**形状

中度可信度描述已自动生成**

# Lung Cancer Detection

Zhen Li, Hong Chen, Hongyi Liu, Jasper Fu

Department of Sever Institute, Washington University in St. Louis

T81-574: Foundations of Analytic

Prof. Shanti Greene

December 14, 2022

# Content

[Cover Page 1](#_Toc121947486)

[Content 2](#_Toc121947487)

[Problem Statement 3](#_Toc121947488)

[Background 3](#_Toc121947489)

[Data Sources 4](#_Toc121947490)

[Program objective 4](#_Toc121947491)

[Data sources and data cleaning 5](#_Toc121947492)

[Method 7](#_Toc121947493)

[Model Improvement 9](#_Toc121947494)

[Process 10](#_Toc121947495)

[Results 13](#_Toc121947496)

[Findings 14](#_Toc121947497)

[References 16](#_Toc121947498)

# Problem Statement

Lung cancer is the most lethal cancer in China, accounting for one-third of all cancer deaths. It took 7.15 million lives in 2020, and the figure is rising every year. The prognosis for advanced lung cancer is dismal. Furthermore, early lung cancers may not show noticeable symptoms in patients. This would cause the patient to go blind and allow the tumor to develop. It is too late to treat cancer once significant symptoms arise. As a result of this phenomena, we proposed and designed a novel lung cancer detection model for the blood test because blood circulates throughout our bodies and contains a plethora of information about us. By examining several signs in blood tests, the model can forecast the chance of acquiring lung cancer.

# Background

The most recent cancer burden figures for 2020 have been released by the International Agency for Research on Cancer (IARC), and they contain updated incidence and mortality rates for 36 different cancer types in 185 different nations. 0.36 was the mortality rate.

According to the survey, China has the greatest estimated number of cancer-related fatalities and new cases worldwide, with lung cancer topping the list of the top 10 cancer types. Lung cancer caused 715,000 fatalities and 816,000 new cases in China in 2020. In China, lung cancer is the primary cause of both new cases and fatalities (Qin, 2021). Early detection of lung cancer by screening can increase the effectiveness of diagnosis and treatment, lower the cost of care, and ultimately increase patient survival rates and quality of life. Early-stage lung cancer, however, may not cause any symptoms at all. Coughing and weariness are the most typical symptoms, and they are frequent in healthy individuals (Birring, 2005).

A few different types of cancer products are still in the early stages of clinical research and analysis in China. Cancer screening technology is still in the early stages of research and development in the field of cancer screening. While lung cancer and pneumonia both have higher-risk populations, early screening permeability is poor. Typically, the patient will do better the sooner the condition is identified. Consider lung cancer as an illustration. The IARC divides the spread of cancer into one to four stages, and if the cancer is caught in the first stage, the five-year chance of survival can be many times higher than in the later stages.

As a result, early detection of lung cancer by screening can increase the effectiveness of both diagnosis and therapy. The late stage of diagnosis is directly connected to the high death rate of lung cancer. Early detection can increase survival rates, decrease lung cancer fatalities, and enable patients to obtain top-notch care more promptly. As a result, we are thinking about developing a screening model that might be included in frequent and routine physical exams.

Additionally, China has a substantial market for early cancer and lung cancer detection. The early tumor screening market size in China is anticipated to increase from $18.4 billion in 2019 to $28.9 billion in 2030, a CAGR of 4.2%. The number of people in China who are at high risk of developing lung cancer, as determined by CSCO, will be 95.8 million by 2023. As a result, the potential market space for lung cancer screening will amount to 5.74 billion US dollars, or around 26.5% of the market share for early cancer screening (Qin, 2021).

# Data Sources

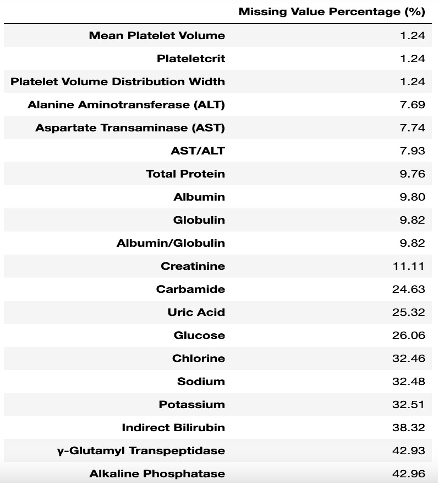
## Program objective

Background screening for lung cancer is challenging, and a definitive diagnosis of lung cancer is quite expensive. Early detection of lung cancer or ongoing monitoring of pertinent indications are, in our opinion, essential. The goal of this research is to develop and evaluate a basic lung cancer prediction model using data from blood tests. The methodology is intended to assist hospitals, hospitals, and healthcare organizations in quickly and inexpensively identifying and highlighting probable lung cancer patients so that they may be notified to undergo additional testing. Currently, blood test reports from many hospitals automatically flag biochemical markers that are outside of the normal range to notify patients and aid in diagnosis; using our model, we want to also include certain lung disease alerts. Additionally, we hope that the detection procedure made possible by these models is affordable and trustworthy and that the only signs needed are the blood tests that are performed as part of a person's routine health checkup.

## Data sources and data cleaning

For practical and business purposes, we want our lung cancer prediction algorithm to be based on blood test results. We also intended for our data collection to be comprehensive and realistic in the choice of data sources because the principles of lung illness are intricate and specialized. Negotiations resulted in our being granted access to real medical records from hospitals on the mainland. The general population, patients with pneumonia, and patients with lung cancer were the three diagnostic groups whose data sets were chosen from the hospital's database. One of the main problems of this project was dealing with the over 800,000 lines of raw data that we got since they were all in database format because they were all taken from databases and each line of data only included information about one blood test item for one person.

We manually deleted unnecessary data, including patient identifiers, access dates, and most of the columns with a high percentage of null values, after making a prolonged effort to combine all the data pertaining to the same individual into a single row. The various diagnostic categories are identified as 0-normal, 1-pneumonia, and 5-lung cancer patients when a column for diagnostic data is included. There are now 6501 rows and 50 columns of data. To more reasonably deal with the residual null values in the data, we counted the percentage of the hollow values in each column. The following table shows the 20 columns with the highest percentage of hollow values:



We then remove the last nine columns with the greatest percentage of NAs based on this list and fill the remaining NAs with an average. After that, by plotting the distribution of ages, we find that there is a large portion of people whose age is under 10. And the data shows that the proportion of unhealthy people ranges from 0 to 20. Therefore, we can ensure that it is an imbalanced dataset. We believe that uneven sampling is the cause of the data imbalance. At the same time, the physical examination of infants mainly focuses on blood tests, while adults will experience more medical information and no longer focus on blood tests, so the proportion of cases under the age of 10 is larger. So, we drop the column of age information.

图表, 直方图

描述已自动生成

图表, 条形图, 直方图

描述已自动生成

Out of the enormous amount of raw data, only 6355 rows and 40 columns of data were kept for additional machine learning.

# Method

In our project, there are three categories of patients: healthy individuals, patients with pneumonia, and patients with lung cancer. We performed two rounds of machine learning to train our models. First round between healthy and unhealthy individuals, including both patients with pneumonia and lung cancer. Second round between patients with pneumonia and patients with lung cancer.

To achieve our goal, we use three types of algorithms for each round of the diagnostic model: Logistic Regression, Decision Tree, and Gradient Boos. The models are trained with the sklearn package in Python. Each type of model is trained with the training set, and we divide the training set and test set into 8:2, and the hyperparameters are tuned by using GridSearchCV which is based on 10-fold cross-validation results. We also use StandardScaler to normalize the features to make the model more accurate.

We believe that medically, the pathology of different lung diseases is complex and different, so the training and testing of the two sets of models are completely independent. It is worth mentioning that we spent a lot of time optimizing the code; a lot of loops were used to improve efficiency, and at this stage our code only needs to enter data once to automatically fill in three models and list the results we want.

Both results run by each round of machine learning are meaningful, and we are more concerned about how many unhealthy people are accurately predicted to have pneumonia or cancer, so the recall value is our credential for selecting and tuning the models.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Predict health or not (0 or 1/5) | Predict pneumonia or lung cancer (1 or 5) | Actual theoretical Recall after Round2 (Round1 \* Round2) |
| Logistic Regression | 96.64% | 80.94% | 78.22% |
| Decision Tree | 98.35% | 82.47% | 81.37% |
| Gradiesnt Boost | 99.57% | 89.00% | 88.62% |

Table 2

For predicting lung health or not, Gradient Boost has the highest recall value, but most of the models have a recall value close to 1. This is reasonable and necessary because it should be simple and accurate to determine whether a person's lungs are healthy based on blood test results with reference to the actual situation. However, the goal of our model is to distinguish people who have lung cancer from those who have pneumonia, so that the first round of prediction is only to screen the data for the second round of subdivision of the disease.

As for predicting pneumonia and lung cancer, the accuracy value of all models has decreased due to the extreme complexity of the causative agent of lung disease, with most of them staying at 80%. However, these recall values assume that 100% of people with lung disease were predicted in the previous round, which does not exist in practice, so the accuracy value of the first round should be considered and the actual theoretical recall value of the models should be the product of the two rounds. From Table 2, the model with the highest actual theoretical recall value is Gradient Boost, which reaches 88.62%.

# Model Improvement

By calculating the VIFs, we find that VIFs are very high, that means there exist multicollinearity (predictor variables are highly correlated with each other).The following fig. shows the VIF of every variable (top 10).

图形用户界面, 文本, 应用程序

描述已自动生成

Because the collinearity will affect our interpretation of the model, making model coefficients unstable. Also, some variables are not necessary, so we decide to use **factor analysis** to drop unimportant variables and reduce the VIF. Finally, we can reduce the number of attributes. The goal of factor analysis is to reduce many variables into a smaller number of factors that can explain the relationships between the variables.

## Process

1. Determine the number of factors to extract. By plotting the scree plot, we chose the point at which the eigenvalues begin to level off.

# Create factor analysis object and perform factor analysis

fa = FactorAnalyzer()

fa.fit(x1)

ev, v = fa.get\_eigenvalues()

# Create scree plot using matplotlib

plt.scatter(range(1,x1.shape[1]+1),ev)

plt.plot(range(1,x1.shape[1]+1),ev)

plt.title('Scree Plot')

plt.xlabel('Factors')

plt.ylabel('Eigenvalue')

plt.grid()

plt.show()

图表, 折线图, 直方图

描述已自动生成

1. Extract the factors. Identify the underlying factors that explain the relationships between the variables.

# Create scree plot using matplotlib

plt.scatter(range(1,x1.shape[1]+1),ev)

plt.plot(range(1,x1.shape[1]+1),ev)

plt.title('Scree Plot')

plt.xlabel('Factors')

plt.ylabel('Eigenvalue')

plt.grid()

plt.show()

1. Rotate the factors. Making the factors more interpretable by maximizing the variance within each factor.

fa = FactorAnalyzer()

fa.fit(x1)

eigen\_values, vectors = fa.get\_eigenvalues()

fa = FactorAnalyzer()

fa.set\_params(n\_factors=14, rotation='varimax')

fa.fit(x1)

loadings = fa.loadings\_

1. Interpret the factors. For each factor, we find the most important original variable for each factor and keep it.

We reduce the number of predictors from 41 to 14 after completing factor analysis. But going through the deleted predictors, we found some predictors should have relationship with the outcome, like “Platelet Volume Distribution Width (PDW)”, according to research “a high PDW means there is a great variation in size, which may be associated with vascular (blood vessel) disease or certain cancers." (Roswell Park Comprehensive Cancer Center) As a result, we reinstate this predictor. We use 15 variables to calculate VIF again, and we can see that VIF is decreasing a lot:

图形用户界面, 文本

中度可信度描述已自动生成

Use those variables to fit the logit regression model again, and we get the result:

图片包含 文本

描述已自动生成

After we decrease the number of variables from 39 to 15, We can see the R-square has not been affected so much, so there indeed exist many variables that do not contribute much to the model.

# Results

After model evaluation, the best-performing model was selected. Based on this model importance report and some other prediction results, we derived some meaningful findings. Based on the model's performance, we discuss the prospect of commercial application of the modified model. Accordingly, we believe that our model has some limitations and room for improvement.

# Findings

图形用户界面, 文本, 应用程序

描述已自动生成图形用户界面, 文本, 应用程序

描述已自动生成

The Gradient model behaves best in the process of analysis, and the outcomes fit our hypothesis. About ranking importance, age is the most important factor in prediction. It's necessary to remind the elderly at home to get a physical examination on time. The ensemble classifiers can improve the model's performance compared with traditional classification models. In two sections, Gradient Boost performs best among three models, including decision trees, logistic regression, and Gradient. After the first exploration, it’s vital to process data while decreasing the number of x variables. We try to utilize VIF to reduce the number of variables from forty to fifteen. When we draw conclusions, it’s easier to make predictions when people have fewer tests to carry out. The predictive model enables people to get a diagnosis in the hospital without a doctor being involved. With accuracy, it is always much easier to differentiate whether a person is healthy or not but more difficult to conclude the severity of a patient's illness without a deeper examination.

One of the important objectives of the merging of medicine and big data in the digital era is efficient analysis and mining of clinical information and bio-omics data using machine learning. Various models and classifiers have been successfully used in lung cancer surgery to aid surgeons in making diagnosis and treatment decisions as well as to advance precision medicine. Machine learning is currently being used in lung cancer and medical research both at home and abroad, but it is still a long way from practical use. Even though the number of data sources and dimensions is growing, there is still a need for a high-quality, standardized lung cancer clinical database that can efficiently integrate data from several dimensions. Currently, the great majority of prediction and diagnostic research is limited to a specific patient population.

Following studies will need to optimize the algorithm to test the model's applicability outside of the training population. In this project, a preliminary predictive model for lung cancer based on blood test data Hope Cup developed. To help medical structures quickly and inexpensively scan large numbers of people to identify potential lung cancer patients among them. To achieve this goal, we use a two-round machine learning approach, i.e., first to distinguish whether an individual is healthy and then to distinguish whether an individual has pneumonia or lung cancer. A large amount of real medical clinical data was used to train 3 machine learning models that utilized the knowledge in this term, and we also used hyperparameter tuning and other methods to improve the accuracy of the models. Finally, Gradient was evaluated and compared to have the best overall performance. In this regard, we analyzed the output of Gradient and obtained some meaningful findings, and at the end of the project, the commercial application scenarios, limitations, and enhancement directions of the model were objectively discussed.

# References

Birring SS, Peake MD. (2005). Symptoms and the early diagnosis of lung cancerThorax 2005;60:268-269.

Qin Z. (2021). China Early Tumor and Lung Cancer Detection Analyst Report[J]. leadleo, 2021/08.

*Understanding Your Blood Test Results*. (2020, June 30). Roswell Park Comprehensive Cancer Center. https://www.roswellpark.org/cancertalk/202006/understanding-your-blood-test- results