Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Metabolic reprogramming-related gene classifier distinguishes malignant from the benign pulmonary nodules

Yongkang Huang[a,1], Na Li[b,1], Jie Jiang[c,1], Yongjian Pei[a], Shiyuan Gao[a], Yajuan Qian[a], Yufei Xing[a], Tong Zhou[a], Yixin Lian[a,**], Minhua Shi[a,*]

[a] *Department of Respiratory and Critical Care Medicine, the Second Affiliated Hospital of Soochow University, 1055 Sanxiang Road, Suzhou, 215004, Jiangsu, China*
[b] *Department of Respiratory and Critical Care Medicine, the Fourth Affiliated Hospital of Soochow University, 9 Chongwen Road, Suzhou, 215004, Jiangsu, China*
[c] *Department of Thoracic Surgery, the Affiliated Brain Hospital of Nanjing Medical University, 264 Guangzhou Road, Nanjing, 210003, Jiangsu, China*

ARTICLE INFO

Keywords:
Pulmonary nodules
Classifier
Metabolic reprogramming
Early diagnosis

ABSTRACT

The current existing classifiers for distinguishing malignant from benign pulmonary nodules is limited by effectiveness or clinical practicality. In our study, we aimed to develop and validate a gene classifier for lung cancer diagnosis. To identify the genes involved in this process, we used the weighted gene co-expression network analysis to analyze gene expression datasets from Gene Expression Omnibus (GEO). We identified the three most relevant modules associated with malignant nodules and performed functional enrichment analysis on them. The results indicated significant involvement in metabolic, immune-related, cell cycle, and viral-related processes. All three modules showed enrichment in metabolic reprogramming pathways. Based on these genes, we intersected genes from the three modules with metabolic reprogramming-related genes and further intersected with differentially expressed genes to get 78 genes. After machine learning algorithms and manual selection, we finally got a nine-gene classifier consisting of SEC24D, RPSA, PSME3, PSMD8, PSMB7, NCOA1, MED12, LPCAT1, and AKR1C3. Our developed and validated classifier-based model demonstrated good discrimination, with an area under the curve (AUC) of 0.763 in the development cohort, 0.744 in the internal validation cohort, and 0.718 in the external validation cohort, and outperformed previous clinical models. Moreover, the addition of nodule size improved the predictive capability of the classifier. We further verify the expression of the gene in the classifier using TCGA lung cancer samples and found eight of the genes showed significant differential expression in lung adenocarcinoma while all nine genes showed significant differential expression in lung squamous carcinoma. Our findings underscore the significance of metabolic reprogramming pathways in patients with malignant pulmonary nodules, and our gene classifier can assist clinicians in differentiating benign from malignant pulmonary nodules in clinical settings.

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: ykhuang@alu.suda.edu.cn (Y. Huang), 18211290001@fudan.edu.cn (N. Li), 1114546650@qq.com (J. Jiang), sdfeypyj@163.com (Y. Pei), gsy@suda.edu.cn (S. Gao), pennyqqq@163.com (Y. Qian), 15850158593@163.com (Y. Xing), zhoutonghxk@163.com (T. Zhou), lianyxsoochow@163.com (Y. Lian), shiminhuahxk@163.com (M. Shi).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e37214
Received 25 July 2023; Received in revised form 2 July 2024; Accepted 29 August 2024
Available online 31 August 2024

1. Introduction

Until 2020, lung cancer held the highest global prevalence among cancer types, a position now surpassed by breast cancer. Despite this shift, lung cancer remains the leading cause of cancer-related deaths, with an estimated death toll of almost 1.8 million [1]. This can be attributed to the subtle or nonspecific early symptoms of the disease, which often lead to a diagnosis at an incurable stage [2]. Early diagnosis is critical for improving outcomes in lung cancer patients, prompting a variety of research efforts aimed at addressing this important concern.

Low-dose CT (LDCT) scan is a well-established approach for patients at a high risk of lung cancer. It has demonstrated promising efficacy in detecting susceptible nodules that might indicate lung cancer. LDCT screening trials have demonstrated a reduction in mortality [3,4], leading to its current recommendation by guidelines in most countries for high-risk individuals [5–7]. In some areas, it is even offered as an optional annual physical examination component for employees. However, as CT scans have become more common, the prevalence of pulmonary nodules has also risen, causing a high incidence of indeterminate nodules and resulting anxiety. Only 2.1–5.5 % of cases with lung nodules are malignant [7,8], making it challenging to distinguish malignant from benign nodules. In most cases, differential diagnosis often entails repeated CT follow-ups or invasive examinations, which are largely dependent on expert judgment by clinicians and can sometimes yield inconclusive results, especially for moderate-risk nodules. Thus, there is a need for non-invasive or minimally invasive, easy-to-use, and accurate methods to help clinicians differentiate between malignant and benign nodules.

Carcinogenesis is a multistep process involving dynamic changes in gene expression. These changes can affect cellular behavior and reveal disease features, making them potential biomarkers. Peripheral blood samples are widely used for transcriptome profiling in order to identify valuable diagnostic and prognostic markers [9]. This approach is minimally invasive and advantageous when disease-relevant tissues are inaccessible [10,11]. In this study, we analyzed gene expression datasets from blood samples of patients with pulmonary nodules. We applied weighted gene co-expression network analysis (WGCNA) to identify modules associated with malignant nodules and explored the genetic mechanisms of lung carcinogenesis. We also developed a gene classifier related to metabolic reprogramming using machine-learning algorithms. We showed that the classifier-based model exhibited good accuracy and calibration.

2. Materials and methods

2.1. Participants

In this study, we aimed to investigate the genetic mechanisms underlying lung carcinogenesis and develop a classifier to

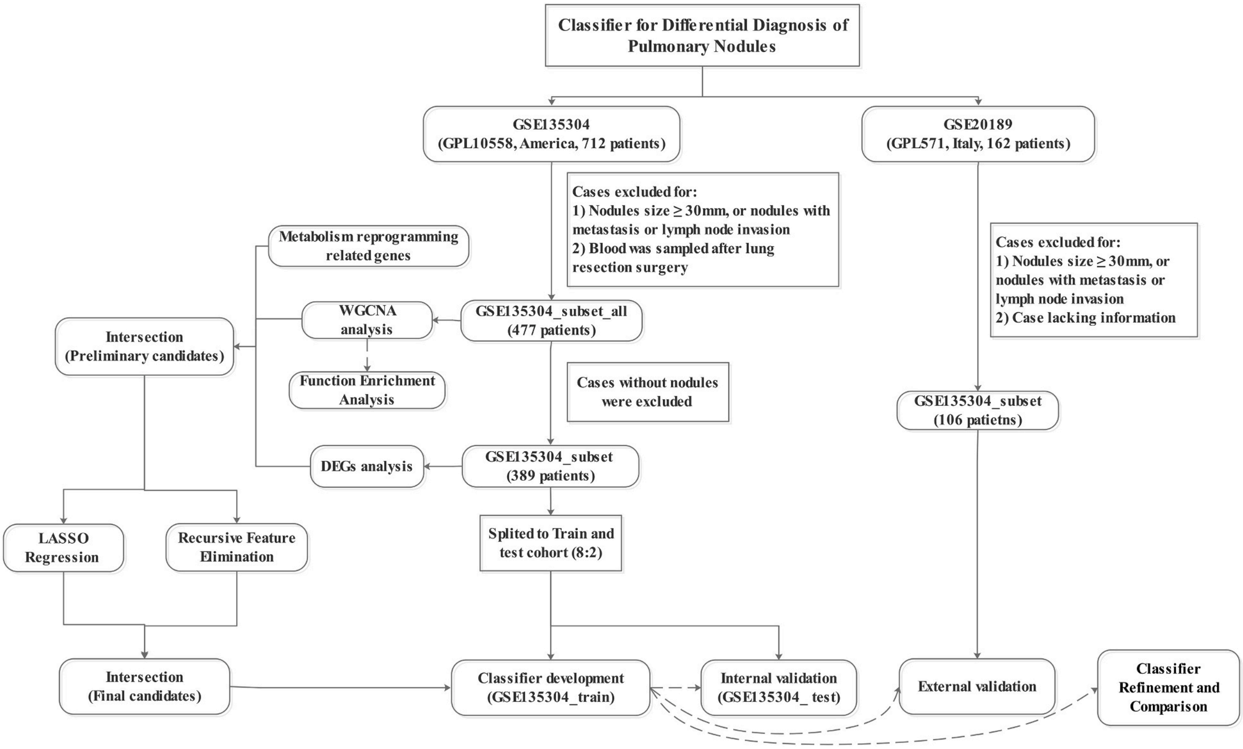
Classifier for Differential Diagnosis of Pulmonary Nodules
GSE135304 (GPL10558, America, 712 patients)
GSE20189 (GPL571, Italy, 162 patients)
Cases excluded for:
1) Nodules size ≥ 30mm, or nodules with metastasis or lymph node invasion
2) Blood was sampled after lung resection surgery
Cases excluded for:
1) Nodules size ≥ 30mm, or nodules with metastasis or lymph node invasion
2) Case lacking information
Metabolism reprogramming related genes
WGCNA analysis
GSE135304_subset_all (477 patients)
Intersection (Preliminary candidates)
Function Enrichment Analysis
Cases without nodules were excluded
GSE135304_subset (106 patietns)
DEGs analysis
GSE135304_subset (389 patients)
LASSO Regression
Recursive Feature Elimination
Splited to Train and test cohort (8:2)
Intersection (Final candidates)
Classifier development (GSE135304_train)
Internal validation (GSE135304_ test)
External validation
Classifier Refinement and Comparison

Fig. 1. Study flow.

differentiate between malignant and benign nodules. To this end, our study included patients with a positive pulmonary nodule (of no more than 30 mm on a CT scan) without metastasis or lymph node invasion (stage I of the AJCC classification). The flowchart of this study is shown in Fig. 1.

To identify appropriate gene expression datasets, we searched the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) using the keywords “lung cancer” and “blood”. We included datasets that met the following criteria: (1) involving human patients with nodules as study participants; (2) diagnosing malignant nodules through pathology; (3) using expression profiles based on whole blood; and (4) sampling blood before lung resection surgery. Based on the inclusion criteria, two expression profiling datasets (GSE135304 and GSE20189) were used in this study. Table S1 summarizes the characteristics of the dataset.

GSE135304 (Platform: GPL10558) comprised 712 patients of diverse races (Caucasian, African American, and others) from America, providing whole-blood gene expression data from individuals with malignant (MN), benign (BN) nodules, and those without nodules (NN). From this dataset, cases were filtered based on specific criteria: 1) blood sampled after lung resection surgery ($n = 90$); 2) nodules larger than 30 mm on CT scan, or nodules with metastasis or lymph node invasion ($n = 145$). A total of 477 cases were selected (termed GSE135304_subset_all), including 216 cases of malignant nodules, 173 cases of benign nodules, and 88 cases without nodules.

GSE20189 (Platform: GPL571) included 162 Caucasians from Italy, and after similar filtering, 106 were chosen for external validation (termed GSE20189_subset). The validation group comprised 26 patients with stage I lung adenocarcinoma (LUAD) and 80 non-cancer controls, predominantly composed of older individuals who were current smokers.

2.2. *WGCNA and functional enrichment analysis*

WGCNA is a computational algorithm that identifies gene co-expression patterns. It groups genes with similar expression patterns across samples and conditions into modules containing functionally related genes [12].

In this study, we performed WGCNA on GSE135304_subset_all to identify genes associated with lung cancer, which were then subjected to Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses to understand the molecular mechanisms of lung carcinogenesis. WGCNA was conducted using the WGCNA package [12] in R, based on the top 5000 genes ordered by the Median Absolute Deviation. We set the scale-free R-squared cutoff to 0.8, cut height to 0.25, and the minimum module size to 30. We ranked the modules with a significant ($P < 0.05$) correlation with malignant nodules based on the absolute value of the correlation coefficient.

GO biological process and KEGG enrichment analyses were conducted using the clusterProfiler package [13] in R based on the top three relevant gene modules obtained from WGCNA. A p-value <0.05 and q-value <0.05 were considered statistically significant.

2.3. *Differentially expressed genes and metabolic reprogramming-related genes*

Patients in the GSE135304_subset_all dataset were divided into two groups based on the presence of malignant nodules: a benign nodule group and malignant nodule group (termed GSE135304_subset). Differentially expressed genes (DEGs) between the two groups were identified using the limma package [14] in R, with a cutoff of the sum of two times the standard deviation and the mean absolute logFC value.

Metabolic reprogramming-related genes were obtained from the supplementary study of Peng et al. [15], who curated gene sets of seven metabolic super-pathways based on Reactome annotations [16], including genes of amino acid metabolism, carbohydrate metabolism, energy integration, lipid metabolism, nucleotide metabolism, tricarboxylic acid cycle, and vitamin and cofactor metabolism.

2.4. *Development and validation of the classifier*

To identify the most relevant genes associated with malignant nodules, we employed machine learning algorithms, such as least absolute shrinkage and selection operator (LASSO) regression and recursive feature elimination with cross-validation (RFE). These algorithms were based on genes from the top three relevant modules identified using WGCNA. LASSO regression was performed using the glmnet package [17] in R, with the lambda set according to 4-fold cross-validation. RFE was conducted using the caret package [18] in R, with a random forest classifier used to perform feature selection, similarly employing 4-fold cross-validation.

To prevent data overfitting, we obtained the intersection of the genes identified by the two aforementioned machine learning algorithms. We further integrated stepwise logistic regression and manual selection to obtain a classifier. The rms package was used to develop a nomogram model based on this classifier. The discrimination and calibration of the model were assessed using the area under the receiver operating characteristic curve (AUC) and calibration curve, respectively. Additionally, we evaluated the practicality of the model using decision curve analysis (DCA) using the rmda package [19] in R.

2.5. *Gene expression profile of genes in lung cancer tissues*

To investigate the gene expression profile in lung cancer tissues, LUAD and lung squamous cell carcinoma (LUSC) samples were obtained from TCGA-LUAD and TCGA-LUSC, respectively. We acquired gene expression and phenotype data from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) and utilized FPKM for comparison between normal and tumor samples.

2.6. Statistical analysis

Continuous variables in this study were reported as either mean ± standard deviation or median (interquartile range), depending on whether they followed a normal distribution. The R language and the corresponding packages mentioned above were used for statistical analysis and visualization. We considered a p-value less than 0.05 as statistically significant.

3. Results

3.1. Clinical features of patients

Two expression profiling datasets (GSE135304 and GSE20189) were used. After filtering, as described previously, subsets containing the population of interest were obtained. The patients of GSE135304_subset were mostly smokers or former smokers (95 %), had a mean age of 66 ± 9 years, and were predominantly female (58 %) and Caucasian (83 %). The median nodule size was 12 mm (IQR 8–17 mm). These patients were divided into two sub-cohorts in an 8:2 ratio using the "createDataPartition" function from the caret package for classifier development (train cohort) or internal validation (test cohort). The demographic data were balanced between the two sub-cohorts (Table 1). The GSE20189_subset comprised 26 cases of stage I lung cancer and 80 controls, primarily consisting of smokers, intended for external validation.

3.2. WGCNA analysis and functional enrichment analysis

With the soft threshold equal to 6 ($\beta = 6$) as per the soft-threshold power network topology analysis (Fig. 2A), weighted gene co-expression networks were constructed based on GSE135304_subset dataset, and 21 modules were generated (Fig. 2B). The lavender, green, and cyan modules were identified as the three most relevant modules that correlated with malignant nodules (Fig. 2C). Consequently, we extracted the genes belonging to these three modules for functional enrichment analysis. GO biological process enrichment analysis showed that 42, 78, and 27 terms were enriched in the lavender, green, and cyan modules, respectively, suggesting that the correlated genes were mainly related to metabolism, immune-related processes, cell cycle and viral-related processes (Table S2A).

KEGG pathway analysis showed similar results, with 13, 19, and seven enriched terms in the lavender, green, and cyan modules, respectively. These included metabolism-related pathways, pathways of neurodegenerative diseases, viral carcinogenesis, and signaling pathways regulating the pluripotency of stem cells (Table S2B). The significant GO terms and KEGG pathway analyses were sorted in ascending order of q-values, and the top 10 terms are shown in Fig. 2D & E.

3.3. Gene selection and classifier development

Given that metabolism-related terms were enriched in all three modules, and considering metabolic reprogramming is a hallmark feature of tumors occurring early in tumorigenesis, we sought to develop a gene classifier focused on metabolism to assist in diagnosing malignant nodules.

The genes in the three modules were extracted, merged, intersected with metabolic reprogramming-related genes, and further intersected with the 812 DEGs based on the GSE135304_subset (Fig. 3A). Finally, 78 genes were selected (Fig. 3B). These genes were then subjected to LASSO regression and RFE to identify the most relevant variables. As a result, sets of 21 (Fig. 3C and D) and 35 genes (Fig. 3E) were chosen. The two gene sets intersected (Fig. 3F) to generate a 14-gene set. Logistic stepwise regression and manual selection were performed to obtain a nine-gene classifier consisting of SEC24D, RPSA, PSME3, PSMD8, PSMB7, NCOA1, MED12, LPCAT1, and AKR1C3.

Table 1
Patient characteristics.

Variable	Overall N = 389	BN N = 216	MN N = 173	Train cohort N = 312	Test cohort N = 77
Group					
BN	216/389 (56 %)	–	–	173/312 (55 %)	43/77 (56 %)
MN	173/389 (44 %)	–	–	139/312 (45 %)	34/77 (44 %)
Smoking	369/389 (95 %)	208/216 (96 %)	161/173 (93 %)	294/312 (94 %)	75/77 (97 %)
Gender					
Female	226/389 (58 %)	114/216 (53 %)	112/173 (65 %)	181/312 (58 %)	45/77 (58 %)
Male	163/389 (42 %)	102/216 (47 %)	61/173 (35 %)	131/312 (42 %)	32/77 (42 %)
Age	66 (9)	64 (9)	68 (9)	65 (9)	67 (9)
Race					
African American	48/389 (12 %)	23/216 (11 %)	25/173 (14 %)	38/312 (12 %)	10/77 (13 %)
Caucasian	324/389 (83 %)	182/216 (84 %)	142/173 (82 %)	257/312 (82 %)	67/77 (87 %)
Other	17/389 (4.4 %)	11/216 (5.1 %)	6/173 (3.5 %)	17/312 (5.4 %)	0/77 (0 %)
Nodule Size (mm)	13 (7)	9 (5, 12)	17 (13, 20)	13 (7)	13 (6)

Abbreviation: BN, benign nodule; MN, malignant nodule. Data were expressed by n/N (%) or Mean (SD).

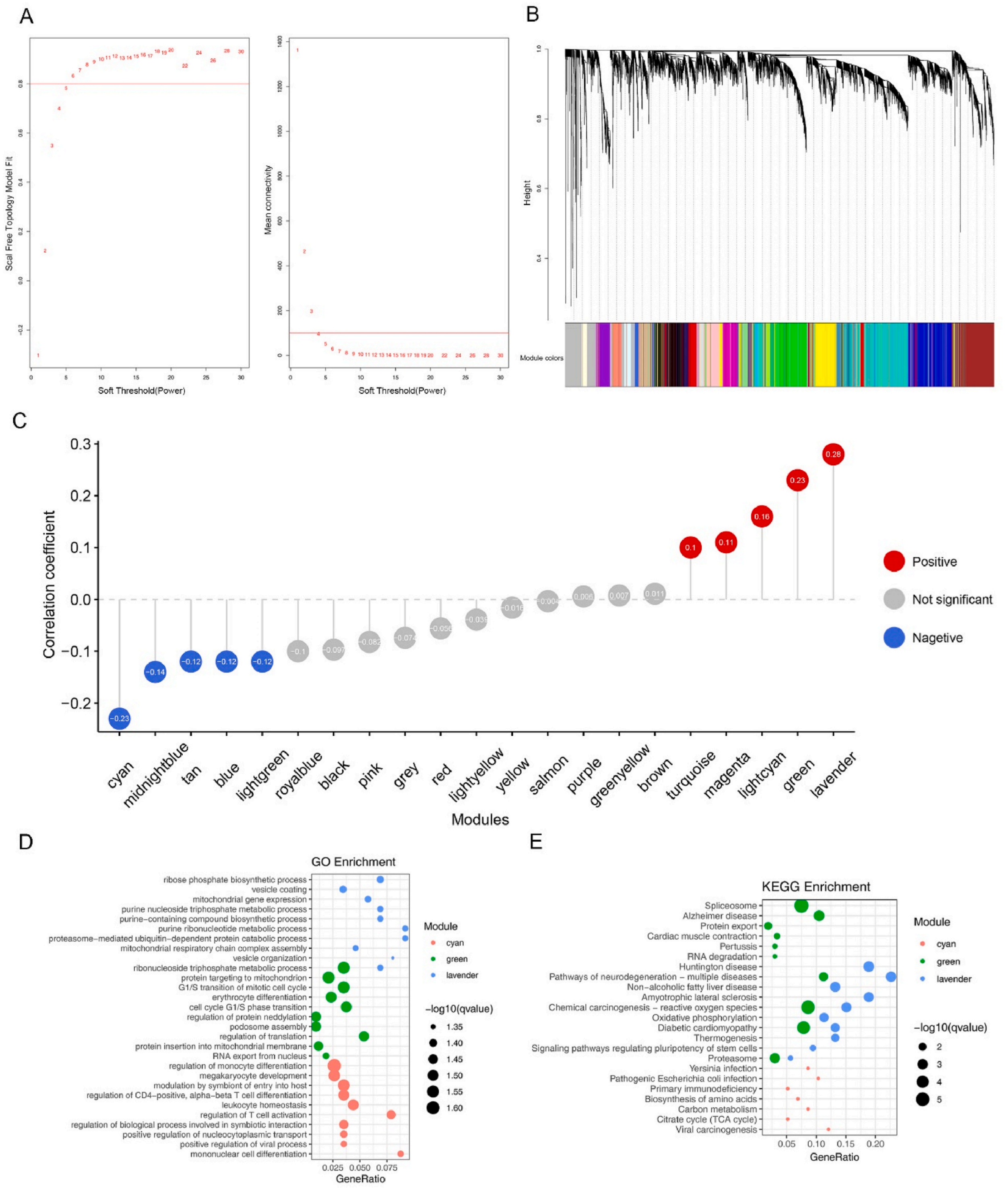

Fig. 2. WGCNA and enrich analysis.
(A) Network topology analysis identified a suitable soft-threshold power of 6. (B) Cluster dendrogram of WGCNA depicting 21 modules labeled by distinct colors. (C) Correlation between modules and pulmonary malignant nodules. (D) Top 10 GO terms associated with the three most relevant modules related to malignant nodules. (E) Top 10 KEGG pathways associated with the three most relevant modules related to malignant nodules.

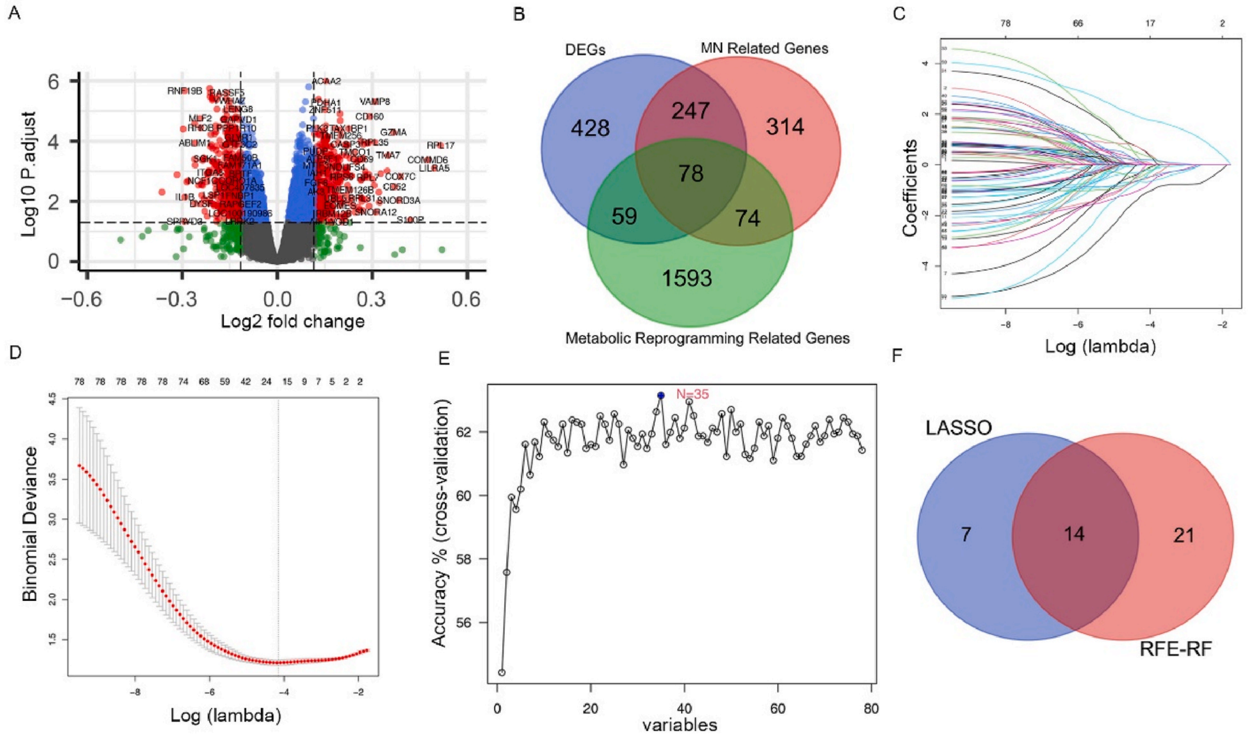

Fig. 3. The feature selection.
(A) Volcano plot of differential expression genes (DEGs) identified by limma, with red dots indicating significant DEGs. (B) Venn diagram of the overlap between DEGs, malignant nodules (MN) related genes and metabolic reprogramming related genes, resulting in 78 genes. (C)LASSO coefficient profiles of the 78 genes. (D) Cross-validation for tuning parameter selection in the LASSO model. (E) Recursive feature elimination with random forest (RFE-RF) of the 78 genes. (F) Venn diagram of the overlap between LASSO and RFE-RF selected genes, resulting in 14 genes.

3.4. Nomogram model development and validation

We developed a model using a classifier and evaluated its diagnostic performance. A nomogram model was developed and validated internally (Fig. S1). It exhibited good discrimination with an AUC of 0.763 in the training cohort and 0.744 in the test cohort (Fig. 4A). As illustrated in Fig. 4B & C, the Hosmer–Lemeshow goodness-of-fit test indicated that the predicted values of the model were not significantly different from the observed values. The DCA diagram showed that there was a net benefit greater than 0 between the risk thresholds of 0.01–0.78, indicating that the model had good clinical practicability (Fig. 4D).

Subgroup analyses were performed to investigate whether the model efficacy differed across subgroups of the study population. Stratification by nodule size revealed that the AUC of the nomogram for nodules measuring 0–10 mm, 10–20 mm, and 20–30 mm were 0.72, 0.714, and 0.675, respectively (Fig. 4E). We also divided the population into Caucasian and non-Caucasian subgroups to determine whether the efficacy of the model was influenced by race. The discrimination of the model in Caucasians (AUC: 0.75) was similar but marginally lower than that in non-Caucasians (AUC: 0.79) (Fig. 4F).

To confirm the reliability of our classifier, we performed validation using an additional dataset (GSE20189_subset). The results demonstrated that our classifier-based model had a comparable discriminatory ability, as indicated by an AUC value of 0.718 (Fig. 4A). Moreover, the calibration curve depicted in Fig. S2 shows good agreement between the predicted and observed values, indicating that the model was well-calibrated.

3.5. Addition of clinical variables increased the efficacy of the classifier

Certain predictors such as nodule size and age have been widely acknowledged as valuable for differentiating between benign and malignant nodules in clinical practice. Incorporating such predictors into a classifier can potentially enhance performance. Thus, we developed refined models that integrated the classifier with nodule size or age. Compared to the previous model, both models yielded a noticeable improvement in the AUC (Fig. 5A, Fig. S3A), especially the model that incorporating nodule size, achieving values of 0.891 and 0.815 in the development and validation cohorts, respectively. This surpassed the model based solely on nodule size, which achieved 0.864 and 0.793 in the development and validation cohorts, respectively (Fig. S4). Moreover, calibration curves demonstrated a good concordance between the model's predictions and actual observations (Figs. S3B and S3C).

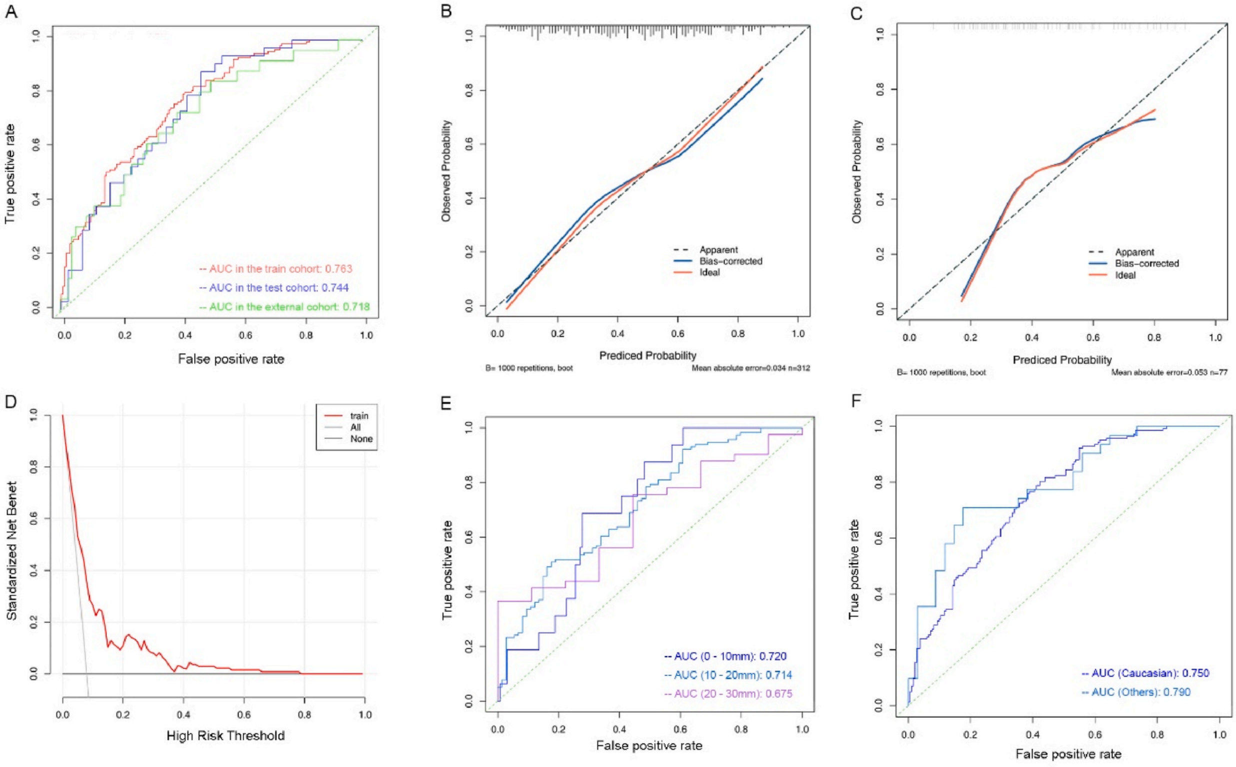

Fig. 4. Development and validation of the classifier based model.
(A) ROC curves for the classifier in the train, test, and external cohorts, with corresponding AUC values of 0.763, 0.744, and 0.718, respectively. (B) Calibration plots demonstrating agreement between predicted and observed probabilities in the train cohort. (C) Calibration plots demonstrating agreement between predicted and observed probabilities in the test cohort. (D) Decision curve analysis indicating a net benefit across risk thresholds of 0.01–0.78 for the classifier. (E) Subgroup analysis: ROC curve for the classifier stratified by nodule sizes. (F) Subgroup analysis: ROC curve for the classifier stratified by race.

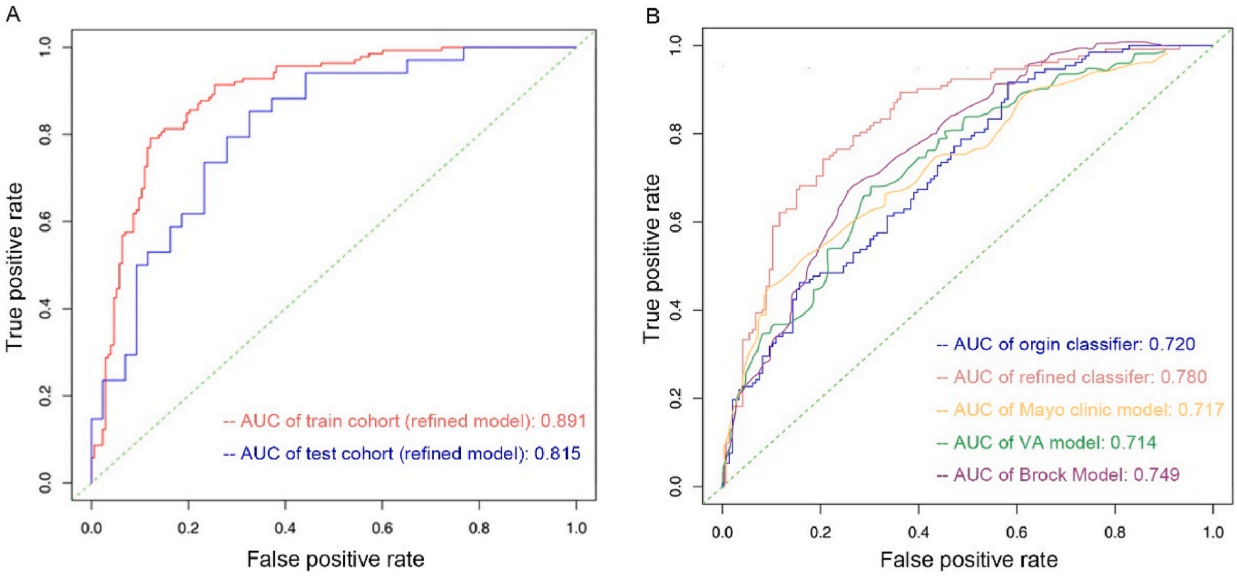

Fig. 5. Classifier refinement and comparison.
(A) ROC curve of the refined classifier (with added nodule size) in the train and test cohorts, with AUC values of 0.891 and 0.815, respectively. (B) Comparison of the performance metrics of the original and refined models.

3.6. *Model based on the classifier outperformed previous clinical models*

We compared the performance of our classifier-based model with several previously published clinical models, such as the Mayo Clinic model [20], Brock model [21], and Veteran's Affairs lung cancer risk clinical (VA) model [22]. According to Kossenkov et al. [23], the AUC of the three models for patients with nodules measuring 6–20 mm in the dataset were 0.717, 0.749, and 0.714, respectively. We evaluated the performance of our classifier-based model for these patients and obtained an AUC value of 0.72, which outperformed the Mayo Clinic and VA models but was slightly lower than that of the Brock model. However, when we integrated the nodule size classifier into our model, we obtained a comprehensive model with an AUC of 0.78, outperforming all three clinical models (Fig. 5B).

3.7. *Validation of classifier gene expression in lung cancer tissues*

We further analyzed the gene expression of the classifier in TCGA database to investigate its potential role in lung cancer. Among the nine genes comprising the classifier, seven genes—AKR1C3, MED12, PSMB7, PSMD8, PSME3, RPSA, SEC24D—showed upregulation in both lung adenocarcinoma and lung squamous carcinoma, while LPCAT1 and NCOA1 exhibited downregulation. This consistent pattern of expression was observed in both the subset of stage I patients (Fig. 6A) and the entire dataset (Fig. 6B). Notably, four genes—LPCAT1, MED12, PSME3, SEC24D—displayed distinct expression profiles between peripheral blood mononuclear cells (PBMC) and TCGA lung tissues. Despite this, eight out of nine genes showed significant differential expression in lung adenocarcinoma (all P < 0.05, except LPCAT1), and all nine genes were significantly differentially expressed in lung squamous carcinoma (all P < 0.05). These findings underscore the potential pivotal role of these genes in the initiation and progression of lung tumorigenesis.

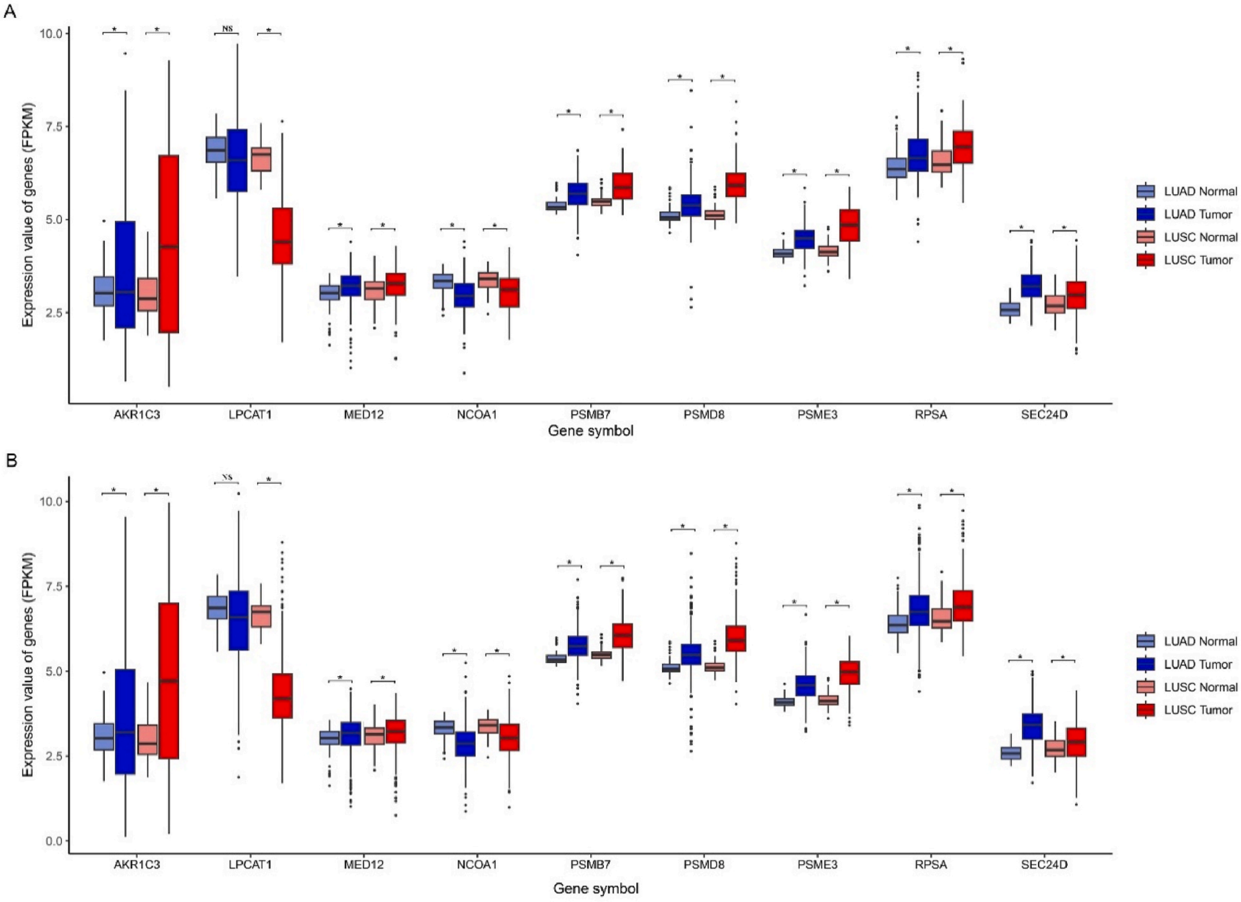

Fig. 6. Validation of classifier in lung cancer tissues.
Gene expression profile in lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) samples. (A) stage I samples exclusively, (B) all samples.
Abbreviation: NS, not significant; *, P < 0.05.

4. Discussion

Tumor metabolic reprogramming encompasses a series of metabolic changes in tumor cells, such as aerobic glycolysis, which enables rapid proliferation and growth, even in microenvironments with limited oxygen and nutrition. This phenomenon has been considered as one of the hallmarks of tumors [24,25]. In this study, we utilized available datasets from the GEO database and filtered suitable cases to analyze the gene expression patterns of PBMC in patients with malignant lung nodules to elucidate the underlying mechanisms of early lung cancer. WGCNA and functional enrichment analyses revealed that the genes in the module most relevant to malignant lung nodule-related genes were mainly involved in metabolism-related processes, immune-related terms or pathways, and the cell cycle. The metabolism-related terms and pathways were significantly more frequent. Subsequently, a classifier containing nine metabolism-related genes was developed, and the model based on this classifier demonstrated good discrimination and calibration. Collectively, we developed a classifier based on metabolic reprogramming genes to distinguish malignant nodules, which could potentially benefit clinical practice.

Numerous models have been developed to aid the differential diagnosis of benign and malignant nodules. The classic and influential example of a traditional model is the Mayo model developed by Swensen et al. [20] in 1997 based on 419 patients from the Mayo Clinic in the United States. This is the first model to accurately predict the probability of malignancy in pulmonary nodules, and its features are readily accessible in clinical practice. Consequently, the Mayo model has been frequently implemented in clinical settings. Subsequently, the Brock [21] and VA [26] models were introduced, and excellent accuracy was demonstrated. In addition to traditional models that rely on clinical and imaging characteristics, gene expression models have become more prevalent in recent years with the advent of precision medicine. For example, Fortunato et al. [27] explored the role of immunosuppressive systemic immunity in lung carcinogenesis and created a molecular blood-based immune signature classifier to aid in the early detection of lung cancer. However, the model was based on PBMC from a small sample size, and although another study demonstrated high accuracy in using PBMC as a diagnostic biomarker [26], it requires rapid centrifugation of the blood within hours of sampling to maintain sample consistency and RNA integrity, which limits its widespread use in multicenter studies and general laboratory-based clinical practice. Ambrosi et al. [28] also developed an eight-target signature with an AUC of 0.92 based on 20 early-stage lung cancer samples (stages I-III) and a control group comprising asymptomatic individuals ($n = 27$) and those with benign lung nodules ($n = 3$). However, the signature was not validated externally, and the mixture of populations and small sample sizes may have undermined the confidence of the signature. Similarly, the prediction models developed by Xing et al. [29], based on DNA methylation biomarkers and radiological characteristics, and the classifier developed by Lin et al. [30], which integrates plasma biomarkers and radiological characteristics, both demonstrate promising discrimination. However, they use mixed patient cohorts, with only 26.9 % and 26.1 % of the training cohort case groups at stage I, respectively. Additionally, the model developed by Xing et al. [29] was not externally validated.

In our study, we demonstrated that the metabolic reprogramming-based classifier model was more accurate than the Mayo and Brock models and superior to the VA model when nodule size was added as a classifier. Our model also outperformed other gene expression-based models. The RNA samples used in this study were whole blood RNA isolated using the PAXgene method and collected in PAXgene tubes. PAXgene RNA is stable for up to 5 days at 15–25 °C and for years at −20 to −70 °C, enabling sample collection in any clinical setting without requiring special equipment and transfer to a central facility for testing (as is required in other blood tests) [31–33]. Moreover, the classifier was developed based solely on patients with malignant pulmonary nodules before lung resection surgery, with those with benign nodules serving as controls. This innate advantage assists the classifiers in distinguishing between benign and malignant nodules. Furthermore, the classifier was externally validated and comprised only nine genes, compared to the classifier developed by Kossenkov et al. [23]. Moreover, the inclusion of the nodule size variable was straightforward. The small number of features makes incorporation into clinical practice easier and more cost-effective.

Our study demonstrated that the classifier-based model exhibited good consistency in the ethnic subgroup analyses. However, in the nodule size subgroup analysis, we observed a decrease in accuracy in the 20–30 mm range. This could be attributed to the varying sample sizes within each size stratum (162, 163, and 62 samples in the 0–10 mm, 10–20 mm, and 20–30 mm subgroups, respectively). Nonetheless, in larger nodules, most imaging signs of malignancy on CT scans, such as the lobulated sign, spiculated sign, vacuole sign, air bronchogram, pleural indentation, and vessel convergence [34], have become more frequent in this group of patients, making it less difficult to distinguish between benign and malignant nodules clinically. Several predictors of malignant nodules are well established, including larger nodule size, older age, and smoking history [35]. Our study demonstrated that integrating nodule size as a classifier improved discrimination and calibration. These findings offer novel insights for future studies aimed at developing and refining these models.

Our study had certain limitations. First, the patient data analyzed were predominantly of Caucasian ethnicity sourced from Europe and America, raising doubts about the generalizability of our findings to other populations. Second, although we observed an improvement in the efficacy of the classifier with the incorporation of clinical variables, the external validation of our model was not feasible because of the unavailability of the corresponding data in the external dataset. Third, the control group for the external validation comprised non-cancer individuals who were current smokers at high risk of lung cancer; however, they did not undergo a CT scan, thus leaving their pulmonary nodule status unknown.

Nonetheless, our study represents a valuable step forward in nodule classification, providing a foundation for future research to build upon our results.

5. Conclusion

Patients with malignant nodules exhibit alterations in energy-related metabolic pathways. Gene classifiers based on tumor metabolic reprogramming can aid in the clinical identification of benign and malignant nodules.

Funding statement

This work was supported by the Competitive discipline lift project of the Second Affiliated Hospital of Soochow University (grant number XKTJ-XK202007); the Science and Technology Development Program of Suzhou (grant number SYSD2019104); the Science Pre-research Foundation of the Second Affiliated Hospital of Soochow University (grant number SDFEYHT2224; SDFEYJC2329); and the Medical science research funding project of Beijing Medical and Health Foundation (grant number YWJKJJHKYJJ-HX05).

Ethical statement

Not applicable.

Data availability statement

The datasets used in the study were from online repositories, the names of which and accession number(s) can be found in the article.

CRediT authorship contribution statement

Yongkang Huang: Conceptualization, Data curation, Writing – review & editing. **Na Li:** Data curation, Formal analysis, Writing – review & editing. **Jie Jiang:** Data curation, Writing – original draft, Writing – review & editing. **Yongjian Pei:** Formal analysis, Writing – review & editing. **Shiyuan Gao:** Data curation, Formal analysis, Writing – review & editing. **Yajuan Qian:** Validation, Writing – review & editing. **Yufei Xing:** Validation, Writing – review & editing. **Tong Zhou:** Visualization, Writing – review & editing. **Yixin Lian:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing. **Minhua Shi:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e37214.

References

[1] H. Sung, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, Ca-Cancer J. Clin. 71 (3) (2021) 209–249.
[2] A.A. Thai, et al., Lung cancer, Lancet. 398 (10299) (2021) 535–554.
[3] H.J. de Koning, et al., Reduced lung-cancer mortality with volume CT screening in a randomized trial, N. Engl. J. Med. 382 (6) (2020) 503–513.
[4] D.R. Aberle, et al., Reduced lung-cancer mortality with low-dose computed tomographic screening, N. Engl. J. Med. 365 (5) (2011) 395–409.
[5] R.S. Wiener, et al., An official American Thoracic Society/American College of Chest Physicians policy statement: implementation of low-dose computed tomography lung cancer screening programs in clinical practice, Am. J. Respir. Crit. Care Med. 192 (7) (2015) 881–891.
[6] P.J. Mazzone, et al., Screening for lung cancer: CHEST guideline and expert panel report, Chest 153 (4) (2018) 954–985.
[7] J.K. Field, et al., The UK Lung Cancer Screening Trial: a pilot randomised controlled trial of low-dose computed tomography screening for the early detection of lung cancer, Health Technol. Assess. 20 (40) (2016) 1–146.
[8] A. McWilliams, et al., Probability of cancer in pulmonary nodules detected on first screening CT, N. Engl. J. Med. 369 (10) (2013) 910–919.
[9] M.K.R. Kalikiri, et al., Technical assessment of different extraction methods and transcriptome profiling of RNA isolated from small volumes of blood, Sci. Rep. 13 (1) (2023) 3598.
[10] S.N. Lone, et al., Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments, Mol. Cancer 21 (1) (2022) 79.
[11] I. Martins, et al., Liquid biopsies: applications for cancer diagnosis and monitoring, Genes 12 (3) (2021) 349.
[12] P. Langfelder, et al., WGCNA: an R package for weighted correlation network analysis, BMC Bioinf. 9 (2008) 559.
[13] T. Wu, et al., clusterProfiler 4.0: a universal enrichment tool for interpreting omics data, Innovation 2 (3) (2021) 100141.
[14] M.E. Ritchie, et al., Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47.
[15] X. Peng, et al., Molecular characterization and clinical relevance of metabolic expression subtypes in human cancers, Cell Rep. 23 (1) (2018) 255–269.
[16] A. Fabregat, et al., The reactome pathway knowledgebase, Nucleic Acids Res. 44 (D1) (2016) D481–D487.
[17] J. Friedman, et al., Regularization paths for generalized linear models via coordinate descent, J. Stat. Software 33 (1) (2010) 1–22.
[18] Kuhn, et al., Building predictive models in R using the caret package, J. Stat. Software 28 (5) (2008) 1–26.
[19] M. Brown, rmda: Risk Model Decision Analysis (2018).
[20] S.J. Swensen, et al., The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules, Arch. Intern. Med. 157 (8) (1997) 849–855.
[21] K. Chung, et al., Brock malignancy risk calculator for pulmonary nodules: validation outside a lung cancer screening population, Thorax 73 (9) (2018) 857–863.
[22] M.K. Gould, et al., A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules, Chest 131 (2) (2007) 383–388.

[23] A.V. Kossenkov, et al., A gene expression classifier from whole blood distinguishes benign from malignant lung nodules detected by low-dose CT, Cancer Res. 79 (1) (2019) 263–273.

[24] N.N. Pavlova, et al., The hallmarks of cancer metabolism: still emerging, Cell Metabol. 34 (3) (2022) 355–377.

[25] D. Hanahan, Hallmarks of cancer: new dimensions, Cancer Discov. 12 (1) (2022) 31–46.

[26] M.K. Showe, et al., Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease, Cancer Res. 69 (24) (2009) 9202–9210.

[27] O. Fortunato, et al., Development of a molecular blood-based immune signature classifier as biomarker for risks assessment in lung cancer screening, Cancer Epidemiol. Biomarkers Prev. 31 (11) (2022) 2020–2029.

[28] S. D Ambrosi, et al., Combinatorial blood platelets-derived circRNA and mRNA signature for early-stage lung cancer detection, Int. J. Mol. Sci. 24 (5) (2023) 4881.

[29] W. Xing, et al., A prediction model based on DNA methylation biomarkers and radiological characteristics for identifying malignant from benign pulmonary nodules, BMC Cancer 21 (1) (2021) 263.

[30] Y. Lin, et al., A classifier integrating plasma biomarkers and radiological characteristics for distinguishing malignant from benign pulmonary nodules, Int. J. Cancer 141 (6) (2017) 1240–1248.

[31] V. Chai, et al., Optimization of the PAXgene blood RNA extraction system for gene expression analysis of clinical samples, J. Clin. Lab. Anal. 19 (5) (2005) 182–188.

[32] M.M. Fricano, et al., Global transcriptomic profiling using small volumes of whole blood: a cost-effective method for translational genomic biomarker identification in small animals, Int. J. Mol. Sci. 12 (4) (2011) 2502–2517.

[33] S. Debey, et al., A highly standardized, robust, and cost-effective method for genome-wide transcriptome analysis of peripheral blood applicable to large-scale clinical trials, Genomics 87 (5) (2006) 653–664.

[34] T.L. Mohammed, et al., The imaging manifestations of lung cancer, Semin. Roentgenol. 40 (2) (2005) 98–108.

[35] Z. Wu, et al., Lung cancer risk prediction models based on pulmonary nodules: a systematic review, Thorac Cancer 13 (5) (2022) 664–677.