

The Relationship between Lung Cancer Prevalence and Air Quality and Other Factors

Wenze Jin *

Northeast Forestry University, Aulin college, Harbin, 150000, China

* Corresponding Author Email: 284510993@nefu.edu.cn

Abstract. Lung cancer poses a significant global health threat. This study employs machine learning techniques for in-depth analysis of a lung cancer patient dataset. Five models were developed, and following a thorough exploration of variable relationships, the study utilized KNN, SVM, and multiple logistic regression methods to construct these models. The accuracy metric served as a performance measure, ultimately identifying the multiple logistic regression model as the most effective, boasting an accuracy of 0.9655. The findings of the multiple logistic regression model revealed that for every one-unit increase in smoking level, the risk of severe lung cancer increased. The study's results emphasize the importance of residing in areas with good air quality and reducing smoking levels to mitigate the risk of severe lung cancer. But none of the models are perfect predictors of lung cancer severity. Therefore, it is necessary to establish more comprehensive clinical data and complex analytical techniques to improve diagnostic performance. This model lays a foundation for better judging the probability of lung cancer by data-driven prediction.

Keywords: Machine learning; prediction model; lung cancer.

1. Introduction

Lung cancer is the second most common cancer in the world and causes the most cancer-related deaths [1, 2]. Based on available data, the World Health Organization estimates that 239,340 new cases of lung cancer will occur in 2023. The estimated number of deaths is as high as 127,070, or 20.8% of all cancer deaths [3]. Lung cancer ranks among the top causes of cancer-related fatalities in 93 countries. Compared to other cancers, lung cancer patients still have a low probability of surviving more than five years, with a relative survival rate of only 25.4 percent [4]. This undoubtedly poses a serious threat to the whole society. At present, researchers have conducted a large number of studies on lung cancer and have roughly divided it into third categories: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), non-small cell lung cancer accounts for about 80% of lung cancer [5]. A third, less common type of lung cancer is called carcinoid. By understanding the pathogenic factors that cause lung cancer, people can prevent the occurrence of lung cancer or detect lung cancer as early as possible. In this way, the survival rate of lung cancer patients or potential patients can be increased. Fortunately, with the continuous development of medical technology, the death rate of lung cancer patients is decreasing year by year.

Smoking and air pollution are the two main causes of lung cancer [6]. According to the researchers, there are at least two carcinogens in the polluted air: PAHs and n-nitroso compounds. The presence of particulate matter can promote lung cancer, especially PM. It has a high PAH concentration which can induce oxidative stress and aromatics receptor activation in human bronchial epithelial cells. Aromatics receptor activation allows cancer cells to invade normal cells and metastasize to other parts of the body [7]. Smoking directly increases the risk of lung cancer. Because smoke releases nitrosamines and other known carcinogens [8]. Air pollution and smoking work together as a cause of lung cancer. In addition to these two factors, age, gender, diet and other factors can also have an impact on the probability of getting lung cancer [9]. If lung cancer can be detected when it is less severe, the survival rate of lung cancer patients will be greatly improved.

At present, the factors considered in the clinical practice model are limited, and the prediction accuracy needs to be improved. On the other hand, the rapid development of machine learning technology offers the possibility of building some more accurate prediction models for lung cancer

severity. Lung cancer caused by smoking is increasing every year and will continue to increase in the future [10]. If the specific probability of disease can remind people of their physical conditions and prevent them in advance, it is possible to reduce the number of people suffering from lung cancer. Therefore, this paper aims to establish a lung cancer risk prediction model based on machine learning algorithm. In addition, this paper selected some potential related factors and compared the predictive performance of various algorithms in order to establish a more accurate and appropriate prediction model for lung cancer severity. Data on 152 lung cancer patients came from the Kaggle website and included 24 related characteristics such as air pollution, alcohol consumption, dust allergies, genetics and smoking status. The model was established by Logistic regression, K-nearest neighbor and support vector machine. Part of the dataset is used as a test dataset to examine the strengths and weaknesses of each model.

2. Methodology

2.1. Data Source

The Breast cancer patients' data used in this paper comes from the UC Irvine Machine Learning Repository website. The original data was saved in .CSV format.

2.2. Data Introduction

The data used in this article contains 569 instances and 10 variables without any missing values. All the 10 variables are represented in the Table 1 and all of them are numeric variables, meaning different kinds of important measured numbers.

Table 1. Different types of variables

Term	Type	Value range
Index	[1] Numeric	1-152
Patient.Id	Numeric	152
Age	Numeric	14-73
Gender	Categorical	Male/Female
Air.Pollution	Categorical	1-8
Alcohol.use	Categorical	1-8
Dust.Allergy	Categorical	1-8
OccuPational.Hazards	Categorical	1-8
Genetic.Risk	Categorical	1-7
Chronic.Lung.Disease	Categorical	1-7
Balanced.Diet	Categorical	1-7
Obesity	Categorical	1-7
Smoking	Categorical	1-8
Passive.Smoker	Categorical	1-8
Chest.Pain	Categorical	1-9
Coughing.of.Blood	Categorical	1-9
Fatigue	Categorical	1-9
Weight.Loss	Categorical	1-8
Shortness.of.Breath	Categorical	1-9
Wheezing	Categorical	1-8
Swallowing.Difficulty	Categorical	1-8
Clubbing.of.Finger.Nails	Categorical	1-9
Frequent.Cold	Categorical	1-7
Dry.Cough	Categorical	1-7
Snoring	Categorical	1-7
Level	Categorical	1-3

Figure 1 shows the gender distribution of this dataset. From the bar chart, we can see that there is little difference in the ratio of men to women in this data set, which is a relatively balanced data set.

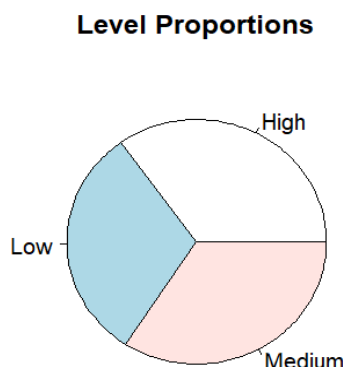


Fig 1. Proportion of lung cancer severity

Since there are 26 variables in Table 1, the following exploration and analysis will be quite complicated. Too many variables will make the model too complex and improve the accuracy of the model's predictions very little. Therefore, it is necessary to screen out some variables that are not relevant to the purpose of the study, or have mutual influence. First, the variables index and patient id that are not relevant to the purpose of the study are screened out. Then the correlation analysis of the remaining variables was performed.

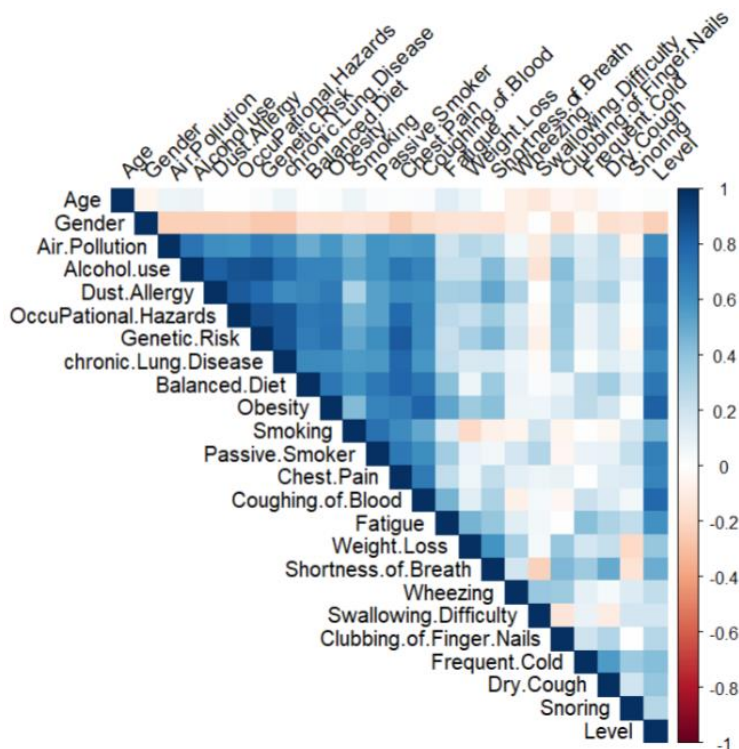


Fig 2. The correlation of each variable

According to the thermal map of variable correlation in Figure 2, it is not difficult to see that there is no collinearity problem among variables. According to Figure 3, it can be seen that Wheezing, Swallowing Difficulty and Clubbing of Finger Nails have nothing to do with the purpose of the study or the severity of lung cancer. Therefore, we removed these three variables to facilitate subsequent analysis.

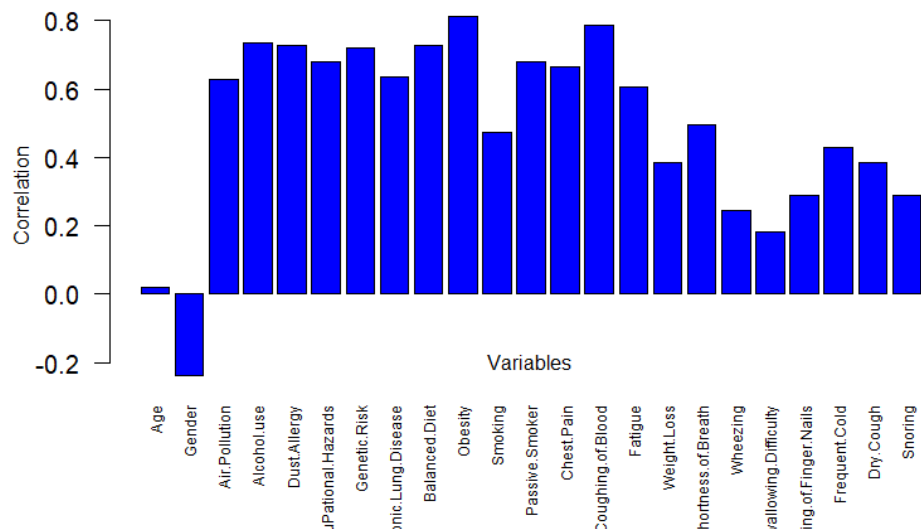


Fig 3. Each variable was associated with the severity of lung cancer

2.3. Method Introduction

The method used in this study is to model fit each variable. Methods used include K Nearest Neighbors (KNN), Support vector machines (SVM), decision tree, random forest and multiple logistic regression. The first step of all method is to divide the data set used for fitting into the training data set and the prediction data set. Then the model is fitted to the training data. The results are then verified with the prediction data set, and finally the percentage of predictions that are correct for each model is calculated. Then accuracies are compared with each other to arrive at the best model.

3. Results and Discussion

3.1. Index Selection

After screening, a total of 21 variables were used for model fitting, all of which were common indicators affecting the severity of lung cancer. Through machine learning, this paper constructs a total of five different categories of models. Because the accuracy of the model's prediction can most intuitively reflect the quality of a model. Therefore, this paper takes the accuracy of the model as the only index to compare the quality of the model.

3.2. Decision Tree Model

The first model to be constructed is the decision tree model. By judging each variable, the decision tree model can easily classify the severity of lung cancer in each patient.

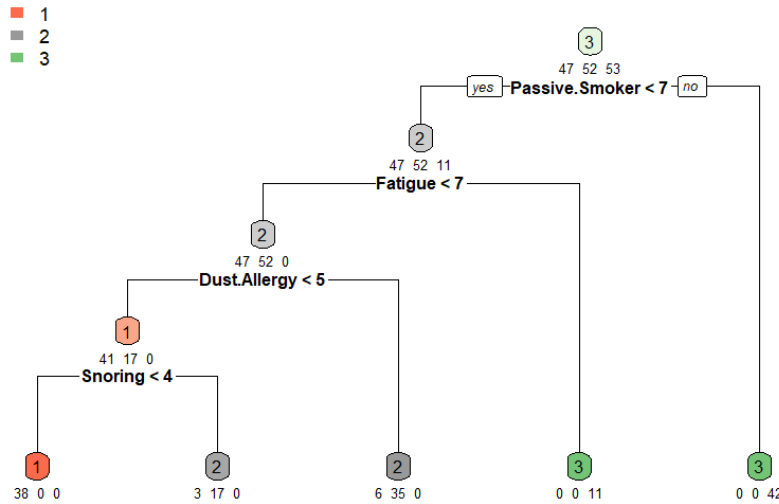


Fig 4. Decision Tree Model

As shown in the figure 4, 1, 2 and 3 respectively represent the mild, medium and high severity of lung cancer. Figure 4 intuitively shows the decision-making process of the decision tree model. The decision tree model demonstrates remarkable performance with a prediction accuracy as high as 0.9407, underscoring its excellent capabilities. This outstanding accuracy holds significant practical importance, particularly in predicting the physical condition of lung cancer patients. Notably, the model achieves this high accuracy while utilizing only four key features as the nodes of the tree.

This succinct selection of key features implies that, in practical application, doctors can make highly accurate predictions regarding the severity of lung cancer by focusing on these specific characteristics. The streamlined nature of the decision tree, with its concise set of crucial attributes, enhances the model's interpretability and usability in real-world scenarios. This focused selection of key features signifies that, in practical applications, doctors can make highly accurate predictions regarding the severity of lung cancer by concentrating on these specific attributes. The streamlined characteristics of the decision tree, coupled with its high interpretability, empower doctors to adapt or omit a characteristic variable, allowing them to tailor the model to better align with the nuances of specific cases.

Nevertheless, the decision tree model exhibits a drawback in its prolonged processing time when confronted with a substantial volume of datasets, be it during the fitting or subsequent prediction stages. To address this issue, the adoption of the random forest algorithm becomes imperative—a refined iteration of the decision tree algorithm.

3.3. Random Forest Model

The accuracy of the random forest model, at 0.8965, falls slightly short of the decision tree model. Nevertheless, as an upgraded algorithm derived from the decision tree model, the random forest exhibits notable advantages, particularly in its ability to swiftly process high-dimensional data samples. Unlike its predecessor, this model proves resilient against the impact of outliers or overfitting, owing to its random sampling and iterative fitting process on the dataset. One of the distinctive strengths of the random forest model lies in its interpretability. It excels in providing an intuitive showcase of feature importance, offering a clear understanding of the variables influencing the predictions.

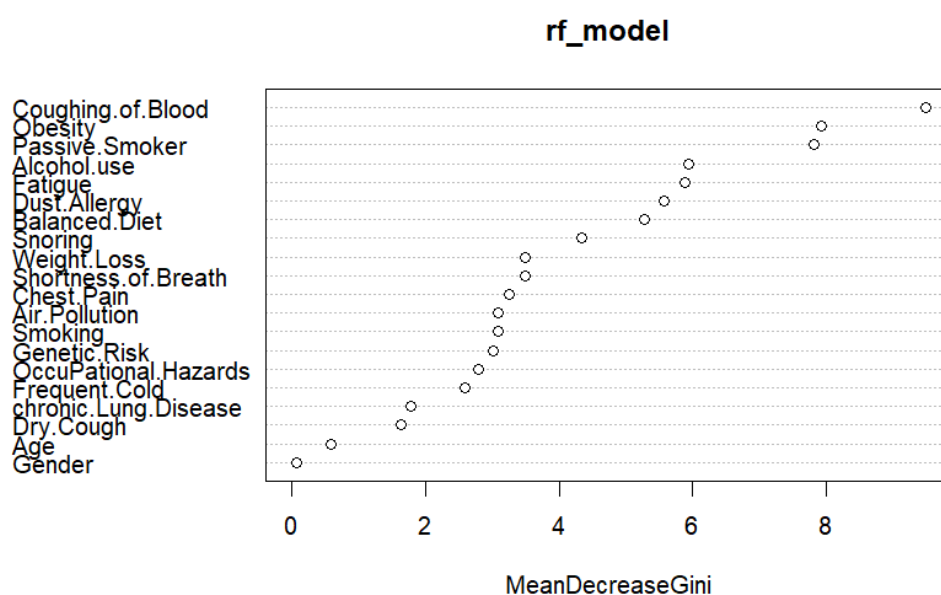


Fig 5. The importance of each feature

As illustrated in Figure 5, it is clear that the degree of blood coughing plays a crucial role in assessing the severity of lung cancer. Both smoking and air pollution exhibit nearly equivalent impacts on the degree of lung cancer. Surprisingly, the influence of passive smoking surpasses that

of active smoking, exerting a more pronounced effect on the severity of lung cancer. This observation may be attributed to the fact that passive smokers are unable to excrete the harmful substances produced by smoking immediately after inhaling them. Another plausible explanation might be that individuals exposed to passive smoke are not as frequently immersed in smoke, potentially leading to a heightened sensitivity or reaction to the occasional exposure.

3.4. KNN Model

For a large number of data, in order to pursue efficiency, K-Nearest Neighbors (KNN) algorithm is also a good choice. This algorithm classifies new observations based on the labels of nearby observations, endowing it with a resistance to noise. The inherent anti-noise property means that a few outliers in the dataset have minimal impact on the overall predictive results. However, the efficacy of the KNN algorithm heavily relies on the selection of the parameter 'k,' representing the number of neighbors considered during classification. Since KNN concentrates on points in proximity to the new observation, the labels of surrounding points and the quantity of points observed become pivotal for accurate predictions. The optimal 'k' value significantly influences the model's accuracy, and fine-tuning this parameter through cross-validation can maximize predictive performance.

Visually depicted in Figure 6 are the variations in model accuracy corresponding to different 'k' values. Ultimately, at a 'k' value of 6, the KNN algorithm attains its peak accuracy at 0.8621. Despite this success, it's essential to acknowledge a limitation inherent in the KNN algorithm—its localized focus, which leads to a tendency to overlook the broader context.

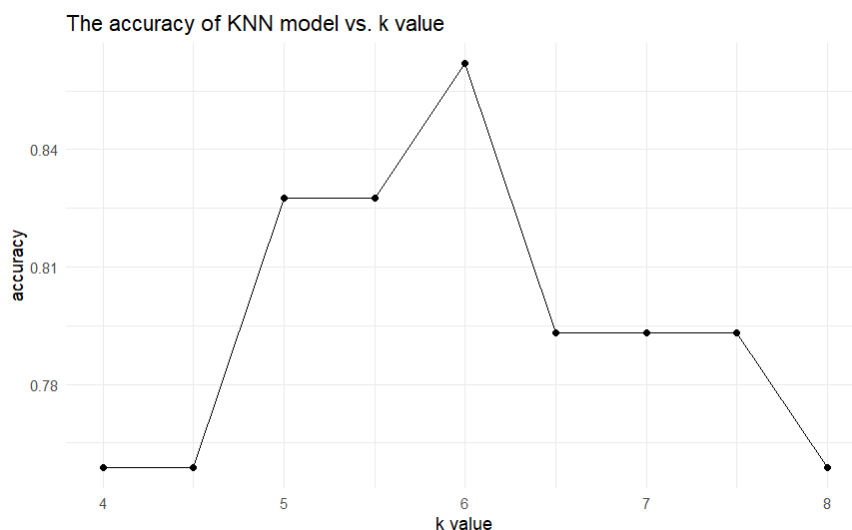


Fig 6. The relationship between k and accuracy

3.5. SVM Model

The SVM (Support Vector Machine) model stands out for its ability to robustly classify each class with strength, concurrently ensuring high overall correctness. With an impressive prediction accuracy of 0.862, the SVM model showcases its proficiency in handling complex datasets. Moreover, the model boasts few algorithmic shortcomings, further enhancing its appeal in practical applications.

SVM's strength lies in its capacity to find optimal hyperplanes that distinctly separate different classes, even in high-dimensional spaces. This capability minimizes classification errors and results in a high degree of accuracy. Additionally, the inherent structure of SVM contributes to its generalization performance, making it adept at handling diverse datasets.

3.6. Multiple Logistic Regression Model

The final model is the multiple logistic regression model. Compared with other models, logistic regression model has the advantage that it can form concrete formulas. The formula, delineated below, offers a clear representation of the predicted probabilities for different severity levels of lung cancer:

Table 2. Parameters of different indicators

Term	Symbol	Level 2	Level 3
Intercept	[2] null	-359.7429	-536.4836
Dust.Allergy	x	51.246913	-7.136176
OccuPational.Hazards	y	-47.110590	-6.363911
Chronic.Lung.Disease	z	37.78174	26.98468
Smoking	a	0.4358495	-7.4655138
Fatigue	b	-0.227953	14.693419
Shortness.of.Breath	c	16.41578	28.10051
Snoring	d	42.84619	42.43494
Passive.Smoker	e	9.375989	30.195792
Air.Pollution	f	-9.466555	24.845067

$$\log(p/1-p) = -359.7429 + 51.246913 \times x - 47.110590 \times y + 37.78174 \times z + 0.4358495 \times a - 0.227953 \times b + 16.41578 \times c + 42.84619 \times d + 9.375989 \times e - 9.466555 \times f, \text{ if Level} = 2. \quad (1)$$

$$\log(p/1-p) = -536.4836 - 7.136176 \times x - 6.363911 \times y + 26.98468 \times z - 7.4655138 \times a + 14.693419 \times b + 28.10051 \times c + 42.43494 \times d + 30.195792 \times e + 24.845067 \times f, \text{ if Level} = 3. \quad (2)$$

This model achieves an impressive accuracy of 0.9655 upon testing, signifying its robust predictive performance (Table 2). At the same time, according to the coefficient of each variable in the model, people can intuitively feel the impact of the change of the value of each variable on the severity of lung cancer. For example, taking smoking as an example, the logarithmic probability of moderate lung cancer severity increases by 29.56,304 and the logarithmic probability of high lung cancer severity increases by 38.95725 for every 1 increase in smoking degree. This direct correlation between variable changes and probability shifts serves as a practical tool to guide individuals in understanding and mitigating the risk factors associated with worsening lung cancer severity. Overall, this logistic regression model not only offers high accuracy but also facilitates a nuanced understanding of the influential factors.

3.7. Model Selection

According to the comparison of accuracy, the multiple logistic regression model shows the best performance (Table 3).

Table 3. Accuracy comparison

	Decision Tree	Random Forest	KNN	SVM	Multiple logistic regression
Accuracy	0.9407	0.8965	0.8621	0.8965	0.9655

4. Conclusion

This study underscores the importance of adopting healthy behaviors to minimize the risk of severe lung cancer. Recommendations include reducing smoking, avoiding second-hand smoke, and residing in areas with good air quality whenever possible. In instances of poor air quality, it is advised to limit outdoor activities, and if necessary, wearing a mask can help mitigate the inhalation of PM2.5. These practices, as suggested by the model, can contribute to a decrease in the likelihood of developing severe lung cancer. Furthermore, the model provides individuals with a valuable tool to assess their risk and potentially seek timely medical attention for early intervention, thereby improving the chances of successful treatment. However, it's crucial to acknowledge that the uneven distribution of individuals living in different air quality levels within the dataset may limit the model's general applicability. Future studies should focus on collecting more diverse clinical data to enhance

the model's universality and ensure its relevance across a broader range of situations. This approach will contribute to the model's effectiveness in providing personalized insights and guidance for individuals from various backgrounds and living conditions.

References

- [1] Nash J, Brims F. International standards of care in thoracic oncology: A narrative review of clinical quality indicators. *Lung Cancer*, 2023, 186: 107421.
- [2] Sung H, Ferlay J, Siegel R L, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a Cancer Journal for Clinicians*, 2021, 71(3): 209-249.
- [3] Maise Al Bakir, Huebner A, Carlos Martinez Ruiz, et al. The evolution of non-small cell lung cancer metastases in TRACERx. *Nature*, 2023, 616: 534-542.
- [4] Hafiza Padinharayil, Varghese J, Mithun Chacko John, et al. Non-small cell lung carcinoma (NSCLC): Implications on molecular pathology and advances in early diagnostics and therapeutics. Working paper, 2022.
- [5] Traub B, Link K H, Kornmann M. Curing pancreatic cancer. *Seminars in Cancer Biology*, 2021, 76: 232-246.
- [6] Berg C D, Schiller J H, Boffetta P, et al. Air pollution and lung cancer: A Review by International Association for the Study of Lung Cancer Early Detection and Screening Committee. *Journal of Thoracic Oncology*, 2023.
- [7] Samet J M. Particulate Air Pollution and Mortality. *Epidemiology*, 1995, 6(5): 471-472.
- [8] Jia X, Sheng C, Han X, Li M, Wang K. Global burden of stomach cancer attributable to smoking from 1990 to 2019 and predictions to 2044. *Public Health*, 2023, 226: 182-189.
- [9] Loomis D, Grosse Y, Lauby-Secretan B, et al. The carcinogenicity of outdoor air pollution. *The Lancet Oncology*, 2013, 14(13): 1262-1263.
- [10] Liu Y, Lu L, Yang H, et al. Dysregulation of immunity by cigarette smoking promotes inflammation and cancer: A review. *Environmental Pollution*, 2023, 339: 122730.