Heliyon 10 (2024) e34585

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Prediction and verification of benignancy and malignancy of pulmonary nodules based on inflammatory related biological markers

Zexin Zhang [a,1], Wenfeng Wu [a,1], Xuewei Li [b], Siqi Lin [a], Qiwei Lei [a], Ling Yu [b], Jietao Lin [b], Lingling Sun [b], Haibo Zhang [c], Lizhu Lin [b,d,*]

[a] Guangzhou University of Chinese Medicine, Guangzhou, China
[b] The First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China
[c] The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China
[d] Guangdong Clinical Research Academy of Chinese Medicine, Guangzhou, China

ARTICLE INFO

Keywords:
Pulmonary nodules
Inflammation-related biological markers
Immune cells infiltration
Support vector machine
Nomogram

ABSTRACT

Objective: Inflammation plays an important role in the transformation of pulmonary nodules (PNs) from benign to malignant. Prediction of benignancy and malignancy of PNs is still lacking efficacy methods. Although Mayo or Brock model have been widely applied in clinical practices, their application conditions are limited. This study aims to construct a diagnostic model of PNs by machine learning using inflammation-related biological markers (IRBMs).

Methods: Inflammatory related genes (IRGs) were first extracted from GSE135304 chip data. Then, differentially expressed genes (DEGs) and infiltrating immune cells were screened between malignant pulmonary nodules (MN) and benign pulmonary nodule (BN). Correlation analysis was performed on DEGs and infiltrating immune cells. Molecular modules of IRGs were identified through Consistency cluster analysis. Subsequently, IRBMs in IRGs modules were filtered through Weighted gene co-expression network analysis (WGCNA). An optimal diagnostic model was established using machine learning methods. Finally, external dataset GSE108375 was used to verify this result.

Results: 4 hub IRGs and 3 immune cells showed significantly difference between MN and BN, C1 and C2 module, namely PRTN3, ELANE, NFKB1 and CTLA4, T cells CD4 naïve, NK cells activated and Monocytes. IRBMs were screened from black module and yellowgreen module through WGCNA analysis. The Support vector machines (SVM) was identified as the optimal model with the Area Under Curve (AUC) was 0.753. A nomogram was established based on 5 hub IRBMs, namely HS.137078, KLC3, C13ORF15, STOM and KCTD13. Finally, external dataset GSE108375 verified this result, with the AUC was 0.718.

Conclusion: SVM model established by 5 hub IRBMs was able to effectively identify MN or BN. Accumulating inflammation and immune dysfunction were important to the transformation from BN to MN.

[*] Corresponding author. The First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China.
E-mail address: linlizhu@gzucm.edu.cn (L. Lin).

[1] The author contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e34585
Received 28 April 2023; Received in revised form 11 July 2024; Accepted 11 July 2024
Available online 21 July 2024

1. Background

Abbreviations

Pulmonary nodules	PNs
Inflammatory related genes	IRGs
Differentially expressed genes	DEGs
Benign pulmonary nodule	BN
Malignant pulmonary nodules	MN
Inflammatory related biological markers	IRBMs
Weighted gene co-expression network analysis	WGCNA
Support vector machines	SVM
Area Under Curve	AUC
Low dose spiral CT	LDCT
Tumor microenvironment	TME
Neutrophil to lymphocyte ratio	NLR
Platelet to-lymphocyte ratio	PLR
Systemic immune inflammation index	SII
Principal Component Analysis	PCA
Receiver Operating Characteristic	ROC
Decision Curve Analysis	DCA
T cells regulatory	Tregs
Human proteinase 3	PRTN3
Neutrophil elastase	ELANE
Neutrophil serine proteases	NSPs
Chronic obstructive pulmonary disease	COPD
Vulvar squamous cell carcinoma	VSCC
Insulin receptor substrate-1	IRS-1
Progression-free survival	PFS
Small cell lung cancer	SCLC
Monocyte to lymphocyte ratio	MLR

Pulmonary nodules (PNs) is the quasi-circular focus with the diameter less or equal than 3 cm, with the imaging features of real or sub real shadows and without atelectasis, enlargement of hilar lymph nodes and pleural effusion [1]. Generally, the detective incidence of PNs was about 14 %–35.5 % in normal human [2]. Although most of PNs were benignant, some of them still were evolved into pulmonary cancer. The research showed that diameter of PNs was revealed to the incidence of deterioration. For example, the probability of 4∼6 mm PN was 0.5 %, the probability of 7∼10 mm PN was 1.7 % and the probability of PNs greater than 10 mm was 10 %–50 % [3]. In present, CT regular follow-up was suitable for the patients with low risk, while surgery or minimally invasive treatment was suitable for the patients with high risk [4]. With the improvement of people's safety awareness and the widespread employment of low dose spiral CT (LDCT), the detective rate of PNs was increased [5]. However, there was still existing circumstances that some patients were escaped diagnosis or misdiagnosed owing to the latent of PNs in iconography [6]. Although recent studies have developed some models used for the diagnosis of the probability of PNs, such as Mayo or Brock models, their applying conditions are usually limited [7,8]. Therefore, to develop a prediction model of malignant lung nodules (MN) and benign lung nodule (BN) is urgent.

As everyone knows, inflammation is a response of human body to tissue damage, which generally causes the changes of cells and immune responses thus accelerating cell proliferation and tissue repair [9]. Tumor immunology think that chronic inflammation was closely related to every stages of cancers, such as cells proliferation, invasion, angiogenesis and so on [10]. Most importantly, chronic inflammation enable tumor microenvironment (TME) to form a new circumstance that supporting tumor cells proliferation [11]. TME was a complex and highly heterogeneous dynamic integrated system [12]. In this system, infiltrating immune cells play dual roles. On the one hand, infiltrating immune cells kill tumor cells and inhibiting the development of cancer. On the other hand, infiltrating immune cells can promote growth of tumor or increase the risk of carcinogenic [13]. The studies have showed that the levels of inflammatory cytokine was increased and the immune function was decreased in lung cancer or pulmonary ground glass nodules comparing to BN [14]. Tian et al. [15]found that high levels of systematic inflammatory markers were related to positive PNs and lung cancer, such as NLR (neutrophil to lymphocyte ratio), PLR (platelet to-lymphocyte ratio) and SII (systemic immune inflammation index). Therefore, accumulating inflammation and immune dysfunction might be profit to BN transform to MN.

Machine Learning is an interdisciplinary science, which is able to identify patterns and trends in data and constantly learn from previous experience [16]. With the increasing complexity and quantity of clinical data, the application of machine learning can improve the quality of data interpretation in diagnosis, thus improving the efficiency and accuracy of diagnosis, which has a high reference value for clinicians. In present, the widespread machines learning methods includes RF, SVM, XGB and GLM [17–22].

Therefore, in this study, we identified novel inflammatory related biological markers (IRBMs) and used for the establishment of diagnostic model using machine learning methods. Finally, external dataset was used to verify this results.

2. Methods and materials

1. Type of the study

The transformation of BN to MN driven by inflammation is a key factor in tumor development. In this study, we used a bioinformatics analysis method to identify hub IRGs and IRBMs. Based on the hub IRBMs, 4 machine models including SVM were used to construct a prediction model of BN and MN. Finally, the ROC curve and DCA decision curve are used to evaluate the reliability of the model.

2. Datasets Acquiring and Processing

The flow of this study was showed in Fig. 1. The researches were carried out by MECHREVO Z2 Air Series GK5CP5X and windows 10 system. First of all, the chip data and platform file of GSE135304 and GSE108375 were acquired from GEO database (https://www.ncbi.nlm.nih.gov/geo/), visiting time: 19 October 2022. Then, we used Perl language to add annotation to the probe matrix. Subsequently, GSE135304 was served as the training group, while GSE108375 was served as the testing group. The patients in the groups were divided into malignant nodules (MN) and benign nodules (BN) according to the clinical manifestations.

3. Screening of Differentially expressed genes (DEGs) between MN and BN

First of all, the inflammatory related genes (IRGs) were downloaded from GeneCards database (The Human Gene Database) (https://www.genecards.org/) by setting the key word as “inflammation”, visiting time: 19 October 2022. In order to improve the confidence level and acquired more important IRGs, we simultaneously set the relevance score as “>15”. Then, the expression of IRGs were extracted from the training group, and DEGs were screened by differentially expressed analysis through limma package of R software 4.2.0. Finally, we displayed this DEGs on heat map and boxplot by the pheatmap package, reshape2 package and ggurb

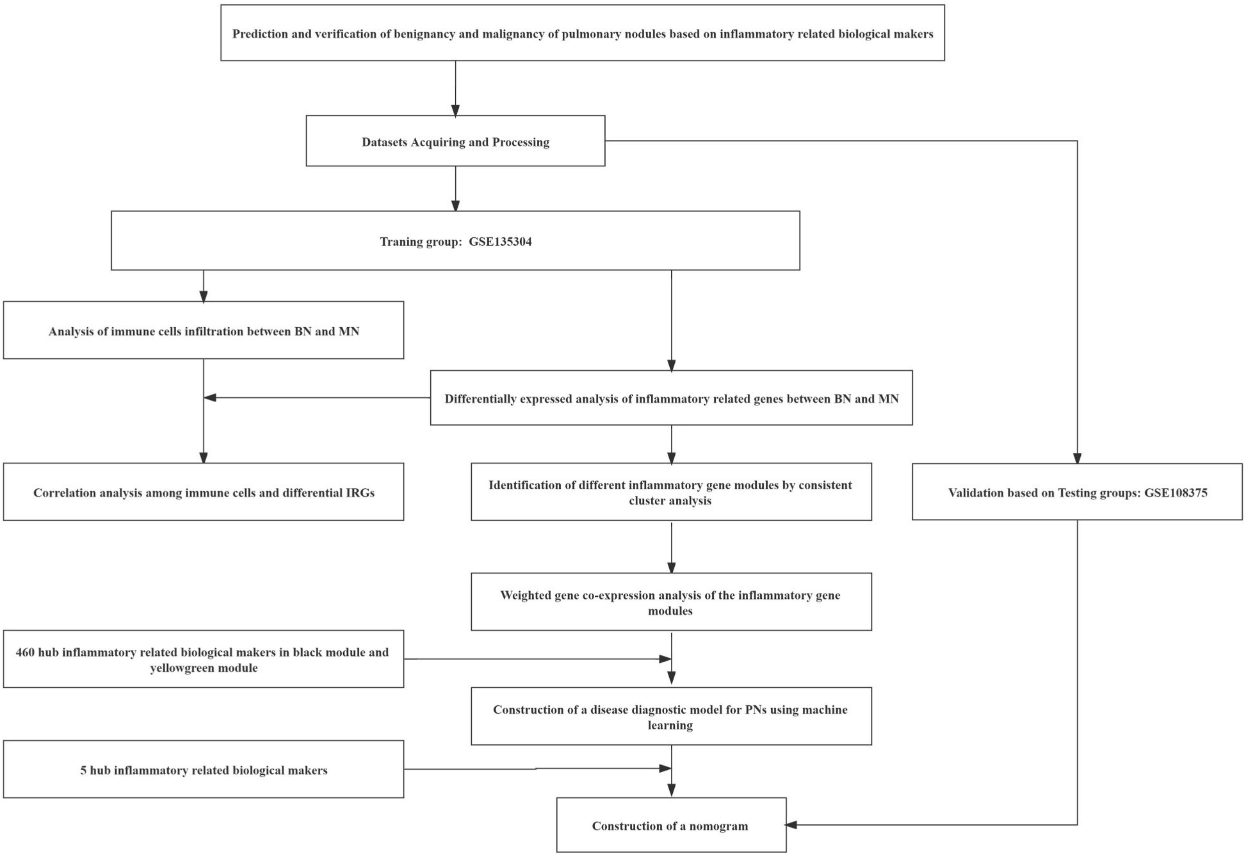
Prediction and verification of benignancy and malignancy of pulmonary nodules based on inflammatory related biological makers
Datasets Acquiring and Processing
Traning group: GSE135304
Analysis of immune cells infiltration between BN and MN
Differentially expressed analysis of inflammatory related genes between BN and MN
Correlation analysis among immune cells and differential IRGs
Identification of different inflammatory gene modules by consistent cluster analysis
Validation based on Testing groups: GSE108375
Weighted gene co-expression analysis of the inflammatory gene modules
460 hub inflammatory related biological makers in black module and yellowgreen module
Construction of a disease diagnostic model for PNs using machine learning
5 hub inflammatory related biological makers
Construction of a nomogram

Fig. 1. Flow of this study.

package of R software 4.2.0.

4. Immune cells infiltration between MN and BN

First of all, CIBERSORT algorithm was used to calculate the immune cells infiltration between MN and BN. Then, we displayed this differentially expressed immune cells by reshape2 package and gggurb package of R software 4.2.0. Moreover, corrplot package was used to perform correlation analysis among DEGs and immune cells.

5. Identification of IRGs molecular modules by Consistency cluster analysis

In order to distinguish different IRGs molecular modules, we used ConsensusClusterPlus package of R software 4.2.0 to carry out consistency cluster analysis according to the expression of IRGs. Furthermore, we also used limma package of R software 4.2.0 to detect differentially expressed IRGs between IRGs molecular modules. Simultaneously, Principal Component Analysis (PCA) was used to test whether we were able to distinguish different patient groups based on the expression of IRGs.

6. Weighted gene co-expression network analysis between IRGs molecular modules

In order to further screen hub IRBMs with the highest significant difference and correlation between IRGs molecular modules, Weighted gene co-expression network analysis was performed by WGCNA package of R language 4.2.0. In this step, we used WGCNA analysis to identify different gene modules between IRGs molecular modules, and this gene modules with the highest significant different and correlation were retained and used for further analysis.

7. Establishment of predictive model of malignant and benign pulmonary nodules by machine learning methods

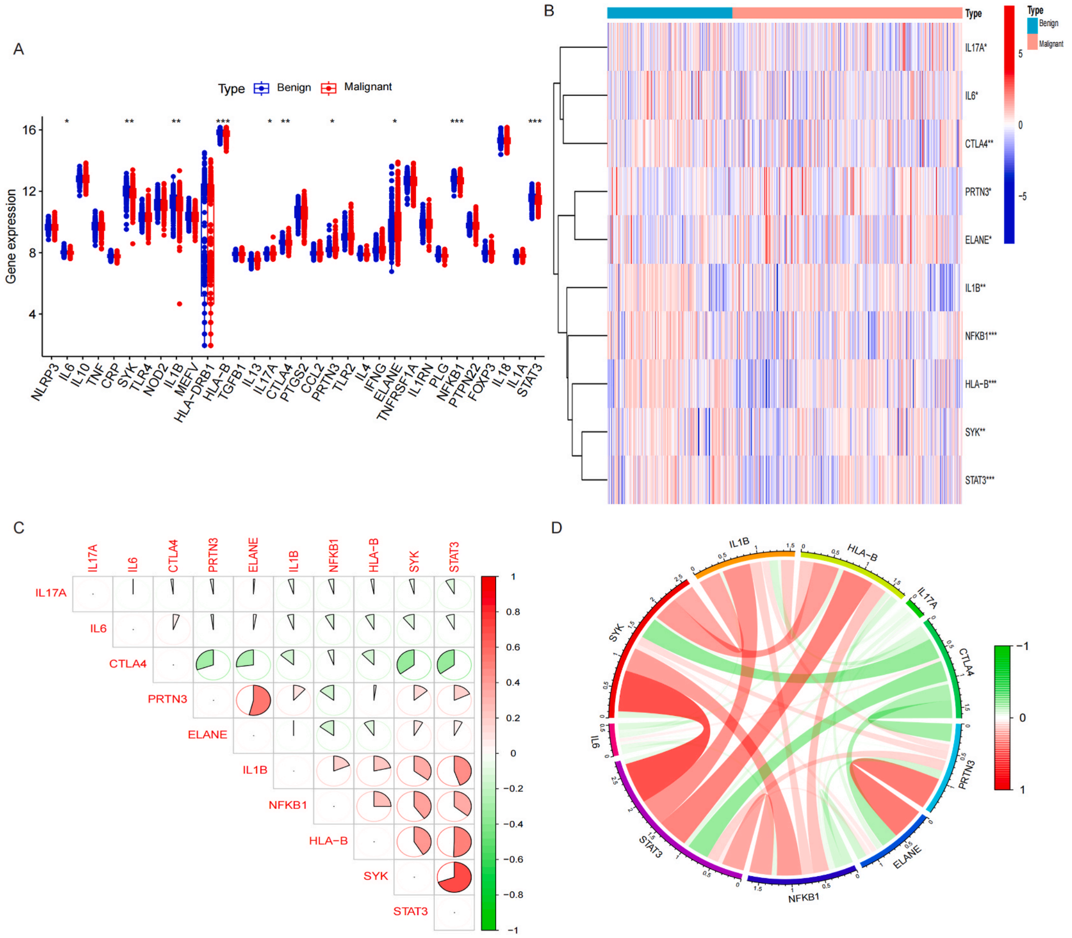

Fig. 2. Screening of Differentially expressed genes (DEGs) between MN and BN. A-B. Box plot and Heatmap showed that 10 IRGs has significant difference between MN and BN. C-D. Correlation analysis showed that PRTN3 and ELANE was negatively related to CTLA4 and NFKB1.

In order to establish of predictive model of MN and BN, we extracted hub IRBMs from the WGCNA analysis screened gene modules. Subsequently, caret package, DALEX package, randomForest package, kernlab package and xgboost package using the methods of repeatedcv、svmRadial、xgbDART and glm to establish RF, SVM, XGB and GLM machine models. Then, we combined Receiver Operating Characteristic (ROC) and residual to choose an optimal model. These 5 IRBMs with the highest important scores were recognized as the hub IRBMs and used for the construction of nomogram.

8. Construction of nomogram based on 5 IRBMs

In order to construct a model used for the diagnosis of MN, we used rms package and rmda package of R language to establish a nomogram. Every IRBMs will be given an independent score in the nomogram. The incidence of MN will be calculated by summing all the scores and come out a total score. Furthermore, we used Decision Curve Analysis (DCA) curve to test the efficacy of the nomogram. If the DCA curve of the model was separated far away from the DCA curve of all, suggesting that the model was reliable. Finally, we also used simulation curve to test the fitting degree of the model.

9. Verification of the nomogram by external dataset

The external dataset GSE108378 was added gene symbol by Perl language. Then, pROC package was used to verify the efficacy of the model. The AUC under ROC curve was greater than 0.6 will be recognized as strong reliability.

3. Results

1. Datasets Acquiring and Processing

GSE135304 included 220 BN samples and 404 MN samples. GSE108375 included 151 BN samples and 164 MN samples. Subsequently, the dataset GSE135304 was standardized as the training group. The patients with PNs in the training group were divided into BN and MN groups.

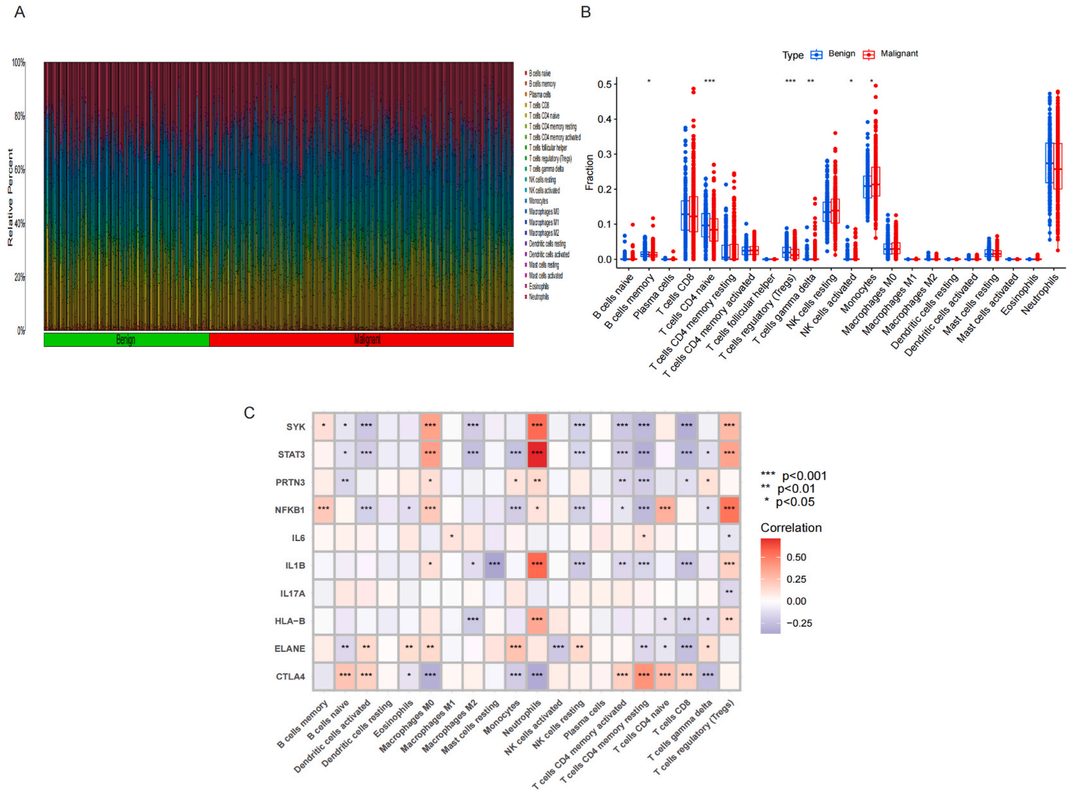

Fig. 3. Immune cells infiltration between MN and BN. A-B. Monocytes and T cells gamma delta of MN were significantly higher than BN, while B cells memory, T cells CD4 naïve, T cells regulatory (Tregs) and NK cells activated of MN were significantly lower than BN. C. CTLA4 and NFKB1 were positively correlated to T cells CD4 naïve, and negatively correlated to Monocytes. While PRTN3 and ELANE were positively correlated to Monocytes, and ELANE were negatively correlated to T cells CD4 naïve and NK cells activated.

2. Screening of Differentially expressed genes (DEGs) between MN and BN

33 IRGs were acquired from GeneCards database **(Supplement table 1). The differentially expressed analysis showed that 10 of 33 IRGs were significantly different between BN and MN groups, of which the expression of PRTN3, ELANE were higher in MN than BN, while the expression of IL6, SYK, IL1B, HLA-B, IL17A, CTLA4, NFKB1 and STAT3 were higher in BN than MN.** Fig. 2A-B**.** Correlation analysis showed that PRTN3 and ELANE was negatively related to CTLA4 and NFKB1. Fig. 2C-D.

3. Immune cells infiltration between MN and BN

CIBERSORT algorithm analysis showed that Monocytes and T cells gamma delta of MN were significantly higher than BN, while B cells memory, T cells CD4 naïve, T cells regulatory (Tregs) and NK cells activated of MN were significantly lower than BN. Fig. 3A-B. The analysis of correlation showed that CTLA4 and NFKB1 were positively correlated to T cells CD4 naïve, and negatively correlated to Monocytes. While PRTN3 and ELANE were positively correlated to Monocytes, and ELANE were negatively correlated to T cells CD4 naïve and NK cells activated. Fig. 3C.

4. Identification of IRGs molecular modules by Consistency cluster analysis

Consistency cluster analysis of the expression of IRGs showed that the internal consistency was higher than 0.8 and less promiscuous existed in different modules when the number of module was 2. Fig. 4. Then, 243 samples were assigned to C1 module, and 161 samples were assigned to C2 module. The differentially expressed analysis showed that CTLA4 and NFKB1 were upregulated in C1

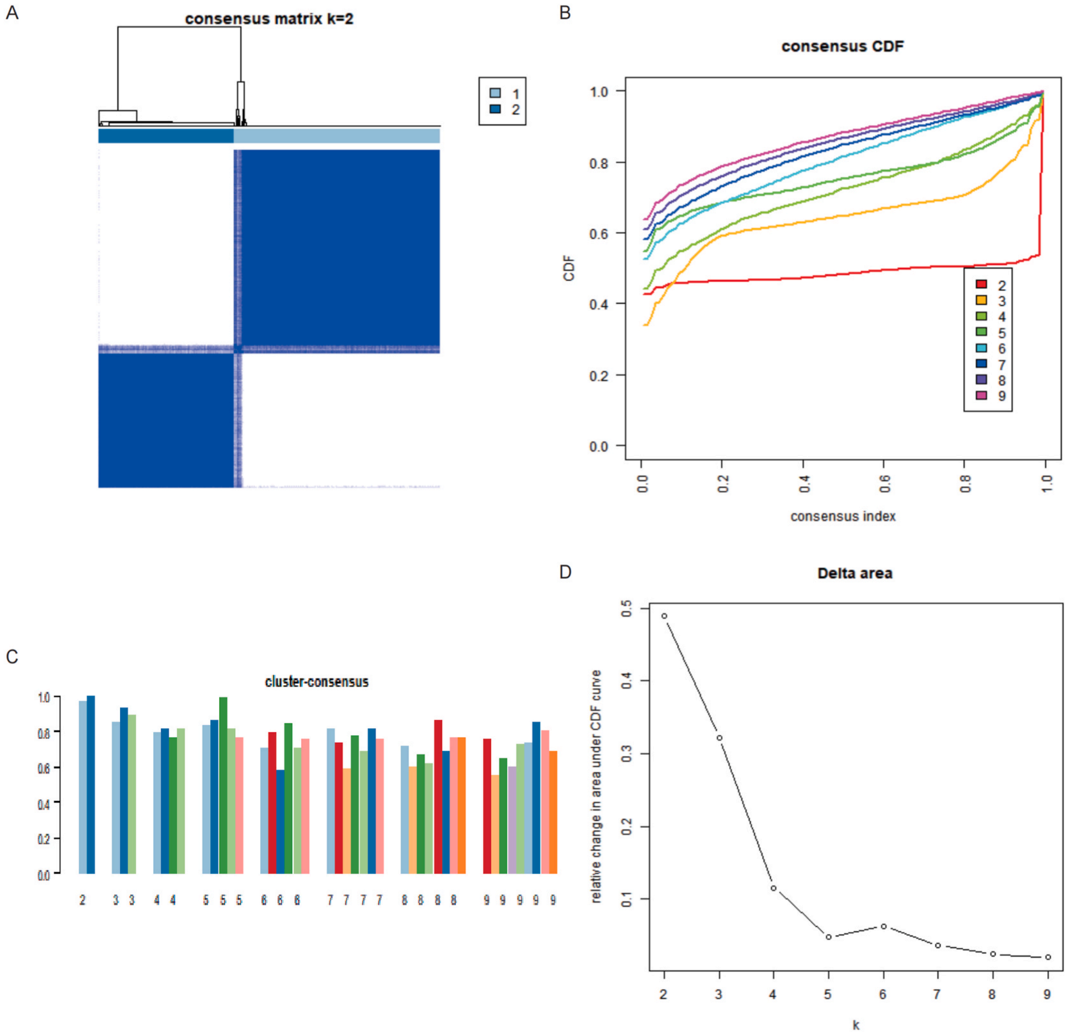

Fig. 4. Identification of IRGs molecular modules by Consistency cluster analysis. A. consensus matrix k = 2. B. consensus index changed with k value. C. Cluster consensus. D. Relative change in area under CDF curve.

module than C2 module, while PRTN3 and ELANE were downregulated in C1 module than C2 module. Fig. 5A-B. PCA analysis proved that patients with PNs were able to distinguish according to the expression of IRGs. Fig. 5C. The immune cells infiltration showed that B cells naïve, T cells CD8, T cells CD4 naïve, T cells CD4 memory resting, NK cells resting, NK cells activated, Monocytes, Mast cells resting and Eosinophils were significantly different between C1 and C2 modules. Fig. 6A-B.

5. Weighted gene co-expression network analysis between IRGs molecular modules

The results showed that WGCNA analysis between IRGs molecular modules divided into 12 gene modules when the $\beta = 13$ (Fig. 7A-B) of which the black module and yellowgreen module has the highest significantly difference and correlation with the p value 2e-08 and 4e-08, and the correlation 0.28 and 0.27. Fig. 7C-D. Therefore, we acquired and integrated the genes of these two modules and gained 460 IRBMs for the further analysis.

6. Establishment of predictive model of malignancy and benignancy pulmonary nodules by machine learning methods

The results showed that the AUC of the RF model was 0.749, the AUC of the SVM was 0.753, the AUC of the XGB was 0.694, the AUC of GLM was 0.500. Fig. 8A. The residual of models showed that SVM model was second to the XGB model. Fig. 8B. Therefore, based on the above results, the SVM model was the optimal model. Finally, the 5 hub IRBMs with highest importance was used for the establishment of nomogram, namely HS.137078, KLC3, C13ORF15, STOM and KCTD13. Fig. 8C.

7. Construction of nomogram based on 5 IRBMs

The results of nomogram showed that we were able to diagnose MN with the 90 percent of probability if the scores of 5 IRBMs

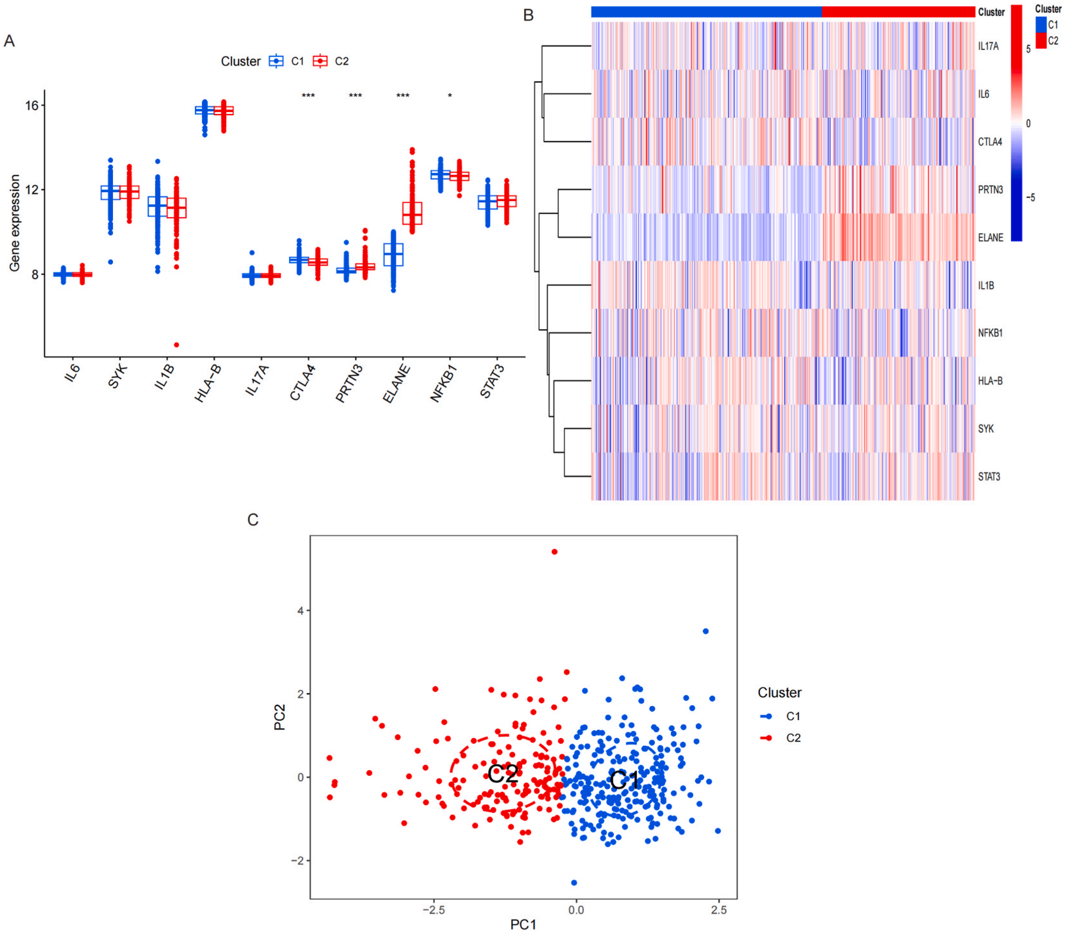

Fig. 5. Screening of Differentially expressed genes (DEGs) between C1 and C2 IRGs modules. A-B. Box plot and Heatmap showed that 4 IRGs has significant difference between C1 and C2 IRGs modules. C. PCA analysis showed that patients with PNs were able to distinguish according to the expression of IRGs.

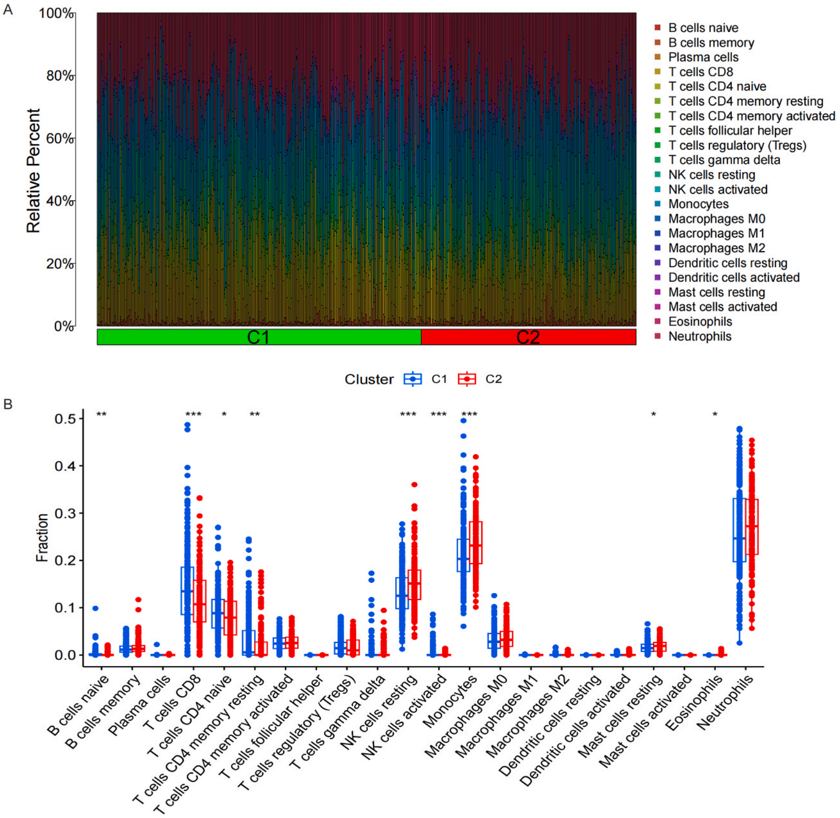

Fig. 6. Immune cells infiltration between C1 and C2 IRGs modules. A-B. B cells naïve, T cells CD8, T cells CD4 naïve, T cells CD4 memory resting, NK cells resting, NK cells activated, Monocytes, Mast cells resting and Eosinophils were significantly different between C1 and C2 modules.

ranking to 180. Fig. 9A and Fig. 9C. The DCA curve of model and simulation curve showed that the nomogram was reliable. Fig. 9B.

8. Verification of the nomogram by external dataset

The external dataset GSE108375 was used to verify the efficacy of the model. The GSE108375 included 164 MN samples and 151 BN samples. Because GSE108375 lacked STOM, we only used the remaining 4 IRBMs for detection. The results showed that the AUC of GSE108375 was 0.718, which proved the efficacy of the model. Fig. 10.

4. Discussion

Although Mayo and Brock model have been widely applied in the assessment of pulmonary nodules (PNs), they usually should evaluate the conditions of diameter, size, numbers of PNs and so on, which was strictly restricted their applying due to most of primary PNs lacking some clear features. Therefore, developing an optimal model to evaluate the benignancy and malignancy of pulmonary nodules is urgent.

In response to this issue, many studies developed some strategies. For example, Li et al. [23]predicted benignancy and malignancy of PNs by detecting DNA methylation in alveolar lavage fluid, with an AUC of up to 0.93. Fan et al. [24]developed a MicroRNA model for distinguishing NSCLC and BNs. He ea tl [25] conducted a bi-directional queue and used combination of clinical, imaging and cell-free DNA methylation to construct a prediction model for PNs. Although these models seem to have shown good predictive performance, the difficulty and cost of their detection in serum samples remain a major obstacle. In addition, as upstream markers, the pathways they regulate are still very complex. Therefore, using markers of functional expression detected by transcriptomics or proteomics to predict the malignancy of PNs will be more convenient and reliable.

As a matter of fact, an increasing number of studies have showed that inflammation was closely related to the development and occurrence of cancers. Chronic inflammation will cause damage to normal cells and tissues, and may induce gene mutations and cancers occurrence. Same to pulmonary nodules (PNs), some researches indicated that high levels of inflammation may exist in MN than BN. Shen et al. [10] discovered that proinflammatory factor elevated significantly in MN than BN and more patients with chronic inflammatory disease history in MN by retrospecting 100 patients with PNs. Tian et al. [15] found that the high levels of inflammation was related to the risk of positive PNs detection and lung cancer transformation. Systematic inflammatory factors such as NLR, PLR and

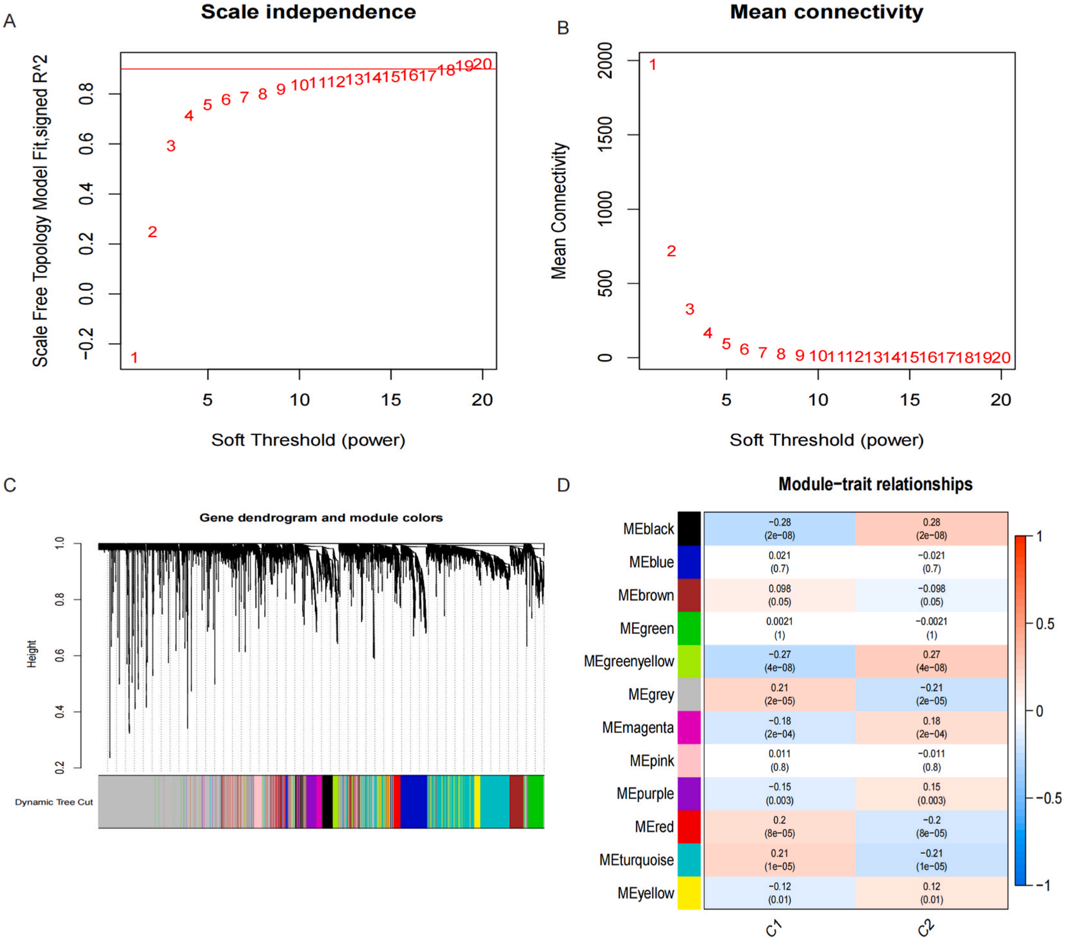

Fig. 7. Weighted gene co-expression network analysis between IRGs molecular modules. A-B. Scale independence and Mean connectivity changed with the soft threshold. C. Gene dendrogram and module colors. D. Module-trait relationships. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

SII were associated with positive PNs and lung cancer by retrospecting 96,476 patients with PNs in Chinese people. Fan et al. identified that five miRNA ratios were higher in NSCLC than BNs, and 13 miRNA ratio were up-regulated in NSCLC than pulmonary inflammation group [24]. Lai et al. [26]found that CST1 levels in serum was a reliable marker to distinguish early NSCLC and BNs. Notably, CST1 levels was higher in NSCLC than BNs, which may achieve transformation from BNs to MNs through promote inflammation. It's obvious that inflammation and immune dysfunction play a pivotal role in PNs. Therefore, to dig out novel IRBMs that use for the diagnosis of PNs is urgent.

In this study, 10 IRGs were significantly different between the BN and MN groups. In the 10 IRGs, PRTN3 and ELANE were expressed highly in MN group than in BN group, while IL6, SYK, IL1B, HLA-B, IL17A, CTLA4, NFKB1 and STAT3 were expressed lowly in MN group than in BN group. To be noted, we also found that PRTN3, ELANE, CTLA4 and NFKB1 showed significantly different between C1 and C2 module. Therefore, this four genes were recognized as the hub IRGs in the development and occurrence of PNs.

Human proteinase 3 (PRTN3) and Neutrophil elastase (ELANE) belongs to the family of neutrophil serine proteases (NSPs), which was released by activated neutrophils at the sites of inflammation, and are mediators of innate immune responses to microbial threats [27,28]. PRTN3 and ELANE were reported to be involved in many inflammatory diseases, such as chronic obstructive pulmonary disease (COPD) [29], sepsis [30] and so on. The levels of PRTN3 and ELANE generally reflect the degrees of inflammation. Of note, in some researches, PRTN3 and ELANE also showed significant roles in cancers. Agnieszka Fatalska et al. [31] found that PRTN3 was elevated in vulvar squamous cell carcinoma (VSCC) than in normal female. However, PRTN3 seemed wasn't relevant to the infection of hrHPV. Therefore, they believed prevention of bacterial infection may serve as a potential therapeutic strategy. Hu et al. [32] found that a low P4HA2 and high PRTN3 expression related to the poor prognosis of pancreatic cancer patients, and PRTN3 may serve as a potential therapeutic target of immunotherapy in pancreatic cancer. A McGarry Houghton et al. [33] found that neutrophil elastase, which was encoded by ELANE, was able to degrade insulin receptor substrate-1 (IRS-1) and accelerate proliferation of tumor cells. Intriguingly, we also observed that expression of PRTN3 was positively highly relevant to ELANE, which suggested that the interaction of PRTN3 and ELANE may further promote inflammation in MN.

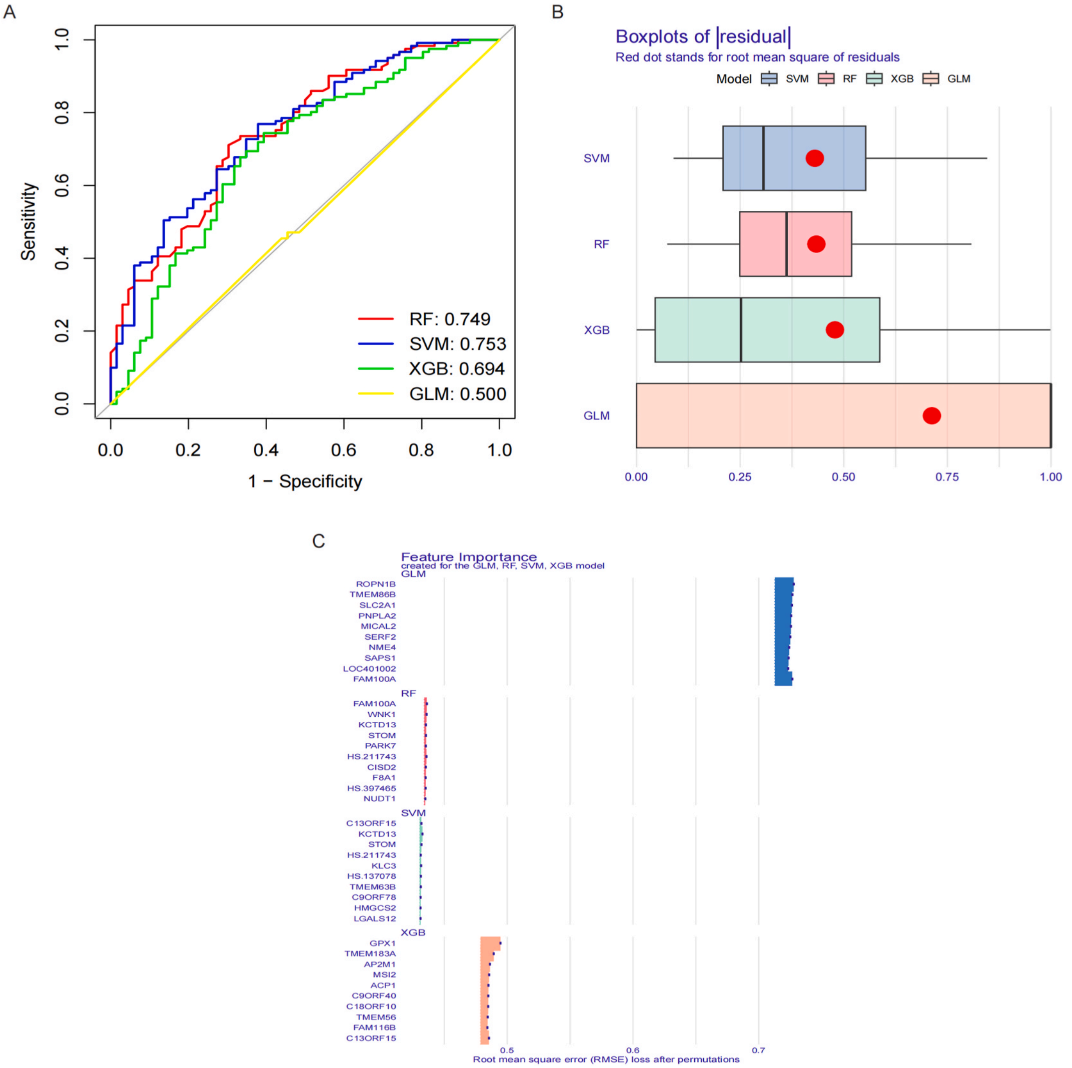

Fig. 8. Establishment of predictive model of malignancy and benignancy pulmonary nodules by machine learning methods. A. ROC curve showed that the AUC of four models was 0.749, 0.753, 0.694 and 0.500, separately. B. Boxplots of residual. C. Feature importance created for the GLM, RF, SVM and XGB models.

NFKB1 is a member of the NF- kB family. Although NFKB1 was generally recognized as an inflammatory promoter [34], the recent study showed that the p50 homodimers can serve as a tumor suppressor that playing anti-inflammatory roles in liver cancer [35]. C.L. Wilson et al. [35] defined that NFKB1 was able to transcriptionally restrain the neutrophil chemokine network through the expression of the p50 homodimers. Yelena Kravtsova-Ivantsiv et al. [36] proved that the overexpression of KPC-1 repressed the proliferation of tumor through generation of p50. Moreover, p50 homodimers was able to downregulate p65 and then take place of tumorigenic p65. As a leukocyte differentiation antigen, CTLA4 can compete B7 molecules with CD28 on the surface of T cells, thus playing an immunosuppressive role [37]. Although high levels of CTLA4 generally reflect an immunosuppressive state, interestingly, we observed that high levels of CTLA4 in BN and CTLA4 was positively correlated to T cells CD4 and NK cells activated and negatively with monocytes. This results may suggest that a lower levels of inflammation and a higher levels of immunity ability in preventing BN from transforming to MN.

The analysis of immune cells infiltration showed that T cells CD4 naive, NK cells activated and Monocytes were three the hub immune cells that both significantly different between in BN and MN, C1 and C2 modules. Moreover, we surprisingly found that CTLA4 and NFKB1 were positively correlated to T cells CD4 naïve, and negatively correlated to Monocytes. While PRTN3 and ELANE were positively correlated to Monocytes, and ELANE were negatively correlated to T cells CD4 naïve and NK cells activated.

T cells CD4 is one of the important components of human immune cells, which represents the intensity of immune system. They help the body respond to infection and disease by coordinating and regulating immune responses. T cells CD4 naïve are CD4$^+$ T cells that have not been activated or differentiated.They can be activated and develop into a variety of functionally specific CD4$^+$ T cells, such as Th1,Th2,Th17 and Treg [38]. Liu et al. [39] found that T cells CD4 naïve was higher in healthy control than in advanced

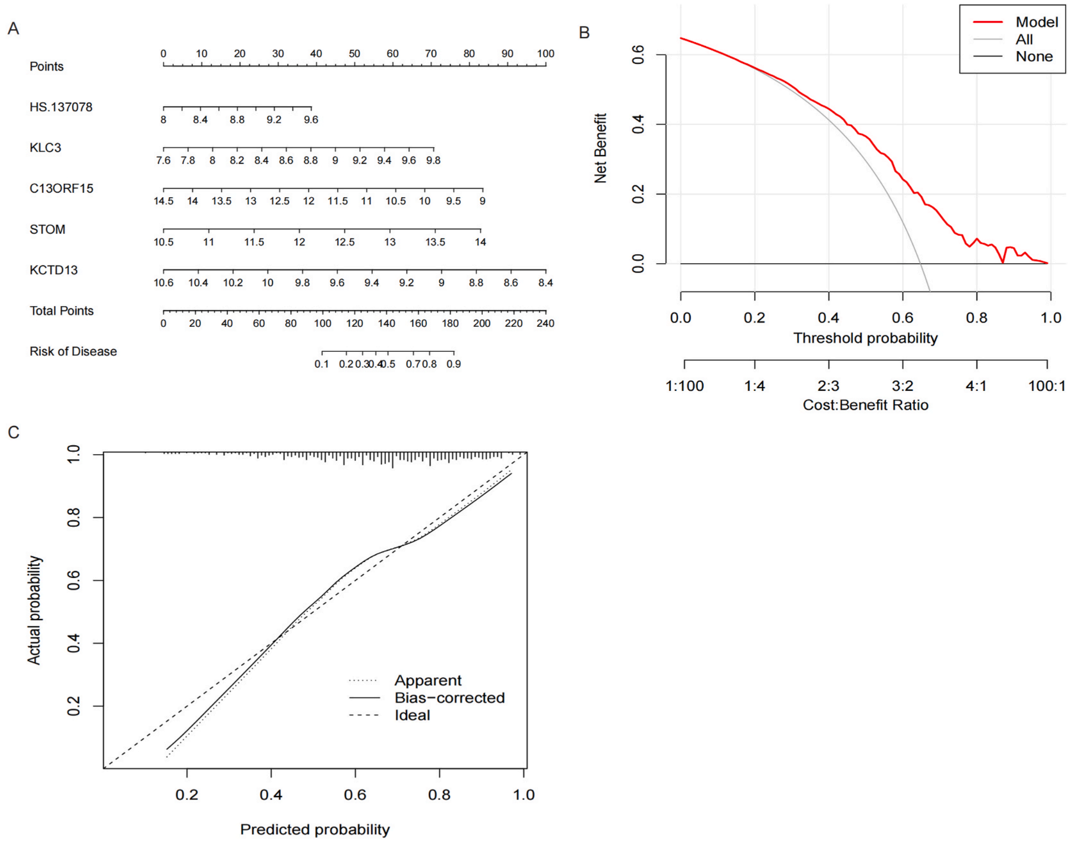

Fig. 9. Construction of nomogram based on 5 IRBMs. A. Prediction of risk of disease according to the scores of 5 IRBMs. B. Net benefit of different decisions. C. Confidence of fit between actual probability and predicted probability.

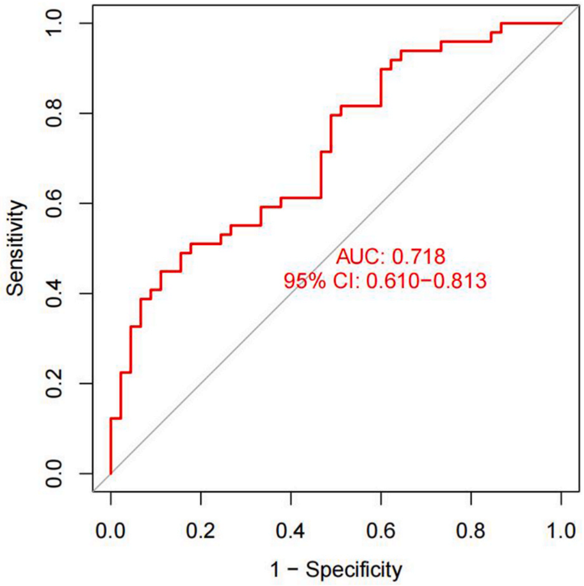

Fig. 10. Verification of the nomogram by external dataset.

NSCLC, which reflected a favorable prognosis. Yang et al. [40] found that decreased T cells CD4 naïve and CD4+naïve/memory ratios in NSCLC than healthy volunteers, and this was correlated to a poor progression-free survival (PFS). NK cells is another important immune cells in human body, which is able to kill invaders indiscriminately, playing a pivotal role in inhibiting tumors. The concept of NK cells therapy in cancer was proposed for ages. Rosenberg et al. [41] proved that NK cells were able to be activated by peripheral blood lymphocytes thus dissolving autologous cancer cells. Ding et al. [42] developed a therapeutic strategy of NK cell-based adoptive immunotherapy, which was able to prolong the survival times of small cell lung cancer (SCLC). Monocytes are affected by cytokines and inflammatory chemokines to recruit to corresponding sites, and then differentiate into tumor associated macrophages [43]. However, tumor associated macrophages are able to promote angiogenesis and tumor growth. Therefore, the high levels of Monocyte to lymphocyte ratio (MLR) have been proved to be associated to the poor prognosis of patients [44].

5 novel IRBMs were finally identified by machine learning methods, namely HS.137078, KLC3, C13ORF15, STOM and KCTD13, of which 4 IRBMs has been studied. However, these 4 novel IRBMs weren't been reported in PNs or lung cancer. KLC3 was reported to be related to the motion of sperm, which was highly expressed in the testis of rat [45]. Claudia Cava et al. [46] identified novel biomarker KLC3 was specific for HER-2 and was related to the poor prognosis in breast cancer. Christiane M. Robbins et al. [47] used next generation sequencing to detect mutation of 3508 exon from 577 cancer related genes and found that C13ORF15 mapping within the deleted regions of 8p22, 13q13.1, 13q14.11, 10q23.31 and 13q14.12 in advanced prostate cancer. Sandra N. Schlick et al. [48] found that C13ORF15 was involved in the transformation process of EBV. Chen et al. [45,49] proved that STOM over-expressed in non-fusogenic JEG-3 choriocarcinoma placental cells, which was able to trigger syncytium formation and upregulate β-hCG for cell fusion. Monique D Appelman et al. [50] proved that STOM was able to interact with sodium taurocholate cotransporting polypeptide thus modulating bile salt transport. KCTD13 was reported to be relevant to many mental diseases, such as autistic features [51], schizophrenia [52] and so on. Christine Ochoa Escamilla et al. [53] proved that KCTD13 was involved in the regulation of neuronal function and play a significant role in brain size and neurogenesis. Although these 4 IRBMs hasn't been studied in PNs or lung cancer, they may showed a promising prospect.

5. Conclusion

In this study, we identified a total of 4 hub IRGs, namely PRTN3, ELANE, CTLA4 and NFKB1, which are closely related to inflammation and immune regulation, and may be key targets affecting the benign and malignant transformation of PNs. In addition, we also identified 5 novel IRBMs, namely HS.137078, KLC3, C13ORF15, STOM and KCTD13. Although 4 of them have been reported, their relationship with PNs or lung cancer is still unclear, and they have good research prospects. We used these 5 novel IRBMs to build a machine learning model and found that it can provide an effective prediction for the diagnosis of benign and malignant PNs, and we look forward to further verification in clinical practice.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data of this study were acquired from GEO database and Genecards database.

Funding

This work was funded by the Key R&D Plan of Guangzhou Municipal Bureau of Science and Technology (SL2022B01J10007); Key R&D Plan of the Ministry of Science and Technology of China (2022YFC3500203); Basic Research Program of Guangzhou Municipal Bureau of Science and Technology (2023A03J0300); Guangdong Provincial Basic and Applied Basic Research Fund (2022B1515230003);

Basic Theoretical Research of Chinese Medicine of Guangzhou University of Chinese Medicine (Jiebang Guashuai Lizhu Lin project, Grant No. Project [2022]-7).

ZZ: Conceptualization, Methodology, Writing original draft. WW: Writing original draft, Investigation. LX: Writing original draft. LS: Writing original draft. LQ: Investigation. YL: Investigation. LJ: Investigation. SL: Investigation. ZH: Writing–review & editing. LL: Writing–review & editing. All authors read and approved the final manuscript.

CRediT authorship contribution statement

Zexin Zhang: Writing – original draft, Methodology, Conceptualization. **Wenfeng Wu:** Investigation. **Xuewei Li:** Writing – original draft. **Siqi Lin:** Writing – original draft. **Qiwei Lei:** Investigation. **Ling Yu:** Investigation. **Jietao Lin:** Investigation. **Lingling Sun:** Investigation. **Haibo Zhang:** Writing – review & editing. **Lizhu Lin:** 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.

Acknowledgments

Not applicable.

References

[1] M.K. Gould, Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines, Chest 143 (5 Suppl) (2013) e93S–e120S.

[2] M.E. Callister, British Thoracic Society guidelines for the investigation and management of pulmonary nodules, Thorax 70 (Suppl 2) (2015) ii1–ii54.

[3] P.J. Mazzone, L. Lam, Evaluating the patient with a pulmonary nodule: a review, JAMA 327 (3) (2022) 264–273.

[4] Y. Lv, B. Ye, Advances in diagnosis and management of subcentimeter pulmonary nodules, Zhongguo Fei Ai Za Zhi 23 (5) (2020) 365–370.

[5] Q.X. Liu, A noninvasive multianalytical approach for lung cancer diagnosis of patients with pulmonary nodules, Adv. Sci. 8 (13) (2021) 2100104.

[6] D. Liu, Pulmonary nodules/lung cancer comprehensive management mode: design and application, Zhongguo Fei Ai Za Zhi 23 (5) (2020) 299–305.

[7] 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.

[8] A. McWilliams, Probability of cancer in pulmonary nodules detected on first screening CT, N. Engl. J. Med. 369 (10) (2013) 910–919.

[9] N. Singh, Inflammation and cancer, Ann. Afr. Med. 18 (3) (2019) 121–126.

[10] Shen, C., et al., Establishment of a Malignancy and Benignancy Prediction Model of Sub-centimeter Pulmonary Ground-Glass Nodules Based on the Inflammation-Cancer Transformation Theory. (2296-858X (Print)).

[11] L. Budisan, Links between infections, lung cancer, and the immune system, Int. J. Mol. Sci. 22 (17) (2021).

[12] B. Arneth, Tumor microenvironment, Medicina (Kaunas) 56 (1) (2019).

[13] A. Merlo, Reverse immunoediting: when immunity is edited by antigen, Immunol. Lett. 175 (2016) 16–20.

[14] C.Y. Liu, A comparative study on inflammatory factors and immune functions of lung cancer and pulmonary ground-glass attenuation, Eur. Rev. Med. Pharmacol. Sci. 21 (18) (2017) 4098–4103.

[15] T. Tian, Associations of Systemic Inflammation Markers with Identification of Pulmonary Nodule and Incident Lung Cancer in Chinese Population, 2024, p. 7634 (Electronic)).

[16] G.S. Handelman, eDoctor: machine learning and the future of medicine, J. Intern. Med. 284 (6) (2018) 603–619.

[17] L.M. Pehrson, M.B. Nielsen, C. Ammitzbøl Lauridsen, Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the LIDC-IDRI database: a systematic review, Diagnostics 9 (1) (2019).

[18] R.C. Deo, Machine learning in medicine, Circulation 132 (20) (2015) 1920–1930.

[19] M.E. Ozer, P.O. Sarica, K.Y. Arga, New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines, OMICS 24 (5) (2020) 241–246.

[20] C. Yan, PESM: predicting the essentiality of miRNAs based on gradient boosting machines and sequences, BMC Bioinf. 21 (1) (2020) 111.

[21] Y. Liang, W. Zheng, W.S. Lee, Nonlinear associations between medical expenditure, perceived medical attitude, and sociodemographics, and older adults' self-rated health in China: applying the extreme gradient boosting model, Healthcare (Basel) 10 (1) (2021).

[22] J. Lee, Comparative analysis on machine learning and deep learning to predict post-induction hypotension, Sensors 20 (16) (2020).

[23] L. Li, Diagnosis of pulmonary nodules by DNA methylation analysis in bronchoalveolar lavage fluids, Clin Epigenetics 13 (1) (2021) 185.

[24] L. Fan, Evaluation of serum paired MicroRNA ratios for differential diagnosis of non-small cell lung cancer and benign pulmonary diseases, Mol. Diagn. Ther. 22 (4) (2018) 493–502.

[25] J. He, Accurate classification of pulmonary nodules by a combined model of clinical, imaging, and cell-free DNA methylation biomarkers: a model development and external validation study, Lancet Digit Health 5 (10) (2023) e647–e656.

[26] Y. Lai, Identification and validation of serum CST1 as a diagnostic marker for differentiating early-stage non-small cell lung cancer from pulmonary benign nodules, Cancer Control 29 (2022) 10732748221104661.

[27] Meyer-Hoffert, U. and O. Wiedow, Neutrophil Serine Proteases: Mediators of Innate Immune Responses. (1531-7048 (Electronic)).

[28] Korkmaz, B., et al., Neutrophil elastase, proteinase 3, and cathepsin G as therapeutic targets in Human Diseases. (1521-0081 (Electronic)).

[29] N.S. Gudmann, Lung tissue destruction by proteinase 3 and cathepsin G mediated elastin degradation is elevated in chronic obstructive pulmonary disease, Biochem. Biophys. Res. Commun. 503 (3) (2018) 1284–1290.

[30] Zhang, S., et al., Time Series Gene Expression Profiles Analysis Identified Several Potential Biomarkers for Sepsis. (1557-7430 (Electronic)).

[31] Fatalska, A.A.-O., et al., Inflammatory Proteins HMGA2 and PRTN3 as Drivers of Vulvar Squamous Cell Carcinoma Progression. *LID - 10.3390/cancers13010027 [doi] LID - 27.* (2072-6694 (Print)).

[32] Hu, D.A.-O., et al., Low P4HA2 and High PRTN3 Expression Predicts Poor Survival in Patients with Pancreatic Cancer. (1502-7708 (Electronic)).

[33] Houghton, A.M., et al., Neutrophil Elastase-Mediated Degradation of IRS-1 Accelerates Lung Tumor Growth. (1546-170X (Electronic)).

[34] DiDonato, J.A., M. Mercurio F Fau - Karin, and M. Karin, NF-κB and the Link between Inflammation and Cancer. (1600-065X (Electronic)).

[35] C.L. Wilson, NFκB1 Is a Suppressor of Neutrophil-Driven Hepatocellular Carcinoma, 2024, p. 1723 (Electronic)).

[36] Kravtsova-Ivantsiv, Y., et al., KPC1-mediated Ubiquitination and Proteasomal Processing of NF-Kb1 P105 to P50 Restricts Tumor Growth. (1097-4172 (Electronic)).

[37] Korman, A.A.-O., S.C. Garrett-Thomson, and N.A.-O. Lonberg, The Foundations of Immune Checkpoint Blockade and the Ipilimumab Approval Decennial. (1474-1784 (Electronic)).

[38] J. Zhu, H. Yamane, W.E. Paul, Differentiation of effector CD4 T cell populations (*), Annu. Rev. Immunol. 28 (2010) 445–489.

[39] Liu, C., et al., Smoking History Influences the Prognostic Value of Peripheral Naïve CD4+ T Cells in Advanced Non-small Cell Lung Cancer. (1475-2867 (Print).

[40] Yang, P., et al., Peripheral CD4+ Naïve/memory Ratio Is an Independent Predictor of Survival in Non-small Cell Lung Cancer. (1949-2553 (Electronic)).

[41] Rosenberg Sa Fau - Eberlein, T.J., et al., Development of Long-Term Cell Lines and Lymphoid Clones Reactive against Murine and Human Tumors: a New Approach to the Adoptive Immunotherapy of Cancer. (39-6060 (Print)).

[42] Ding, X., et al., Cellular Immunotherapy as Maintenance Therapy Prolongs the Survival of the Patients with Small Cell Lung Cancer. (1479-5876 (Electronic)).

[43] Wang, Y., et al., Targeted Therapy of Atherosclerosis by a Broad-Spectrum Reactive Oxygen Species Scavenging Nanoparticle with Intrinsic Anti-inflammatory Activity. (1936-086X (Electronic)).

[44] Bilen, M.A.-O., et al., The Prognostic and Predictive Impact of Inflammatory Biomarkers in Patients Who Have Advanced-Stage Cancer Treated with Immunotherapy. (1097-0142 (Electronic)).

[45] Junco, A., et al., Kinesin Light-Chain KLC3 Expression in Testis Is Restricted to Spermatids. (6-3363 (Print)).

[46] Cava, C., et al., Identification of Long Non-coding RNAs and RNA Binding Proteins in Breast Cancer Subtypes. (2045-2322 (Electronic)).

[47] Robbins, C.M., et al., Copy Number and Targeted Mutational Analysis Reveals Novel Somatic Events in Metastatic Prostate Tumors. (1549-5469 (Electronic)).

[48] Schlick, S.N., et al., Upregulation of the Cell-Cycle Regulator RGC-32 in Epstein-Barr Virus-Immortalized Cells. (1932-6203 (Electronic)).
[49] Chen, T.W., et al., Over-expression of Stomatin Causes Syncytium Formation in Nonfusogenic JEG-3 Choriocarcinoma Placental Cells. (1095-8355 (Electronic)).
[50] Appelman, M.D., et al., The Lipid Raft Component Stomatin Interacts with the Na(+) Taurocholate Cotransporting Polypeptide (NTCP) and Modulates Bile Salt Uptake. LID - 10.3390/cells9040986 [doi] LID - 986. (2073-4409 (Electronic)).
[51] Madison, J.M., et al., Regulation of Purine Metabolism Connects KCTD13 to a Metabolic Disorder with Autistic Features. (2589-0042 (Electronic)).
[52] Degenhardt, F., et al., Identification of Rare Variants in KCTD13 at the Schizophrenia Risk Locus 16p11.2. (1473-5873 (Electronic)).
[53] Escamilla, C.O., et al., Kctd13 Deletion Reduces Synaptic Transmission via Increased RhoA. (1476-4687 (Electronic)).