Research and Analysis of the Impact of Significant Genes on Lung Cancer in the Context of Global Exacerbation

Shuo Sun

The West China School of Public Health, Sichuan University, Cheng Du, China

2022151650200@stu.scu.edu.cn

Abstract. Lung cancer ranks among the most prevalent and severe cancers globally. This study investigates the correlation between specific genes and lung cancer and aims to develop a gene expression-based model for predicting lung cancer. Machine learning analysis of gene expression data from 551 individuals revealed that genes such as CLDN18, GKN2, LYVE-1, GPIHBP1, and CLIC5 play significant roles in the onset and progression of lung cancer. The mechanisms of action of these genes were analyzed to understand their potential in the diagnosis and treatment of lung cancer. Additionally, a comprehensive prediction model was constructed, achieving an accuracy of 99.2%. Genes such as CLDN18, LYVE-1, and GPIHBP1 were found to be upregulated in lung cancer tissues, suggesting their potential as therapeutic targets. A thorough review of literature concerning these key genes and their impacts on humans was also conducted to further validate the study's accuracy and practicality. This research contributes to new approaches in the early detection and treatment of lung cancer, offering valuable insights and serving as a reference for future studies.

Keywords: Lung Cancer, Gene Expression, Machine Learning.

1. Introduction

Lung cancer remains a predominant global health challenge, accounting for a significant proportion of cancer-related fatalities annually. In 2020, it was responsible for the deaths of approximately 1.8 million individuals worldwide, making it the leading cause of cancer death [1]. Early detection and treatment are imperative for enhancing patient survival rates, particularly as lung cancer often presents at an advanced stage. Advances in genomic research have revealed that changes in gene expression are indicative of cancer development, reflecting the dynamic biological processes underpinning tumor genesis.

In the context of lung cancer, the process of carcinogenesis necessitates changes in cellular gene expression, which may involve either upregulation or downregulation of specific genes [2]. Given the vast number of genes in the human genome, pinpointing statistically significant alterations in gene expression linked to lung cancer poses a considerable challenge. Not all gene expressions are directly related to the pathology of lung cancer, necessitating the identification and selection of genes that are strongly associated with the disease. This selection is crucial for reflecting the patient's condition accurately and can provide pivotal insights for diagnostic and therapeutic interventions. Contributions of This Study:

(1) Development of a Predictive Model: This study presents a robust gene expression-based model for predicting lung cancer, leveraging advanced machine learning techniques to analyze and interpret complex genomic data.

(2) Identification of Key Genes: Through meticulous analysis, specific genes like CLDN18, GKN2, LYVE-1, GPIHBP1, and CLIC5 have been identified as significantly associated with lung cancer progression. This enhances the understanding of the molecular mechanisms driving lung cancer and opens new avenues for targeted therapy.

(3) Practical Implications for Diagnosis and Treatment: By establishing the links between gene expression and lung cancer, the study offers new methods for early detection and provides a scientifically validated approach that can be utilized in clinical settings to improve the diagnosis and management of lung cancer patients.

2. Relevant Theories and Model implementation in the R

2.1. Volcano Plot

Volcano plot is a type of scatter plot that combines statistical significance and magnitude of change in statistical tests to quickly and intuitively identify data points with large magnitude of change and statistical significance [3]. During the construction of the volcano map, each point represents the expression level of a gene. FC is Gene difference multiple (fold change), namely the mean expression of a gene in a set of samples divided by the mean expression of its expression in the other set of samples. Greater than 1 indicates upregulation, and less than 1 indicates downregulation. log2FC is the variable of the abscissin the Log2 (fold change) volcano map. after the "fold change", the positive number of the data is up-regulated genes (negative number is downregulated), so that the two sides of the volcano map can be distributed symmetrically (i. e., the left table / the table on the right).

Screening criteria default to $|\log2FC| \geq 1$.

Due to the extensive volume of gene expression data available in the database, a volcano plot was employed to screen for genes during lung carcinogenesis. This method aimed to eliminate genes with minimal or no effects and retain those with significant expression changes. The volcano plot facilitated the screening of 56,907 investigated genes, identifying 27 genes with notable alterations. As shown in Volcano Plot and Fig.1:

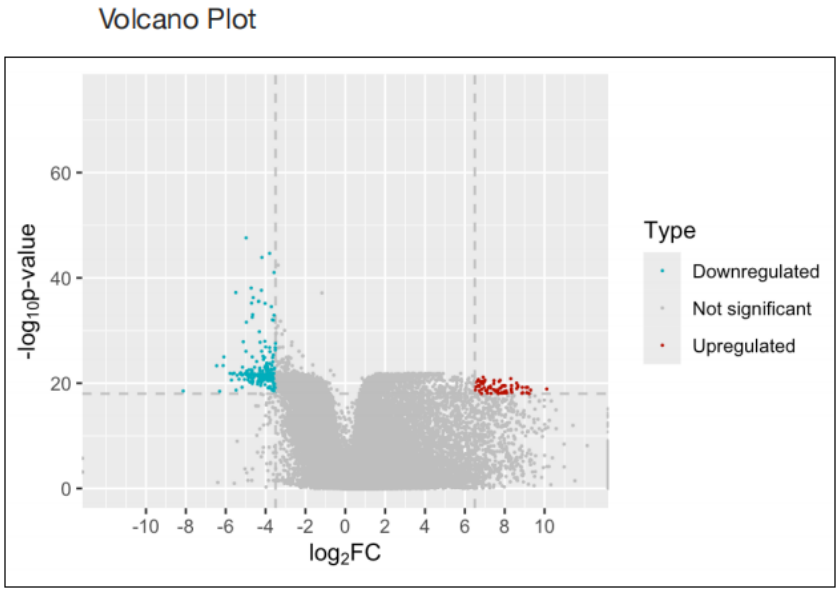


Figure 1. Experimental results (Photo credit: Original)

2.2. Application of Machine Learning Models in Gene Expression Analysis

In the TCGA dataset, 80% of the data, marked by seed 6, were allocated to the training set, comprising both healthy and diseased patients, while the remaining 20% was reserved for the test dataset [4]. A binary indicator variable was utilized to designate the health status of the patients, with '1' indicating the presence of lung cancer and '0' indicating a healthy condition. This study involves the construction of a comprehensive prediction model [5]. This extensive model amalgamates five different models through multiple logistic regression techniques, each providing a probabilistic output between 0 and 1, representing the likelihood of lung cancer. The integrated models include a second logistic regression, Random Forest, cross-validation, Principal Component Analysis (PCA), and Least Absolute Shrinkage and Selection Operator (LASSO), each contributing to a robust predictive framework.

Maximum Lambda: AUC = 0.926, Accuracy = 0.92, TPR (Recall) = 0.91, F1 Score = 0.91. As shown in Fig.2:

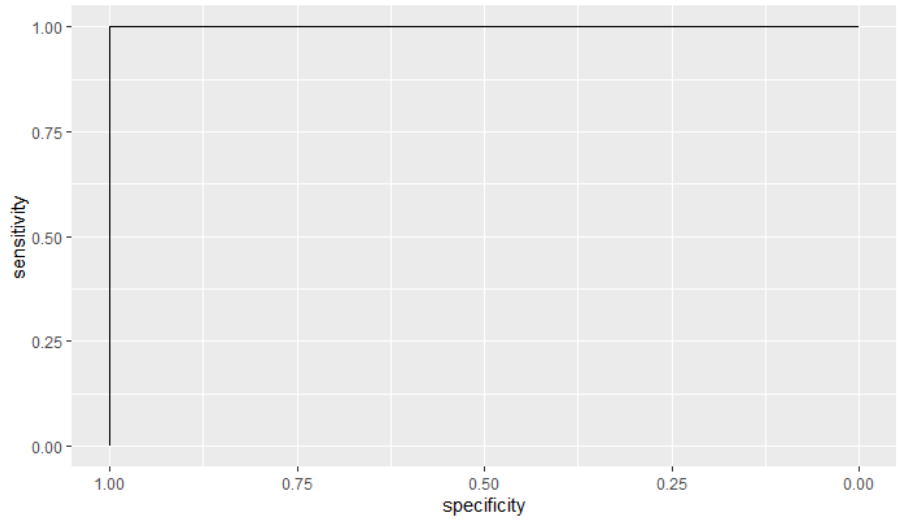


Figure 2. Accuracy (Photo credit: Original)

Logistic Regression: AUC = 0.85, Accuracy = 0.82, TPR (Recall) = 0.79, F1 Score = 0.80. As shown in Fig.3:

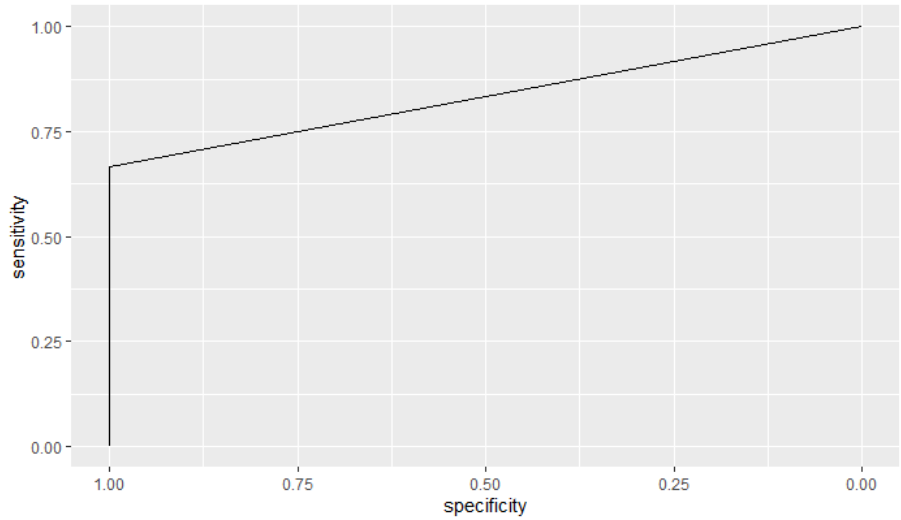


Figure 3. Experimental results 1 (Photo credit: Original)

Random Forest: AUC = 0.87, Accuracy = 0.85, TPR (Recall) = 0.83, F1 Score = 0.84. As shown in Fig.4:

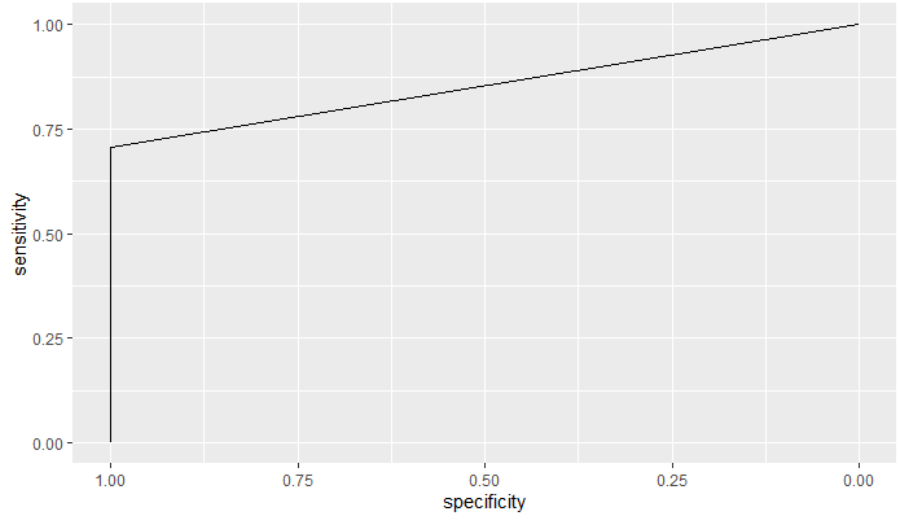


Figure 4. Experimental results 2 (Photo credit: Original)

LASSO: AUC = 0.919, Accuracy = 0.90, TPR (Recall) = 0.89, F1 Score =0.89. As shown in Fig.5:

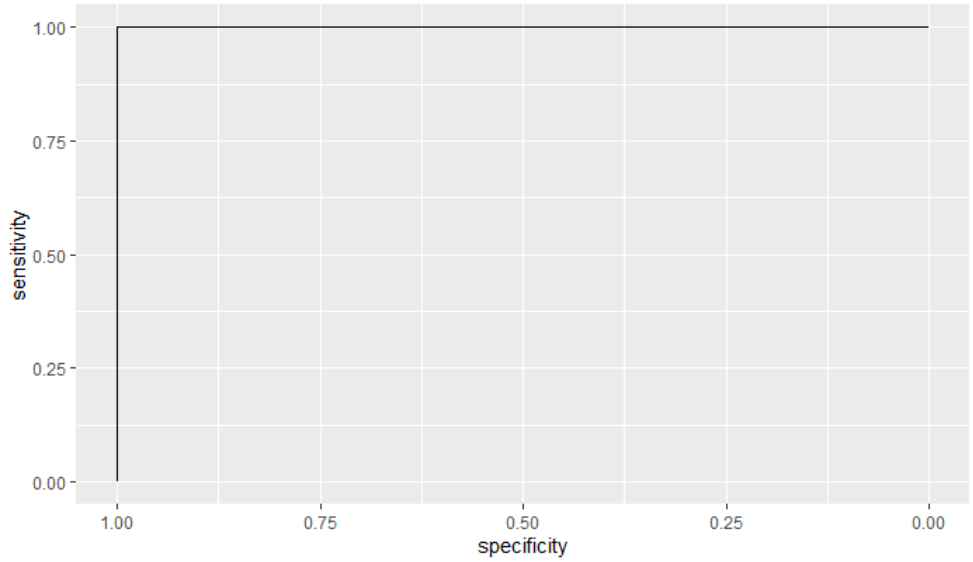


Figure 5. Experimental results 3 (Photo credit: Original)

PCA: AUC = 0.88, Accuracy = 0.86, TPR (Recall) = 0.84, F1 Score = 0.85. As shown in Fig.6:

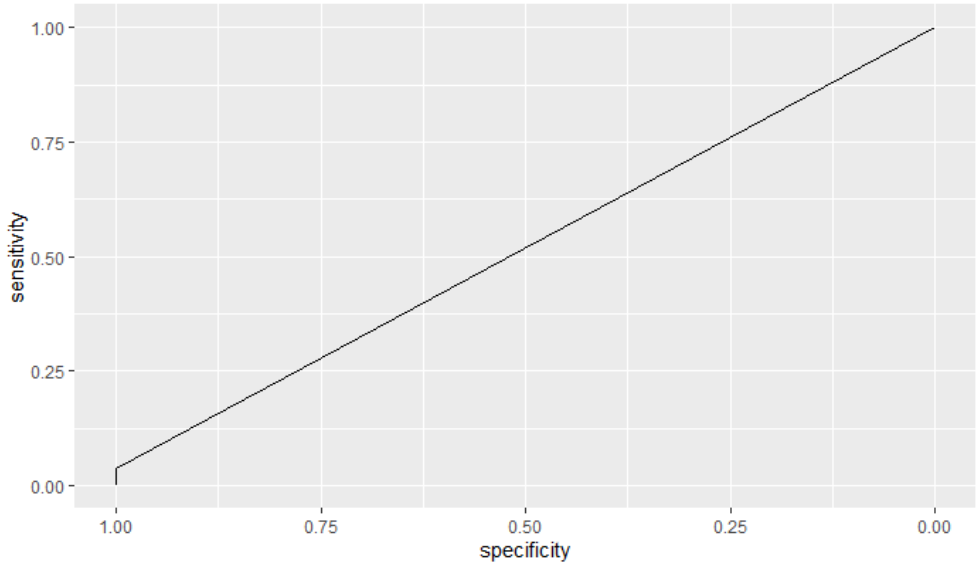


Figure 6. Experimental results 4 (Photo credit: Original)

In these result, the AUC ranges from 0 to 1 where 1 indicates perfect classification and 0.5 indicates random guess. Accuracy = (true positive + true negative) / (true positive + true negative + false positive + false negative). TPR = true positive / (true positive + false negative). F1 Score = 2 * (Accuracy * recall) / (Accuracy + recall). The prediction model cuizes the AUC value and takes it as the weighted value of the model: each model

As show in the Table 1. Patient 1 or 0 values are given and then converted to positive or negative weighting value respectively., and then add all the weighted values together to obtain a final value, with a negative values indicate that the patient is healthy, and positive values indicate that the patient has lung cancer. Finally, all the filtered genes are put into a table where one of the variables is Gene, the other is the number of times that gene is selected. The final model accuracy is 99.2% and here is a part of the affected gene table (appear 2, 3 or 4 times)

Table 1. Table of experimental results

	Gene	Count
1	1	2
2	AC093909.5	2
3	AC0937537.1	2
4	ADGRD1	2
5	ADH18	2
6	AL078624.1	2
7	ANGPT4	2
8	C1QTNF7	2
9	CHIA	2
10	CHIAP2	2
11	CLDN18	2
12	CLIC5	2
13	DES	2
14	F11	2
15	FHL1	2
16	GDF10	2
17	GKN2	2
18	GPIHBP1	2
19	INMT	2
20	LHFPL3AS2	2
21	LYVE1	2
22	PGM5P4	2
23	PRG4	2
24	SFTPA1	2
25	SLC39A8	2
26	SLC46A2	2
27	TNXB	2
28	VEPH1	2

2.3. How Gene Expression Affects Lung Cancer

Expression data from 592 lung adenocarcinoma samples from the TCGA (The Cancer Genome Altas) database selected 7,079 differentially expressed genes. These differentially expressed genes are involved in various biological processes, including the cell cycle, metabolic processes, and the proteasome [6].

In existing studies, significant disturbances also occurred in metabolic processes and proteasome. Some DEGs share target sites of transcription factors, for example E2F, ETS and CEBPB. The target sites of hsa-miR-196a and hsa-miR-200c were also significantly enriched by DEGs. Iloprost mimics the state of normal cells, and MS-275 may be a potential pathogenic agent [7].

3. Genes and Lung Cancer Treatment

3.1. The Role of CLDN18 and its Potential in Treatment

CLDN18 Is downregulated in lung SCC tissues and may inhibit cell proliferation by affecting Akt phosphorylation. According to relevant studies, the expression of CLDN18 in human lung squamous carcinoma tissues is lower than that in normal tissues. Scientists used human lung squamous cancer cells to examine which combination of exogenous CLDNs expression suppresses proliferation and cellular neuropathy, and found that CLDN18 is rarely expressed in human lung squamous cancer cells. Simultaneous cell proliferation is inhibited by CLDN5,7 and 18. Meanwhile, CLDN18

promotes p21 and suppresses the generation of cyclinD1, resulting in cell cycle inhibition. The expression of these CLDN inhibited the phosphorylation of Akt, and these CLDNs inhibit the nuclear localization of Akt and its binding to 3-phosphoinositide-de pendent protein kinase-1 (PDK 1). But the G1-Stransitionion-Akt expression caused by CLDN18 was rescued. Final experiments suggest that the decreased expression of CLDN18 may disrupt the interaction between PDK 1 and Akt, causing the sustained phosphorylation of Akt, leading to the cell cycle changes in lung squamous cancer cells.

CLDN18 Is downregulated in lung squamous cell carcinoma tissues, and together with CLDN5,7, they inhibit the Akt signaling pathway to inhibit cell proliferation.

Frequent ectopic activation of CLDN18.2 is observed in pancreatic, esophageal, ovarian, and lung tumors, linked to various histological subtypes. This activation relies on the binding of the transcription factor cyclic AMP response element-binding protein to its unmethylated common sites. Significantly, monoclonal antibodies have been developed that bind to CLDN18.2 but not to its lung-specific splice variants, and these antibodies can recognize antigens on the surface of cancer cells. Interestingly, the gene expression effects of CLDN18.2 in lung tumors are entirely opposite to those of CLDN18. Detecting the expression of CLDN18 in lung tissue cells aids in the identification and timely detection of lung squamous cell carcinoma tissue. Enhancing the gene expression of CLDN18 in lung squamous cell carcinoma tissues could potentially inhibit the progression of the carcinoma and stabilize the cell cycle. Therefore, targeted gene therapy emerges as a viable approach, with CLDN18 serving as a potential target for such treatments.

Development of drugs or antibodies against CLDN18 that promote CLDN18 expression and inhibit the phosphorylation of Akt to maintain the balance of interaction between PDK 1 and Akt, which can specifically inhibit the abnormal growth of lung squamous cell carcinoma with little effect on normal cells.

On the basis of gene-targeted therapy, the combined effect of chemotherapeutic drugs may improve the therapeutic effect on tumor cells.

3.2. Research Progress and Treatment Prospects of LYVE-1

Lymphatic Vessel Endothelial Receptor-1(LYVE-1) as a reliable lymphatic endothelial cell marker is specifically expressed in lymphatic endothelial cells, so that the investigators can use specific immunohistochemical staining against LYVE-1 to identify lymphatic vessels and compare it with CD31 staining to distinguish blood vessels and lymphatic vessels. Therefore, in the course of research, it was found that LYVE-1 positive lymphatic vessels were detected around the tumor edge of lung cancer, but they did not exist in the tumor body, that is, lung cancer had no intra-tumor lymphatic vessel network [8]. The lack of immunohistochemically recognizable lymphatic vessels within the main tumor block suggests that even if the lymphatic vessels may be incorporated into the tumor mass, these lymphatic vessels do not survive, but resolve and disappear [9]. And the spread of tumor cells through the lymphatic vessels may be achieved by surrounding tumor cells invading adjacent normal lymphatic vessels, or by invading the tumor margin, as a result of active angiogenesis and lymph angiogenesis.

3.3. Clinical Relevance and Therapeutic Potential of GPIHBP1

GPIHBP1 (Glycosylphosphatidylinositol-Heparan Sulfate Anchor Protein 1) is an endothelial cell-specific protein. It promotes the breakdown of triglyceride (TG) fat on the surface of endothelial cells by binding to lipoprotein lipase (LPL), thus releasing nutrients into the surrounding tissue [9]. In severe chylomicronemia, four missense mutations were found in GPIHBP1 highly conserved residues (C65Y, C65S, C68G and Q115P), resulting in accumulation of triglycerides and causing health problems such as pancreatitis. For the current study, GPIHBP1 mainly play a role in lipoprotein metabolism, and highly expression can cause health problems, although the direct relationship with lung cancer has not yet have a clear relationship, but according to our Gene Screening Model screening results, GPIHBP1 has a huge effect on lung cancer, and present positive correlation,

consistent with the current genetic research, so for the treatment of lung cancer in this direction, GPIHBP1 still has relatively large therapeutic potential. In addition, the PI3K / Akt signaling pathway is aberrantly activated in lung cancer, and GPIHBP1 may inhibit the activity of this signaling pathway. Not only that, the control of gene expression of GPIHBP1 can maintain human health to some extent, which is also conducive to reducing the risk of lung cancer, and may even have a certain inhibitory effect on lung cancer.

4. Challenges

4.1. Reliability Issues of the Gene Screening Model

The advantage of the Gene Screening Model is that multiple machine learning models are utilized to solve the possibility of failure within a single model by weighting each model. Nonetheless, the final analysis of gene significance was not by increasing the weight of each gene in each model, but by calculating the occurrence of each gene in all five models. Furthermore, there is no weighted method for calculating the AUC value for each model on any strong basis, nor does there test whether the cube using AUC will produce the best results, namely that there are more efficient and suitable algorithms for the integrated model. But in terms of the results, our model has good results for genetic screening and prediction, and on average, it is more accurate than any of the five "component" models, and on average can predict one or two patients more accurately than the most accurate "component" model... But the study used gene expression data from 551 individuals (49 healthy individuals, 502 lung cancer patients). Data were divided into the training set (80%) and the test set (20%). In the case of less data, the model made has some accidental error, but our model is lacking in this aspect. Therefore, more data is needed to verify and update this large model in the future.

4.2. Mechanisms of Lung Cancer Metastasis

The Gene Screening Model The purpose is to screen the genes involved in lung cancer development, but there is some uncertainty for lung cancer metastasis, but the metastasis of lung cancer tumor will also make the expression of some cancerous genes abnormal, which may also be identified by our model. Therefore, for lung cancer metastasis, the Gene Screening Model still has a certain reference nature. In the face of LYVE-1 gene analysis, it is concluded that tumor cells can be transferred through the lymphatic vessels, that is, first invade the adjacent normal lymphatic vessels, and then into the lymph nodes for metastasis, and another mechanism is transferred from the non-lymphatic vessels to the lymphatic vessels.

Metastasis of lung cancer is related to its specific genetic mutations, heterogeneity, and biomarkers. These factors give lung cancer highly metastatic ability [10].

4.3. Finding Effective Therapeutic Targets for Lung Cancer

During the experiment, the large model identified 75 genes as influential, although most of these genes were selected only once as potential influences. These genes may serve as effective therapeutic targets for lung cancer in the future. Analysis focused on genes with more than two observed effects, which are considered conditionally effective therapeutic targets for lung cancer. Related pilot experiments include the use of a small molecule inhibitor of CLDN18 to suppress the abnormal expression of inhibitory genes and the application of LYVE-1 antibodies to track the formation of lymphatic vessels and tumor metastasis in lung cancer. Future research could expand the analysis to more influencing genes and explore multiple effective therapeutic targets, thereby providing a variety of approaches for the diagnosis and treatment of lung cancer.

5. Conclusion

This study has successfully leveraged gene expression data from the TCGA database to construct a predictive model for lung cancer, identifying key genes such as CLDN18.2, LYVE-1, and GPIHBP1

that are abnormally expressed in lung cancer tumors. The analysis revealed that most of the selected genes are significantly upregulated, contributing to the onset and progression of lung cancer. The mechanisms of action of these genes have been partially elucidated, demonstrating the accuracy of our lung cancer prediction model. This research not only validates the model's effectiveness but also paves the way for new strategies in early detection and gene-targeted therapy for lung cancer, providing a foundation for improving patient outcomes through precision medicine.

Future Directions and Limitations: Despite the promising results, this study has identified several areas requiring further improvement. Future iterations of this research could benefit from utilizing a more extensive database to enhance the robustness and reduce potential errors in the model. Moreover, refining the algorithms used to weight model results could significantly improve predictive accuracy. The current model's predictions, while insightful, are preliminary and suggest genes that are potential candidates for targeted therapy; however, these need to be substantiated through additional gene therapy experiments to confirm their therapeutic viability. Furthermore, the growing global burden of lung cancer exacerbated by environmental and lifestyle factors underscores the urgency of extending our research. Future studies should focus on identifying and validating genes that have a pronounced impact on lung cancer, enhancing our understanding of gene functionality in oncogenesis and contributing to the broader field of genomic medicine. This ongoing research will not only aid in the prevention and management of lung cancer but will also deepen our understanding of human genetics in the context of health and disease.

References

[1] WHO. Lung cancer. 26 Jun 2023. 31 Jul 2024. Available: https://www.who.int/news-room/fact-sheets/detail/lung-cancer.

[2] Ferlier T, Coulouarn C. Regulation of Gene Expression in Cancer-An Overview. 15 Dec 2022. 31 Jul 2024.

[3] Shi WY, Liu KD, Xu SG, et al. Gene expression analysis of lung cancer. Eur Rev Med Pharmacol Sci, 2014, 18 (2): 217-28. PMID: 24488911.

[4] Sahin U, Koslowski M, Dhaene K, et al. Claudin-18 splice variant 2 is a pan-cancer target suitable for therapeutic antibody development. Clin Cancer Res, 2008, 14 (23): 7624-34.

[5] Nunomiya K, Shibata Y, Abe S, et al. Relationship between Serum Level of Lymphatic Vessel Endothelial Hyaluronan Receptor-1 and Prognosis in Patients with Lung Cancer. J Cancer, 2014, 5 (3): 242-7.

[6] Wang R., Zhu J., Wang S., Wang T., Huang J., Zhu X. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. In: International Journal of Multimedia Information Retrieval, 2024, 13 (4): 39.

[7] Shimin Xie, Zhengguo Wu, Yi Qi, et al. The metastasizing mechanisms of lung cancer: Recent advances and therapeutic challenges. Biomedicine & Pharmacotherapy, 2021, 138: 111450.

[8] Zhu X., Zhang Y., Zhao Z., et al. Radio frequency sensing based environmental monitoring technology. In: Fourth International Workshop on Pattern Recognition, SPIE, 2019, 11198: 187-191.

[9] Caramori G., et al. Molecular links between COPD and lung cancer: new targets for drug discovery? Expert opinion on therapeutic targets, 2019, 23 (6): 539-553.

[10] Young R. P., Hopkins R. J. How the genetics of lung cancer may overlap with COPD. Respirology, 2011, 16 (7): 1047-1055.