SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

*Molecular & Cellular Biomechanics* 2024, 21, 105.
https://doi.org/10.62617/mcb.v21.105

Article

# Diagnosis and correlation analysis of lung cancer based on multi-parameter regression of respiratory volatile organic compounds

**Lishan Qin[1,†], Yunzhen Wang[2,†], Fei Wang[1,*], Ziyi Zhu[2,*], Raojun Luo[2], Guojun Lv[1], Haibin Cui[1]**

[1] State Key Laboratory of Clean Energy Utilization (Zhejiang University), Hangzhou 310027, China
[2] Department of Thoracic Surgery, Sir Run Run Shaw Hospital, School of Medicine, Hangzhou 310016, China
*** Corresponding authors:** Fei Wang, wangfei@zju.edu.cn; Ziyi Zhu, zhuziyicn@zju.edu.cn
[†] The two authors contribute equally to the article.

**CITATION**

Qin L, Wang Y, Wang F, et al. Diagnosis and correlation analysis of lung cancer based on multi-parameter regression of respiratory volatile organic compounds. Molecular & Cellular Biomechanics. 2024; 21: 105. https://doi.org/10.62617/mcb.v21.105

**ARTICLE INFO**

Received: 11 April 2024
Accepted: 26 April 2024
Available online: 30 May 2024

**Abstract:** Lung cancer is a prevalent and life-threatening disease worldwide. The primary diagnostic approach for lung cancer is the utilization of low-dose spiral CT scans. However, repeated scans can expose patients to harmful radiation. Consequently, there is growing interest in exploring alternative methods such as the analysis of exhaled volatile organic compounds (VOCs) for lung cancer detection. In this study, we employed a gas chromatography-mass spectrometry analyzer to identify and quantify a total of 108 VOCs of lung cancer patients. Our objective is to investigate the correlation between VOCs in exhaled breath and lung cancer. Through the application of orthogonal partial least squares-discriminant analysis and correlation analysis, we identified several VOCs, including acetone, ethanol, isopropanol, and ethyl acetate, which exhibited a strong association with lung cancer. Unlike the use of a single marker, our study employed a multi-parameter regression method, resulting in superior accuracy. A diagnostic model based on the neural network algorithm was established, demonstrating an accuracy of 93.5% after screening, surpassing the accuracy before screening at 81%. Furthermore, we optimize the model by incorporating the gender factor, leading to an accuracy exceeding 96%. Numerous studies have demonstrated that the analysis of VOCs in exhaled breath holds significant potential for effectively distinguishing lung cancer patients from healthy individuals. These findings emphasize the potential of respiratory analysis as a novel diagnostic tool for early detection of lung cancer.

**Keywords:** volatile organic compounds; multi-parameter regression method; neural network

## 1. Introduction

Currently, the global condition of lung cancer is extremely serious. According to the World Health Organization, over 1.8 million people die from lung cancer annually worldwide, including many non-smokers. Lung cancer is one of the most prevalent and lethal cancers globally [1]. The incidence and mortality rates of lung cancer are on the rise in many countries, particularly in developing nations [2,3]. The primary risk factors for lung cancer include smoking, air pollution, radiation, genetic factors, and occupational exposure [4]. Due to the relatively insidious onset of lung cancer, symptoms such as hemoptysis, chest pain, and dizziness often indicate that the tumor has progressed to a late stage [5]. Therefore, early detection of lung cancer is crucial for patient prognosis. According to statistics, the 5-year survival rate of early-stage IA patients is 90%, while that of late-stage IV patients is less than 5% [6]. Currently, the most important screening method worldwide is low-dose spiral CT, which is 4–10 times more sensitive than conventional chest x-rays for early lung cancer detection [7,8]. The National Lung Cancer Screening Trial in the United States has demonstrated that low-dose spiral CT screening can reduce lung cancer mortality in high-risk populations by 20% [9,10]. However, low-dose spiral CT screening for lung cancer also poses certain problems [11]. The judgment of benign and malignant pulmonary nodules involves subjective and objective factors of the attending doctor, leading to a high false positive rate [12]. Consequently, there is an urgent need to establish a more economical, non-invasive, efficient, universal, high-throughput, and accurate screening method for lung cancer [7].

Expiratory diagnosis is a current hot topic in medical research. This diagnostic technique detects the presence of certain diseases by analyzing the type and concentration of VOCs in exhaled breath [13,14]. It offers several advantages, including non-invasiveness, low cost, high sensitivity and specificity, low false positive rate. Expiratory diagnosis is a promising method for screening and diagnosing lung cancer. Currently, research on exhalation markers for lung cancer is gaining momentum, primarily focusing on hydrocarbons, alcohols, aldehydes, ketones, aromaticity, and esters [15]. In 1985, Gordon et al. conducted a pioneering study on detecting VOC biomarkers in exhaled breath, analyzing 12 lung cancer patients and 17 control samples using computer-aided chromatography-mass spectrometry (GC-MS) [14]. In 1999, Phillips et al employed GC-MS to screen 22 VOC-specific compounds in the exhaled breath of lung cancer patients [16]. In 2016, Saalberg identified 77 VOCs from the literature on lung cancer markers published between 1985 and 2015 [14]. **Table 1** shows lung cancer exhalation markers reported in the literature and screened from lung cancer patients and healthy people. Due to differences in detection technologies and analytical methods, there is still no recognized marker with sufficient specificity.

**Table 1.** Lung cancer breath testing performance and biomarkers.

| Author, year | Sensitivity | Specificity | Exhalation markers |
|---|---|---|---|
| Bousamra, 2014 [17] | 87.9 | 77.5 | 1. 2-butanone, 2. 3-hydroxy-2-butanone, 3. 2-hydroxyacetaldehyde, 4. 4 hydroxyhexenal |
| Ligor, 2015 [18] | 63.5 | 72.4 | 1. Butane, 2. 2methylbutane, 3. 4-methyloctane, 4. propane, 5. 2-pentanone, 6. propionaldehyde, 7. 2,4-dimethylheptane, 8. propylene |
| Nisreen, 2016 [19] | 87 | 82 | 1. Heptane, 2. Decane, 3. 2-methylpentane, 4. 2-ethyl-1-hexanol, 5. Propanal, 6. Valeraldehyde, 7. Acetone |
| Sanchez, 2017 [20] | 32 | 87.6 | 1. Hexanal, 2. Heptanal, 3. Octanal, 4. Nonanal, 5. Propionic acid, 6. Nonanoic acid |
| Wang, 2018 [21] | 80.8 | 84 | 1. 3-ethyltoluene, 2. 1,2,3-trimethylbenzene, 3. n-propylbenzene, 4. propylcyclohexane, 5. indene, 6. 1-methyl-3-propylbenzene, 7. o-xylene, 8. 4-Methyl-2-pentanone, 9. 5-methylindole, 10. methylcyclohexane |

Several methods are currently used for breath detection, including electronic nose, spectral analysis, and GC-MS [22]. Electronic nose technology has limitations such as a short service life and insufficient detection accuracy [23–25]. Spectral analysis is limited by the light source and can only measure a limited number of VOCs. In contrast, GC-MS has high sensitivity, and a low detection limit [26]. When combined with a pre-concentration device, it can provide qualitative and quantitative measurements of hundreds of VOCs at parts-per-billion (ppb) levels [27]. The operation is simple, and the method is well-established, making it an ideal detection method for breath VOCs [28]. Currently, the challenge in the clinical application of breath analysis for the diagnosis of lung cancer lies in several factors. On one hand, due to the diversity and complexity of physiological and pathological states in humans, no single marker can distinguish healthy individuals from lung cancer patients. On the other hand, there is a lack of quantitative correlation between breath biomarkers and diseases, metabolic abnormalities, and human body parameters. The relationship between lung cancer and markers is complex and usually non-linear, making it difficult to identify lung cancer through markers. For instance, studies have found that lung cancer patients have higher exhaled acetone levels compared to controls, suggesting that acetone may be a potential biomarker for lung cancer [29]. However, increased exhaled acetone is also associated with several other diseases, such as diabetes, alcoholic hepatitis, as well as heart failure. Therefore, screening breath markers for lung cancer and establishing a correlation model between markers and diseases will be key to the clinical application of breath diagnosis of lung cancer.

In this study, a diagnostic model for lung cancer was constructed using multiple VOC indicators. Compared to using a single VOC as an indicator, the diagnostic model exhibited better specificity and sensitivity, while avoiding errors caused by a single indicator. Furthermore, the non-linear relationship between multiple indicators and lung cancer was filtered out using the orthogonal partial least squares-discriminant analysis (OPLS-DA) method, which eliminated poorly correlated feature variables. A neural network algorithm was then employed to establish a correlation model for lung cancer diagnosis, and the stability of the model was verified through experimental validation. These findings provide a novel approach to developing a more accurate and reliable diagnostic model for lung cancer.

## 2. Materials and methods

### 2.1. Study population

All participants in this study were from the Department of Thoracic Surgery, Sir Run Run Shaw Hospital, Zhejiang University. The subjects consisted of 75 lung cancer patients and 50 healthy controls. Patients included in this study must meet the following criteria: (I) All cases were diagnosed as lung cancer by histopathology and CT imaging, and had clear TNM staging information; (II) No cancer-related surgery, radiotherapy, chemotherapy and other cancer-related treatments were received before admission. (III) No other lung diseases; no heart, liver, kidney and other organ dysfunction; (IV) no infection or pregnancy, no blood disease; the healthy subjects included in the study must meet the following criteria: (I) good health, no previous medical history and no family history of lung cancer; (II) individuals without other lung diseases, heart, liver, kidney and other organ dysfunction; (III) no infection or pregnancy, no blood disease; prior to sample collection, all participants provided informed consent to ensure that they were fully aware of the purpose, procedures and potential risks of sample collection. The collection procedure followed the requirements of medical ethics, respected the rights of participants, avoided unnecessary physical and mental harm, and ensured that each participant understood and supported the sampling process. Subject information is shown in **Table 2**.

**Table 2.** Characteristics of the patients and healthy controls.

| **Characteristic** | **Patients** | **Healthy controls** |
|---|---|---|
| Age (year) | 59.8 ± 10.4 | 52.3 ± 15.3 |
| gender | | |
| male | 35 | 26 |
| female | 40 | 24 |
| BMI | 23.2 ± 2.8 | 22.8 ± 2.6 |
| Clinical stage | | |
| II | 5 | |
| III | 21 | |
| IV | 49 | |
| History of COPD | no | no |
| History of pulmonary infection | no | no |
| family history of lung cancer | no | no |

## 2.2. Exhaled breath collection and sampling

Due to individual differences in breathing patterns, diet, and other factors, the composition of exhaled breath can show significant variations. To obtain repeatable and reliable breath samples, the sampling procedure was as follows: subjects were asked not to brush their teeth or eat prior to sampling, rinse their mouth with water, and breathe in a quiet and well-ventilated environment for 5 min without engaging in any strenuous exercise. A 1L Tedlar bag and blow nozzle were used, with the gas bag inlet valve turned 1/2 turn counterclockwise to open it. The subjects were instructed to blow into the bag with moderate force, and to strictly avoid any backward suction. After the gas was collected, the gas bag inlet valve was turned 1/2 turn clockwise to close it. A mark was made on the white label outside the gas bag, and the patient information was recorded. Environmental background air samples were also collected as a control.

## 2.3. GC-MS equipment

The detection instrument is a gas chromatography mass spectrometer produced by Agilent, and the model is 5977A-7890B. And ENTECH's three-layer cold sink pre-concentrator, model ENTECH7200. The GC-MS instrument parameters are set as follows. Chromatographic conditions: The column type was DB-5 ms (60 m× 320 μm × 1 μm), the forward sample port was connected to the mass spectrometry, and the constant current was 2 mL/min. The temperature of the forward sample port was 230 ℃, the carrier gas was He, the split ratio was 10:1, and the split flow rate was 20 mL/min. Column temperature program: the initial temperature of 35 ℃ for 5 min, 5 ℃/min to 150 ℃ for 7 min, then at a rate of 10 ℃/min to 200 ℃ and maintained for 4min, the end of the analysis. The analysis took a total of 44 minutes, and the use temperature of the cold sink was 50 ℃. Mass spectrometry conditions: ion source temperature of 230 ℃, quadrupole stability of 150 ℃, acquisition mode of selective ion scanning (SIM), EMV voltage of 1221.9 V, emission current of 34.6 μA, emission energy of 70.0 eV.

The standard substances were TO-15 and PAMS, a total of 108 substances, and a mixed standard gas of 1 ppb was configured.

### 2.4. Statistical analysis

In this study, two statistical methods, orthogonal partial least squares discriminant analysis (OPLS-DA) and Spearman correlation coefficient, were used to verify the correlation between volatile organic compounds in exhaled breath and disease. In the OPLS-DA analysis, we focus on the importance projection (VIP) value of the variable, and a VIP value higher than 1 is considered to be statistically significant. At the same time, we use the Spearman correlation coefficient as a non-parametric measure. When the ρ value is greater than 0.8, we think it is statistically significant.

### 2.5. The construction of multi-parameter regression model

In this study, we adopt a complex neural network architecture, which consists of an input layer, two hidden layers and an output layer. The input layer preliminarily processes the original data and transmits the processed information to the hidden layer. These hidden layer neurons are activated using the Leaky ReLU function. This unique activation function allows neurons to pass some information after being activated, helping the network avoid potential inactive neuron problems. After these two layers of processing, the data is finally collected to the output layer. In the output layer, we use the Sigmoid activation function, which can map the output of the neuron to between 0 and 1, so that it can be interpreted as probability. In order to evaluate the performance of the model, we use the binary cross entropy loss to measure the consistency between the model output probability and the real label. Finally, in order to optimize our network parameters, we choose adaptive moment estimation as the optimizer. This optimizer combines the advantages of the adaptive gradient algorithm and root mean square propagation algorithm, which can adaptively adjust the learning rate and improve the speed and stability of our model learning.

## 3. Results

### 3.1. Data dimension reduction and feature extraction

In this study, pre-concentrator and GC-MS were employed to analyze 125 respiratory samples, aiming to investigate the differences in VOCs exhaled by lung cancer patients and healthy individuals. A total of 108 VOCs were identified, including 44 hydrocarbons, 12 ketones, 26 alcohols, 9 aldehydes, 8 esters, and 9 acids. The contents and types of these VOCs are of great significance in the early diagnosis and treatment of lung cancer.

To establish a reliable model, the OPLS-DA method was utilized with 108 VOCs as independent variables and disease status as dependent variables to distinguish patients and healthy individuals. The fitting indices of independent variables (R2X), dependent variables (R2Y), and model prediction (Q2) were 0.891, 0.924, and 0.769, respectively. Both R2 and Q2 values were over 0.5, indicating that the fitting results were acceptable. The R2X and R2Y values of the model indicate a strong correlation between independent and dependent variables, while the Q2 values indicate a good prediction ability of the model. In addition, 200 permutation experiments were conducted, and the intersection of the Q2 regression line and the vertical axis was found to be −0.19, proving that the model was not over-fitting, and the reliability and robustness of the model were also verified, as illustrated by **Figure 1**. These results suggest that our model can be used to distinguish lung cancer patients from healthy individuals and has good predictive ability.

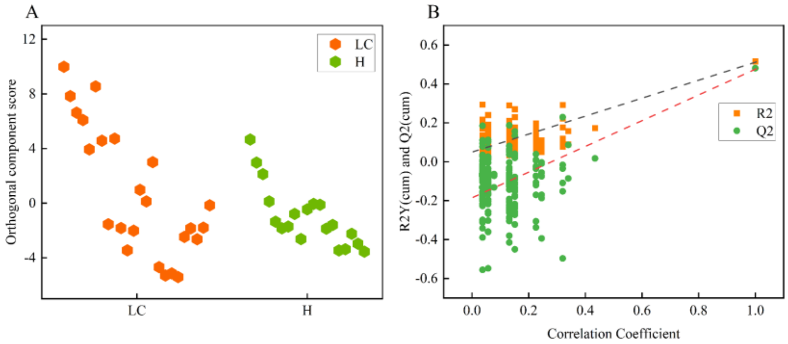

**Figure 1.** OPLS-DA pretreatment results. (**A)** OPLS-DA scatter plot; (**B)** permutation test.

In OPLS-DA, the VIP value is a crucial metric for evaluating the contribution of each independent variable to the model. The calculation of the VIP value is based on the projection direction in the OPLS-DA model. The projection direction divides the independent variables into two parts: the part related to the dependent variable and the part not related to the dependent variable. The VIP value indicates the importance of each independent variable in the dependent component. The greater the VIP value, the greater the contribution of the independent variable to distinguish between different categories of samples. The significance of the VIP value lies in its ability to help us determine which independent variables are most important in distinguishing between different categories of samples. By using the VIP value, we can select the most discriminative independent variables to improve the predictive ability and reliability of the model. Additionally, the VIP value can be used for feature selection, which involves selecting the most representative and discriminative independent variables from a large number of independent variables to simplify the model and improve its interpretability. The VIP values of six substances were greater than 1, as shown in **Table 3**.

**Table 3.** VOCs with significant VIP value.

| **VOCs name** | ethanol | acetone | ethyl acetate | isopropyl alcohol | cis-1,2-dichloroethene | 2,3,4-Trimethylpentane |
|---|---|---|---|---|---|---|
| **VIP value** | 6.8 | 4.3 | 2.7 | 2.1 | 1.1 | 1 |

### 3.2. Analysis of relationship

The present study aimed to investigate the distribution of VOCs in all breath samples, with a particular focus on discerning differences between healthy individuals and lung cancer patients. Notably, the concentrations of acetone and methyl ethyl ketone were found to be more concentrated in healthy individuals, while more dispersed in patients with lung cancer (**Figure 2**).

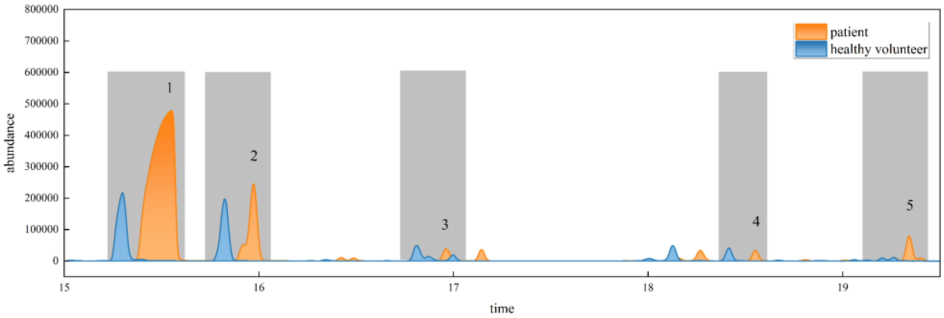

**Figure 2.** Comparison of VOCs content in exhaled breath between healthy people and lung cancer patients.

1. acetone; 2. lsopropanol; 3. disulfide carbon; 4. trans-1,2-dichloroethylene; 5. methyl ethyl ketone.

Before conducting correlation analysis, it is important to check whether the data conforms to normal distribution. Normality tests can use various methods, such as the Shapiro-Wilk test, the Kolmogorov-Smirnov test, and the Anderson-Darling test. If the data conforms to a normal distribution, the parameter correlation analysis method, such as the Pearson correlation coefficient, can be used. Pearson’s correlation coefficient measures the strength and direction of a linear relationship between two continuous variables. It assumes that the data follows a normal distribution and that the relationship between the two variables is linear. However, if the data does not conform to normal distribution or if the relationship between the two variables is not linear, using the Pearson correlation coefficient may result in inaccurate results. Non-parametric correlation analysis methods, such as Spearman correlation coefficient, can be used if the data does not conform to normal distribution. The Spearman correlation coefficient is a non-parametric method that does not require the assumption that the data conforms to normal distribution or that the relationship between two variables is linear. It calculates the strength and direction of the relationship between two variables by converting the data to rank. The Spearman rank correlation coefficient is applicable to non-normal distribution data and non-linear relationship data.

The Shapiro-Wilk test is suitable for data sets with small sample sizes (less than 50) and has higher sensitivity for normally distributed data. Therefore, the Shapiro-Wilk test was used to analyze the normality of the 108 VOCs. Of the 108 VOCs, 104 VOCs did not conform to normal distribution. In the present study, Spearman's correlation coefficient was utilized as an indicator to evaluate the relationship between VOCs and Disease Risk [30], which is a robust approach for assessing non-linear relationships (**Figure 3**). The findings revealed a significant association between certain VOCs and Disease Risk, including commonly occurring organic compounds such as acetone, ethanol, and ethyl acetate (**Table 4**).

$$\rho = \frac{\sum_{i=1}^{N} (R_i - \bar{R})(S_i - \bar{S})}{\left[\sum_{i=1}^{N} (R_i - \bar{R})^2 \sum_{i=1}^{N} (S_i - \bar{S})^2\right]^{\frac{1}{2}}} = 1 - \frac{6\sum d_i^2}{N(N^2 - 1)} \tag{1}$$

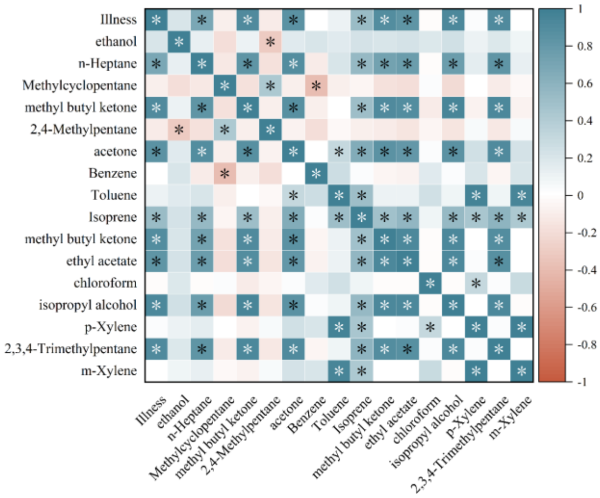

**Figure 3.** Heat map of correlation between VOCs in exhaled breath and disease.

**Table 4.** $\rho$-value of Spearman correlation analysis.

| **VOCs name** | acetone | ethanol | isopropyl alcohol | ethyl acetate | n-Heptane | 2,3,4-Trimethylpentane | methyl butyl ketone |
|---|---|---|---|---|---|---|---|
| **$\rho$-value** | 0.95 | 0.93 | 0.96 | 0.91 | 0.87 | 0.86 | 0.89 |

### 3.3. Establishment and optimization of lung cancer diagnosis model

In this study, we employed a statistical significance level ($p > 0.8$) and a variable importance index (VIP > 1) to screen for VOCs related to lung cancer. After analysis, we identified four VOCs, namely ethanol, acetone, isopropanol, and ethyl acetate, which exhibited significantly higher levels in the exhaled breath of lung cancer patients compared to healthy individuals (**Table 5**). These VOCs hold potential diagnostic value for lung cancer.

**Table 5.** Comparison of four VOCs content.

| **VOCs** | **Patients** | **Healthy controls** |
|---|---|---|
| ethanol | 3037 ± 2966 | 817 ± 621 |
| acetone | 1663 ± 1507 | 565 ± 357 |
| isopropanol | 354 ± 450 | 134 ± 94 |
| ethyl acetate | 1058 ± 1067 | 627 ± 367 |

A confusion matrix is a tool used to evaluate the performance of classification models. It includes four metrics: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). TP refers to the number of samples correctly identified as positive by the classifier, while FP refers to the number of samples incorrectly identified as positive. TN refers to the number of samples correctly identified as negative, and FN refers to the number of samples incorrectly identified as negative. These metrics can be used to calculate various performance indicators such as accuracy, recall, and precision, which help assess the model's classification ability and optimize its performance. Specificity, sensitivity, precision, and accuracy are commonly used to evaluate the non-invasive monitoring performance of lung cancer [31]. Specificity refers to the probability that a sample is correctly identified as negative when it is indeed negative. Sensitivity, also known as recall, refers to the probability that a sample is correctly identified as positive when it is indeed positive. Accuracy represents the proportion of correct predictions among the total number of samples. Precision represents the proportion of true positive samples among all samples predicted as positive.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \tag{2}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{4}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \tag{5}$$

**Figure 4** illustrates the accuracy of the neural network-based model before and after screening. The accuracy of the VOCs model noticeably improved from 81% to 93.5% following the screening process. This improvement can be attributed to the potential interference that arises when all VOCs are simultaneously used as inputs to the model. It is likely that the presence of multiple VOCs simultaneously affected the accuracy of the model negatively. However, through the screening process, the model was optimized by minimizing the interference and refining the input variables, resulting in a significant enhancement in accuracy.

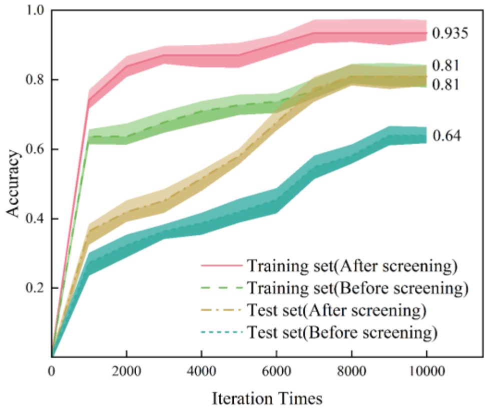

**Figure 4.** Comparison of model accuracy before and after screening.

In order to study the performance differences between single-marker and multi-marker combination models, four VOCs, ethanol, acetone, isopropanol, and ethyl acetate, were used to establish diagnostic models. As shown in the **Figure 5**, the accuracy of the model using a single marker is less than 70%, of which the accuracy of acetone and ethyl acetate is only 58% and 57%. This shows that the use of a single marker cannot accurately distinguish lung cancer patients. The use of multiple markers can identify lung cancer patients from multiple dimensions and greatly improve the accuracy of the model.

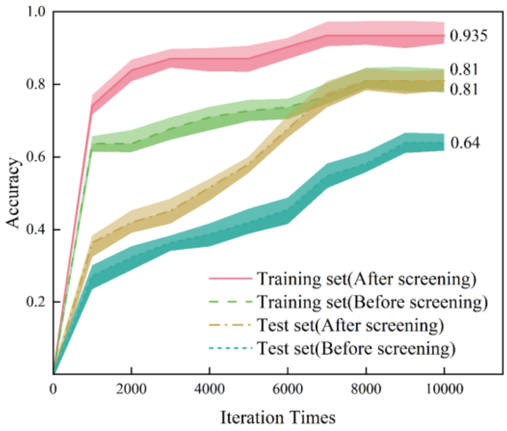

**Figure 5.** Comparison of accuracy of single marker and multi-marker combination models.

To select the most suitable algorithm for lung cancer diagnosis, we utilized ethanol, acetone, isopropanol, and ethyl acetate as input variables for the diagnosis model. Four machine learning algorithms, namely neural network, decision tree, support vector machine tree, and random forest, were employed to establish the diagnosis model. Through comparative analysis, we found that the neural network algorithm-based diagnosis model was the best, with a training set accuracy of 93.5% and a test set accuracy of 81% (**Figure 6**). These results demonstrate that the neural network algorithm has high accuracy and reliability in the diagnosis of lung cancer. Compared to other algorithms, the neural network algorithm can handle non-linear relationships and high-dimensional data more effectively, thereby improving the accuracy and reliability of the diagnosis model.

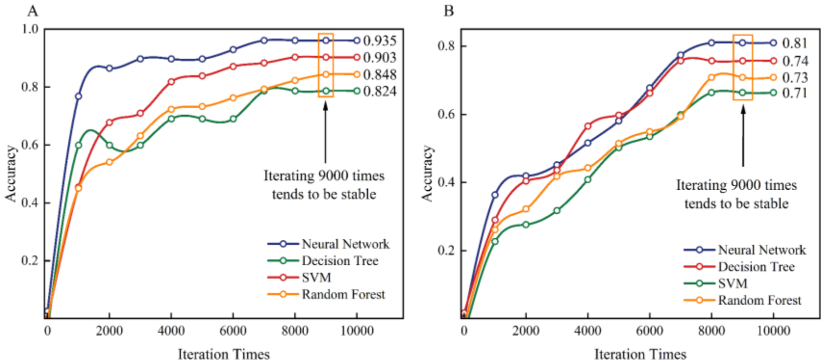

**Figure 6.** Comparison of the effects of four types of models. (**A)** training sets; (**B)** testing set.

Due to the close correlation between the concentration of VOCs in exhaled breath and sex, it is essential to exclude the influence of sex in the diagnosis of lung cancer. To ensure the accuracy and reliability of diagnostic results, fractional analysis of patients of different sexes is required, and the influence of gender factors must be considered when establishing diagnostic models.

In this study, healthy individuals and patients were categorized by gender, and the distribution of four VOCs—ethanol, acetone, isopropanol, and ethyl acetate - in healthy men, healthy women, male patients, and female patients were analyzed. The results showed that in the healthy group, the mean value of exhaled ethanol in male subjects was 685 ± 201 ppb, acetone was 506 ± 401 ppb, isopropanol was 117 ± 106 ppb, and ethyl acetate was 519 ± 386 ppb. In female subjects, the mean value of exhaled ethanol was 1114 ± 206 ppb, acetone was 698 ± 220 ppb, isopropanol was 171 ± 49 ppb, and ethyl acetate was 867 ± 175 ppb (**Figure 7**). Overall, the levels of these four VOCs in the breath of healthy women were significantly higher than those in healthy men.

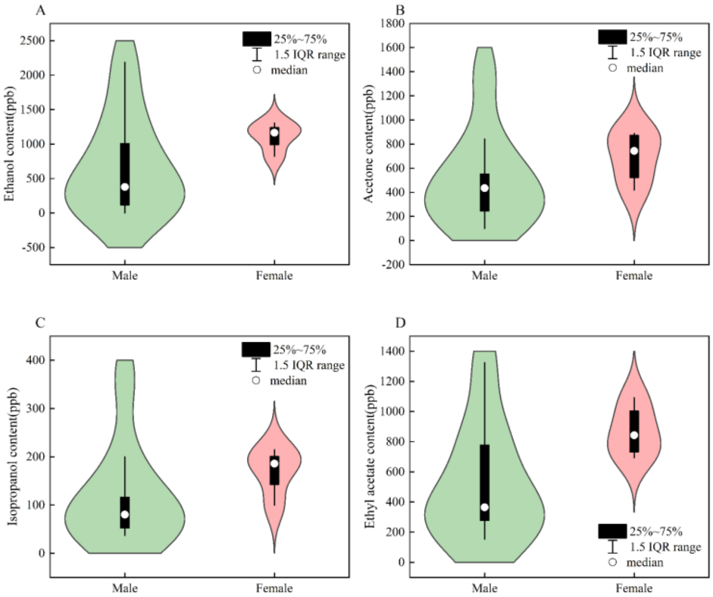

**Figure 7.** Comparison of four VOCs content distribution in exhaled breath of healthy groups of different genders. (**A)** ethanol; (**B)** acetone; (**C)** isopropanol; (**D)** ethyl acetate.

In contrast, the results showed that in lung cancer patients, the mean value of exhaled ethanol in male patients was 2291 ± 1676 ppb, acetone was 2806 ± 2013 ppb, isopropanol was 708 ± 596 ppb, and ethyl acetate was 1768 ± 1210 ppb. The mean value of exhaled ethanol in female patients was 874 ± 657 ppb, acetone was 1265±885 ppb, isopropanol was 228 ± 310 ppb, and ethyl acetate was 1150 ± 1109 ppb (**Figure 8**). Overall, the levels of these four VOCs in the breath of male patients were significantly higher than those in female patients. These findings suggest that gender plays a significant role in the distribution of VOCs in exhaled breath, and this effect differs between healthy individuals and patients. Therefore, gender should be considered as an input in diagnostic models.

After incorporating gender as a factor into the model, a diagnostic model was established based on a neural network algorithm. As shown in **Figure 9**, the efficiency of the training set is increased to 96.8%, and the efficiency of the test set is increased to 85% This indicates that gender is an important factor for the diagnostic model, significantly improving its accuracy and reliability. Further research will be conducted to investigate other factors that may affect the diagnosis results and improve the performance of the model.

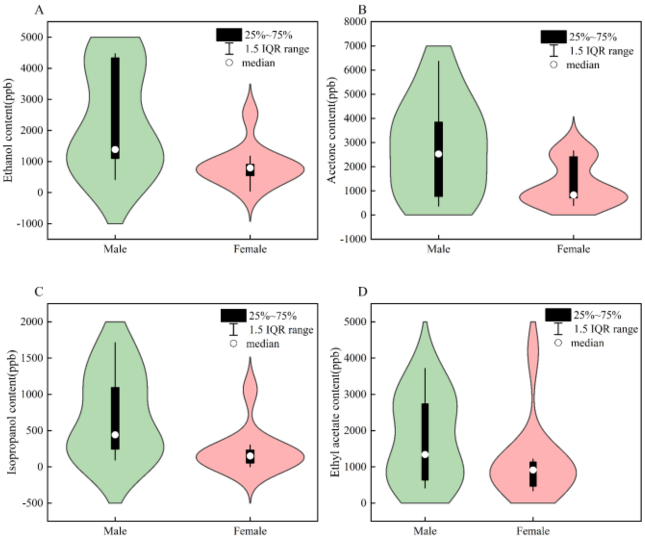

**Figure 8.** Comparison of the distribution of four VOCs in the exhaled breath of lung cancer patients of different genders. **(A)** ethanol; **(B)** acetone; **(C)** isopropanol; **(D)** ethyl acetate.

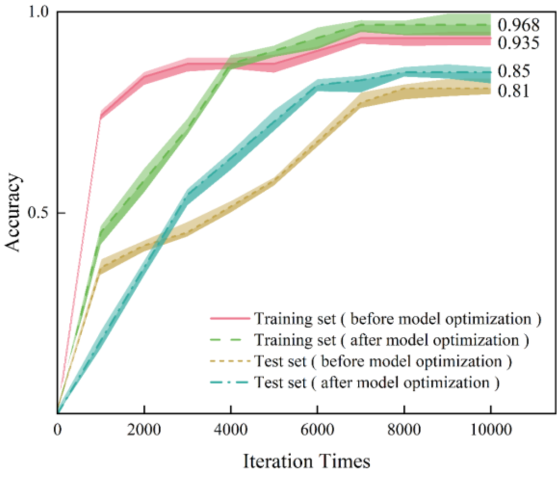

**Figure 9.** Comparison of model accuracy before and after optimization.

To further confirm the performance of our model, we have initiated an additional data collection process. This includes exhaled samples from 10 patients diagnosed with lung cancer and 10 healthy participants, maintaining a balanced gender ratio with each group consisting of 5 men and 5 women. According to the diagnostic results of the model, all the patients in the lung cancer cohort were accurately identified with the disease. However, in the healthy cohort, one participant was mistakenly diagnosed with lung cancer. Consequently, the overall diagnostic accuracy of this Confusion matrix of model diagnosis results model reached 95%. This reiterates the superior capacity and potential utility of our model in the detection of lung cancer.

## 4. Discussion

In this study, 75 lung cancer patients and 50 healthy subjects were used to explore the differences in the content of volatile organic compounds in exhaled breath. A total of 108 VOCs were identified, including 44 hydrocarbons, 12 ketones, 26 alcohols, 9 aldehydes, 8 esters, and 9 acids. It can reflect the patient's physiological condition more comprehensively. The results showed that there was a significant correlation with the disease in the four substances of ethanol, acetone, isopropanol and ethyl acetate. The diagnostic model established by the neural network showed excellent performance, with an accuracy of 96%, a sensitivity of 95.8%, and a specificity of 96.7%.

By using solid-phase microextraction (SPME) combined with GC-MS, poli measured the content of straight-chain aldehydes C3-C9 in exhaled breath. Non-small cell lung cancer (NSCLC) patients had higher levels of aldehydes compared to a control group of asymptomatic non-smokers [32]. Ligor found that eight volatile organic compounds, such as butane, 2-methylbutane, propane, and 4-methyloctane, were potential biomarkers for lung cancer by using an artificial neural network model [33]. Current research mostly focuses on a single type of substances in exhaled breath, such as alkanes and aldehydes. Although effective results can be obtained in some cases, there are also some shortcomings. Metabolic processes in living organisms are very complex, and changes in a type of substance often only reflect part of the physiological process, making it impossible to comprehensively assess disease status. This study detected hydrocarbons, ketones, alcohols, aldehydes, esters, acids, and other substances in breath. Compared with previous studies, the detection content was more comprehensive. By combining it with machine learning, significant improvements in accuracy, sensitivity, and specificity are achieved, enabling more accurate diagnosis of patients and requiring fewer markers to achieve it. This innovation promotes the promotion of respiratory diagnostic methods and provides strong support for progress in the field of early detection and diagnosis of lung cancer.

In order to capture the diversity of the patient population more comprehensively, future research plans will focus on expanding the age range and geographical scope of the subjects. This will help to establish a more widely applicable model and provide more representative research conclusions. At the same time, we plan to further increase the sample size to strengthen the statistical effectiveness of the experiment and ensure the reliability of the research results.

## 5. Conclusions

In this study, we employed Gas chromatography-mass spectrometry to quantify 108 VOCs in respiratory samples obtained from lung cancer patients. Through statistical significance level and variable importance index screening, we identified four VOCs closely associated with lung cancer: ethanol, acetone, isopropanol, and ethyl acetate. After the screening process, the accuracy of our diagnostic model increased from 81% to 93.5%. To determine the most suitable machine learning algorithm, we established a lung cancer diagnosis model based on four commonly used algorithms: neural network, decision tree, support vector machine tree, and random forest. The neural network algorithm-based diagnosis model demonstrated the highest accuracy, achieving 93.5% on the training set and 81% on the test set.

Considering the influence of gender on the content of VOCs in exhaled breath samples, we included the gender factor in the model input to mitigate its impact on the diagnosis results. Through model optimization, the accuracy of the diagnostic model further improved, reaching 96.8% on the training set and 85% on the test set. Moving forward, we will continue to investigate factors affecting the exhaled breath diagnosis of lung cancer, aiming to enhance the accuracy of the diagnostic model and provide better support for early diagnosis and treatment of lung cancer.

**Author contributions:** Study conception and design, FW, ZZ, YW and LQ; data collection, RL, YW, ZZ, HC and GL; analysis and interpretation of results, LQ; draft manuscript preparation, final approval of manuscript, LQ, YW, FW, ZZ, RL, GL and HC. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was financially supported by Zhejiang Province Natural Science Foundation Project (TGY24H160037).

**Availability of data and materials:** The data pertaining to the present study are accessible upon request, allowing interested parties to obtain the relevant information for further examination.

**Ethical approval:** The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by the Medical Ethics Committee of Sir Run Run Shaw Hospital Affiliated to Zhejiang University (Approval NO: 2023-0488; Acceptance Number: 2023-768-01). All participants agreed to participate in the study and signed informed consent.

**Conflict of interest:** The authors declare no conflict of interest.

## References

1. Li Y, Wu X, Yang P, et al. Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis. Genomics, Proteomics & Bioinformatics. 2022; 20(5): 850-866. doi: 10.1016/j.gpb.2022.11.003
2. Svoboda E. Research round-up: Allergies. Nature. 2020; 588(7836): S2-S3. doi: 10.1038/d41586-020-02776-6
3. Brunetti A, Altini N, Buongiorno D, et al. A Machine Learning and Radiomics Approach in Lung Cancer for Predicting Histological Subtype. Applied Sciences. 2022; 12(12): 5829. doi: 10.3390/app12125829
4. Tan WL, Jain A, Takano A, et al. Novel therapeutic targets on the horizon for lung cancer. Lancet Oncol. 2016; 17(8): e347-e362. doi: 10.1016/S1470-2045(16)30123-1
5. Wang Y, Chen E. Interventional bronchoscopic treatment of lung cancer. Laparoscopic, Endoscopic and Robotic Surgery. 2022; 5(2): 52-56. doi: 10.1016/j.lers.2021.09.005
6. Dhaware BU, Pise AC. Lung cancer detection using Bayasein classifier and FCM segmentation. 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT). Published online September 2016. doi: 10.1109/icacdot.2016.7877572
7. Rahouma KH, Mabrouk SM, Aouf M. Lung Cancer Diagnosis Based on Chan-Vese Active Contour and Polynomial Neural Network. Procedia Computer Science. 2021; 194: 22-31. doi: 10.1016/j.procs.2021.10.056
8. Bevilacqua V, Brunetti A, Guerriero A, et al. A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images. Cognitive Systems Research. 2019; 53: 3-19. doi: 10.1016/j.cogsys.2018.04.011
9. Song MA, Benowitz NL, Berman M, et al. Cigarette Filter Ventilation and its Relationship to Increasing Rates of Lung Adenocarcinoma. JNCI: Journal of the National Cancer Institute. 2017; 109(12). doi: 10.1093/jnci/djx075
10. Aerts HJWL. The Potential of Radiomic-Based Phenotyping in Precision Medicine. JAMA Oncology. 2016; 2(12): 1636. doi: 10.1001/jamaoncol.2016.2631
11. Cao W, Duan Y. Current Status of Methods and Techniques for Breath Analysis. Critical Reviews in Analytical Chemistry. 2007; 37(1): 3-13. doi: 10.1080/10408340600976499
12. Dandıl E. A Computer-Aided Pipeline for Automatic Lung Cancer Classification on Computed Tomography Scans. Journal

of Healthcare Engineering. 2018; 2018: 1-12. doi: 10.1155/2018/9409267

13. Kamysek S, Fuchs P, Schwoebel H, et al. Drug detection in breath: effects of pulmonary blood flow and cardiac output on propofol exhalation. Analytical and Bioanalytical Chemistry. 2011; 401(7). doi: 10.1007/s00216-011-5099-8
14. Saalberg Y, Wolff M. VOC breath biomarkers in lung cancer. Clinica Chimica Acta. 2016; 459: 5-9. doi: 10.1016/j.cca.2016.05.013
15. Fuchs P, Loeseken C, Schubert JK, et al. Breath gas aldehydes as biomarkers of lung cancer. International Journal of Cancer. 2010; 126(11): 2663-2670. doi: 10.1002/ijc.24970
16. Phillips M, Cataneo RN, Cummin ARC, et al. Detection of Lung Cancer With Volatile Markers in the Breatha. Chest. 2003; 123(6): 2115-2123. doi: 10.1378/chest.123.6.2115
17. Bousamra M, Schumer E, Li M, et al. Quantitative analysis of exhaled carbonyl compounds distinguishes benign from malignant pulmonary disease. The Journal of Thoracic and Cardiovascular Surgery. 2014; 148(3): 1074-1081. doi: 10.1016/j.jtcvs.2014.06.006
18. Ligor T, Pater Ł, Buszewski B. Application of an artificial neural network model for selection of potential lung cancer biomarkers. Journal of Breath Research. 2015; 9(2): 027106. doi: 10.1088/1752-7155/9/2/027106
19. Shehada N, Cancilla JC, Torrecilla JS, et al. Silicon Nanowire Sensors Enable Diagnosis of Patients via Exhaled Breath. ACS Nano. 2016; 10(7): 7047-7057. doi: 10.1021/acsnano.6b03127
20. Callol-Sanchez L, Munoz-Lucas MA, Gomez-Martin O, et al. Observation of nonanoic acid and aldehydes in exhaled breath of patients with lung cancer. Journal of Breath Research. 2017; 11(2): 026004. doi: 10.1088/1752-7163/aa6485
21. Wang M, Sheng J, Wu Q, et al. Confounding effect of benign pulmonary diseases in selecting volatile organic compounds as markers of lung cancer. Journal of Breath Research. 2018; 12(4): 046013. doi: 10.1088/1752-7163/aad9cc
22. Smith D, Wang T, Sulé-Suso J, et al. Quantification of acetaldehyde released by lung cancer cells in vitro using selected ion flow tube mass spectrometry. Rapid Communications in Mass Spectrometry. 2003; 17(8): 845-850. doi: 10.1002/rcm.984
23. Amann A, Spanel P, Smith D. Breath Analysis: The Approach Towards Clinical Applications. Mini-Reviews in Medicinal Chemistry. 2007; 7(2): 115-129. doi: 10.2174/138955707779802606
24. Di Natale C, Macagnano A, Martinelli E, et al. Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors. Biosens Bioelectron. 2023; 18: 1209-1218. doi: 10.1016/S0956-5663(03)00086-1
25. Yu H, Xu L, Cao M, et al. Detection volatile organic compounds in breath as markers of lung cancer using a novel electronic nose. Proceedings of IEEE Sensors 2003 (IEEE Cat No03CH37498). doi: 10.1109/icsens.2003.1279164
26. O’Neill HJ, Gordon SM, O’Neill MH, et al. A computerized classification technique for screening for the presence of breath biomarkers in lung cancer. Clinical Chemistry. 1988; 34(8): 1613-1618. doi: 10.1093/clinchem/34.8.1613
27. Phillips M, Gleeson K, Hughes JMB, et al. Volatile organic compounds in breath as markers of lung cancer: a cross-sectional study. The Lancet. 1999; 353(9168): 1930-1933. doi: 10.1016/S0140-6736(98)07552-7
28. Peng G, Tisch U, Adams O, et al. Diagnosing lung cancer in exhaled breath using gold nanoparticles. Nature Nanotechnology. 2009; 4(10): 669-673. doi: 10.1038/nnano.2009.235
29. Sorocki J, Rydosz A. A Prototype of a Portable Gas Analyzer for Exhaled Acetone Detection. Applied Sciences. 2019; 9(13): 2605. doi: 10.3390/app9132605
30. Xiao C, Ye J, Esteves RM, et al. Using Spearman’s correlation coefficients for exploratory data analysis on big dataset. Concurrency and Computation: Practice and Experience. 2015; 28(14): 3866-3878. doi: 10.1002/cpe.3745
31. Wisanwanichthan T, Thammawichai M. A Double-Layered Hybrid Approach for Network Intrusion Detection System Using Combined Naive Bayes and SVM. IEEE Access. 2021; 9: 138432-138450. doi: 10.1109/access.2021.3118573
32. Poli D, Goldoni M, Corradi M, et al. Determination of aldehydes in exhaled breath of patients with lung cancer by means of on-fiber-derivatisation SPME–GC/MS. Journal of Chromatography B. 2010; 878(27): 2643-2651. doi: 10.1016/j.jchromb.2010.01.022
33. Ligor T, Pater Ł, Buszewski B. Application of an artificial neural network model for selection of potential lung cancer biomarkers. Journal of Breath Research. 2015; 9(2): 027106. doi: 10.1088/1752-7155/9/2/027106