Research Article

Feature Optimization of Exhaled Breath Signals Based on Pearson-BPSO

Lijun Hao,^{1,2} Min Zhang,¹ and Gang Huang^{ID}³

¹School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

²Medical Instrumentation College, Shanghai University of Medicine & Health Sciences, Shanghai 201318, China

³Shanghai Key Laboratory of Molecular Imaging, Shanghai University of Medicine & Health Sciences, Shanghai 201318, China

Correspondence should be addressed to Gang Huang; huanggg@sumhs.edu.cn

Received 24 September 2021; Revised 3 November 2021; Accepted 10 November 2021; Published 3 December 2021

Academic Editor: Fazlullah Khan

Copyright © 2021 Lijun Hao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Feature optimization, which is the theme of this paper, is actually the selective selection of the variables on the input side at the time of making a predictive kind of model. However, an improved feature optimization algorithm for breath signal based on the Pearson-BPSO was proposed and applied to distinguish hepatocellular carcinoma by electronic nose (eNose) in the paper. First, the multidimensional features of the breath curves of hepatocellular carcinoma patients and healthy controls in the training samples were extracted; then, the features with less relevance to the classification were removed according to the Pearson correlation coefficient; next, the fitness function was constructed based on K-Nearest Neighbor (KNN) classification error and feature dimension, and the feature optimization transformation matrix was obtained based on BPSO. Furthermore, the transformation matrix was applied to optimize the test sample's features. Finally, the performance of the optimization algorithm was evaluated by the classifier. The experiment results have shown that the Pearson-BPSO algorithm could effectively improve the classification performance compared with BPSO and PCA optimization methods. The accuracy of SVM and RF classifier was 86.03% and 90%, respectively, and the sensitivity and specificity were about 90% and 80%. Consequently, the application of Pearson-BPSO feature optimization algorithm will help improve the accuracy of hepatocellular carcinoma detection by eNose and promote the clinical application of intelligent detection.

1. Introduction

Hepatocellular carcinoma is a malignant tumor with a high incidence rate and high mortality rate, which seriously endangers the quality and life of our lives. According to the statistics, in 2018, the mortality rate of hepatocellular carcinoma accounted for 8.2% of the total number of cancer cases in the world [1]. More than half of the world's Hepatocellular carcinoma cases and deaths occur in China [2]. Studies have found that the early symptoms of hepatocellular carcinoma are particularly insignificant. Many patients with hepatocellular carcinoma have entered the middle and late stages of the disease when they are diagnosed. Therefore, to reduce the mortality, it is anxious to improve the early diagnosis and screening of hepatocellular carcinoma.

Disease can lead to metabolic changes, which can result in different exhalation gases. The result showed that there are significant differences in 6 VOCs in the exhaled gases of hepatocellular carcinoma patients compared with healthy controls ($P < 0.05$) [3]. The emerging expiratory detection technology in recent years can detect the exhalation gas of human body and determine the disease, which can be widely used in early clinical examination [4]. As an expiratory detection device, the eNose device records the exhaled gas by the response of its internal sensors to different exhaled gases. It is worth noting that the device does not detect the specific gas composition; it just records the overall response curve of the gas. Researchers need to build models through a large number of different volunteer gas response curves and find the potential relationship between disease and expiratory

response, to realize the purpose of disease diagnosis based on eNose.

In the process of constructing a disease classification model using exhalation signals, feature extraction is the first step. In order not to lose information that may affect the accuracy of the test, we usually extract as many features as possible [5]. However, this may cause redundancy of features, increase in computation, and decrease in computation speed and accuracy. Therefore, the study of feature optimization algorithms has received increasing attention. At present, principal component analysis (PCA) is widely used in feature optimization. The algorithm is a linear dimension reduction algorithm, which transforms high-dimensional data into low-dimensional data by matrix compression. The algorithm is fast in calculation and low in complexity, but it is not very good for dimensionality reduction of complex nonlinear data [6]. A binary particle swarm optimization (BPSO) algorithm proposed by Ebert in 1997 can be used to solve the problem. In recent years, the algorithm has been continuously improved and applied to feature selection by researchers, such as BPSO based on GA, BPSO based on average fitness, BPSO combined with bacterial algorithm, and immune algorithm [7–9]. In these researches, the improvement of BPSO mostly focuses on the design of fitness function in particle swarm optimization algorithm, without considering the characteristics of feature selection in the actual classification problem. Therefore, the maximum performance of the algorithm cannot be exploited. Some researchers propose to use SVM-RFE method to select part of the initial population of the particle swarm optimization algorithm to reduce the search space of particles. The algorithm can effectively improve the accuracy of classification and recognition [10]. However, the dimension of SVM-RFE algorithm is different each time, which leads to the randomness of the final optimization result.

Moreover, due to the uncertainty of the feature optimization factor, the optimization algorithm is not suitable for generalization to feature optimization of new samples.

An improved feature optimization algorithm Pearson-BPSO based on the traditional BPSO algorithm and Pearson correlation coefficient was proposed in the paper. The purpose of this study is to test the feasibility of the new algorithm. We applied three different algorithms, including PCA, BPSO, and Pearson-BPSO to optimize features. Then, we evaluate the three optimization algorithms by the performance of two different classification models based on support vector machine (SVM) and random forest (RF), respectively. The comparison of the results has shown that the new feature optimization algorithm was beneficial to improve the accuracy of the classifier. The next section, i.e., Section 2, following the introduction, is based on the materials and methods. After this, Section 3 contains the results. The discussion is actually in Section 4, and Section 5 is the concluding section.

2. Materials and Methods

2.1. Signal Acquisition. The eNose device, also known as an artificial olfactory system, can simulate the biological olfactory systems through the combination of gas sensor and

pattern recognition technology. Its basic principle is to use a gas sensor to simulate olfactory sensory nerve cells in a biological olfactory systems and use a computer or special chips to process the collected information, to achieve the purpose of identifying gas or odor [11].

In the project, the response curve of exhaled gas with sensors was collected by an eNose device named ILD.3000, which was designed by the UST Sensors GmbH Company of Germany. As shown in Figure 1, three different gas sensors RS1, RS2, and RS3 are the core of the hardware system [12]. Different people exhaled different gas composition, and thus the sensor response curves will also be distinct. 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 three gas sensors in the device are the GGS1000 series sensor, which is sensitive to combustible gases; the GGS3000 series sensor, which can detect hydrocarbons, especially for C_1 , C_2 , ..., C_8 ; and the GGS7000 series sensor, which can detect NO_2 [13]. The controllable temperature sensor R_t is used to provide a suitable temperature environment to improve the response ability of the sensor to the gas.

As shown in Table 1, during the study, the expiratory data of 121 volunteers were collected in Renji Hospital, including 69 patients with hepatocellular carcinoma and 52 healthy controls. All expiratory data were collected voluntarily. During the collection process, the disposable exhalation nozzle was used, which was completed in vitro without any interventional device and no harm to the human body. The inclusion criteria of the volunteers were that the patients must be of primary liver cancer, no other metastatic cancer, no respiratory diseases, and no history of smoking, and drinking in the past three months. After that, the collection should be carried out in case of fasting.

As shown in Figure 2, in the test, we could simultaneously obtain three response curves (indicated by three different colors: yellow, grey, and orange) of exhaled gas and a temperature curve by the eNose device, corresponding to three different sensors. The temperature varies from $280^\circ C$ to $420^\circ C$, and the response curve represents the response resistance of the sensor to different gases. Compared with the temperature, the resistance value varies widely.

2.2. Signal Preprocessing. As shown in Figure 2, the values and amplitudes of the curves for the three sensors collected in a single time varied greatly. To facilitate the comparison, the normalized processing method was used firstly to reduce the magnitude without changing the waveform state. Each curves were transformed by formula (1) into relative values within the range of $[0, 1]$, to simplify the subsequent analysis.

$$y_k(i, j) = \frac{y_k(i, j) - \min(y_k(i))}{\max(y_k(i)) - \min(y_k(i))}, \quad (1)$$

where $y_k(i)$ represents the i -th sample curve collected by a certain sensor, and the sample length is 60. k in $y_k(i)$ can be A, B, and C, representing three sensors, respectively; i in $y_k(i)$ represents the i -th sample; $y_k(i, j)$ means that the j -th

FIGURE 1: Construction of electronic nose.

TABLE 1: The basic information of all volunteers.

	Male	Female	Total number	Age
Hepatocellular carcinoma	58	11	69	55.58 + 10.63
Healthy control	35	17	52	51.60 + 14.57
Total number	93	28	121	--

FIGURE 2: Signal curves collected by different sensors of the eNose device.

point of the i -th sample curve. The value of j is from 0 to 59, and $\min(y_k(i))$, $\max(y_k(i))$ represents the minimum and maximum of the signal, respectively.

2.3. Features Extraction. After the signal curve was normalized, as many features as possible for each curve were then extracted. In the study, we extracted time features, frequency-domain features, statistical features for each curve, and relevant features between the three curves obtained by different sensors. The 15 time-domain features were maximum value and corresponding position, minimum value and corresponding position, mean, peak-to-peak, rectified mean,

variance, standard deviation, waveform factor, pulse factor, peak factor, margin factor, and area. The 14 frequency-domain features included center of gravity frequency, frequency variance, root mean square difference, spectrum, and power spectrum calculated by various methods. The 10 statistical features were extreme deviation, median, quantile, and plurality, coefficient of variation, skewness, kurtosis, autocorrelation coefficient, and information entropy. In addition, the two-by-two correlations between the three sensor signals were calculated, and three features were obtained. Finally, for one breath test of each volunteer, we combine all features of the three curves, and 2082 dimensional high-dimensional features could be achieved.

2.4. Feature Optimization. In the classification task, the initial analysis of the sample feature base is to extract the features that are most significant for distinguishing different categories from the original data, while discarding those features that do not contribute much to the classification. Thus, feature optimization actually removes irrelevant factors and reduces their interference with the classification. The selection of the most optimal feature set can effectively reduce the dimensionality of the feature space. Therefore, feature optimization can reduce the computational effort and increase the computational speed.

As shown in Figure 3, the traditional BPSO feature optimization algorithm was improved, which not only considers the classification accuracy and the number of features, but also makes full use of the feature of classification to consider the correlation between features and categories. Before using BPSO algorithm for feature selection, the correlation between features and categories was firstly calculated. According to Pearson correlation coefficient, the first certain number of features with high correlation was selected. It is important to note that the number of features selected can be set by experience. In the study, the number was taken as a relatively large number 1000. Then, with the optimization objectives of reducing the classification error rate and feature dimensionality of KNN, the fitness function was constructed. Therefore, the optimal feature subset was selected based on BPSO, and the feature optimization operator was determined. The specific flow of the algorithm is shown in Figure 3.

2.4.1. Initial Screening of Features Based on Pearson Correlation Coefficient. Pearson correlation coefficient was proposed and evolved by the British statistician Karl Pearson in the 1880s [14]. The coefficient can be used to measure the correlation (linear correlation) between two variables X and Y , and its value is between -1 and 1.

In the study, the eigenvalue of each sample was regarded as the input variable x , and the label of each sample was regarded as the variable y . The Pearson correlation coefficient could determine the degree of correlation between the label and each feature in a multidimensional feature set by calculating the correlation between the input features and the output labels. Then, according to the Pearson correlation coefficient, the preliminary screening of features could be realized.

The Pearson correlation coefficient was obtained by the following formula:

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \quad (2)$$

where E represents the mathematical expectation, and X and Y represent the input feature and output labels, respectively. The value of correlation coefficient is between -1 and 1. When the value of correlation coefficient is close to 0, there is no correlation between them. When the value of correlation coefficient is close to 1, there is a significant positive correlation between the feature and the label. Similarly, when

the value is close to -1, there is a negative correlation between the input variable and the label. That is, when the value of an input feature rises, the label will be classified change.

2.4.2. Feature Optimization Based on Recognition Error and Feature Dimension. For a $M \times N$ dataset, each row represents a sample. M rows represent M samples, and N columns represent N features of a sample. Feature optimization was essential to find the smallest possible subset of features among these N features. With the new features, the higher correct classification results could be ensured. The subset of features could be regarded as the optimized features.

By calculating the optimized conversion factor, more sample features could be optimized. The main steps [15, 16] are as follows:

Step 1. Set the features after initial screening as particles, the feature dimension as the dimension of particles, and the initial number of particles to 300. The positions of the particles and the individual optimal positions were randomly initialized using binary encoding.

Step 2. The fitness function for feature selection is constructed based on the classification error rate and the optimized feature dimension, as shown in the following formula:

$$\text{fitness}(\mathbf{i}) = \mathbf{k}_1 \times \text{error}(\mathbf{i}) + \mathbf{k}_2 \times \frac{\text{Wei du}(\mathbf{i})}{\mathbf{D}}, \quad (3)$$

where $\text{fitness}(\mathbf{i})$ is the fitness obtained based on particles; $\text{error}(\mathbf{i})$ is the error rate of classifier recognition after feature selection based on particles; $\mathbf{D}$ is the original feature dimension; $\text{Wei du}(\mathbf{i})$ is the feature dimension selected based on particles; $\mathbf{k}_1$ and $\mathbf{k}_2$ are the weights of classifier recognition error rate and feature dimension optimization, which can be taken as 0.8 and 0.2, respectively.

Step 3. The fitness value of each particle is calculated according to Step 2, and the individual and global dynamic factors and inertia weights are updated according to the fitness value, as formulae (4) to (6):

$$\mathbf{c}_1 = \text{rand} \times \left(2.4 - 1.4 \times \frac{\text{iter}}{\text{TNum}} \right), \quad (4)$$

$$\mathbf{c}_2 = \text{rand} \times \left(0.9 + 1.6 \times \frac{\text{iter}}{\text{TNum}} \right), \quad (5)$$

$$\mathbf{w} = \mathbf{w}_{\max} - (\mathbf{w}_{\max} - \mathbf{w}_{\min}) \times \frac{\text{iter}}{\text{TNum}}. \quad (6)$$

Among them, $\mathbf{c}_1$ and $\mathbf{c}_2$ are the dynamic factors of individual adjustment and global adjustment, $\mathbf{w}$ is the inertia weight, rand is a random number [0, 1], iter is the number of iterations, TNum is the preset number of iterations, $\mathbf{w}_{\max}$ and $\mathbf{w}_{\min}$ are the maximum inertia weight and minimum inertia weight, respectively.

FIGURE 3: Flow diagram of the improved Pearson-BPSO feature optimization algorithm.

According to formulae (4) to (6), the iterative update value of velocity can be further calculated, such as the following formula:

$$\mathbf{v}(\mathbf{i} + 1) = \mathbf{w} \times \mathbf{v}(\mathbf{i}) + \mathbf{c}_1 \times (\mathbf{p}(\mathbf{i}) - \mathbf{x}(\mathbf{i})) + \mathbf{c}_2 \times (\mathbf{g} - \mathbf{x}(\mathbf{i})), \quad (7)$$

$\mathbf{v}(\mathbf{i} + 1)$ is the updated velocity value, $\mathbf{w}$ is the dynamic inertia weight, $\mathbf{v}(\mathbf{i})$ is the last velocity value, and $\mathbf{x}(\mathbf{i})$ is the current position; $\mathbf{p}(\mathbf{i})$ is the individual optimal position, and $\mathbf{g}$ is the global optimal position.

Step 4. Multiple iterations are performed, and the particle positions are updated binarized (0, 1) according to the velocity definition condition using formulae (8) and (9):

$$\text{sig}(\mathbf{v}_i) = \frac{1}{1 + e^{-\mathbf{v}_i}}, \quad (8)$$

$$\mathbf{x}_{ij}(\text{iter} + 1) = \begin{cases} 0, & \text{if}(\text{rand} \geq \text{sig}(\mathbf{v}_{ij}(\text{iter} + 1))), \\ 1, & \text{otherwise.} \end{cases} \quad (9)$$

In formula (8), a sigmoid function is used to map the velocity to the interval [0, 1] as a probability, and this probability is the probability that the particle will take a value of 1 next.

Also, $X_{ij}(\text{iter} + 1)$ in formula (9) is the absolute probability of a change in position.

Step 5. Determine whether the maximum number of iterations has been reached. If the number of iterations has reached **TNum**, then the optimized feature subset was got according to the optimal position in the population history, and the optimal position record will be used as the feature optimization conversion operator; otherwise, return to Step 3.

3. Results

After feature optimization, it is necessary to evaluate the effect of optimization by quantitative methods. In the paper, we evaluate the effect of feature optimization on the performance of the classifier.

First of all, we obtain the feature optimization operator based on the training exhalation sample features and implement the feature optimization for the test samples. The specific steps are as follows.

First, the collected two types of samples, totaling 121 breath signals containing healthy control and hepatocellular carcinoma patients, were divided into a training set and a test set after multidimensional feature extraction. Then, the training set

was used to determine the feature optimization operator. The specific method was as follows: using tag value 1 (representing hepatocellular carcinoma patients) and tag value 0 (representing healthy controls) to construct a tag array. Taking it as the dependent variable y and the high-dimensional sample feature array as the variable x , the interrelationships between sample features and categories were calculated by Pearson correlation analysis. Thus, the sample feature groups were sorted by the absolute values of Pearson correlation coefficients, and the top 1000-dimensional features were retained. Furthermore, the fitness function was constructed by KNN classification error rate and feature dimension, and the optimal subset was achieved based on BPSO. And meanwhile, the feature optimization conversion factor was obtained. Afterwards, feature optimization was performed on the test set using the feature optimization operator derived in the above steps.

Once the feature optimization was completed, the next step was to build the classifier.

Two different classifiers were constructed to obtain a more respectable evaluation. One was a classifier built based on the support vector mechanism (SVM classifier), and the other was a classifier built based on the random forest method (RF classifier).

Here, we applied two different classifiers to classify and detect the optimized features processed by three various optimization methods. By comparing the performance of the classifiers, we found that the Pearson-BPSO is more effective in classification compared to the other two traditional feature optimization methods, PCA and BPSO.

3.1. Performance Comparison of Pearson-BPSO and BPSO. To compare the feature optimization effect of the improved Pearson-BPSO and the traditional BPSO, the search for the optimal subset of features and the determination of the feature transformation factor were performed based on the above two algorithms, respectively.

Figure 4 shows the adaptation curves of the two BPSO algorithms in 100 iterations. The horizontal coordinate in the figure represents the number of iterations, and the maximum setting is 100; the vertical coordinate is the fitness value, and the smaller value means better optimization performance. Among them, Figure 4(a) shows the adaptation curve of the improved Pearson-BPSO algorithm. Based on the optimized fitness, feature dimension could be reduced to 251, and the adaptation value could be lower than 0.045. Figure 4(b) shows the adaptation curve of the traditional BPSO algorithm, with the optimized feature dimension of 712 and the optimal adaptation value of about 0.08.

3.2. Classification Performance. The performance of the feature mapping optimization algorithm can be reflected by the performance of the classifier. Here, we calculated the performance of the SVM classifier and RF classifier to compare the performance of the optimization algorithms.

After the feature extraction of the original samples, the positive and negative samples were divided into 10-fold and combined into 10 groups of sample data. Onefold of the data (about 7 cases of liver and 5 cases of control) was taken as the

test sample each time, and the rest of the samples were taken as the training samples. Then, the feature transformation factors calculated by the improved feature optimization algorithm Pearson-BPSO, the traditional BPSO, and the PCA optimization algorithm were used to optimize and reduce the feature dimension of the training samples and test samples, respectively, to obtain different optimized feature datasets. Furthermore, the classifiers were constructed based on SVM and RF, respectively, and the classification performance was calculated for each time. The process was repeated ten times, and different onefold data was taken as test samples in turn, and the classification performance was calculated separately for each time. Finally, the average of each performance was obtained as the performance metrics of the two classifiers under the three different feature optimization algorithms, as shown in Tables 2 and 3.

From Table 2, we found that the best accuracy was 86.03, and the best sensitivity was 90.79, when the Pearson-BPSO feature optimization was applied. From Table 3, we found that the best accuracy was 90, and the best sensitivity was 94.83, when the Pearson-BPSO feature optimization was applied.

The performance indicators in Table 2 and 3 include the following: Acc is used to measure the accuracy of the classifier in correctly classifying samples. Sens represents the sensitivity of the classifier in recognizing hepatocellular carcinoma samples. Spec is the specificity of the classifier in recognizing normal samples. F-score represents the comprehensive performance of the classifier, and the higher the F-score value, the better the performance of the classifier.

4. Discussion

According to the mechanism of breath testing, due to pathological reasons, the metabolism of hepatocellular carcinoma patients will change, and the composition of exhaled gas will also change. Therefore, classification and recognition of exhalation data of patients with hepatocellular carcinoma and healthy people were the most important work of intelligent detection of hepatocellular carcinoma. In the study, we distinguished hepatocellular carcinoma by constructing a dichotomous model to distinguish the breath signals of hepatocellular carcinoma patients and healthy individuals.

The findings from the reviewed studies were consistent with other studies that have shown that volatile breath biomarkers can discriminate persons with malignant solid tumors from noncancer control subjects [17]. However, there is no clear conclusion on the types of volatile marker gases for hepatocellular carcinoma. The present study is based on the fact that the collection device used can respond to a large number of volatile exhaled gases, including possible hepatocellular carcinoma specific exhaled gases among them [18]. We do not need to know the specific type of gas; we just need to record the overall response of the exhaled gases containing some specific gases. Here, we attempted to construct a dichotomous classifier using the different characteristics of the integrated response curves of exhaled gas in healthy individuals and hepatocellular carcinoma patients. However, to improve the validity of the exhaled gas response detection, we need to use GC-MS to further

FIGURE 4: Adaptation curves. (a) Pearson-BPSO. (b) BPSO.

TABLE 2: Performance of SVM classifier based on three different feature optimization algorithms.

Algorithms	Acc	Sens	Spec	F-score
Pearson-BPSO	86.03	90.79	78.64	0.878
BPSO	77.69	87.98	65.38	0.811
PCA	75.19	86.63	73.56	0.795

TABLE 3: Performance of RF classifier based on three different feature optimization algorithms.

Algorithms	Acc	Sens	Spec	F-score
Pearson-BPSO	90	94.83	85.17	0.841
BPSO	84.42	87.80	81.40	0.800
PCA	82.49	90.90	76.33	0.832

compare and analyze exhaled gas from patients and healthy individuals, and to determine the specific gas species. Thus, a highly sensitive eNose specifically designed to detect some diseases then can be designed.

In addition, the devices used in the study have not been applied to the clinic, and there are no clear international standards for the manner and criteria of exhaled gas collection. The amount and sources of clinical data utilized in the data analysis are relatively limited. Some advanced intelligent algorithms [19], such as deep learning, which were based on big data, cannot be utilized. Hence, the establishment of a breath database is an essential step to advance the clinical application of eNose research. It depends on the establishment of international unified standards for breath collection. The collection criteria include the type of gas collected, the collection method, the patient's age, gender, diet or not, and even race, and other more comprehensive information needs to be collected.

The study is still in the exploratory stage, the amount of data collected is limited, and the results of clinical analysis

may be one-sided. Due to the limited nature of the sample, traditional machine learning algorithms were used in the study for classification, i.e., signal preprocessing, and feature extraction and classification model construction to distinguish hepatocellular carcinoma patients from healthy individuals. In the specific work, because the human exhalation signal collected by the eNose device has a large interindividual and individual variability at different moments, which cannot visually and effectively distinguish the data of hepatocellular carcinoma patients from other health data, the signal is firstly subjected to feature extraction. The extraction of signal curves' features helps discover more potential information. However, high-dimensional features may lead to the degradation of classification accuracy and slow computation, so the optimization of features becomes a hot topic of research.

A way to measure whether feature optimization is more effective is to feed the same features, optimized by different algorithms, into the same classifier and test the classification performance. In the study, the tenfold cross-validation method was used in the data analysis, taking into account that the performance of the constructed classifiers varied when different training samples were used, and the average performance was used as the final measure. In addition, based on the randomness of selection when dividing samples, an imbalance between positive and negative samples in training and testing samples may occur. To keep the samples consistent, a stratified screening method is used. That is, the positive and negative samples are divided by tenfold separately, and the divided data is then further combined into training and testing samples.

In addition, to evaluate the generalization ability of the optimization algorithm, the KNN classifier used in the algorithm for optimizing the fitness was avoided when selecting the classifier, and the SVM classifier and RF classifier were chosen instead. From the tables, we found that the classifiers applying the improved optimization algorithm

Pearson-BPSO all outperformed the other optimization algorithms.

Although there is still much work to be done in this study, we can obtain the following conclusion from the experimental results. First, it is meaningful and feasible that the eNose device can identify hepatocellular carcinoma from healthy controls. However, there are still many difficulties to overcome in the clinical application of eNose. Secondly, the improved feature optimization algorithm is indeed beneficial to improve the detection performance to some extent, as shown, by 10 times the average results of the two different classifiers. From Tables 2 and 3, we can find both classifiers applying the improved optimization algorithm Pearson-BPSO all outperforming the other optimization algorithms.

5. Conclusion

Different gases have different response curves, which lead to different sensor measurement signals and different exhalation signals. After collecting the human exhalation signal through an electronic nose, it is difficult to distinguish the data of hepatocellular carcinoma patients from other health data intuitively and effectively because of the great differences between individuals and at different times. The extraction of waveform features is helpful in finding more potential information. However, high-dimensional features may lead to the decline of classification accuracy and slow calculation, so feature optimization has become a research hotspot.

In the paper, an improved feature optimization algorithm, Pearson-BPSO, is proposed based on binary particle swarm optimization (BPSO) for the “two-classification” task of distinguishing hepatocellular carcinoma patients and healthy people by breath. Based on the Pearson coefficient of the relationship between the quantifiable feature and the label, the algorithm preliminarily sifts the features, optimizes the feature set to minimize the KNN classification recognition rate and feature dimension, improves the classification accuracy of the algorithm, and reduces the amount of data. Compared with the traditional BPSO algorithm and PCA algorithm, this algorithm improves the classification performance to a certain extent and is conducive to improving the classification accuracy and detection speed of electronic nose detection.

6. Future Work

In the next step, we can further analyze the correlation between the features and then effectively combine it with this method to search the optimal subset more directionally and improve the accuracy of the classifier. We will also apply more advanced algorithms and continuously optimize the improved feature optimization method Pearson-BPSO to achieve more stable and better classification results [20].

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was financially supported by the Key Program of National Natural Science Foundation of China (Grant no. 81830052) and Shanghai Municipal Education Commission (Class II Plateau Disciplinary Construction Program of Medical Technology of SUMHS, 2018–2020). 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] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] L. James and Global burden of disease cancer collaboration, “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: a systematic analysis for the global burden of disease study,” *Journal JAMA Oncology*, vol. 4, no. 11, pp. 1553–1568, 2018.
- [3] A. S. Bannaga, H. Tyagi, E. Daulton, J. A. Covington, and R. P. Arasaramnam, “Exploratory study using urinary volatile organic compounds for the detection of hepatocellular carcinoma,” *Molecules*, vol. 26, no. 9, p. 2447, 2021.
- [4] M.-A. Galen, A.-M. Lou-Anne, G. David et al., “Breath metabolomics provides an accurate and noninvasive approach for screening cirrhosis, primary, and secondary liver tumor,” *Journal of Hepatology communications*, vol. 4, no. 7, pp. 1041–1055, 2020.
- [5] X. Yu, D. Zhan, L. Liu, H. Lv, L. Xu, and J. Du, “A privacy-preserving cross-domain healthcare wearables recommendation algorithm based on domain-dependent and domain-independent feature fusion,” *IEEE Journal of Biomedical and Health Informatics*, vol. 9, pp. 89344–89359, 2021.
- [6] C. D. TONG and X. H. SHI, “Mutual information based on PCA algorithm with application in process monitoring,” *Journal of CIESC*, vol. 66, no. 10, pp. 4101–4106, 2015.
- [7] Y. J. Lü and D. Y. Li, “Feature selection algorithm based on label correlation,” *Pattern Recognition and Artificial Intelligence*, vol. 33, no. 8, pp. 716–722, 2020.
- [8] B. Liu, M. D. Cai, and Y. C. Bo, “A feature extraction and classification algorithm based on PSO-CSP-SVM for motor imagery EEG signals,” *Journal of Central South University (Science and Technology)*, vol. 51, no. 10, pp. 2855–2866, 2020.
- [9] L. MA, C. W. LU, Q. H. GU, and S. L. RUAN, “Particle swarm optimization with search operator of improved pigeon-inspired algorithm,” *Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 10, pp. 909–920, 2018.
- [10] J. LIN, L. Xu, and L. LIU, “Feature selection method based on SVM-RFE and particle swarm optimization,” *Journal of Chinese Computer Systems*, vol. 36, no. 8, pp. 1865–1868, 2015.
- [11] C. G. Waltman, T. A. T. Marcelissen, and J. G. H. van Roermund, “Exhaled-breath testing for prostate cancer based on volatile organic compound profiling using an

- electronic nose device (aeonose): a preliminary report,” *European Urology Focus*, vol. 6, no. 6, pp. 1220–1225, 2020.
- [12] L. J. Hao and G. Huang, “Research on non-invasive detection method of liver cancer by respiratory based on electronic nose,” *Journal of Transducer and Microsystem Technologies*, vol. 39, no. 4, pp. 46–48, 2020.
 - [13] V. Andreas, W. Katharina, K. Tobias et al., “Detecting cannabis use on the human skin surface via an electronic nose system,” *Journal of Sensors*, vol. 14, pp. 13256–13272, 2014.
 - [14] J. Y. Xie, Z. Z. Wu, and Q. Q. Zheng, “An adaptive 2D feature selection algorithm based on information gain and person correlation coefficient,” *Journal of Shaanxi Normal University (Natural Science Edition)*, vol. 48, no. 6, pp. 69–81, 2020.
 - [15] Z. B. Zhu, J. F. Tao, and H. L. Ge, “Passive sonar target classification and recognition technique based on BPSO-KNN algorithm,” *Journal of Technical Acoustics*, vol. 38, no. 2, pp. 219–223, 2019.
 - [16] J. D. Han, L. Sun, and S. L. Wang, “Android malware application detection method based on BPSO-NB,” *Journal of Computers and Modernization*, vol. 4, pp. 109–113, 2017.
 - [17] B. Swanson, L. Fogg, W. Julion, and M. T. Arrieta, “Electronic nose analysis of exhaled breath volatiles to identify lung cancer cases: a systematic review,” *Journal of the Association of Nurses in AIDS Care*, vol. 31, no. 1, pp. 71–79, 2020.
 - [18] G. Danila, C. Sare, D. A. Mario, R. Veronica, S. Antonio, and B. Maurizia, “An E-nose for the monitoring of severe liver impairment a preliminary study,” *Journal of Sensor*, vol. 19, no. 17, Article ID 3656, 2019.
 - [19] M. Yu, T. Quan, Q. Peng, X. Yu, and L. Liu, “A model-based collaborate filtering algorithm based on stacked auto encoder,” *Neural Computing & Applications*, vol. 290, 2021.
 - [20] W. Hamdy, I. Elansary, A. Darwish, and A. E. Hassanien, “An optimized classification model for COVID-19 pandemic based on convolutional neural networks and particle swarm optimization algorithm,” *Journal of Digital Transformation and Emerging*, vol. 3, 2021.