Prediction Of Lung Cancer and Analysis of Lung Cancer Related Gene Expressions Through Logistic Regression, PCA Random Forest, And LASSO Regression

Jiang, Ryan Ming

Shanghai American School, Shanghai, China

ryan01px2026@saschina.org

Abstract. Research Background: Lung cancer is frequently associated with the expression levels of certain genes, where both over-expression and under-expression may indicate the presence of the disease. It was hypothesized that specific genes might exhibit a more pronounced correlation with lung cancer. To explore this, gene expression data from patients with and without lung cancer were obtained and initially screened using volcano plots to identify genes with significant differential expression. **Study Contributions:** The refined dataset was subjected to comprehensive analysis using five distinct methodologies: principal component analysis (PCA), random forest, logistic regression, least absolute shrinkage and selection operator (LASSO), and a method employing lambda cross-validation for selecting influential genes. Each model's performance was assessed by its area under the curve (AUC) value, which then informed a weighting system to prioritize the findings. A weighted table was subsequently developed to finalize the diagnosis of lung cancer. The synthesis of these approaches not only enhanced the accuracy of the diagnostic model, confirmed to be precise in at least 96 percent of cases, but also led to the identification of 75 genes significantly associated with lung cancer. Among these, CLDN18, GKN2, LYVE-1, GPIHBP1, and CLIC5 were determined to be the most closely linked to the disease.

Keywords: Up-regulated genes; Down-regulated genes; Area under Curve (AUC); Cross Validation; Regression.

1. Introduction

Lung cancer claimed 1.8 million lives in 2020, making it the leading cause of cancer death globally [1]. Early detection of lung cancer could dramatically increase survival rates. One method for early detection is through monitoring gene expression changes, as these changes can indicate the presence of lung cancer due to their critical role in biological processes [2]. However, not all genes contribute equally to the development or detection of the disease; therefore, identifying the most influential genes is crucial for effective diagnosis and treatment.

The aim of this study is to pinpoint the genes that significantly influence lung cancer manifestation in patients by employing various machine learning techniques to analyze gene expression data. Utilizing a dataset encompassing 551 individuals—49 healthy and 502 with lung cancer—this study compares the expression levels of 56,907 genes to discern those most indicative of lung cancer. Study Contributions:

- Utilized an extensive dataset to perform a comparative analysis of gene expression between individuals diagnosed with lung cancer and healthy controls.
- Employed advanced machine learning techniques, including principal component analysis (PCA), random forest, logistic regression, and least absolute shrinkage and selection operator (LASSO), to identify key genes associated with lung cancer.
- Implemented lambda cross-validation to refine the selection process of influential genes, ensuring robustness and reliability in the predictive model.
- Developed a predictive model capable of diagnosing lung cancer with high accuracy based on the expression profiles of the identified genes.
- The findings provide a foundation for future research into targeted therapies and diagnostic tests that could enhance early detection and treatment outcomes for lung cancer patients.

2. Relevant Theories

2.1. Theoretical Background on Gene Expression and its Relation to Cancer

In order for tumor progression to occur, normal processes inside a cell need to change. Due to this, genes can become up-regulated or down-regulated [3]. Thus, gene expression can become a sign that an individual may have lung cancer, or a sign that the individual does not have cancer. However not every gene is correlated with whether the patient has or doesn't have cancer, making certain genes more influential in deciding whether a patient has lung cancer. Additionally given the large number of genes that humans possess; it is difficult to test every gene when using gene expression to identify lung cancer. As a result, it is necessary to identify which genes are more influential in deciding whether a patient has lung cancer. Furthermore, knowing the genes that are more influential could also help doctors more easily identify patients with lung cancer.

2.2. Overview of Analytical Techniques Used in Gene Expression Studies

One common method to analyze gene expression is to use differential gene expression (DGE) analysis. This analysis can be done in many ways, such as a visual analysis of heat maps or in the form of volcano plots [4]. Another method to analyze gene expression data is using a LASSO regression with cross validation to select genes after conducting a DGE analysis in order to identify influential genes [5].

3. Application Research

Wide availability of data resources of gene expression in the public domain allows for greater analysis and research to be done in the field of cancer research. This analysis of gene expressions allows for identifying molecular subtypes and predicting patient outcome. The analysis of gene expression can also help with identifying more influential cancer genes, resulting in more targeted treatment in patients with cancer [6].

4. Methodology and Materials

The materials that were used in this study include an openly available data table and version 2024.04.2-764 of Rstudio.

To start, we read the Excel data table into Rstudio. Then 80 percent of the data was split proportionately from healthy and ill patients, with the seed 6, into a training data set, while the other 20 percent of the data was kept as a testing data set. A binary system of identification was then created to identify the patients who are healthy or have lung cancer using the numbers 0 and 1, where 0 represented a patient that is healthy and 1 representing a patient with lung cancer. Following that, we calculated the fold change of the gene expressions and then used the Wilcoxon test to find the p-value. After the data was standardized, a volcano plot was created to select genes that were shown to have positive or negative changes from the control variables. After that, five models were created to select the most influential genes as well as predicting the patients' health. Before making the first model, we created a logistic regression that simply identified the 20 highest absolute value coefficients and selected the 20 genes with those coefficients for further use. A K-fold cross validation would also be used. The first model was a logistic regression model that uses the previous regression model's 20 genes to predict patient health. The second model was a random forest prediction model built on the 20 genes selected from the first logistic regression. For the random forest model, the number of trees was set to 500 in order to reduce error. The third model used a PCA on the genes selected by the volcano plot. In this model we selected the 5 most important principal components and selected the 5 most influential genes in each of the 5 selected principal components to serve as our most influential genes. In model 4, we used LASSO to pick out the most influential genes while using the lambda with lowest MSE value and an alpha of 0.3 in order to select the elastic net to account for

multicollinearity. The fifth model uses lambda cross validation to select the most influential genes. For each model, the influential genes that were selected were also used to predict patient health with the model it was selected in. As a result, we were able to obtain the AUC value of every model. Then a larger weighted prediction model was created by adding the weighted results of all of the 5 prediction models together. The weight used for a prediction model was its own AUC value cubed. Every model gives a 1 or 0 value for the patient which gets converted into a positive or a negative weighted value respectively. All the weighted values are then added together to get a final value, with a negative value meaning that the patient is healthy and a positive value meaning that the patient has lung cancer. Finally, all the filtered genes are collected into a table to check for how many times they were selected as influential.

5. Experimental Results and Visualization

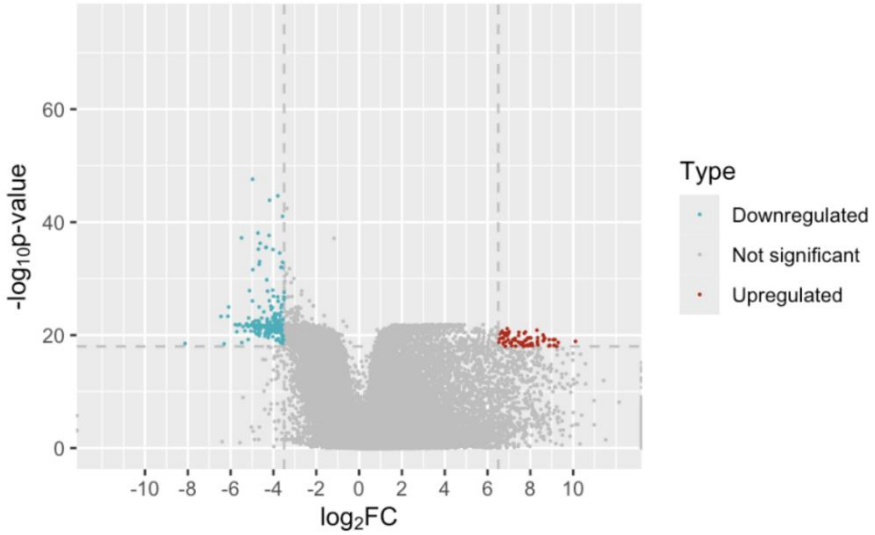

Fig. 1 Volcano Plot of Gene Expressions in Patients Withand Without Lung Cancer (Photo credit: Original).

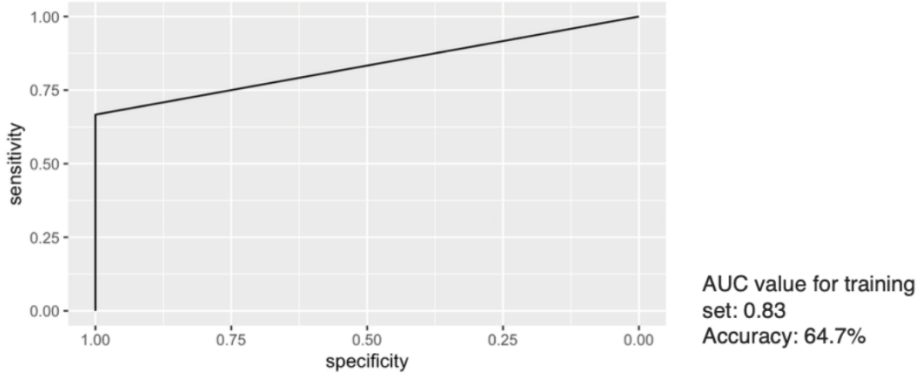

Fig. 2 AUC Graph of Model 1 (Photo credit: Original).

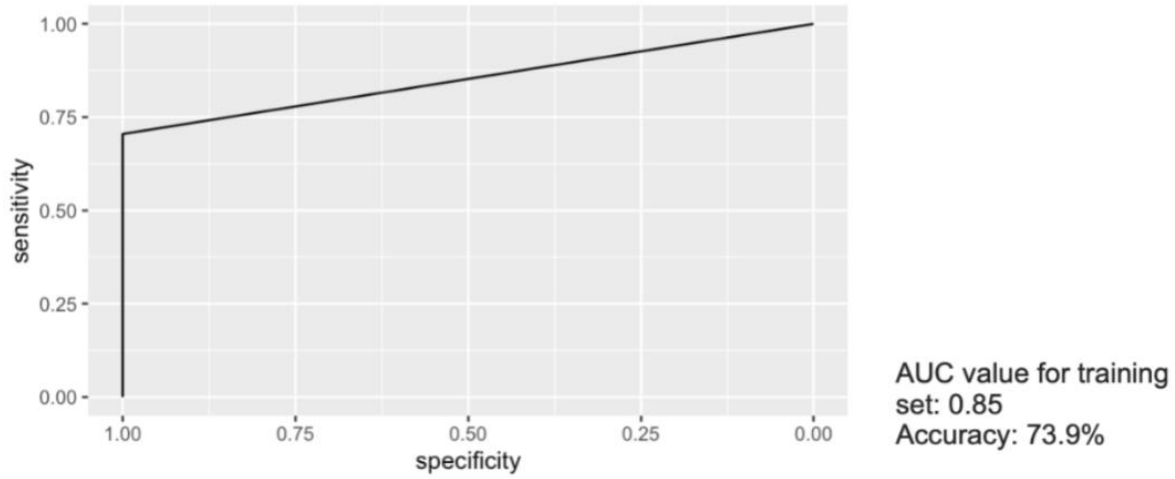

Fig. 3 AUC Graph of Model 2 (Photo credit: Original).

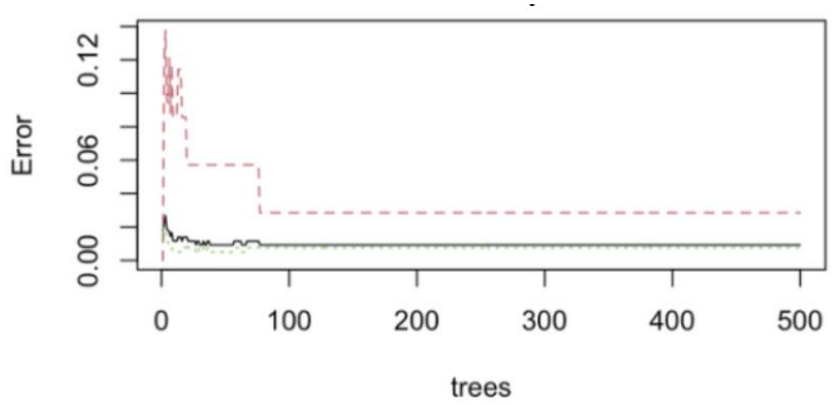

Fig. 4 Error and Trees Graph for Model 2 (Photo credit: Original).

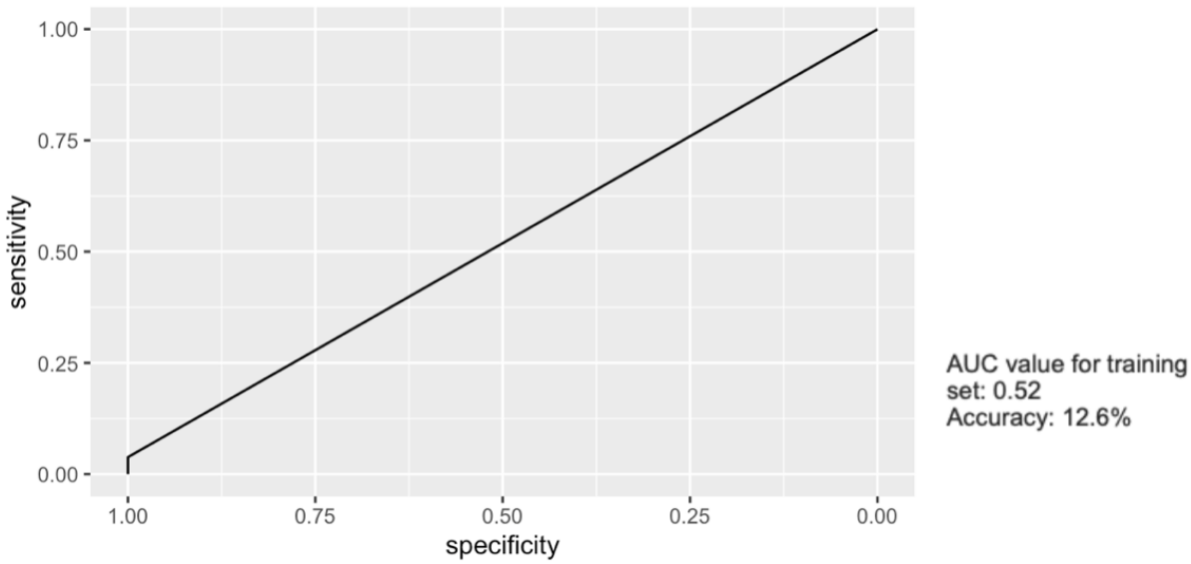

Fig. 5 AUC Graph of Model 3 (Photo credit: Original).

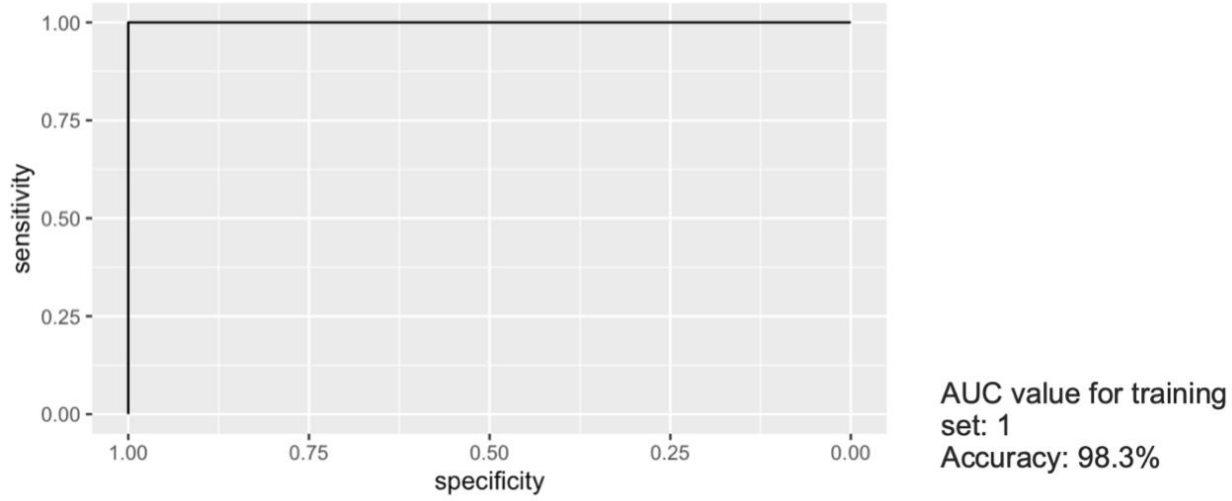

Fig. 6 AUC Graph of Model 4 (Photo credit: Original).

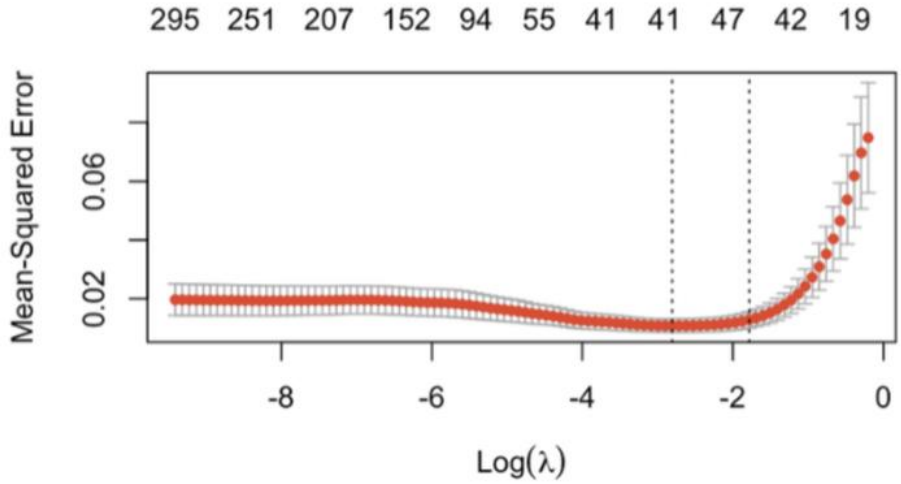

Fig. 7 Lambda and MSE Graph for Model 4 (Photo credit: Original).

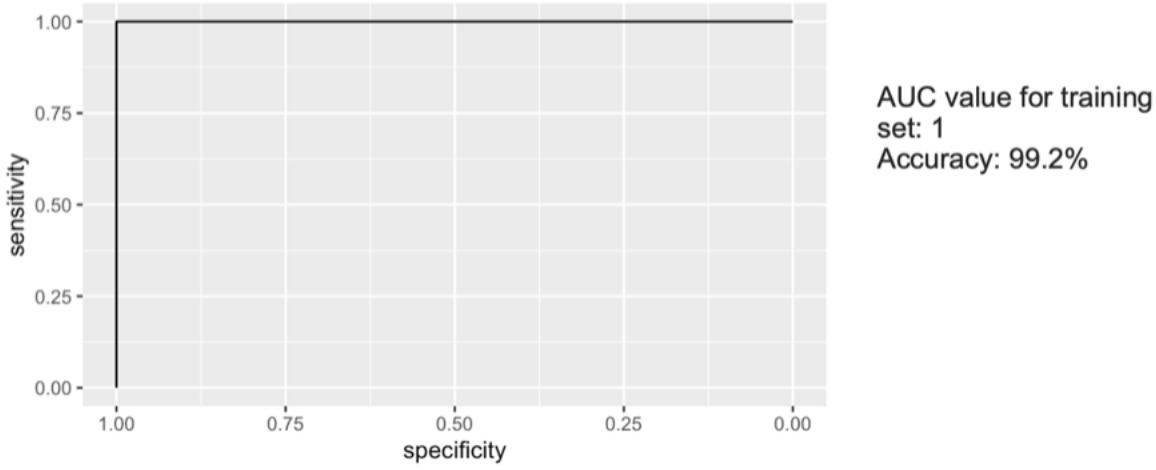

Fig. 8 AUC Graph of Model 5 (Photo credit: Original).

6. Discussion

As show in the Fig. 1 to the Fig.8. After analyzing the average accuracy of all the models, this study found that the accuracy of the final model was higher than each of the 5 modes used to construct

the final model. On average the model has an accuracy of at least 96 percent on the testing data set, however, for the run shown above, it had a 99.2 percent accuracy. Additionally, on average, the most accurate model apart from the final model would be model 5, followed by model 4, then followed model 2, then by model 3, and then by model 1. From the table collected on the selected influential genes of various models, this study identifies 75 influential genes. This number includes only the genes that were identified as influential by more than one model. In these 75 genes, the only gene that was identified by 4 models as influential was the gene CLDN18. On the other hand, the gene GKN2, LYVE-1, GPIHBP1, and CLIC5 were found to be influential in 3 different models. Lastly the genes FTPA1, PRG4, SLC39A8, AC097537.1, AL078624.1, AC093909.5, ADGRD1, ADH1B, ANGPT4, CHIA, CHIAP2, DES, F11, FHL1, GDF10, LHFPL3AS2, 1, SLC46A2, TNXB, VEPH1, INMT, C1QTNF7, PGM5P4 were each identified to be influential by 2 models. In these 75 influential genes, some were identified to be positively correlated with patients with lung cancer while others were identified as negatively correlated with patients with lung cancer.

When analyzed in other studies, the gene CLDN18 is shown to have ectopic activation in various forms of cancers including lung cancer and is seen as a candidate for treating and identifying cancer within patients [7]. In another study, the gene CLDN18 has also been shown to have been a therapeutic target due to its consistent expression within gastric and lung tumor cells [8]. Similarly, other genes found in our study also appear to have a strong connection with cancer. GKN2 is shown to have the ability to increase apoptosis and reduce cancer cells when it is up-regulated, however, this also reflects the opposite [9]. This means that if GKN2 was down-regulated, it would point to a higher likelihood that the patient has cancer. Similarly, another study also shows that GKN2 is significantly down-regulated in samples of lung cancer cells when compared to that of normal samples [10]. Another study demonstrates the relationship between the gene expression of the gene LYVE-1 in cancer formation, showing that it plays a key role in tumor formation and metastasis [11]. The gene expression of the gene GPIHBP1 was also shown to have a relation with the presence of cancer within patients, with down-regulated genes being shown to occur within patients with cancer [12]. Last but not least, the CLIC5 gene was shown to be able to be another promising target of cancer therapy and detection although this particular gene may not be the most promising target to use [13].

However, this study also possesses many strengths and weaknesses. A strength of the study lies in the usage of multiple machine learning methods in order to find our result so that issues within singular models could be addressed through the weight of another model. However, this approach also has issues such as ineffective models potentially dragging down the accuracy rate of the final prediction model. Our future improvements that should be considered for this study is finding the influential genes through the weight of the model it occurred in rather than simply using the number of times that the gene was selected by a model. This allows genes that occurred in more accurate models to be more likely to be seen as influential. Additionally, the method of weighting models based on cubing the AUC value should also be improved since the idea of cubing the AUC value has no basis except for helping more accurately reflect the accuracy value of the models. A likely better way to do this would be through modeling the impact of the power of the AUC value on the average accuracy of the final prediction model. Lastly, the number of patient data available for this study was also limiting since there were only the data for 49 healthy patients which meant that after splitting out 20 percent of the data for the testing set, there would not be a lot of patients left for training the models.

7. Conclusion

This study has identified five genes—CLDN18, GKN2, LYVE-1, GPIHBP1, and CLIC5—as the most influential in the detection of lung cancer, with an additional 23 genes showing a strong association with the disease. Based on these findings, we recommend a focused analysis and further research on these 28 genes as potential biomarkers for the detection and possible treatment of lung cancer. This approach could lead to significant advancements in early diagnostic processes and therapeutic strategies, enhancing the precision of lung cancer interventions.

For future research, we advocate the use of a larger patient dataset and the development of a more sophisticated method for weighing various analytical approaches. These improvements are expected to increase the reliability of the results and ensure a higher degree of accuracy in gene-based lung cancer studies. Such enhancements will not only strengthen the validity of our findings but also broaden the potential for discovering new insights into the molecular mechanisms underpinning lung cancer, potentially leading to breakthroughs in its management and treatment.

References

[1] WHO. Lung cancer. 2023-06-26. https://www.who.int/news-room/fact-sheets/detail/lung-cancer.

[2] Otálora-Otálora B. A., López-Kleine L., Rojas A. Lung Cancer Gene Regulatory Network of Transcription Factors Related to the Hallmarks of Cancer. 2023-01-05.

[3] Ferlier T., Coulouarn C. Regulation of Gene Expression in Cancer-An Overview. 2022-12-15.

[4] McDermaid A., Monier B., Zhao J., Liu B., Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. Briefings in bioinformatics, 2019, 20(6): 2044–2054.

[5] Yu S. H., Cai J. H., Chen D. L., Liao S. H., Lin Y. Z., Chung Y. T., Tsai J. J. P., Wang C. C. N. LASSO and Bioinformatics Analysis in the Identification of Key Genes for Prognostic Genes of Gynecologic Cancer. Journal of personalized medicine, 2021, 11(11): 1177.

[6] Creighton C. J. Gene Expression Profiles in Cancers and Their Therapeutic Implications. Cancer journal (Sudbury, Mass.), 2023, 29(1): 9–14.

[7] Zhu, X., Huang, Y., Wang, X., Wang, R.: 'Emotion recognition based on brain-like multimodal hierarchical perception', Multimedia Tools and Applications, 2024, 83, (18), pp. 56039-56057

[8] Mathias-Machado M. C., de Jesus V. H. F., Jácome A., Donadio M. D., Aruquipa M. P. S., Fogacci J., Cunha R. G., da Silva L. M., Peixoto R. D. Claudin 18.2 as a New Biomarker in Gastric Cancer-What Should We Know? Cancers (Basel), 2024-02-05.

[9] Ouyang J., Pan X., Lin H., Hu Z., Xiao P., Hu H. GKN2 increases apoptosis, reduces the proliferation and invasion ability of gastric cancer cells through down-regulating the JAK/STAT signaling pathway. 2017-02-15.

[10] Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X., Wang, R.: 'A Review of Key Technologies for Emotion Analysis Using Multimodal Information', Cognitive Computation, 2024, 1, (1), pp. 1-27

[11] Hara Y., Torii R., Ueda S., Kurimoto E., Ueda E., Okura H., Tatano Y., Yagi H., Ohno Y., Tanaka T., Masuko K., Masuko T. Inhibition of tumor formation and metastasis by a monoclonal antibody against lymphatic vessel endothelial hyaluronan receptor 1. 2018.

[12] Bao Y., Wang L., Shi L., et al. Transcriptome profiling revealed multiple genes and ECM-receptor interaction pathways that may be associated with breast cancer. 2019-06-06.

[13] Gururaja Rao S., Patel N. J., Singh H. Intracellular Chloride Channels: Novel Biomarkers in Diseases. 2020-02-14.