Application of Multidimensional Features and GRNN in Electronic Nose for Detection of Hepatocellular Carcinoma

Hao Lijun1*, 2, Zhang Min2, Zhu Geng1, Wang Xuan1

1Medical Instrumentation College, Shanghai University of Medicine &Health Sciences Shanghai, Shanghai, 201318, China
2School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology Shanghai, Shanghai, 200093, China
*Corresponding author’s e-mail: haolj@sumhs.edu.cn

Abstract. A model based on multidimensional features and GRNN was designed for electronic nose (eNose) in the paper. It can be applied to distinguish hepatocellular carcinoma from normal controls. Hepatocellular carcinoma patients have altered composition of exhaled gas due to abnormal metabolism. Thus, we can detect them by the exhaled gas. In the paper, the exhaled gas signals of hepatocellular carcinoma patients and health controls were first collected with eNose. And then the features were extracted and the multidimensional combined features were achieved. Furthermore, the PCA method was applied to optimize the features. Next, the classification model based on GRNN was constructed for training and generalization ability testing. Finally, the constructed model was adopted to predict the test and the performance was calculated. The result shows that, with the limited training set, the performance of the GRNN model is better than the BPNN model. The prediction accuracy could reach to 91.3%. Therefore, the proposed model is well suited for the classification detection with small training set and this will contribute to the study of the practical application of the eNose in the clinic.

1. Introduction

In China, hepatocellular carcinoma is one of the three major malignant tumors that endanger the life and health of residents. According to statistics, about 700 thousand people around the world are diagnosed with hepatocellular carcinoma every year. In 2016, about 829 thousand people died of hepatocellular carcinoma [1]. In clinical, the survival rate of primary hepatocellular carcinoma is the lowest, and the five-year survival rate is only 10.1%. This is mainly due to the fact that the specific symptoms of hepatocellular carcinoma are not obvious in the early stage, which leads to the inability to detect the disease and implement treatment in time. Therefore, in order to reduce the mortality, it is eager to improve the early diagnosis and screening of hepatocellular carcinoma.

Clinical studies have shown that metabolism influences the composition of human metabolites, including changes in the composition of exhaled gases. Diseases of the human organism, especially metabolic diseases, are often accompanied by the production of specific odors. For example, the exhaled gas of diabetic patients has a typical fruity taste. Patients with kidney disease have a high content of ammonia in their exhaled breath. And hepatocellular carcinoma patients have significantly higher concentrations of acetaldehyde and heptanal in their blood [2]. Highly volatile aldehydes, when reaching the lungs with the blood circulation, are exhaled through gas exchange, which leads to changes in the
composition of the exhaled gas [3]. Thus, a new method for detecting hepatocellular carcinoma can be explored based on a comparative study for exhaled gas of hepatocellular carcinoma and healthy controls.

The eNose is a kind of olfactory system simulating human body, which can detect many kinds of gas components exhaled by human body through sensor array. The present eNose devices, hardware technology is more mature. However, intelligent algorithms have yet to be studied in depth in order to achieve high precision detection, especially for limited clinical samples [4-8].

Artificial neural networks are a hot research topic that emerged in the field of artificial intelligence in the 1980s. They can simulate biological neural networks to build data processing models. They have shown good properties in pattern classification, prediction estimation, and so on [9-11]. Among them, back propagation neural network (BPNN) is one of the most widely used neural networks. It uses error back propagation algorithm to adjust the weights to achieve any complex nonlinear mapping. But, BPNN has the disadvantages of requiring more parameters, slow convergence, long training time, and easy to fall into local optimization. Generalized regression neural network (GRNN), however, calculates the network output based on the maximum probability principle. It has good nonlinear mapping ability and learning speed. Its algorithm is simple and easy to train. Even if the training set is small, the output of GRNN can converge to the optimal regression surface [9, 10].

The purpose of the study was to construct a model based on GRNN to see if it could distinguish patients with hepatocellular carcinoma and healthy controls based on the breath curves collected by eNose when the training set is small. For the aim, the breath response signals from a total of 120 volunteers, including training samples and testing samples were collected. And the ten-fold crossover method was applied to performance assessment. As a result, it was found the GRNN model outperforms the BPNN model in detecting hepatocellular carcinoma when the sample size is small.

2. System and experiment

The eNose used in the study is ILD.3000 designed by UST Sensors GmbH, Germany, as shown in Figure 1(a). Its main components are three gas sensors and one temperature sensor. The gas sensors are the reactive part of the measuring system where each layer of the sensor possesses different sensitivities and selectivity for a variety of different gases at varying temperatures. The temperature sensor is used to provide an operating temperature to improve the sensitivity of the gas response. The three gas sensors in the device, are the GGS1000 series sensor, which is sensitive for combustible gases; the GGS3000 series sensor, which can detect hydrocarbons, especially for C1, C2 ......C8; and the GGS7000 series sensor, which can detect NO2[4].

![ILD.3000](a) ![Disposable exhalation nozzle](b)

Figure 1. The experiment system

In the study, we collected exhaled gas reaction signals from volunteers through a disposable exhalation nozzle, shown in Figure 1(b). We spent two years collecting breath curves of 120 volunteers at Shanghai Jiao Tong University Renji Hospital. The cases included 69 hepatocellular carcinoma patients and 51 healthy controls. The collection was performed on a voluntary basis and done in vitro using a disposable exhalation nozzle without any harm to the human body. The inclusion criteria for volunteers were that the patients must be hepatocellular carcinoma, no other metastatic cancer evidence, no respiratory disease, and no history of smoking or alcohol abuse in the last three months. The collection was performed under fasting conditions.
### Table 1. Basic information of volunteers

|                      | Hepatocellular carcinoma | Healthy volunteers |
|----------------------|--------------------------|--------------------|
| number               | 69                       | 51                 |
| age (x±s, years)     | 56.32±10.56              | 53.03±15.67        |
| sex (male/female)    | 57/12                    | 31/20              |

3. **Feature extraction**

3.1. **Data preprocessing**

The collection process consists of five stages, corresponding to five waveform segments, as shown in Figure 2. P1, P2 and P4 are blank measurements, which complete the cleaning and measurement preparation of the sensors. P3 and P5 indicate the two tests of the patient's exhaled gas. The three different coloured curves in Figure 2(a) represent the different responses of the three sensors to the same exhaled gas. The blue curve in Figure 2(b) reflects the temperature variation in the measurement. We collected valid data based on the temperature changes, such as the corresponding part marked by the dashed line. In fact, 60 sampling points for each sensor were collected in a test. And we could obtain three expiratory curves for each sample for once test.

![Figure 2: Signal acquisition process and data extraction](image)

3.2. **Extraction of multidimensional feature**

According to 3.1, for all samples, we could obtain a high-dimensional array of $60 \times 3 \times n$. Where, 60 denotes sample points of one sample; 3 represents the number of sensors; n can represent the number of samples. Further, features were extracted independently for three expiratory curves and then the features are combined into a new sample. We first extracted 15 time features for each curve, including: maximum value and position corresponding to the maximum value, minimum value and position corresponding to the minimum value, mean value, peak-to-peak value, rectified mean value, variance, standard deviation, root mean square, waveform factor, peak factor, pulse factor, margin factor, and area under the curve [12]. Then, in the frequency domain, 11 features were extracted for each curve, such as: gravity frequency, frequency variance, mean square frequency, frequency spectrum, cepstrum, and power spectrum calculated by several means including FFT, indirect method, modified direct method, and welch three window function method, respectively [13]. In addition, the statistical features of each curve were also calculated, including polar deviation, median, quantile corresponding to 0.25 and 0.75, plural, coefficient of variation, origin moment, skewness, kurtosis, autocorrelation coefficient, and information entropy for a total of 10 parameters with dimension of 129. Ultimately, for a volunteer sample, 39
features were extracted for an expiratory curve, with a dimension of up to 1068. Thus, for three curves, a set of 3204 dimensional features will be obtained. With the increase of samples, the multidimensional feature dataset will become larger.

3.3. Principal Component Analysis
Principal Component Analysis (PCA) is a matrix compression algorithm that is commonly used for data dimensionality reduction [14]. It can reduce the dimensionality of the matrix while retaining as much information of the original matrix as possible, thus greatly saving space and data volume.

The main steps of the PCA algorithm are as follows.

1. The original data is divided into n rows and m columns matrix $X_{nm}$, where n represents the number of samples and m represents the number of features.

2. Centralize all samples:

$$X_i = \frac{1}{n} \sum x_{ij}, j = 1:m.$$

3. Find out the covariance matrix of the sample, $\text{cov}(XX^T)$

4. For the covariance matrix, find the eigenvalue and the corresponding eigenvector.

$$[\text{eigenvector}, \text{eigenvalue}] = \text{eig}(\text{cov})$$

5. The eigenvectors are rearranged by rows according to the corresponding eigenvalues. The first k rows are selected to form the eigenmatrix $P$ according to the proportion of eigenvalues. Then the new sample $\overline{X}_{nm}$ after dimensionality reduction can be obtained by multiplying the matrix $P$ with the original sample.

6. For new samples $\overline{X}_{new}$, the same dimensionality reduction can be realized by $Y = P\overline{X}_{new}$.

In the study, PCA was used to optimize the high-dimensional features, and only 95% of the features were retained. The original feature array $3204 \times n$ was converted into a matrix of $10 \times n$. The amount of data was greatly reduced and the calculation was more convenient. For a new sample, the same feature optimization can be realized according to step (6).

4. Classification model based on GRNN
GRNN is a forward neural network with tutor learning. It has strong nonlinear mapping ability and learning speed, which is good at prediction when the sample data is small. We constructed a classification model based on GRNN, which consists of an input layer, an implicit layer, a summation layer, and an output layer [11], as shown in Figure 3.

In the model, the sample $\overline{X}_{nm}$ was the input, which will be fed into the input layer. The number of neurons of the input layer in the constructed network was the number of features after the dimensionality reduction of PCA. And its transfer function is the simplest linear function. The implicit layer: was the radial base layer, the number of neurons is equal to the number of training samples n, and the base function is generally a Gaussian function. The summation layer included two nodes, of which the first neuron calculates the algebraic sum of each neuron in the hidden layer and the second neuron calculates
the weighted sum of the implicit layer, and the weight is the expected output value of each training sample. And the output layer get the estimated value of the output y. The construction of GRNN is relatively simple. It can be achieved by simply performing a one-dimensional search for a smooth factor. In the study, we set the range of the smooth factor is 0.1~2 and the step size is 0.1. We determined the optimal parameters by solving step-by-step, and then the network could be realized[12].

5. Result

5.1. Performance Comparison
In the study, we collected exhaled gas samples from 120 volunteers at different times. The samples were independent of each other. We first randomly divided the collected data into training set and testing set. The training set, included breath data samples from 56 hepatocellular carcinoma patients and 41 healthy controls. The testing set, included breath samples from 13 hepatocellular carcinoma patients and 10 healthy controls. Meanwhile, we set the hepatocellular carcinoma patients as positive cases with label value of 1 and the healthy controls as negative cases with label value of 0. The multiple features of training set were then extracted and combined. The optimized features were used as samples to train the classification models of BPNN and GRNN, respectively.

The BPNN model constructed in the study included an input layer, two implicit layers and an output layer. The transfer function between the input layer and the first implied layer, and the transfer function between the two implied layers were logsig function. The gradient descent method trainingdx with momentum and adaptive lr was chosen for the output training function.

In order to objectively evaluate the generalization ability of the training model, the ten-fold crossover method was used for performance evaluation. Table 2 shows the generalization performance of the models constructed using BPNN and GRNN respectively.

| Table 2. Comparison of the performance of two models |
|------------------------------------------------------|
| model       | Accuracy | Sensitivity | F-score |
| BPNN        | 82.49    | 81.67       | 83.43   |
| GRNN        | 84.72    | 89.33       | 82.62   |

5.2. Prediction of testing set
We assumed the 23 samples of the test set as the new samples. We applied the two different models to predict the classification of the same testing set separately. Table 3 shown the prediction performance of BPNN model and GRNN model for the 23 samples. The results shown that the prediction performance of GRNN is better than that of BPNN.

| Table 3. Comparison of prediction using two models |
|---------------------------------------------------|
| model       | Accuracy | Sensitivity | Specificity | Precision | F-score |
| BPNN        | 82.61    | 76.92       | 90          | 90.91     | 83.33   |
| GRNN        | 91.30    | 92.31       | 90          | 92.31     | 92.31   |

6. Conclusion
The four methods of diagnosis in Chinese medicine are "looking, smelling, asking and cutting". Among them, smelling refers to listening with the ears, smelling with the nose, and identifying health conditions through breathing. However, the method relies mainly on the doctor's experience and lacks an objective basis. With the development of technology, researchers have specifically designed the eNose. The device is sensor-based, allowing for more objective and accurate detection of human exhaled gases. However, how to effectively distinguish healthy people from sick people relies on the method of artificial intelligence. It can be said that artificial intelligence replaces the thinking and judgment of doctors to a certain extent. In the paper, we obtained the exhalation signals of two groups of people, hepatocellular carcinoma and healthy controls by eNose, and extracted the multidimensional features. A model based on GRNN was applied to detect of hepatocellular carcinoma. The results show that the classification accuracy, sensitivity, and specificity of GRNN are better than the traditional BPNN with
a small size training set. We can conclude that the GRNN model can be applied in eNose to detect hepatocellular carcinoma when the sample is limited. With the continuous maturity and development of intelligent algorithms, the accuracy of eNose applied to disease detection will become higher, and the generalization ability of classification detection will become better. In the next research, we will devote ourselves to mining more features in the breath curves, constructing more intelligent prediction models, and advancing the application of eNose in clinical. In the future, the development of artificial intelligence technology and the application of engineering technologies in solving practical clinical problems will be beneficial to vigorously promote the development of clinical.

Acknowledgments
This work was supported by Shanghai University of Medical & Health. The data collection work was completed by the team from Shanghai Jiao tong University and Shanghai University of Technology. And the electronic nose department was supported by the German UST company.

References
[1] Fitzmaurice C, Akinyemiju TF, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2016: A systematic Analysis for the Global Burden of Disease Study [J]. JAMA Oncol. 2018, Jun 2.
[2] ZHANG Yingping, WANG Lin, LIU Chan, et al. The screening of VOCs in the breath of diffuse large B-cell lymphoma patients [J]. Acta Universitatis Medicinalis Anhui, 2016, 51(8): 1204-1207.
[3] Galen Miller-Atkins, Lou-Anne Acevedo-Moreno, David Grove, et al. Breath Metabolomics Provides an Accurate and Noninvasive Approach for Screening Cirrhosis, Primary, and Secondary Liver Tumor [J]. Hepatology communications. 2020, 4 (7): 1041-1055.
[4] Andreas Voss, Katharina Witt, Tobias Kaschowitz, et al. Karl-Jurgen Bar. Detecting Cannabis Use on the Human Skin Surface via an Electronic Nose System [J]. Sensors. 2014, 14: 13256-13272.
[5] Bijithra C., Rajan G., Shanmugasundaram P. Chemical and electronic sensors used as a diagnosing tool for diseases-A review [J]. Biomedicine. 2016, 36(1): 5-10.
[6] Lijun H, Fei X. An investigation on Electronic nose diagnosis of hepatocellular carcinoma [C]. CSIP-BMEI 2017, Shanghai: IEEE, 2018: 1-5.
[7] Germanese D, Colantino S, D’Acunto M, et al. An E-Nose for the Monitoring of Severe Liver Impairment: A Preliminary Study [J]. Sensors, 2019, 19(17): 3656.
[8] Swanson B, Fogg L, Julion W, et al. Electronic Nose Analysis of Exhaled Breath Volatiles to Identify Lung Cancer Cases: A Systematic Review [J]. The Journal of the Association of Nurses in AIDS Care: JANAC, 2020, 31(1): 71-79.
[9] Dai, Y., Guo, J., Yang, L., et al. A New Approach of Intelligent Physical Health Evaluation based on GRNN and BPNN by Using a Wearable Smart Bracelet System [C]. Procedia Computer Science. 2019, 147: 519-527.
[10] GUO Qing, WU Ru-qin, XU Cui-feng. The Electrocardiogram Classification Approach Based on Wavelet Transform and FOAGRNN [J]. MICROELECTRONICS & COMPUTER. 2018, 35(6): 31-35.
[11] Ruian Q, Yuankun W; Dong W; et al. Water temperature forecasting based on modified artificial neural network methods: Two cases of the Yangtze River [J]. Science of the Total Environment. 2020, 737: 139729.
[12] LIU Yu, YU Yue, LU Yong-le, et al. Real-time human activity recognition based on time-domain features of multi-sensor [J]. 2017, 25(4): 455-460.