD-L1 expression in lung cancer via PXN/AKT signaling phosphorylation and neutrophil chemotaxis. *J Exp. Clin. Cancer Res.* **22** (1), 202. https://doi.org/10.1186/s13046-024-03122-8 (2024).
53. Sabnis, R. W. Novel bicyclic amines as CDK2 inhibitors for treating metastatic breast Cancer or metastatic lung Cancer. *ACS Med. Chem. Lett.* **13** (2), 198–199. https://doi.org/10.1021/acsmedchemlett.5c00005 (2025).
54. Matsui, Y. et al. Initial AXL and MCL-1 Inhibition contributes to abolishing lazertinib tolerance in EGFR-mutant lung cancer cells. *Cancer Sci* **115** (10), 3333–3345. https://doi.org/10.1111/cas.16292 (2024).

55. Zhao, Y. et al. circRACGAP1 promotes proliferation of Non-Small cell lung Cancer cells through the miR-1296/CDK2 pathway. *Folia Biol. (Praha)*. **70** (2), 104–112. https://doi.org/10.14712/fb2024070020104 (2024).
56. Islam, M. S. et al. Exploiting spirooxindoles for dual DNA targeting/CDK2 Inhibition and simultaneous mitigation of oxidative stress towards selective NSCLC therapy; synthesis, evaluation, and molecular modelling studies. *RSC Med. Chem*. **14** (8), 2937–2958. https://doi.org/10.1039/d4md00337c (2024).
57. Chen, C. L. et al. Polypeptide antibiotic actinomycin D induces Mcl-1 uncanonical downregulation in lung cancer cell apoptosis. *Life Sci*. **15**, 321:121615. https://doi.org/10.1016/j.lfs.2023.121615 (2023).

Acknowledgements
The authors gratefully acknowledge Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) for providing the publicly available datasets used in this study.

Author contributions
All authors made substantial contributions to the development and completion of this study. Z.A. was involved in the conception and design of the study; acquisition, analysis, and interpretation of data; and drafting of the manuscript. M.Z. contributed to the data analysis and revised the manuscript. X.W. contributed to study design, data analysis, and manuscript drafting. All authors provided final approval and agreed to be accountable for all aspects of this study.

Funding
No external funding was received for this study.

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval
Ethical approval was not required for this study, as it utilized de-identified publicly available datasets, and the original research had already obtained appropriate ethical clearance.

Informed consent
Informed consent was not required for this secondary analysis, as the original studies had already obtained written informed consent from participants.

Consent for publication
Not applicable.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-04639-4.

Correspondence and requests for materials should be addressed to X.W.