Analysis of biomarkers in exhaled gas using a fully connected deep neural network

Wenzhu Zhao[1], Pingping Cao[1], Yuanyang Zhu[1], Sheng Liu[1, *], Wangan Song[1], Jianyu Xiao[1], Xue Zou[2], Chengyin Shen[2, 3, *]

[1]College of Computer Science and Technology, Huaibei Normal University, Huaibei, Anhui 235000, China. [2]Key Laboratory of Medical Physics and Technology of Anhui Province, Medical Physics and Technology Center, Hefei Institute of Material Science, Chinese Academy of Sciences, Hefei, Anhui 230031, China. [3]Hefei Cancer Hospital, Chinese Academy of Sciences, Hefei, Anhui 230031, China.

Received: May 19, 2020; **accepted:** July 6, 2020.

Inhaling a complex chemical mixture of combustion compounds in tobacco smoke causes adverse health effects, particularly cancer, cardiovascular, and pulmonary diseases. Therefore, patient smoking can be used as an important basis for clinical diagnosis. In this study, a model for predicting the individual smoking behavior based on an artificial neural network and expiratory mass spectrometry was proposed. The exhalation data of 1,119 volunteers were collected using proton transfer reaction mass spectrometry. The exhalation data contained the information on m/z 20-150. A logic regression algorithm was used to analyze the signal intensity data corresponding to the m/z in each column. The results showed that m/z with a greater correlation with smoking were 42, 67, 83, and 97. A nonlinear relationship model between the characteristic data and smoking behavior was constructed by using a fully connected neural network. Thus, the corresponding smoking behavior prediction model was developed. The model was then trained and tested many times, and the final smoking behavior prediction model with an accuracy of 95.089% was established. Comparing to the traditional questionnaire survey and conventional logistic regression methods, the proposed method has higher reliability, shorter detection period, lower detection cost, and higher accuracy. This method can play an important role in disease prevention, medical diagnostics, and evidence collection.

Keywords: smoking prediction; expiratory mass spectrometry; logic regression; fully connected neural network; medical diagnosis.

Abbreviations: fully connected (FC), back propagation (BP), mass flow controller (MFC), proton transfer reaction mass spectrometry (PTR-MS), volatile organic compounds (VOCs), receiver operating characteristic (ROC).

[*]**Corresponding authors:** Sheng Liu, College of Computer Science and Technology, Huaibei Normal University, Huaibei, Anhui 235000, China. Phone: +86 183 6523 9378. E-mail: liurise@139.com. Chengyin Shen, Key Laboratory of Medical Physics and Technology of Anhui Province, Medical Physics and Technology Center, Hefei Institute of Material Science, Chinese Academy of Sciences, Hefei, Anhui 230031, China. Phone: +86 189 1963 7939. E-mail: chyshen@aiofm.ac.cn.

Introduction

Research shows that smoking seriously endangers all the organs of the human body, causes many diseases, affects the overall health of a person, and reduces life expectancy [1]. Smoking not only leads to various cardiovascular [2] and respiratory diseases [3], but also causes sexual dysfunction and fertility reduction [4]. Smoking can also cause cancer and prevent the body from resisting cancer [5]. In addition, smoking affects the health of bones and teeth [1], leading to series of problems such as inflammation [6] and decreased immune function [7]. The harm caused by smoking is well known to people and has aroused broad concerns globally.

The diseases caused by using tobacco are among the largest global public health threats. Therefore, many health departments over the world have mandated smoking cessation to help smoking addicts quit smoking. In the process of quitting smoking, frequently determination of the smoking status and monitoring the effects of smoking cessation are required. In addition, the medical insurance industry needs to determine whether an insured's illness has been caused by smoking. When a criminal case occurs, cigarette or second-hand cigarette residues and other relevant evidence are often left behind at the scene of the crime. Forensic laboratories need to make an objective determination of smoking and determine the person at the scene of the crime, so as to identify the suspect [8].

The most commonly used method for smoking determination is a questionnaire survey [9], in which reliability often depends on the attitude of the respondents and the spirit of seeking the truth from the facts. The more accurate method to identify whether the tester smokes is to measure nicotine and cotinine levels in the plasma, urine, and saliva of the tester [10, 11]. The concentration of nicotine and cotinine in the biological samples (such as urine, blood, and saliva) of a smoker is at least 2-3 times higher than that in a nonsmoker. Therefore, it has been proved that it is reasonable to use such biological samples as a biomarker to determine smoking [12]. Specific analytical methods include immunoassay [13], high-performance liquid chromatography [14], and gas chromatography [11]. However, these analytical methods are time-consuming and require a large number of samples, specialized laboratories, and trained personnel [8]. In addition, there was a cross reaction and a lack of specificity in immunoassay [15]. In recent years, researchers have attempted to analyze abnormal methylation of smoking-related genes in order to identify smoking behavior [16]. Deoxyribonucleic acid methylation is a potential mechanistic link between current smoking and cancer as well as prenatal cigarette-smoke exposure and the development of adult chronic diseases. The genetic analysis of gene methylation may provide some clues for smoking identification. However, there are too many factors that may cause gene methylation, and smoking is not the only one or even one of the top causes [17]. Therefore, genome sequencing for methylation evaluation cannot be the method for screening the smoking situation.

Breath detection has become a new method for identifying diseases or the metabolic state due to its convenience, speed, hygiene, and non-invasiveness, and has attracted a lot of interest from international researchers. Although acetonitrile exhaled by smokers has been recognized internationally as one of the smoking markers [18], complete and accurate determination of whether an individual smokes or not is difficult by relying solely on acetonitrile. Mass spectrometry data, such as breath detection, also contains a large amount of information on components other than acetonitrile. If a comprehensive analysis is conducted, it is expected to improve the accuracy of smoking determination further. A fully connected (FC) neural network is suitable for handling complex nonlinear relationship data and can automatically learn and extract feature data from large amounts of data.

In recent years, neural networks have quickly become a popular topic for disease analysis and prediction. For example, in medical imaging, neural networks are used to analyze images of breast cell nuclear grading [19], and in disease prediction, a back propagation (BP) neural network was used to predict the trend of the Severe Acute Respiratory Syndrome epidemic [20]. In terms of air quality prediction, a long short-term memory-FC neural network model was used to predict the PM2.5 pollution index at an air-quality monitoring station, and the prediction performance of this model was great [21].

In this study, we applied the breath component

Table 1. Characteristics of the subject groups (Age is quoted as a median).

		Train			Test			
		Smokers	Nonsmokers	Ex-smokers	Smokers	Nonsmokers	Ex-smokers	Total
Male	Age	35(21-80)	30(23-87)	31(56-62)	32.5(22-57)	27(19-63)	41	33(19-87)
	No.	176	445	18	40	120	1	800
Women	Age	65(39-66)	35(23-82)	38.5(33-44)	27	27(20-62)	27.5(27-28)	34(20-82)
	No.	3	251	2	1	60	2	319
All	Age	35(21-80)	35(23-87)	41.5(24-79)	32(22-57)	27(19-63)	28(27-41)	33(19-87)
	No.	179	696	20	41	180	3	1119

mass spectrometry data obtained through breath detection to establish a fast and accurate smoking prediction model based on the FC neural network by processing the input eigenvalues of the nonlinear relationship to predict smoking behavior. The result of this study is expected to be a new reliable non-invasive method for predicting smoking.

Materials and Methods

In this study, breath data were collected from 1,119 volunteers. 80% of breath data (895 volunteers) were used to train the model, and 20% of breath data (224 volunteers) were used as the prediction samples. The personal data and smoking status of the volunteers were recorded including age, gender, whether the volunteer smokes, smoking duration, time since quitting, and the number of cigarettes smoked per day. The specific volunteer information is shown in Table 1.

The framework for predicting the smoking behavior by combining expiratory mass spectrometry with the FC neural network is shown in Figure 1.

Acquisition of exhaled mass spectrometry data

Breath detection was performed by using direct exhalation injection (breath sampling system) combined with proton transfer reaction mass spectrometry (PTR-MS) [22, 23]. The breath sampling system mainly included a disposable mouthpiece, a mass flow controller (MFC), a pressure controller, and two pumps (Pump1 and Pump2). As PTR-MS operates in the full-spectrum scanning mode (*m/z* 20-150, except *m/z* 37), one full scan takes 143 s, which is much longer than the breath duration of human being. Therefore, to ensure that the full-spectrum scanning was completed with only one breath, the following operations were performed on the sampling system. In the beginning of the experiment, the MFC was set to 500 mL/min. The experimental subject began to blow for 7 s, and the exhaled samples were brought into the sampling pipeline by Pump1. At this time, the MFC was closed, and blowing was stopped. Then, the breath sample stored in the sampling pipeline entered the PTR-MS detector slowly under the action of the pressure gradient. By the end of this testing, the whole process time lasted more than 200 s, which met the requirement of full-spectrum scanning. PTR-MS devices included an ion source, a drift tube, and a mass spectrometry detection system [24, 25]. Water vapor entered the ion source and generated H_3O^+ by glow discharge. When the exhaled gas entered the drift tube, if the proton affinity of the volatile organic substance M in the breath was greater than that of water, a proton transfer reaction as Equation (1) occurred with the generated H_3O^+ and MH^+, which entered the mass spectrometry, and the detection was completed.

$$\mathrm{M}^+ + \mathrm{H_3O}^+ \rightarrow \mathrm{MH}^+ + \mathrm{H_2O} \qquad (1)$$

There are many volatile organic compounds (VOCs) in the exhaled gas of people. The target VOCs include acetonitrile, benzene, toluene, and 2-methylpentane. Previous studies have shown that the concentrations of acetonitrile, benzene,

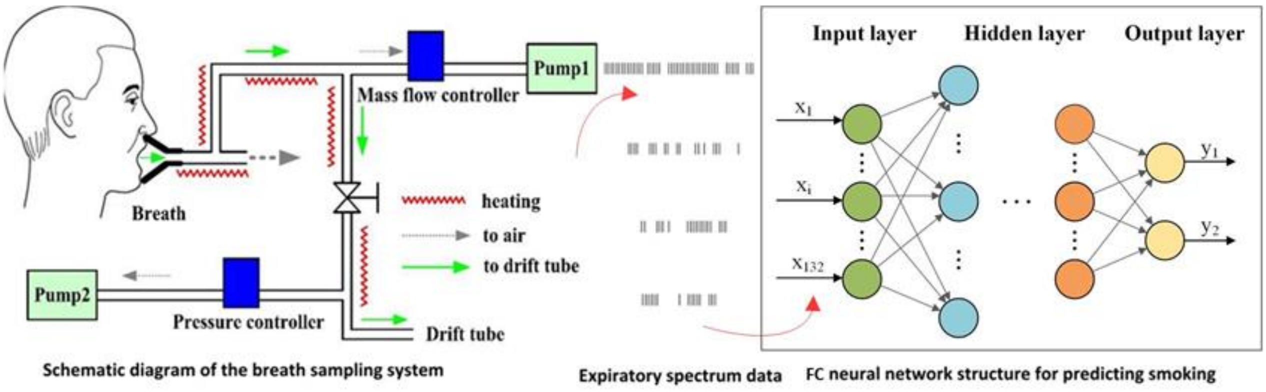

Figure 1. Framework for predicting the smoking behavior by combining exhalation mass spectrometry with a fully connected neural network.

and other target VOCs in the exhaled air of smokers are significantly higher than that of nonsmokers. Thus, these VOCs can be used as biomarkers of smoking breath [18].

Statistical analysis

Advanced statistical analysis including logistic regression classification was applied to select important features to be included in the training of the FC neural network models. A statistical significance test was conducted for the difference between nonsmokers and smokers by using Statistical Product and Service Solutions (SPSS) software (IBM Company, Armonk, New York, USA) through the Mann–Whitney U test.

Data Processing

The number of protons is represented by m, and the number of charges of ions is represented by z. The m/z is the ratio of protons divided by the number of charges, i.e., the charge-to-mass ratio. The horizontal axis in the mass spectrum is m/z. The m/z range of the scan was from 20 to 150, and the corresponding signal intensity was obtained.

According to the specification of the neural network input data, the processed data were numerically processed. The volunteers only had two conditions: smoking or nonsmoking. The smoking and nonsmoking behaviors were indicated by 1 and 0, respectively. The numeric ones and zeros were used as learning labels for the neural networks.

To ensure the prediction accuracy and avoid too long learning time of the neural network, data cleaning was required. First, the logistic regression method was used to calculate the parameters between each column of the data and smoking. Then, the correlation of the data was determined according to the size of the parameters, and the optimal feature dataset was finally obtained after cleaning the interference data.

The classification of logistic regression is fast with less occupied memory, which is suitable for solving two-class problems. The parameters of each feature value can be directly obtained. The larger parameters indicate that the correlation between the feature data and the target classification is higher. A total of 80% of the exhaled m/z data collected from 1,119 volunteers were used as the training samples for the logistic regression model, and the other 20% were used to test the classification accuracy of the logistic regression model. The study was conducted to determine whether the volunteer smokes; hence, the problem was a two-class problem with smoking or nonsmoking.

For a single input $x^{(i)}$, there are:

$$z^{(i)} = w^{T}x^{(i)} + b \tag{2}$$

A predictive function can be obtained:

$$\hat{y}^{(i)} = a^{(i)} = \text{sigmoid}\left(z^{(i)}\right) \quad (3)$$

The corresponding loss function is:

$$L\left(a^{(i)}, y^{(i)}\right) = -y^{(i)} \log\left(a^{(i)}\right) - \left(1 - y^{(i)}\right) \log\left(1 - a^{(i)}\right) \quad (4)$$

where $y^{(i)}$ is the true output corresponding to the input $x^{(i)}$.

By taking the training sample of (5), the cost function is obtained:

$$J = \frac{1}{m} \sum_{i=1}^{m} L(a^{(i)}, y^{(i)}) \quad (5)$$

The minimum cost function was optimized by using the traditional gradient descent method, and the best parameters W and b were obtained. First, forward propagation and BP were performed. Then, the parameter was updated according to the partial derivative of each parameter obtained by BP.

In forward propagation, for input X, the predicted value of logistic regression is:

$$A = \alpha(w^{T}X + b) = (a^{(1)}, a^{(2)}, \cdots, a^{(m-1)}, a^{(m)}) \quad (6)$$

The cost function can be obtained from the training data and the predicted values:

$$J = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log\left(a^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - a^{(i)}\right) \quad (7)$$

In BP:

$$dW = \frac{\partial J}{\partial W} = \frac{1}{m} X(A - Y)^{T} \quad (8)$$

$$db = \frac{\partial J}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} (a^{(i)} - y^{(i)}) \quad (9)$$

The update parameters are:

$$W = W - \alpha * dW \quad (10)$$

$$b = b - \alpha * db \ (\alpha \text{ was the learning rate}) \ (11)$$

BP neural network

Because the exhalation data are one-dimensional, traditional network models such as recurrent neural network and convolutional neural network cannot process the data well. Therefore, the traditional BP neural network was used to construct the smoking prediction model. The BP neural network classification prediction model is a supervised learning network, which is generally divided into an input layer, a hidden layer, and an output layer. Training and test data were input through the input layer, calculated by the hidden layer, and then the result was output by the output layer. The hidden layer is the core of the entire neural network. The algorithm of the hidden layer determines the role of the neural network.

The maximum training times were 1,000, and the learning rate was 0.1. The error was set to 0.001. The classification result was only 0 or 1. When the output was smaller than 0.5 and greater than or equal to 0.5, it was rounded to 0 and 1, respectively. Therefore, the error range of 0.001 did not affect the output. In learning the neural network model, the data corresponding to high weight m/z in exhaled gas were input from the input layer, and then a set of free weights was randomly selected as the initial weight between the input layer and hidden layer. The input of the hidden layer was obtained by using the initial input and initial weight:

$$X_{\text{hidden}} = I_{\text{input}} \cdot W_{\text{input-hidden}} \quad (12)$$

Where W_{hidden} is the input of the hidden layer, I_{input} is the feature data from the input layer, and $W_{\text{input-hidden}}$ is the weight between the input layer and hidden layer, which is updated with the weight of the hidden layer.

The updated weights calculated by the Levenberg–Marquardt algorithm [26-28] were returned to equation (12), and the new X_{hidden} was calculated. Then, the actual output of the hidden layer was calculated by the S activation function. The Sigmoid function is expressed as follows [29]:

$$\mathrm{Sigmoid(x)} = \frac{1}{1+e^{-x}} \quad (13)$$

$$\mathrm{O_{hidden}} = \mathrm{Sigmoid(X_{hidden})} \quad (14)$$

$\mathrm{O_{hidden}}$ is the output value of the hidden layer, and the predicted classification results of the output layer are calculated as follows:

$$\mathrm{O_{output}} = \mathrm{O_{hidden}} \cdot \mathrm{W_{hidden}} \quad (15)$$

Where $\mathrm{W_{hidden}}$ is the updated hidden layer weight.

Fully connected neural network

In order to improve the accuracy and precision of the model further, a smoking prediction model was constructed by using an FC neural network. In FC neural networks, adjacent network layers are FC to each other [30]. The neural network implemented in this research model has a three-layer structure with an input layer, a hidden layer, and an output layer. The hidden layer extracts features and outputs the corresponding probability of smoking and nonsmoking from the output layer. The higher probability is taken as the final prediction output.

In order to enhance the model ability to describe nonlinear features and accelerate training, the rectified linear unit (ReLU) function was used as the activation function instead of the common Sigmoid function. It has been reported that the commonly used sigmoidal element has the problem of vanishing gradient, which is usually accompanied by slow optimization convergence to a local minimum of difference [31]. The modified linear (ReLU) element solves this problem. When it is activated above 0, its partial derivative is 1. The ReLU function is defined as follows [32]:

$$\mathrm{Re}LU(x) = \max(x,0) = \begin{cases} x & if\ x > 0 \\ 0 & otherwise \end{cases} \quad (16)$$

As a one-sided piecewise linear function, the ReLU function changes all negative values to 0, whereas positive values remain unchanged. It does not force sign symmetry or antisymmetry. This operation is called one-sided suppression. Because of its unilateral inhibition, the neurons in the neural network also have sparse activation. The sparse model implemented by ReLU can better mine related features, improve training speed, and enhance fitting accuracy. When the Sigmoid function is used to calculate the activation value, the exponential calculation is needed. When the BP is used to calculate the error gradient, the derivation involves division, and the calculation amount is relatively large. As for the ReLU function, for accelerating the propagation, only the threshold value should be set.

The weight of the hidden layer needs to be constantly updated to ensure the accuracy of the output. However, neural networks using traditional gradient descent algorithms often have the problem of local minimum. When the learning rate is too small or the activation function is very small, the update speed of weights and thresholds will decrease, and the convergence rate of the network will slow down. Thus, in the FC neural network model for predicting smoking, we used the Adam algorithm to calculate the weight of the hidden and output layers, rather than the traditional gradient descent algorithm.

The Adam algorithm is a first-order optimization algorithm, which can replace the traditional random gradient descent process. It can update the neural network weight based on the training data iteratively [33]. The Adam algorithm designs independent adaptive learning rates for different parameters by calculating the first and second moment estimates of the gradient. The Adam algorithm combines the best features of the AdaGrad and RMSProp algorithms. It still provides optimization methods to solve sparse gradient and noise problems. Adam's algorithm tuning parameters are relatively simple, and the default parameters can handle most problems. It solves the problems of low convergence speed of traditional all connected neural networks and falling into the local minimum. The calculation formulas are as follows:

$$t = t + 1 \tag{17}$$

$$g_t = \nabla_\theta f_t(\theta_{t-1}) \tag{18}$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{19}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^{\,2} \tag{20}$$

where β_1, β_2 are the exponential decay rates for the moment estimates.

As m_t and v_t are initialized as 0 vectors, they will bias to 0 vectors if the attenuation rate is small. Therefore, bias correction is needed for m_t and v_t. The average weight of the corrected gradient is $\hat{m}_t$, and the deviation of the corrected gradient is $\hat{v}_t$. The correction formulas are as follows:

$$\hat{m}_t = m_t / (1 - \beta_1^{\,t}) \tag{21}$$

$$\hat{v}_t = v_t / (1 - \beta_2^{\,t}) \tag{22}$$

After the correction, the updated formula of the final Adam algorithm is as follow:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\varepsilon + \sqrt{\hat{v}_t}} \hat{m}_t \tag{23}$$

The good default settings for the tested machine learning problems are $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. All operations on vectors are elementwise.

Results

Mann–Whitney U test

There was no significant difference in the mean values of the two independent populations from which H_0 was assumed as 0 in the Mann–Whitney U test of the two independent samples. After the data were processed by SPSS, the p value was smaller than the significance level of 0.05 ($P < 0.05$). The results showed that there was a significant difference between the expiratory data of nonsmokers and smokers.

Logistic regression classification

Considering only the relationship between smoking and each m/z, the data after the preliminary coarse screening included the smoking status of the volunteers and the intensity corresponding to each m/z. Figure 2 shows the typical exhalation mass spectrometry information for the volunteers, including all mass spectral information for the m/z 20-150 (except 37) range.

The model parameters of smoking or not corresponding to m/z are shown in Figure 3. The model parameters returned by the logistic regression of m/z 38, 39, 40, 41, 42, 45, 55, 57, 67, 68, 69, 83, and 97 were relatively large and were incorporated into the logistic regression model for prediction and classification. The data columns corresponding to m/z 38, 39, 40, 41, 45, 55, and 68 had only one type of predicted result, which could not be correctly classified. Therefore, they were determined as irrelevant eigenvalues. The prediction results of m/z 57 and 69 had lower accuracy and could not be used as eigenvalues. After eliminating the interference data, the data columns corresponding to m/z 42, 67, 83, and 97 were determined as eigenvalues. The results of each feature column are shown in Figure 4.

The exhaled gas of the volunteers was analyzed by using PTR-MS. The information on the PTR-MS mass spectra between smokers and nonsmokers was obtained. The data of the signal intensity corresponding to m/z of each column was analyzed by using the logistic regression algorithm. It was found that the charge mass ratios with a high correlation to smoking were 42, 67, 83, and 97. According to the principle of proton transfer reaction, it is speculated that ions at m/z 42, 67, 83, and 97 may be acetonitrile, isoprene, difluoroethanol, and fluorobenzene, respectively. Other VOCs in the breath and their relationship to smoking should be studied

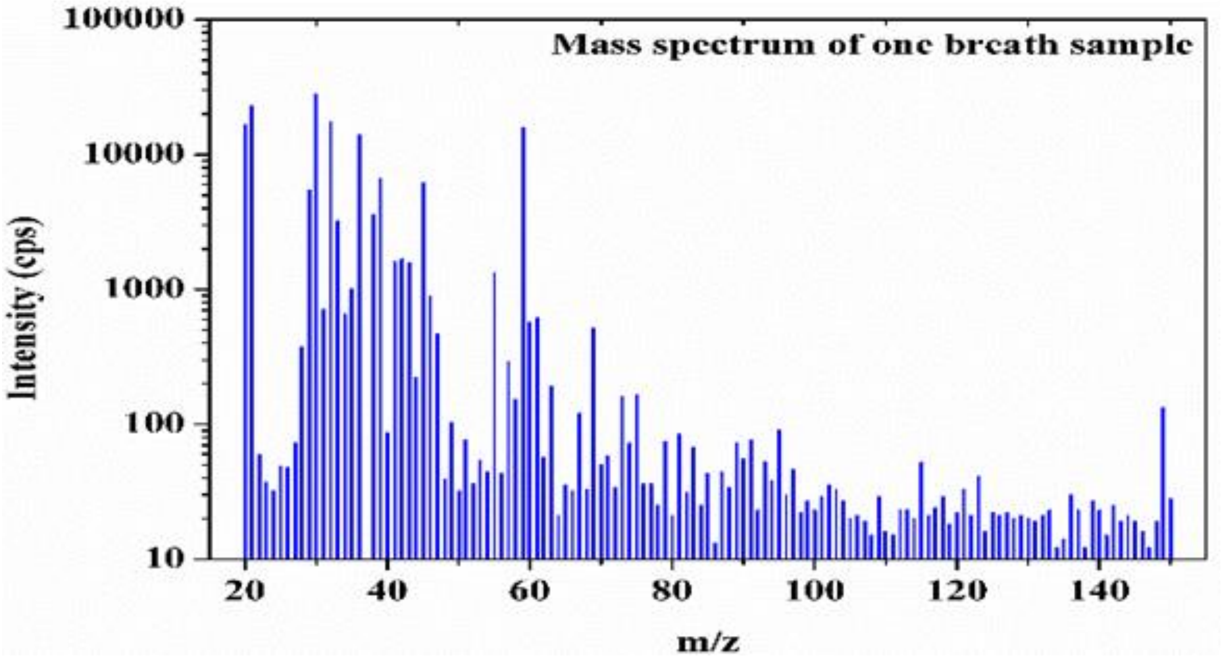

Figure 2. Exhalation spectrum.

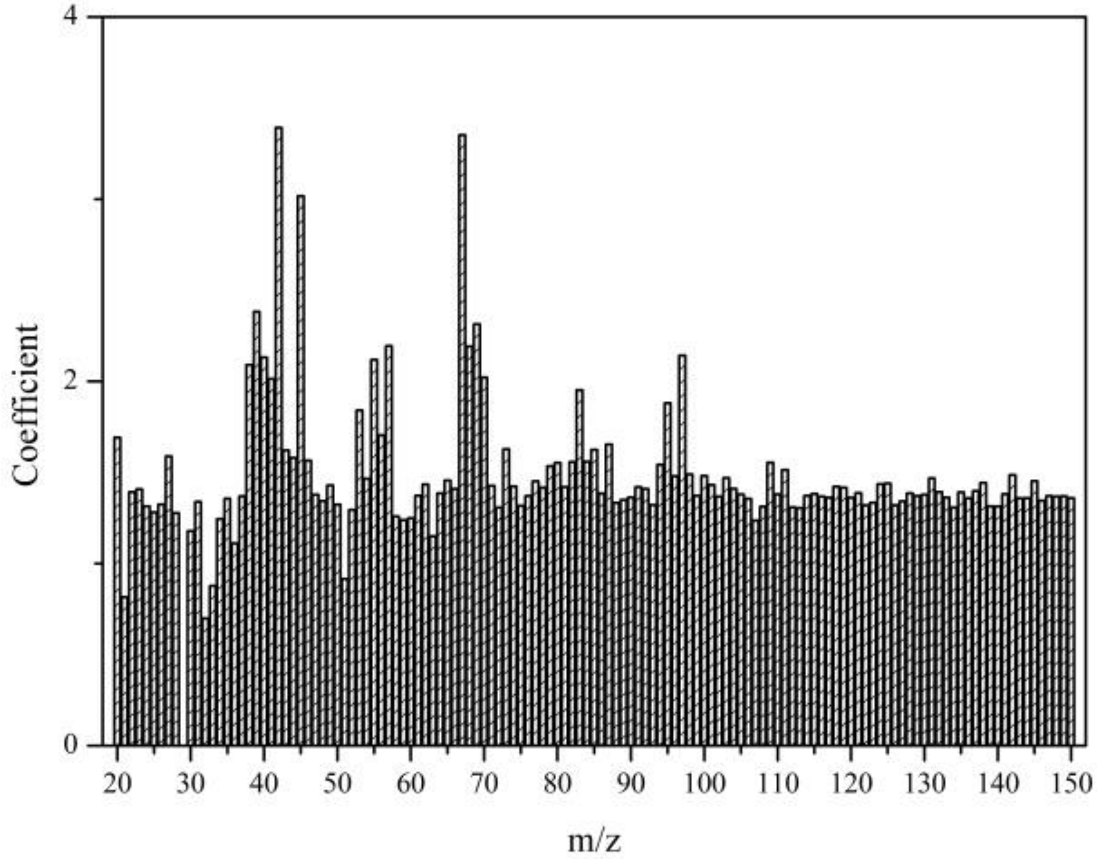

Figure 3. m/z from 20 to 150 corresponding to the parameter of smoking.

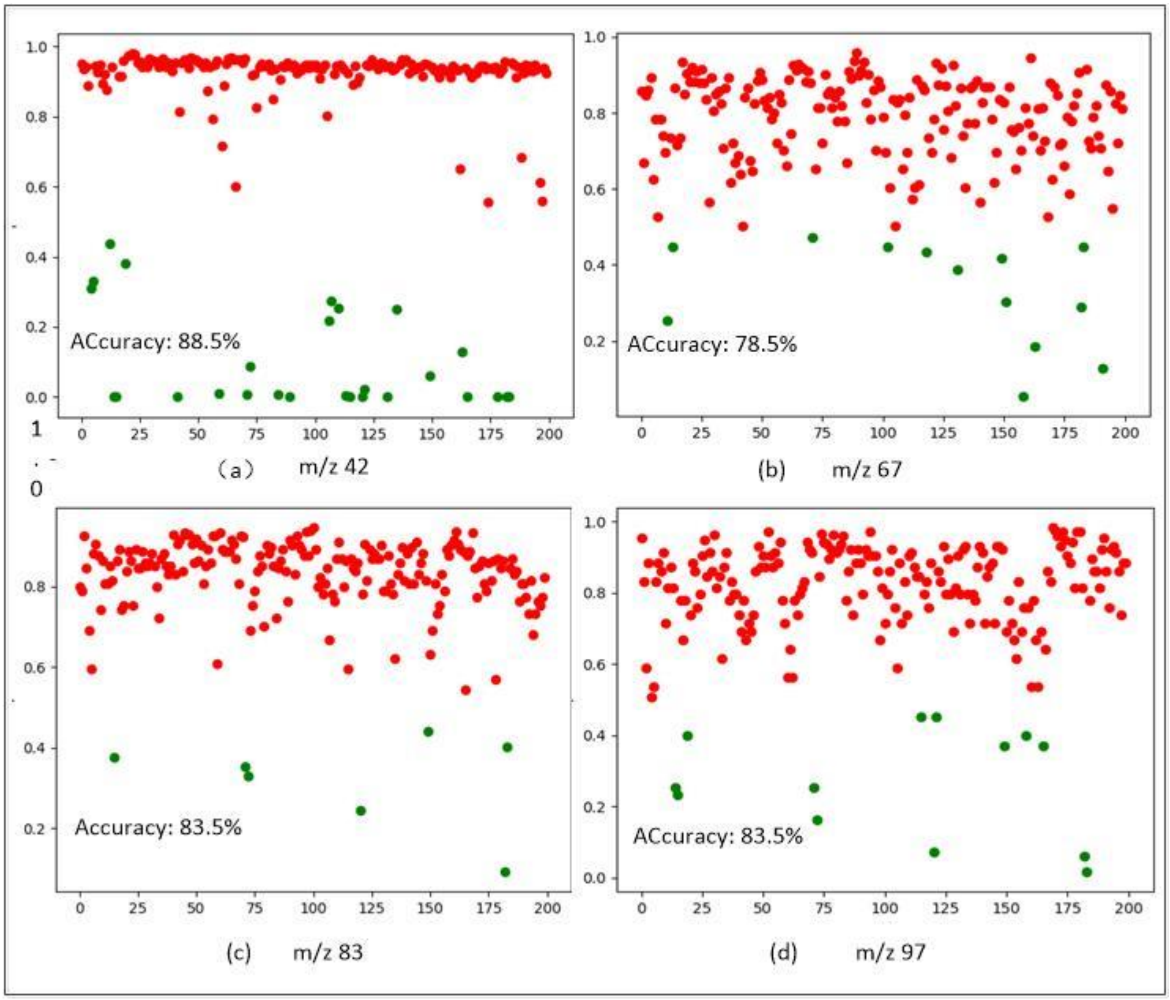

Figure 4. Logical regression prediction results. **(a)** Accuracy of the m/z 42 prediction was 88.5%. **(b)** Accuracy of the m/z 67 prediction was 78.5%. **(c)** Accuracy of the m/z 83 prediction was 83.5%. **(d)** Accuracy of the m/z 97 prediction was 83.5%.

further.

Training process of the BP and FC neural networks

Classification prediction models using the traditional BP and FC neural networks were established by using the Python programming language under the PyTorch framework. The data columns of 39, 42, 45, and 67 in the dataset were input as eigenvalues to the input layer. Of the 1,119 instances, 80% were used as the training set for the FC neural network, while the other 20% were used to test the prediction accuracy of the neural network. The FC neural network model included both negative and positive controls in training and testing. A total of 199 groups of positive samples and 696 groups of negative samples were used in the training, and 44 groups of positive samples and 180 groups of negative samples were used in the testing. The neural network attempted to identify the best network in every training process and learning set, and randomly chose different initial weights and thresholds. Thus, the prediction results were different each time. After 20 predictions of the test data, the minimum accuracy rate of the BP neural network prediction model was 90.179% twice, 91.071% 5 times, 91.518% thrice, 91.964% 7 times, and the optimal accuracy rate of 92.411% thrice. The FC neural network prediction model had a minimum accuracy rate of 93.304% twice, 93.75% thrice, 94.196% 7 times, 94.643%

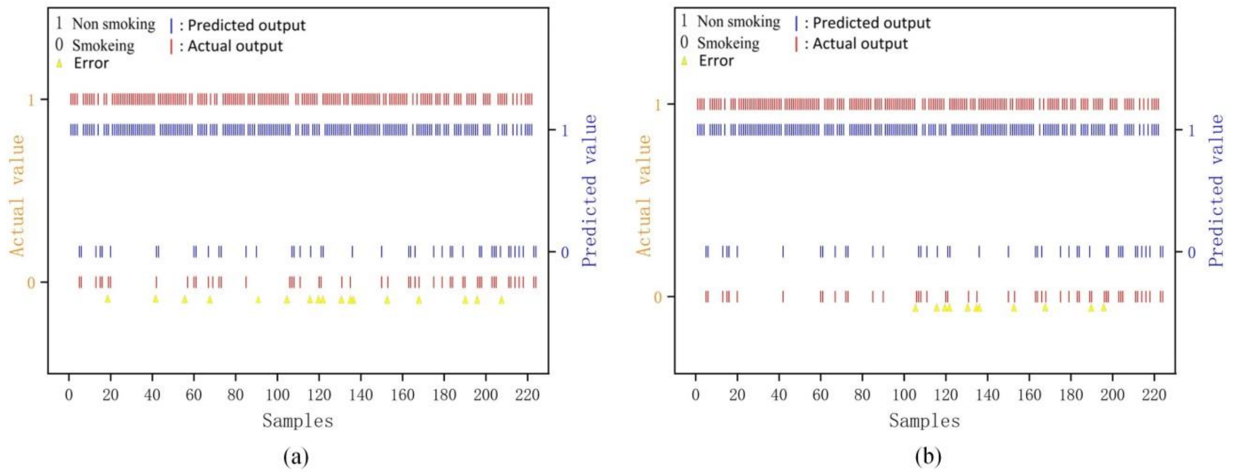

Figure 5. Predicted results of 895 sets of the training data to 224 sets of the test data. **(a)** Back propagation neural network prediction model classification results and the expected output results, with 17 incorrect predictions. **(b)** Fully connected neural network prediction model classification results and the expected output results, with 11 incorrect predictions.

Table 2. Logical Regression and Connected Neural Network classification results.

	Logistic regression algorithm				**Back propagation neural network**	**Fully connected neural network**
	m/z 42	*m/z* 67	*m/z* 83	*m/z* 97		
Accuracy	88.5%	78.5%	83.5%	83.5%	92.411%	95.089%

6 times, and the optimal accuracy rate of 95.089% appeared twice.

Prediction classification results of the BP and FC neural networks

The exhalation mass spectrometry data of 224 volunteers were input into the trained neural network for prediction. The predicted classification results and the expected output results are shown in Figure 5. The accuracies of the BP and FC neural network prediction models were 92.411% and 95.089%, respectively.

By comparing the classification results and the prediction accuracies of the models constructed using the traditional BP and FC neural networks, the latter had higher accuracy and prediction precision. The FC neural network model with an optimal accuracy rate of 95.089% was saved as the final model for smoking prediction. At the 21st training, the minimum variance was 0.029946, minimum gradient was 0.000599, and correlation between the final learning data and the target was 0.83425, which is a high value.

Receiver operating characteristic (ROC) curve

In this study, an FC neural network model combined with expiratory mass spectrometry was used to predict the smoking behavior, and the best calculated sensitivity, specificity, and accuracy were 72.5%, 97.5%, and 95.089%, respectively.

Comparison of Logical Regression and Connected Neural Network

Compared to the logistic regression classification, the classification accuracy of the smoking behavior was improved by using the FC neural network combined with four columns of characteristic values obtained using expiratory mass spectrometry, which was superior to the conventional two-class method. The comparison results were shown in Table 2.

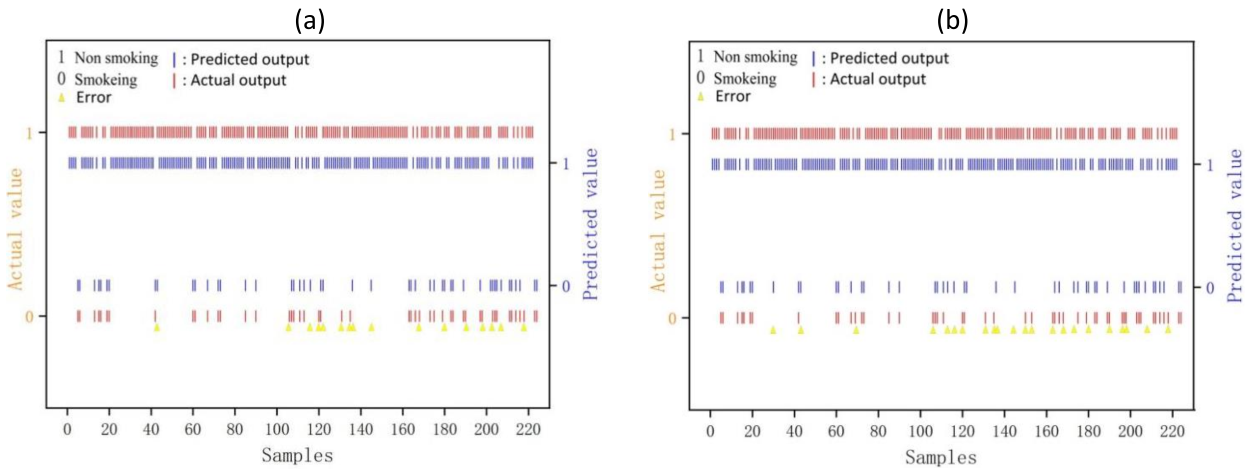

Figure 6. Predicted results after 20 predictions. **(a)** 671 sets of the training data to 224 sets of the test data. The worst accuracy of the prediction results was 90.179% thrice, 90.625% twice, 91.071% twice, 91.964% 5 times, and 92.411% thrice, and the optimal accuracy rate of 92.857% thrice. The prediction results after preserving the neural network training model had an accuracy of 92.857% with 16 erroneous predictions. **(b)** 448 sets of the training data to 224 sets of the test data. The worst accuracy of the prediction results was 86.607% once, 87.054% twice, 87.5% once, 87.946% once, 88.839% 5 times, and 89.286% twice, and the optimal accuracy rate of 90.179% appeared 8 times. The prediction results after preserving the neural network training model having an accuracy of 90.179% with 22 erroneous predictions.

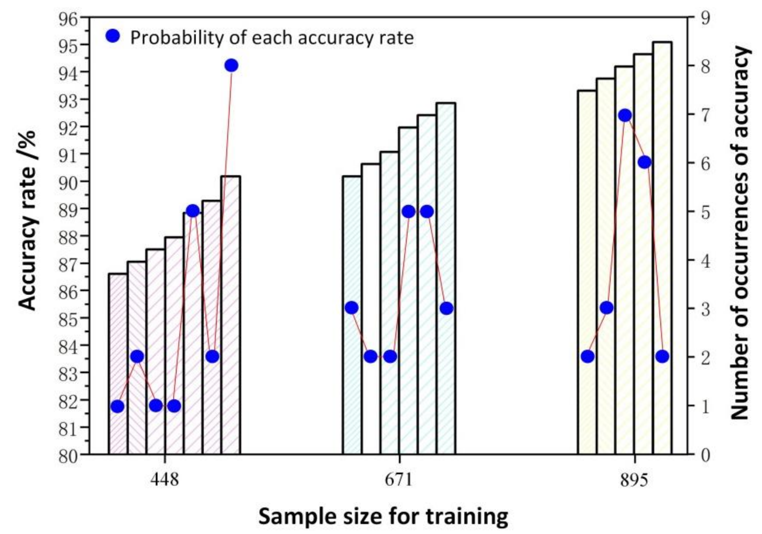

Figure 7. Bar height corresponds to the prediction accuracy of each training dataset, and the blue ellipse corresponds to the number of occurrences of each accuracy. The prediction accuracy of 895 sets of the training data for 224 sets of the test data was the best in terms of both accuracy and optimal accuracy.

Prediction results

To study the influence of the sample size on the accuracy of the prediction results, the first 60% as the first 671 group and the second 40% as the second 448 group of the training data were selected and input into the neural network for training. The trained neural network was used to predict the last 224 groups of the expiratory mass spectrometry data, and the predicted results are shown in Figures 6. The smoking prediction model trained using different training sets was developed, and the prediction accuracy and the corresponding probability of each group of the training data are shown in Figure 7. The training processes also had different performances as shown in Table 3.

Table 3. Comparison of different numbers of the data training processes.

Training sample size	Mean-square error	Gradient	Data dependency
895	0.029946	0.00059599	0.83425
671	0.076588	0.00000697	0.82672
448	0.093298	0.000001992	0.80718

Discussion

The questionnaire method can be used to determine whether an individual smokes, and the results are easy to quantify and analyze statistically. However, survey results are vulnerable to the subjective will of the respondents. The authenticity is not guaranteed, and the accuracy is low. Although the accuracy of the blood test and lung texture analysis in determining the smoking behavior is guaranteed, the detection cycle is long, and the patient bodies will also be harmed. The use of genetic methods to detect smoking requires multiple investigators to conduct multiple experiments and comparisons. The individual smoking behavior is finally discovered by determining the mutation of the smoking-related genes. The accuracy of this method is very high, but the detection is time-consuming, expensive, and labor-intensive.

In the previous study [34], the Mann–Whitney U test was used to test the acetonitrile content in smokers and nonsmokers, and the cut-off point of the acetonitrile concentration was determined by the SPSS data analysis with an receiver operating characteristic (ROC) curve drawn to determine the sensitivity of 79% and the specificity of 91% for predicting smoking. By combining the sensitivity and specificity, the accuracy of this reported method could be estimated at approximately 88.8%. In another study, the contents of cyanide and acetonitrile in exhaled gases of 370 volunteers were classified as markers of smokers and nonsmokers [35]. The ROC characteristic curves were obtained by using SPSS software. The best sensitivity and specificity for distinguishing smokers from nonsmokers were 73.7% and 95.5%, respectively, and the calculated accuracy was approximately 90.7%. In this study, the best calculated sensitivity, specificity, and accuracy for smoking behavior prediction were 72.5%, 97.5%, and 95.089%, respectively. The comparation of different methods showed that there were differences on samples collection and used datasets. Therefore, the accuracy was the only reference, which confirmed that predicting the smoking behavior based on the FC neural network was a feasible method.

Logistic regression is a conventional classification method that is used to handle the regression problem with dependent variables as classified variables. Moreover, it is often used to handle two-class or binomial distribution problems and can also handle multiclassification problems. Therefore, Logical regression was used to classify and predict the optimal dataset in this study.

The smoking prediction model trained using different training sets was investigated in this study. By comparison of the prediction results of different training samples, it was observed that the smoking prediction model became increasingly reliable with an increase in training samples. From the point of view of the training process, the more the training samples, the smaller the minimum mean variance, the larger the gradient, and the stronger the data correlation. From the training results, it was confirmed that the more the training samples, the higher and more stable the prediction accuracy. Therefore, increasing the number of the training samples can improve the accuracy and stability of the smoking prediction models.

Conclusion

In this study, expiratory mass spectrometry data were combined with an FC neural network for predicting smoking behavior. Characteristic values with a high correlation with smoking were obtained through parameter analysis. They were used to classify and predict smoking. The signal intensities corresponding to m/z 42, 67, 83, and 97 were used as the input of the neural network, and the label column data were used as the output value of the training model. The reliability of the smoking prediction model designed in this study was determined according to the performance of the training process of the model, prediction results, and prediction accuracy. The FC neural network combined with expiratory mass spectrometry has the advantages of being non-invasive, fast, and accurate in predicting smoking behavior. Comparing to social smoking surveys, our model is less laborious, has shorter investigation duration, and has greater investigation scope. Concerning disease prevention, smokers can be alerted, and then the diseases caused by smoking can be reduced. The prediction methods proposed in this study may also have potential applications in exhalation-related disease prediction and drug abuse detection, which should be studied further. However, the instrument for collecting breath data by proposed method is relatively bulky, so that the miniaturized instrument can be developed in the future to expand its application field.

Acknowledgement

This paper was funded in part by the National Natural Science Foundation of China (No. 61671434), the National Key R&D Program of China (2016YFC0200200) and the National Natural Science Foundation of China (Nos. 21705152, 21477132 and 81401756)

References

1. Warren GW, Alberg AJ, Kraft AS, Cummings KM. 2014. The 2014 Surgeon General's report: "The health consequences of smoking--50 years of progress": a paradigm shift in cancer care. Cancer. 120(13):1914-1916.
2. Rigotti NA, McDermott MM. 2019. Smoking Cessation and Cardiovascular Disease: It's Never Too Early or Too Late for Action. J Am Coll Cardiol. 74(4):508-511.
3. Erbas B, Knudsen TM, Janson C, *et al*. 2018. Critical age windows in the impact of lifetime smoking exposure on respiratory symptoms and disease among ever smokers. Environ Res. 164:241-247.
4. Centers for Disease Control and Prevention (US); National Center for Chronic Disease Prevention and Health Promotion (US); Office on Smoking and Health (US). 2010. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. Atlanta (GA): Centers for Disease Control and Prevention (US).
5. Park S, Jee SH, Shin H, *et al*. 2014. Attributable fraction of tobacco smoking on cancer using population-based nationwide cancer incidence and mortality data in Korea. BMC Cancer. 14:406.
6. Nikota JK, Stämpfli MR. 2012. Cigarette smoke-induced inflammation and respiratory host defense: Insights from animal models. Pulm Pharmacol Ther. 25(4):257-262.
7. Han S, Jerome JA, Gregory AD, Mallampalli RK. 2017. Cigarette smoke destabilizes NLRP3 protein by promoting its ubiquitination. Respir Res. 18(1):2.
8. Al-Hetlani E, Halámková L, Amin MO, Lednev IK. 2020. Differentiating smokers and nonsmokers based on Raman spectroscopy of oral fluid and advanced statistics for forensic applications. J. Biophotonics. 13:e201960123.
9. Suarez L, Barrett-Connor E. 1984. Interaction between cigarette smoking and diabetes mellitus in the prediction of death attributed to cardiovascular disease. American Journal of Epidemiology. 120(5):670.
10. Massadeh AM, Gharaibeh AA, Omari KW. 2009. A single-step extraction method for the determination of nicotine and cotinine in Jordanian smokers' blood and urine samples by RP-HPLC and GC-MS. J Chromatogr Sci. 47(2):170-177.
11. da Fonseca BM, Moreno IE, Magalhães AR, *et al*. 2012. Determination of biomarkers of tobacco smoke exposure in oral fluid using solid-phase extraction and gas chromatography-tandem mass spectrometry. J Chromatogr B Analyt Technol Biomed Life Sci. 889-890:116-122.
12. Sylwia Narkowicz, Żaneta Polkowska, Jacek Namieśnik. 2012. Analysis of Markers of Exposure to Constituents of Environmental Tobacco Smoke (ETS), Critical Reviews in Analytical Chemistry. 42(1):16-37.
13. Benkirane S, Nicolas A, Galteau MM, Siest G. 1991. Highly sensitive immuno-assays for the determination of cotinine in serum and saliva. Comparison between RIA and an avidin-biotin ELISA. Eur J Clin Chem Clin Biochem. 29(6):405-410.
14. Ramdzan AN, Barreiros L, Almeida MI, Kolev SD, Segundo MA. 2016. Determination of salivary cotinine through solid phase extraction using a bead-injection lab-on-valve approach hyphenated to hydrophilic interaction liquid chromatography. J Chromatogr A. 1429:284-291.

15. Ramdzan A, Almeida M, McCullough M, Segundo M, Kolev, S. 2018. Determination of salivary cotinine as tobacco smoking biomarker. TrAC Trends in Analytical Chemistry. 105:89.
16. Zhang Y, Florath I, Saum KU, Brenner H. 2016. Self-reported smoking, serum cotinine, and blood DNA methylation. Environ Res. 146:395-403.
17. Lee KW, Pausova Z. 2013. Cigarette smoking and DNA methylation. Front Genet. 4:132.
18. Buszewski B, Ulanowska A, Ligor T, Denderz N, Amann A． 2009. Biomed. Chromatogr. 23(5):551-556.
19. Dawson AE, Jr AR, Weinberg DS. 1991. Nuclear grading of breast carcinoma by image analysis. Classification by multivariate and neural network analysis. American Journal of Clinical Pathology. 95(4 Suppl 1):S29.
20. Bai Y, Jin Z. 2005. Prediction of SARS epidemic by BP neural networks with online prediction strategy. Chaos Solitons & Fractals. 26(2):559-569.
21. Zhao J, Deng F, Cai Y. *et al*. 2019. Long short-term memory-Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. Chemosphere. 220:486-492.
22. Shen CY, Li JQ. 2012. Proton Transfer Reaction Mass Spectrometry for On-line Detection of Trace Volatile Organic Compounds in Breath. Chinese Journal of Analytical Chemistry. 40(5):773-777.
23. Zou X, Zhou W, Lu Y. *et al*. 2016. Exhaled gases online measurements for esophageal cancer patients and healthy people by proton transfer reaction mass spectrometry. Journal of Gastroenterology and Hepatology. 31(11):1837–1843.
24. Zou X, Kang M, Li A. *et al*. 2016. Spray Inlet Proton Transfer Reaction Mass Spectrometry (SI-PTR-MS) for Rapid and Sensitive Online Monitoring of Benzene in Water. Analytical Chemistry. 88(6):3144-3148.
25. Zou X, Lu Y, Xia L. *et al*. 2018. Detection of volatile organic compounds in a drop urine by ultrasonic nebulization extraction proton transfer reaction mass spectrometry. Analytical Chemistry. 90:2210-2215.
26. Yamashita N, Fukushima M. 2001. On the Rate of Convergence of the Levenberg-Marquardt Method. Computing. 15:239-249.
27. Budil DE, Lee S, Saxena S. *et al*. 1996. Nonlinear-Least-Squares Analysis of Slow-Motion EPR Spectra in One and Two Dimensions Using a Modified Levenberg–Marquardt Algorithm. Journal of Magnetic Resonance. 120(2):155-189.
28. Fan JY, Yuan YX. 2005. On the Quadratic Convergence of the Levenberg-Marquardt Method without Nonsingularity Assumption. Computing. 74(1):23-39.
29. Tsai CH, Chih YT, Wong WH. *et al*. 2015. A hardware-efficient sigmoid function with adjustable precision for a neural network system. IEEE Transactions on Circuits and Systems II: Express Briefs. 62(11):1073-1077.
30. Wang J, Zhang C. 2018. Software reliability prediction using a deep learning model based on the RNN encoder-decoder. Reliabil. Eng. Syst. Saf. 170:73–82.
31. Maas AL, Hannun AY, Ng AY. 2013. Rectifier nonlinearities improve neural network acoustic models. Proc. Icml. 30(1):3.
32. Barushka A. Hájek P. 2016. Spam filtering using regularized neural networks with rectified linear units. Conference of the Italian Association for Artificial Intelligence. 65-75.
33. Kingma DP, Ba J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv. 1412:6980.
34. Schoberberger R, Groman E. 2004. Quantification of recent smoking behaviour using Proton Transfer Reaction-Mass Spectrometry (PTR-MS). Wiener Klinische Wochenschrift. 116(5-6):21-25.
35. Kushch I, Schwarz K, Schwentner L. *et al*. 2008. Compounds enhanced in a mass spectrometric profile of smokers' exhaled breath versus non-smokers as determined in a pilot study using PTR-MS. Journal of Breath Research. 2(2):026002.