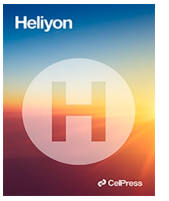
Heliyon 9 (2023) e13633

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

An improved AdaBoost algorithm for identification of lung cancer based on electronic nose

Lijun Hao [a,c,1,*], Gang Huang [b,a,1]

[a] School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China
[b] Shanghai Key Laboratory of Molecular Imaging, Jiading District Central Hospital Affiliated Shanghai University of Medicine and Health Sciences, Shanghai, 201318, China
[c] Medical Instrumentation College, Shanghai University of Medicine and Health Sciences, Shanghai, 201318, China

ARTICLE INFO

Keywords:
Electronic nose
Lung cancer
Enhancing learning
AdaBoost
K-fold cross-validation
Voting
GA

ABSTRACT

The research developed an improved intelligent enhancement learning algorithm based on AdaBoost, that can be applied for lung cancer breath detection by the electronic nose (eNose). First, collected the breath signals from volunteers by eNose, including healthy individuals and people who had lung cancer. Additionally, the signals' features were extracted and optimized. Then, multi sub-classifiers were obtained, and their coefficients were derived from the training error. To improve generalization performance, K-fold cross-validation was used when constructing each sub-classifier. The prediction results of a sub-classifier on the test set were then achieved by the voting method. Thus, an improved AdaBoost classifier would be built through heterogeneous integration. The results shows that the average precision of the improved algorithm classifier for distinguishing between people with lung cancer and healthy individuals could reach 98.47%, with 98.33% sensitivity and 97% specificity. And in 100 independent and randomized tests, the coefficient of variation of the classifier's performance hardly exceeded 4%. Compared with other integrated algorithms, the generalization and stability of the improved algorithm classifier are more superior. It is clear that the improved AdaBoost algorithm may help screen out lung cancer more comprehensively. Additionally, it will significantly advance the use of eNose in the early identification of lung cancer.

1. Introduction

Lung cancer is currently one of the most prevalent and deadly cancers worldwide. According to IARC(International Agency for Research on Cancer), the number of new lung cancer cases in the world in 2020 is 2.2 million, and the number of new lung cancer deaths is 1.8 million [1]. Studies [2–4] have shown that the five-year survival rate of patients with lung cancer in the middle stage is about 60%, while that of patients with advanced lung cancer is even less than 5%. However, the five-year survival rate of early-stage lung cancer patients can grow to more than 90% after treatment [5]. Therefore, early diagnosis of lung cancer is very important.

At present, there are many techniques for screening of lung cancer, including X-ray, computed tomography (CT), positron emission tomography (PET), and magnetic resonance tomography (MRI). However, each has its own disadvantages, such as the radiation risk of

* Corresponding author. School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China.
E-mail address: haolj@sumhs.edu.cn (L. Hao).
[1] The authors contributed equally to the work.

https://doi.org/10.1016/j.heliyon.2023.e13633
Received 26 August 2022; Received in revised form 1 February 2023; Accepted 6 February 2023
Available online 21 February 2023

X-ray; the radiation risk and high false positive rate of CT; the low spatial resolution of MRI. The combined PET/CT diagnosis technique can characterize and stage lung cancer, but it is expensive and cannot play a role in early lung cancer screening [6].

eNose is a smart instrument developed in recent years. It is designed to detect and discriminate among complex odors using an array of sensors. It is a completely noninvasive method and is virtually unlimited with respect to frequency, access, and cost [7,8]. Because it can link specific breath volatile organic compounds (VOCs) or breath-prints (i.e., patterns of VOCs) to the health status [9]. De Vries et al. [10] in 2019 used an eNose to analyze VOCs from lung cancer patients to predict whether patients on immunotherapy will achieve objective remission. The accuracy of its prediction can be as high as 85%. However, eNose devices do not provide information about the specific breath VOCs composition, yet they identify a specific profile or "smell-print" of the overall composition of the breath.

Studies have shown the usefulness of eNose devices for the detection of lung cancer. Unlike traditional gas composition detection techniques, eNose devices build a mathematical diagnostic model to detect lung cancer by VOCs. To effectively distinguish the gas response of lung cancer patients and healthy individuals detected by eNose, researchers have designed a variety of algorithms. Hubers et al. [11] in 2014 applied a combination of principal component analysis (PCA), independent component analysis (ICA), and logistic regression analysis (LR), which could distinguish lung cancer patients from healthy individuals with 80% accuracy but only 48% specificity. Chen lu et al. [12] in 2015 applied logistic regression analysis to effectively distinguish two types of breath samples with 80.6% accuracy and 74% specificity, respectively. Dekel Shlomi et al. [13] 2017 applied a support vector machine (SVM) for the differentiation of two types of breath samples with 79.1% accuracy and specificity up to 88.9%. Maribel et al. [14] in 2019 applied PCA and Fisher's discriminant method to differentiate lung cancer patients from healthy individuals and the accuracy and specificity could be improved to 82.2% and 91%. However, the performance of these algorithms is not yet sufficient to meet clinical needs. In addition, these classifiers are based on small samples. However, its generalization performance has not been tested.

In the paper, an improved AdaBoost (ImAdaBoost) algorithm has been proposed to construct a classifier that can distinguish the breath of lung cancer patients and healthy individuals. Based on the traditional AdaBoost algorithm, we applied the K-fold cross-validation and the voting method to the innovative design of sub-classifiers [15]. And an integrated and enhanced classifier was formed by weighting multiple heterogeneous sub-classifiers [16,17]. Then, we designed experiments to compare the performance of the improved algorithm with other algorithms. Furthermore, experiments were designed to test the stability and generalization performance of the improved algorithm. Through many random tests, the performance fluctuation curve was analyzed, and the stability and generalization performance of the algorithm was tested [18]. Finally, a new test sample set was collected and predicted by the proposed algorithm in the paper.

2. Material and method

2.1. The eNose device

In the present study, the eNose device (shown in Fig. 1) we applied was commercial equipment suitable for general gas detection. The device (ILD.3000, UST Sensors GmbH Company, Germany) is equipped with three electrochemical sensors, a metal oxide semiconductor, and a controllable temperature sensor [19]. Fig. (a) is the hardware system of the device and Fig. (b) shows the collection interface of the device.

The three gas sensors in the device are the core of the device. They are the GGS1000 series sensor, which is sensitive to combustible gases; the GGS3000 series sensor, which can detect hydrocarbons, especially for C1, C2 …...C8; and the GGS7000 series sensor, which can detect NO_2 [20]. The controllable temperature sensor Rt is designed to provide a suitable temperature environment to improve the

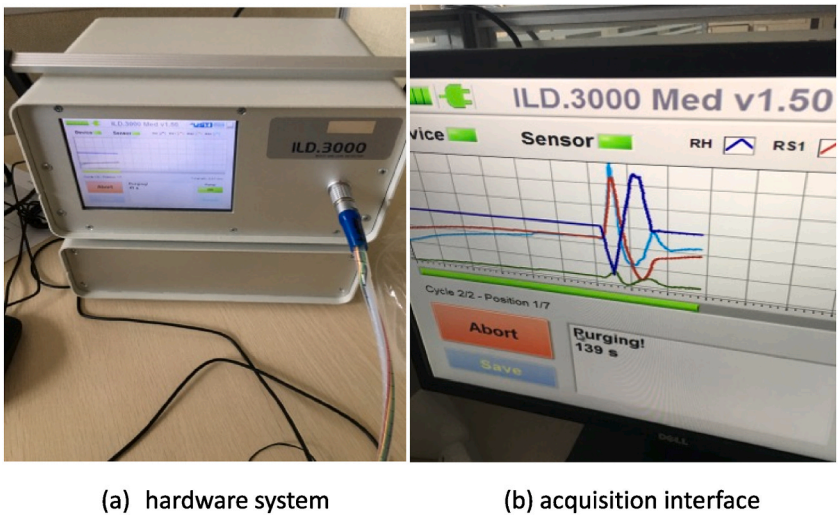


Fig. 1. The eNose device.

response-ability of the sensor to the gas. The temperature variation range is from 200 °C to 400 °C [19].

The whole process to collect breath data from a volunteer with the eNose device goes through 5 stages, including flushing the sensor, patient measurement, and so on. After the system has warmed up, the total time is 17 min and 50 s. During the collection process, only disposable mouthpieces were used and no interventional devices were used, which did not cause any harm to the human body.

2.2. Data acquisition and pre-processing

The training data set is the breath data of 142 volunteers including 91 lung cancer patients and 51 healthy individuals. The information of all volunteers was recorded anonymously. The newly collected test samples include the breath data of 12 lung cancer patients and 10 healthy individuals.

The inclusion criteria for volunteers for both groups comprised: (1) individuals >18 years, able to understand and read the consent form; (2) patients who have primary lung cancer, no other evidence of metastatic cancer; (3) no history of smoking and alcohol abuse in the last three months; (4) in the fasting state. The basic information about the volunteers is shown in Table 1.

The study was approved by the ethical committee of Shanghai Changzheng Hospital (Approval file number 2018SL029). All volunteers were informed of the aim of the study. Instructions were given and verbal consent was obtained from each volunteer before the collection of breath data.

The three curves as shown in Fig. 2 are the three exhaled gas response signals simultaneously collected by the eNose device, which can be noted as $\boldsymbol{y_A}$, $\boldsymbol{y_B}$, and $\boldsymbol{y_C}$.

As shown in Fig. 2, the signals collected by different sensors vary greatly, and the signals collected by the same sensor also vary greatly. To facilitate data comparison and statistical analysis, all signals are normalized here. In normalization, $\max(y_l)$ and $\min(y_l)$ used are respectively taken from the maximum and minimum values of signals collected by all volunteers on all sensors. Then, the signal collected by each sensor can be transformed into the range of [0, 1] through Eq. (1).

$$y_l^{(i)} = \frac{y_k^{(i)} - \min(y_l)}{\max(y_l) - \min(y_l)} \tag{1}$$

where, i can be taken as A, B and C, representing three sensors respectively; i denotes the ith sample collected by a certain sensor. And $\min(y_l)$ and $\max(y_l)$ represent the minimum and maximum values of all sample signals collected by the same sensor, respectively.

After normalizing the signals, their features such as time domain, frequency domain, and statistics are further extracted. In the study, 14 time-domain features (maximum value and corresponding position, minimum value and corresponding position, mean, peak-to-peak value, variance, standard deviation, waveform factor, pulse factor, peak factor, margin factor, and area), 14 frequency-domain features (center of gravity frequency, frequency variance, root mean square difference, frequency spectrum and power spectrum calculated by various methods) and 10 statistical features (polar deviation, median, quantile, plurality, coefficient of variation, skewness, kurtosis, autocorrelation coefficient and information entropy, and correlation between any two sensor signals). By combining all the features extracted from the three sensor signals, a set of high-dimensional (1557) features corresponding to one sample could be obtained.

2.3. Feature optimization

To avoid the dimension disaster and improve the performance of the classifier, PCA and genetic algorithm (GA) were applied to feature optimization, respectively.

PCA is a common feature dimensionality reduction method [21]. Its goal is to map high-dimensional data into low-dimensional space through some kind of linear projection, that is, to replace the original n features with a smaller number of M features. It is expected that the variance of the data is the largest in the projected dimension so that the new M features are not related to each other as much as possible.

GA is a feature optimization algorithm based on natural selection that has emerged in recent years [22]. Feature selection based on GA can be realized through four steps: generation of the initial population, feature selection according to the fitness function, crossover, and mutation. In the study, the fitness function was constructed based on the error rate of the Bayesian classifier. The crossover probability was set to a random number, the mutation probability was set to 0.5, and the number of iterations was set to 20.

Table 1
The basic information of the volunteers.

	Training samples			Test samples		
	Male	Female	Average age (years) ± SD	Male	Female	Average age (years) ± SD
Lung cancer	63	28	55.8 ± 12.1	11	1	62.7 ± 6.6
Control	34	17	51.6 ± 14.6	6	4	55.9 ± 11.2

*In the table, SD means standard deviation.

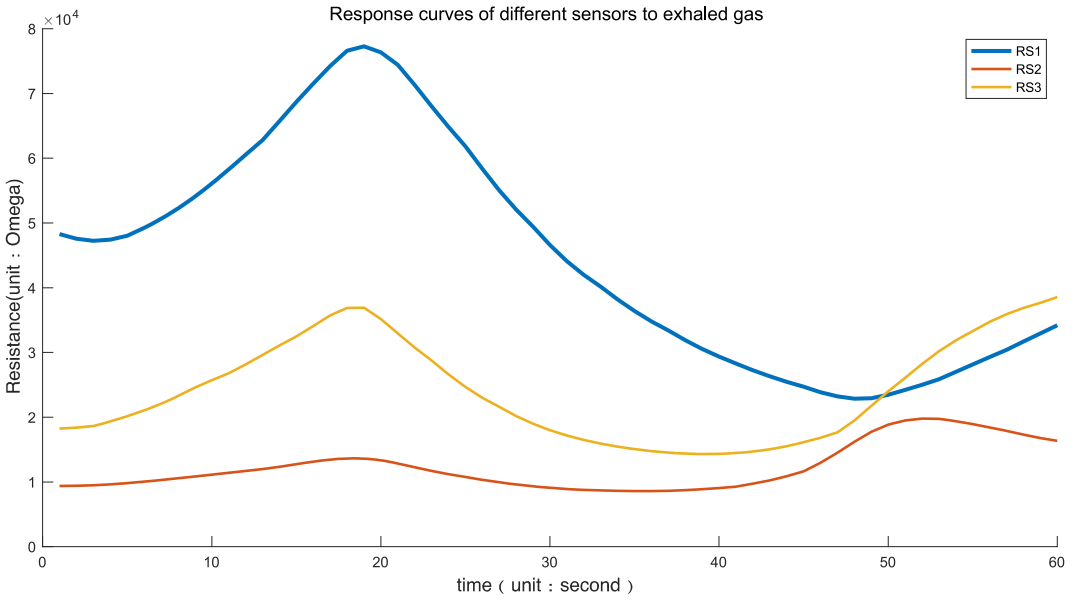


Fig. 2. Breath signals collected by Sensors.

2.4. *Design of classifier*

Lung cancer detection based on VOCs detected by eNose is essentially a classification problem. In the study, a highly robust classification model is first been constructed by the expiratory signals of lung cancer patients and healthy individuals. When a new sample is obtained, it can be fed as input to the classifier to determine whether it is a lung cancer patient or not.

The ensemble learning algorithm is a combinatorial algorithm [23]. The algorithm first constructs a series of sub-classifiers (weak learners) and then aggregates the weak learners by using different strategies to make an overall prediction. The error rate of this aggregated model will be reduced. Moreover, due to the mutual inhibition between the models, the generalization performance is improved and the overfitting phenomenon can be avoided.

At present, the representative integration algorithms include bagging, stacking, and boosting. The bagging algorithm constructs multiple classifiers based on different training samples to predict new samples respectively and obtains the final prediction results by voting [24]. The stacking model is a two-layer hierarchical integrated model framework. The basic idea is to fuse the prediction results of several single models through one model, to reduce the generalization error of a single model [25]. In the first layer, cross-validation is used to predict and generate a new training set as well as a new test set. In the second layer, the new training set is used to build a classifier, and the new test set is predicted to get the final prediction results. The representative boosting algorithm is the AdaBoost algorithm. The algorithm first trains a sub-classifier and adjusts the data distribution according to the training error. Then a new sub-classifier is constructed based on the new data distribution. Repeat this process until the required number of sub-classifiers are

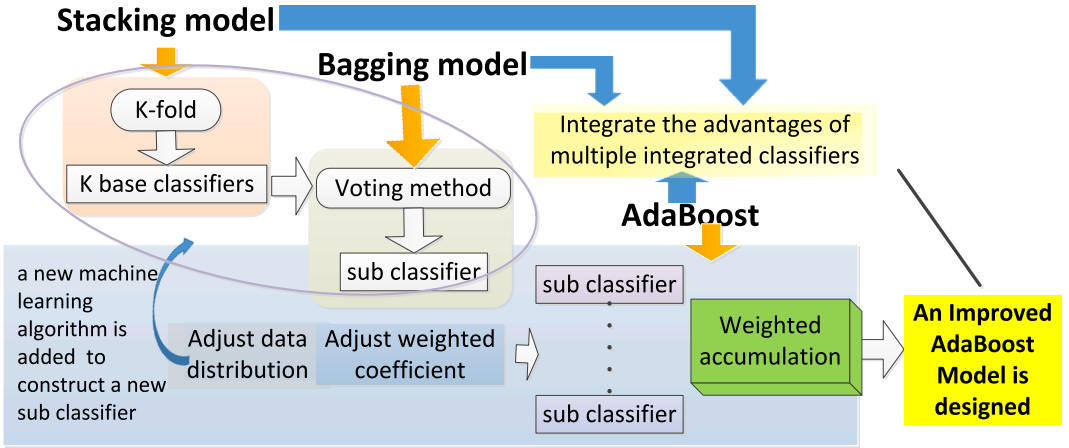


Fig. 3. Design of an ImAdaBoost classification model based on multiple integrated models.

obtained. Finally, these sub-classifiers are weighted and combined to obtain the final prediction results. The three algorithms have their advantages and disadvantages. The bagging algorithm is mainly used to improve generalization performance and solve over-fitting problems. In the stacking model, because the training data used in the two layers are different, the results are more robust. In contrast, the AdaBoost algorithm can improve training accuracy and reduces underfitting.

For small samples, to obtain a high-precision classifier with high generalization performance, we try to fuse the above three commonly used integrated classifiers in our study.

Just as shown in Fig. 3, First, K-fold cross-validation is used to train multiple base learners, and then the sub-classifier is obtained based on the voting method. Furthermore, the coefficients of the sub-classifier are obtained based on AdaBoost theory, and the data distribution of training samples is adjusted. In a new round of training, new machine learning algorithms will be added to train new sub-classifiers After reaching the preset training times, the training is stopped, and all sub-classifiers are weighted and combined to obtain the final prediction result.

3. Theory

The idea of the AdaBoost algorithm is to combine the outputs of multiple "weak" classifiers (sub-classifier) in a weighted manner to produce an effective classification. Its adaptability lies in that the samples misclassified by the previous sub-classifier will be strengthened by a higher weight, and the weighted samples will be used to train the next sub-classifier again. However, in the paper, the design of the sub-classifier has been improved based on cross-validation, voting method, and multiple different classification algorithms.

3.1. Improved design of sub-classifiers

As shown in Fig. 4, the construction of sub-classifiers based on the K-fold cross-validation method consists of three specific steps. In the figure, the training set is denoted as 'TrainSet' and the test set is denoted as 'TestSet'.

In the first step, K-fold cross-validation is applied to the training data ('TrainSet') and k basic classifiers are obtained. First, divide the 'TrainSet' into k sets randomly. Then select one group as the test sample "TestData" and the other (k-1) groups as the training sample "TrainData". Afterward, based on a classification algorithm, a group of classifiers can be obtained according to K-fold cross-validation. At the same time, the basic classifier can predict the prediction of the test sample "TestData" and the test set "TestSet" respectively. Finally, k different base classifiers are obtained. And predictions of k different test sample "TestData" sets will be obtained by these different classifiers. In fact, these k different test samples constitute the training set. In the second step, combine the predictive

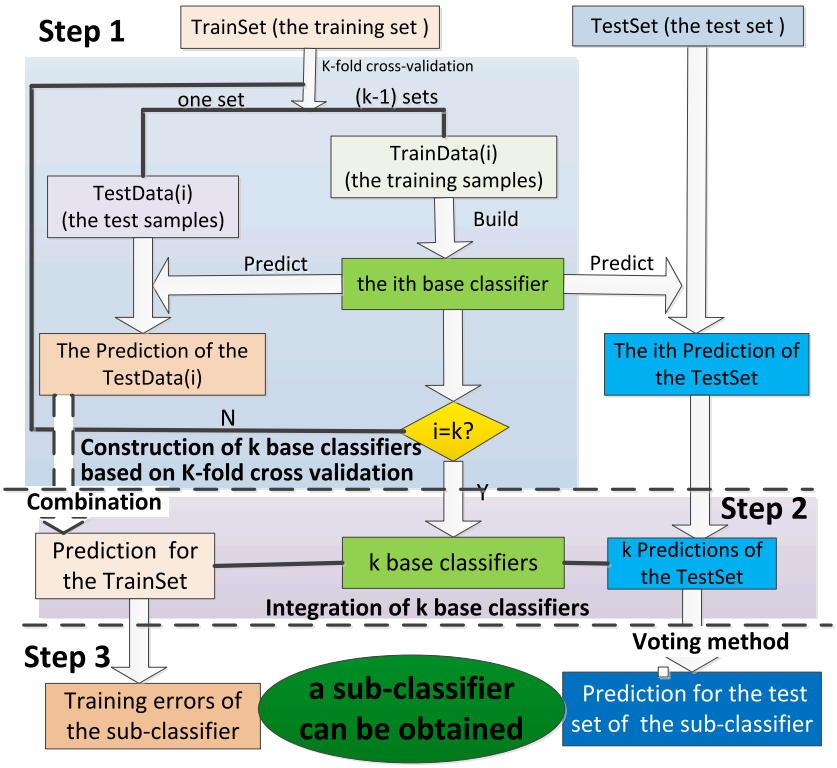


Fig. 4. Construction of a sub-classifier based on K-fold cross-validation and the voting method.

values of the test set to obtain the predictive values of the training samples. For the test set "TestSet", we will also get k-group prediction results. In the third step, we can determine the final prediction result of the sub-classifier to the test set by voting. At the same time, the prediction error of the sub-classifier to the training samples can also be obtained. A sub-classifier is thus constructed.

Following the process described above, more sub-classifiers will be designed. Meanwhile, to combine the advantages of multiple classification algorithms, different algorithms can be selected when building different sub-classifiers. Thus, a plurality of heterogeneous sub-classifiers can be constructed sequentially.

3.2. Building integrated classifier based on AdaBoost

The core of the AdaBoost algorithm is the weighted combination of multiple sub-classifiers [25]. In Section 3.1, we get the training sample error of each sub-classifier. And based on the training error, the weighting coefficients of the sub-classifiers can be obtained and the sample weight will be adjusted. The enhanced classifier is finally obtained by weighting the set of these sub-classifiers.

Compared with the traditional AdaBoost algorithm, the results of the sub-classifier are more complicated to get and need to be obtained from multiple base classifiers by voting. The core steps are as follows.

Step 1, select a machine learning algorithm and divide the training set by K-fold as described in 3.1. One fold of data is selected as the test sample in turn, and k training sessions are performed to obtain k different base classifiers based on the chosen algorithm one by one, forming a base classifier group, which is noted as the i-th sub-classifier.

The predicted value $g_i(j)$ of the sub-classifier group for the training set is obtained by Eq. (2).

$$g_i(j) = \left[g_{i,}^1, g_i^2 \ldots, g_i^k\right] \quad (i = 1 \ldots T, j = 1 \ldots m) \tag{2}$$

where, k is the number of base classifiers. $g_{i,}^1, g_i^2 \ldots, g_i^k$ are the predictions of the k base classifiers for one fold of data in the training set, respectively, which are combined to form the i-th sub-classifier for all samples in the training set. T is the number of sub-classifiers. And m is the number of the training samples.

The training error rate e_i corresponding to the base classifier group then can be calculated by Eq. (3).

$$e_i = \sum_{\mathrm{k}} \mathrm{D}_{\mathrm{i}}^{\mathrm{j}}(\mathrm{k}) \quad \mathrm{k} = 1, 2, \ldots, \mathrm{m}\ \left(g_i(j) \neq y_j\right) \tag{3}$$

where, k traverses all samples in the training set where the predicted value does not match the true value. $\mathrm{D}_{\mathrm{i}}^{\mathrm{j}}(\mathrm{k})$ is the distribution coefficient of sample k in the process of constructing the i-th sub-classifier. Its initial valuer is $D_i^j = \frac{1}{m}$. For the binary classification problem, the error rate is essentially the sum of the weights of these samples.

Step 2, calculate the weight coefficient α_i of the i-th sub-classifier using the exponential function as the loss function by Eq. (4) [26].

$$\alpha_i = \frac{1}{2}\log\left(\frac{1 - e_i}{e_i}\right) \tag{4}$$

Further, use Eq. (5) to adjust the sample weights D_{i+1}^j for training the (i+1)-th sub-classifier.

$$\begin{cases} D_{i+1}^j = D_i^j e^{-\alpha_i y_j g_i(j)} / Dsum \\ Dsum = \sum_{j=1}^{m} D_{i+1}^j \end{cases} \tag{5}$$

where, D_i^j is the training set sample weight corresponding to the ith base classifier group, while D_{i+1}^j is the adjusted training sample weight corresponding to the (i+1)th sub-classifier. α_i is the weight coefficient of the i-th sub-classifier, while y_j and $g_i(\mathrm{j})$ are the true and predicted values of samples j, respectively. *Dsum* is the normalization factor.

Step 3, get the prediction results of each sub-classifier for the test set based on Eq. (6) by the voting method.

$$h_i(l) = \begin{cases} \frac{1}{p}\sum\left(h_i^t(j) < 0.5\right), p > \frac{k}{2} \\ \frac{1}{k-p}\sum\left(h_i^t(j) \geq 0.5\right), \mathrm{p} \leq \frac{k}{2} \end{cases} \tag{6}$$

where, $h_i^t(j)$ is the predicted value of the base classifier t for the sample j to be tested. p is the number of k-base classifiers for which the predicted value of the sample to be tested is less than 0.5. If p is more than half of the number of base classifiers, the output is the average of all predicted probabilities less than 0.5; otherwise, the output probability is the average of all predicted probabilities that greater than or equal to 0.5.

Once the predicted values of T sub-classifiers for the testing samples are obtained, the final predicted values of the classifiers can be achieved according to Eq. (7) by weighing them.

$$(l) = sign\left[\sum_{i=1}^{T} \alpha_i \bullet h_i(l)\right] \quad l = 1, 2 \ldots n \ , \ n \text{ means the number of the test set} \tag{7}$$

where, $h_i(l)$ is a set of predictions for the test set of the i-th sub-classifier. α_i is the weight coefficient of the i-th sub-classifier. $H(l)$ is the prediction result of the integrated classifier.

4. Experiment and results

Identifying lung cancer patients through exhalation is essentially a classification problem. In the study, we expect to distinguish lung cancer patients from healthy individuals by constructing a model. Therefore, the lung cancer group and the healthy group were labeled as 1 and 0 respectively.

To quantitatively evaluate the application of the improved algorithm proposed in this paper for lung cancer detection, a series of tests and experiments were designed.

First, the performance of this algorithm was verified by the cardiovascular public dataset.

Then, collected and pre-processed breath training samples (including 91 lung cancer patients and 51 controls). And features were further extracted from the signals and optimized with PCA and GA, respectively. Thus, multiple feature sets of different dimensions were obtained.

Furthermore, the training samples were randomly divided into a training set and a test set according to a certain ratio. The classifier was constructed using the training set. In the study, six different algorithms, such as SVM, k-Nearest Neighbor method (KNN), random forest (RF), LR, linear discriminant analysis (LDA), and back propagation neural network (BPNN), were selected to design sub-classifiers. In addition, five-fold cross-validation was used to construct a sub-classifier. The specific method is described in Section 3.

In order to get more objective evaluation results, the training samples are randomly divided multiple times. Each division is relatively independent. Different classifiers were constructed based on different training sets. And the average performance of these different classifiers is taken as an evaluation index.

Finally, collect a new set of test samples. Constructed a classifier using the original training set and make predictions on it. Run 20 times and take the average value as the prediction performance.

4.1. Testing of the ImAdaBoost algorithm

The improved algorithm designed in the paper is to improve the performance of classifier detection under small samples. Its performance was tested and compared with other algorithms by the open cardiovascular data set. In each test, 0.48% of data (about 316 pieces) was randomly extracted from the cardiovascular data set as the training set, and then 0.02% of data (about 130 pieces) was taken as the test set. And the training set and the test set were randomly generated and independent of each other. The process was repeated 20 times. The average performance of different algorithms [27–30] based on the cardiovascular public dataset was shown in Table 2.

Compared with the current popular depth learning algorithm, the overall performance of the algorithm is not good. But because of its fast running speed and low requirements for computer configuration, it is helpful to be used in universal screening to quickly assist in disease identification. As seen in Table 2, the accuracy, specificity, and precision of the improved algorithm were obviously better than those of other algorithms in small samples.

After verifying the effectiveness of the algorithm, we further tested the classification performance, generalization performance, and robustness of the algorithm for the identification of lung cancer by breath.

4.2. Determination of feature optimization dimensions and optimization algorithm

Feature selection is an important first step in machine learning. It may optimize the effect and performance of the classification model. Therefore, we first applied the PCA algorithm and the GA algorithm to optimize features in different dimensions. And then, by comparing the average performance 20 times, the appropriate feature optimization algorithm and optimization dimension were

Table 2
Comparison of different algorithms based on the cardiovascular public dataset.

Classifier	Accuracy	Sensitivity	Specificity	Precision	F1-score
ImAdaBoost	71.88	65.92	77.84	75.10	70.07
AdaBoost	66.88	66.92	66.84	67.46	66.71
Bagging	64.96	66.30	63.61	64.82	65.21
Stacking	69.92	68.53	71.30	70.99	69.42
KNN	52.07	69.69	34.46	51.54	59.18
SVM	69.61	64.61	74.61	72.42	67.94
LDA	54.46	53.23	55.69	55.03	53.26
RF	68.84	68.46	69.23	69.46	68.66
LR	66.65	64.46	68.84	68.03	65.95
BP	51.15	98.84	3.46	68.03	80.4

determined. In the test, 20% of the data (including 10 healthy individuals and 18 lung cancers) were taken as the test set and the remaining 80% as the training set.

Tables 3 and 4 show the average performance results after using different methods to optimize features into different dimensions. Ln the tables, Dim means the optimized feature dimensions; accuracy is the percentage of correct detections; sensitivity is the percentage of lung cancer patients that can be detected correctly; while specificity is the percentage of healthy individuals that can be detected correctly, and precision is the percentage of correct judgments as lung cancer. In disease detection, it is necessary to improve sensitivity while ensuring precision. While F1-score is the harmonic average of recall and precision, and it takes into account both the sensitivity and the precision so that the two can reach the highest level at the same time. And AUC is the area under the ROC curve. The larger the AUC, the better the classifier effect [30,31].

As can be seen from the two tables, regardless of the optimization algorithm, the AUC of the classifier was optimal when the feature dimension was optimized to 30. And it was clear that the performance of the classifier based on the GA algorithm was better than that of the PCA algorithm.

Therefore, in the following tests, we selected the GA algorithm to optimize the feature to 30 dimensions and further analyzed other performances of the improved algorithm.

4.3. Determination of the number of training samples

It has been proved that the ImAdaBoost algorithm can be applied to the construction of classifiers with small sample training sets. In the study, we continuously adjusted the ratio of training samples to test samples, expecting to find the optimal number of training samples. In addition, the change in the performance of the classifier could be tested when the number of detected samples increased.

In the experiment, the random division ratios of the training samples were continuously adjusted and tested separately. The ratios of the training set and test set were set as 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, and 9:1, respectively. And then 20 tests were performed in each case. The average performance was compared in Fig. 5.

As shown in Fig. 5. It can be seen that the performance of the classifier gradually becomes better as the number of samples in the training set increases. When the test samples exceeded the training samples, there was a significant decrease in the performance of the classifier as the test samples increased. In particular, the specificity decreased most significantly, with a change of about 30%. However, when the training samples accounted for 60% or more of the total samples, the test performance of the classifier was relatively stable. However, when the training sample accounted for 90%, all performances reached 100%, at which point overfitting may have occurred.

Therefore, in the subsequent evaluation of the classifier performance, we took 80% of the sample for the training sample and 20% for the test sample.

4.4. Classification performance

Further, we compared the performance of the improved AdaBoost algorithms with other integrated classification algorithms. The integrated algorithms include the Stacking model and the Bagging mode, just as shown in Figs. 6 and 7.

Fig. 6 is the 20 times average performance comparison of the four integrated classifiers. Among the five indicators, the improved algorithm had the best accuracy, sensitivity, and F1 score. They were 97.85%, 98.33%, and 98.34% respectively. Fig. 7 is the ROC curves of the four integrated algorithms [31]. And the average AUC of the four integrated algorithms (ImAdaBoost, traditional AdaBoost, Stacking model, and Bagging model)were 0.996,0.989,0.992, and 0.988. Although the precision of the improved algorithm was slightly lower than that of the stacking model, its sensitivity is much higher than that of the Stacking model. In the study, we expect to detect as much lung cancer as possible, which means that the sensitivity of the algorithm should be high. However, at the same time, we do not want the specificity of the algorithm to be so low as to cause unnecessary panic. Therefore, we try to enhance sensitivity as much as possible while ensuring precision. The comparison shows that the performance of the improved algorithm was optimal.

4.5. Generalization performance and robustness

Besides, evaluating whether an algorithm is effective depends not only on the performance but also on the generalization performance and robustness. In the study, the stability and generalization performance of the classifier was tested based on the fluctuation of performance changes in 100 tests. In each test, the training samples were divided randomly and independently.

We analyzed the volatility of the classifier performance by calculating the coefficient of variation of each performance over 100 tests. When the coefficient of variation is greater than 15%, it indicates that the performance of the classifier is unstable [32]. The

Table 3
Average performance comparison of the ImAdaBoost algorithm based on PCA.

Dim	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
10	74.64	77.22	70	82.74	79.44	0.845
20	68.57	63.88	77	85.8	70.51	0.855
30	76.07	70	87	91.26	78.29	0.901
40	77.85	78.88	76	85.57	81.72	0.857

Table 4
Average performance comparison of the ImAdaBoost algorithm based on GA.

Dim	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
10	91.42	93.88	87	92.99	93.37	0.926
20	91.42	92.77	89	94.17	93.25	0.958
30	97.86	98.33	97	98.47	98.34	0.996
40	92.14	98.88	80	90.09	94.25	0.95

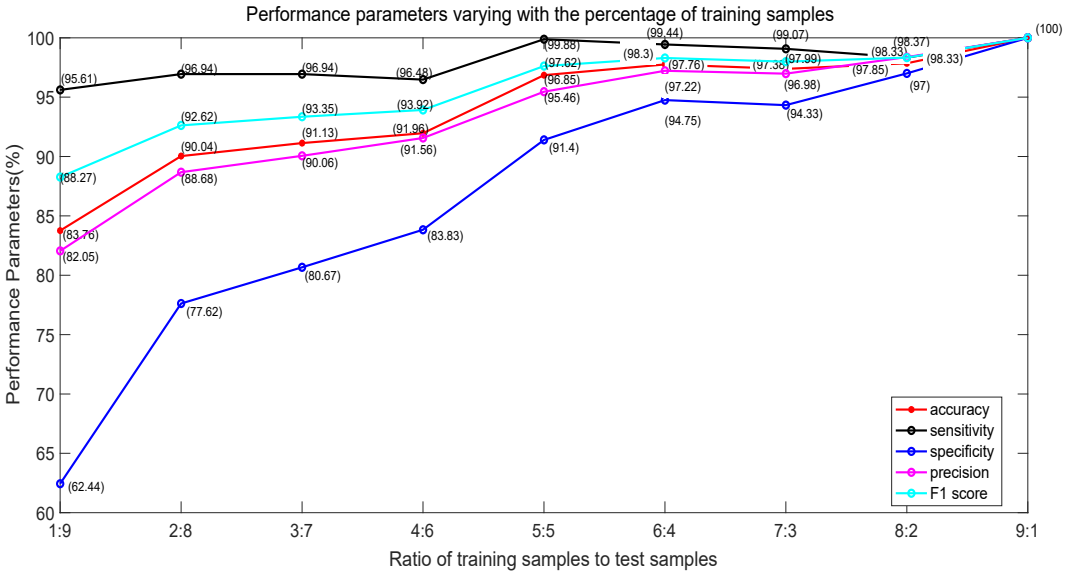


Fig. 5. Performance of classifiers with different division ratios for training data.

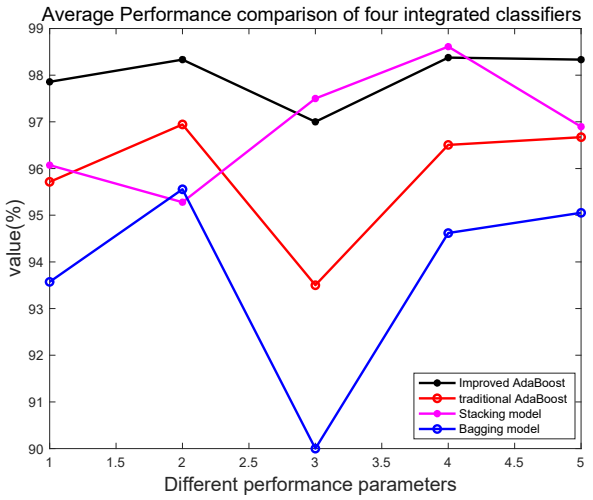


Fig. 6. Average performance comparison of the four integrated classifiers.

smaller the coefficient of variation, the more stable the performance of the classifier is, and the better the generalization performance and robustness of the classifier [33].

In the analysis of the previous section, it was found that the performance of the improved algorithm and the stacking model was significantly better than other integrated classifiers. Table 5 shows the comparison of the coefficients of variation of performance over 100 tests when the two classifiers were applied separately. By comparing the coefficients, it can be found that the classifier based on the ImAdaBoost algorithm had more stable performance. The generalization performance and robustness of the improved algorithm were better than the stacking model.

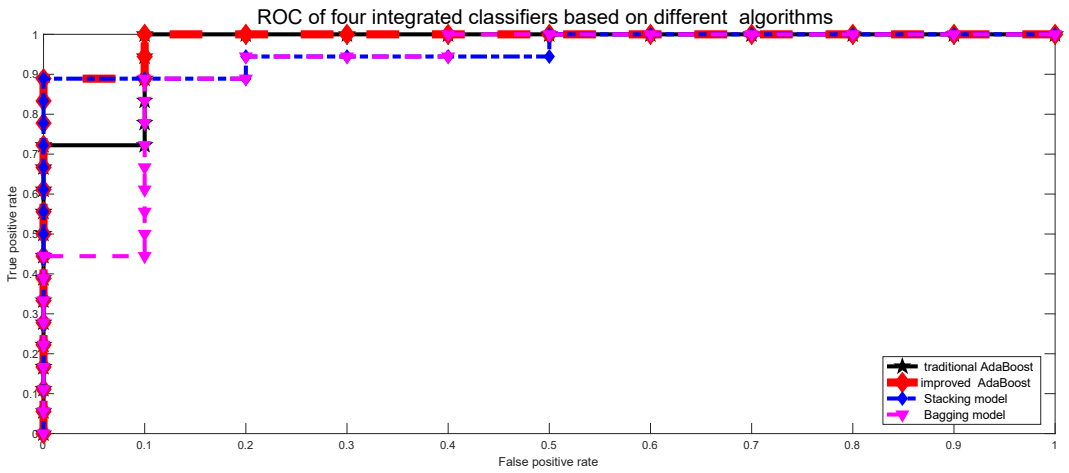


Fig. 7. ROC of the four integrated classifiers.

Table 5
Coefficients of variation of each performance parameter in different classifiers.

classifier	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
ImAdaBoost	2.657	1.891	6.483	3.269	2.005	0.35
Stacking model	2.803	1.827	7.723	3.781	2.078	0.32

4.6. Classifier predictions for the new tests

In order to test the classifier in practice, a new set of test samples was collected. The samples consisted of 9 healthy individuals and 12 lung cancers. The classifier was constructed by randomly using 80% of the training data as the training set. To get more objective results, 20 tests are randomly performed. The training samples were extracted randomly and independently in multiple tests. The average classification discriminatory performance is shown in Table 6. The statistics of false positives and false negatives in 20 tests are given in Fig. 8.

As can be seen from Fig. 8(a), the number of false negatives that occurred was smaller than that of the traditional AdaBoost algorithm and the bagging model. The ability of the improved algorithm to detect lung cancer is comparable to the stacking algorithm. However, the frequency of false positives is higher than that of the traditional AdaBoost algorithm and the stacking model, as shown in Fig. 8(b). The improved algorithm has a weaker ability to detect healthy individuals. However, analyzing the fluctuation changes of AUC, as shown in Fig. 9, it can be found that the performance of the improved algorithm is more stable, with a coefficient of variation of 1.72.

Probably due to the small number of test samples, the advantage of the improved algorithm in identifying lung cancer was not reflected. In the next work, we will test the performance of the improved algorithm by obtaining more samples.

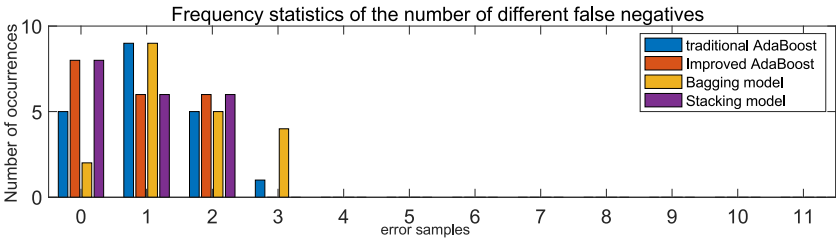
5. Discussion

The eNose device is a new diagnostic instrument developed in recent years, which can be applied in the fields of food testing, environmental monitoring, and medical diagnosis. The device monitors and diagnoses human diseases by collecting VOCs in human breath, which has the advantages of being non-invasive, simple operation, and low examination cost. However, due to the complexity of human metabolism and the diversity of diseases, as well as the impact of different training samples on the performance of classifiers, eNose has not been widely used clinically in disease detection. To improve the performance of the detection algorithm, researchers have focused on three aspects: first, to improve the hardware acquisition equipment, find the characteristic gas related to diseases and optimize the sensors; second, to steadily collect as many samples as possible to provide a basis for the universality of the algorithm; third, to innovate and improve the algorithm to build a more intelligent and robust detection algorithm.

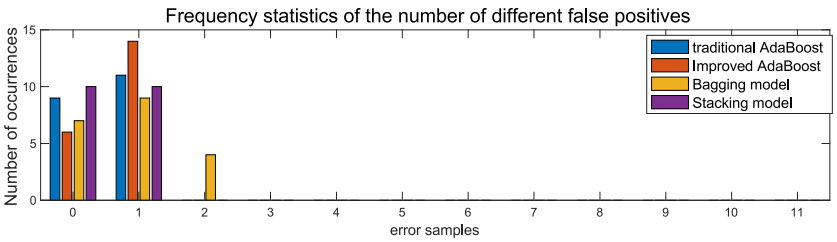
Differentiating and classifying the breath data of lung cancer patients and healthy individuals is the core work of intelligent lung

Table 6
Average performance in 20 tests (%).

accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
92.38	92.5	92.22	94.27	93.17	0.984



(a) Frequency statistics of the false negatives number



(b) Frequency statistics of the false positives number

Fig. 8. Statistics of the number of detection errors for new samples.

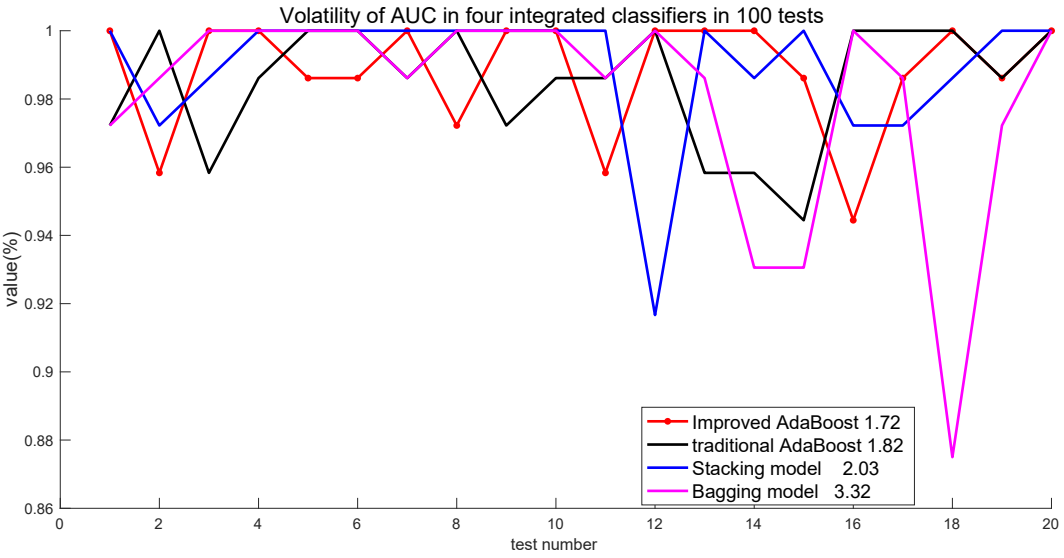


Fig. 9. AUC fluctuations in 20 new sample tests.

cancer detection. The human metabolic mechanism analysis shows that lung cancer patients' exhaled gas composition will change due to the pathology [34]. However, the association of this information, which is implicit in the non-stationary exhalation signal, with the disease is not very obvious. Therefore, it is not an easy task to tap the intrinsic link between the expiratory signal and the disease and build a corresponding discriminative model. The acquisition of large data samples facilitates the construction of universally adapted classifiers, but it is not easy. When the number of samples is certain, it is important to extract more features from the existing collected sample waveforms. But, too many features may also lead to degradation of classification accuracy and slowdown of computation, so optimization of features is also needed. In the study, the runtime of the classifier was greatly improved after sample feature optimization compared to the original feature set. The time for 20 runs was reduced from more than 2 h to about 7 min. In addition, it can also be found from the experimental results in Section 4.2 that changing the feature optimization algorithm and the optimization dimension results in significant differences in the performance of the classifier. However, only two feature optimization algorithms, PCA and GA, were applied in the paper. Whether there are more convenient feature optimization algorithms (e.g., optimization algorithms that do not require exploring the optimal feature dimensions) needs further investigation [35].

In the study, we focus the work on the design and construction of a high-performance classifier.

An ImAdaBoost integrated learning algorithm was designed based on the traditional AdaBoost algorithm. In theory, in the two AdaBoost algorithms, the sub-classifiers are weighted to combine as enhanced classifiers. However, the construction of their sub-classifiers is different. In the ImAdaBoost algorithm, sub-classifiers were obtained based on the 5-fold evaluation and voting method. And in the traditional AdaBoost algorithm, sub-classifiers were trained based on the samples extracted by the bootstrap method.

To improve the performance of the classifier, we first used the integration idea to integrate the advantages of multiple machine learning algorithms. For integrated classifiers, sub-classifiers are their important components. In the study, heterogeneous sub-classifiers were constructed using six traditional machine learning algorithms. The integration process is also a process of continuously reducing the error rate. This has been verified with the cardiovascular public data set. As shown in Table 2, the prediction accuracy of the ImAdaBoost algorithm is better than that of the sub-classifiers. However, since only 12 attributes were collected in the public dataset, this may have resulted in the overall performance of the algorithm not being very good. However, it can be found that the sensitivity and specificity of this algorithm are weaker compared to some algorithms. However, its precision is optimal.

In identifying lung cancer patients and healthy individuals based on the breath signal, the algorithm has a somewhat lower specificity relative to the stacking algorithm, but its sensitivity is greatly improved. The result is that lung cancer patients are less likely to be missed, but healthy individuals are slightly more likely to be misdiagnosed. As an auxiliary screening tool, it is clearly more important to reduce missed diagnoses and more meaningful to screening for disease. Of course, the reason for this phenomenon could also be the small sample size of healthy people. The positive and negative samples are not balanced. In the next work, we will also further test this idea by increasing the number of healthy individuals.

Besides, the performance of the classifier is closely related to training samples. Evaluating whether an algorithm is effective depends not only on the performance but also on the stability of the algorithm [36,37]. To improve the robustness of the algorithm and enhance the generalization performance of the classifier, a K-fold cross-validation approach was then used and several base classifiers were trained with different data samples in turn to build a sub-classifier. Then in the study, the performance and stability of the improved algorithm were analyzed more systematically and comprehensively, which is rare in previous studies. According to the results of 100 random independent experiments, the stability of the classifier performance is even better than the stacking model, as shown in Table 5 in Section 4.5.

In the available studies [38], the accuracy of using eNose to identify lung cancer is around 85%. In comparison, the algorithm proposed in this paper greatly improves the ability of eNose to screen for lung cancer. However, the study in this paper can be further improved, such as enhancing the sample size and sample type and collecting more comprehensive sample information.

To further improve the performance of the breath discrimination algorithm, in future research, we will devote the following study: first, to conduct more collection of multicenter breath samples and reduce the impact of sample imbalance; second, to improve the construction of a highly robust breath discrimination classification algorithm based on small samples and to improve the generalization ability of the algorithm [15,36]; and third, to try the combination of breath detection and other detection techniques to mine multi-omics features to provide more meaningful sample features for the classification algorithm. Besides, we will also apply more feature optimization algorithms, such as the Relief feature optimization algorithm, to optimize the features to improve the performance of the classifier.

6. Conclusion

In the paper, we proposed an improved AdaBoost lung cancer breath algorithm based on integration and reinforcement theory for the task of distinguishing lung cancer patients from healthy individuals' breath samples by eNose. In the improved algorithm, a group of k-base classifiers was first obtained by K-fold cross-validation based on a group of training samples, and then a sub-classifier is obtained further by a voting method. When each sub-classifier was obtained, its weight coefficients would also be obtained synchronously according to the training error, and the data distribution of the training samples would be adjusted. Moreover, more heterogeneous sub-classifiers would be obtained based on multiple base classifier groups. Finally, these sub-classifiers were weighted together to get an integrated enhanced classifier.

In general, the algorithm not only improved the classifier's performance by integrating multiple heterogeneous classifiers, but also improved the classifier's generalization performance with k-fold cross-validation. Compared with the traditional algorithms, the algorithm improves the performance of the classifier to correctly identify lung cancer and healthy individuals. The proposed algorithm will help to promote the non-invasive lung cancer screening method by eNose, and promote the application of clinical disease auxiliary examination method based on eNose. However, to improve the robustness of the method, feature optimization algorithms and intelligent classification algorithms for small samples need to be further studied. To advance the algorithm to clinical applications, more samples of more quantity and types need to be collected, and more tests and confirmatory experiments need to be carried out.

Author contributions

These authors contributed equally to this work.

Hao lijun: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Huang Gang (Co-first author): Conceived and designed the experiments; Contributed materials and analysis tools; Funding Acquisition, and Supervision.

Funding statement

Gang Huang was supported by Construction project of Shanghai Key Laboratory of Molecular Imaging [18DZ2260400], the Key Program of National Natural Science Foundation of China [81830052].

Data availability statement

None.

Declaration of interest's statement

The authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e13633.

References

[1] J. Ferlay, et al., Cancer statistics for the year 2020: an overview, Int. J. Cancer 149 (4) (2021) 778–789.
[2] D. Sun, et al., OA08.03 the 5-year survival rate of postoperative non-small cell lung cancer patients with two different follow-up patterns, J. Thorac. Oncol. 16 (10) (2021) S860–S861.
[3] Q. Pei, et al., Artificial intelligence in clinical applications for lung cancer: diagnosis, treatment and prognosis, Clin. Chem. Lab. Med. 60 (12) (2022) 1974–1983.
[4] M. Zoair, et al., Value of (18)F FDG-PET/CT parameters on long term follow-up for patients with non-small cell lung cancer, Innov. Surg. Sci. 7 (2) (2022) 35–43.
[5] S. Kort, et al., Multi-centre prospective study on diagnosing subtypes of lung cancer by exhaled-breath analysis, Lung Cancer 125 (2018) 223–229.
[6] M. Donaghy, S. Stolberg, S. Grundy, PET-CT before biopsy in lung cancer diagnostic pathways, Lung Cancer 139 (2020).
[7] B. Liu, et al., Lung cancer detection via breath by electronic nose enhanced with a sparse group feature selection approach, Sens. Actuators B Chem. 339 (2021).
[8] K. Chen, et al., Recognizing lung cancer and stages using a self-developed electronic nose system, Comput. Biol. Med. 131 (2021), 104294.
[9] W. Biehl, et al., VOC pattern recognition of lung cancer: a comparative evaluation of different dog- and eNose-based strategies using different sampling materials, Acta Oncol. 58 (9) (2019) 1216–1224.
[10] P.J. Mazzone, et al., Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array, Thorax 62 (7) (2007) 565–568.
[11] A.J. Hubers, et al., Combined sputum hypermethylation and eNose analysis for lung cancer diagnosis, J. Clin. Pathol. 67 (8) (2014) 707–711.
[12] L. Chen, et al., Prediction model of volatile organic compounds in exhaled breath for diagnosis of lung cancer, Tumor 35 (4) (2015) 404–413.
[13] D. Shlomi, et al., Detection of lung cancer and EGFR mutation by electronic nose system, J. Thorac. Oncol. 12 (10) (2017) 1544–1551.
[14] M. Rodriguez-Aguilar, et al., Ultrafast gas chromatography coupled to electronic nose to identify volatile biomarkers in exhaled breath from chronic obstructive pulmonary disease patients: a pilot study, Biomed. Chromatogr. 33 (12) (2019) e4684.
[15] T.T. Nguyen, et al., A weighted multiple classifier framework based on random projection, Inf. Sci. 490 (2019) 36–58.
[16] H.B.F. David, A. Suruliandi, S.P. Raja, Stacked framework for ensemble of heterogeneous classification algorithms, J. Circ. Syst. Comput. 30 (15) (2021).
[17] W. Wang, D. Sun, The improved AdaBoost algorithms for imbalanced data classification, Inf. Sci. 563 (2021) 358–374.
[18] J.H. Morra, et al., Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation, IEEE Trans. Med. Imag. 29 (1) (2010) 30–43.
[19] A. Voss, et al., Detecting cannabis use on the human skin surface via an electronic nose system, Sensors 14 (7) (2014) 13256–13272.
[20] L.j. Hao, M. Zhang, G. Huang, Feature optimization of exhaled breath signals based on pearson-BPSO, Mobile Inf. Syst. 2021 (2021) 1–9.
[21] J. Fu, et al., Pattern classification using an olfactory model with PCA feature selection in electronic noses: study and application, Sensors 12 (3) (2012) 2818–2830.
[22] M.R. Gauthama Raman, et al., An efficient intrusion detection system based on hypergraph - genetic algorithm for parameter optimization and feature selection in support vector machine, Knowl. Base Syst. 134 (2017) 1–12.
[23] M.S. Nawaz, B. Shoaib, M.A. Ashraf, Intelligent cardiovascular disease prediction empowered with gradient descent optimization, Heliyon 7 (5) (2021), e06948.
[24] M.N. Uddin, R.K. Halder, An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach, Inform. Med. Unlocked 24 (2021).
[25] T.T. Nguyen, et al., A Novel 2-stage combining classifier model with stacking and genetic algorithm based feature selection, in: Intelligent Computing Methodologies, 2014, pp. 33–43.
[26] Y.S. Jeon, D.H. Yang, D.J. Lim, FlexBoost- A flexible boosting algorithm with adaptive loss functions, IEEE Access 7 (2019) 125054–125061.
[27] V. Gupta, M. Mittal, KNN and PCA classifier with Autoregressive modelling during different ECG signal interpretation, Procedia Comput. Sci. 125 (2018) 18–24.
[28] E. Kasbohm, et al., Strategies for the identification of disease-related patterns of volatile organic compounds: prediction of paratuberculosis in an animal model using random forests, J. Breath Res. 11 (4) (2017), 047105.
[29] Y.Y. Liu, et al., A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent Re-offending, J. Quant. Criminol. 27 (4) (2011) 547–573.
[30] V. Dominic, D. Gupta, S. Khare, An effective performance analysis of machine learning techniques for cardiovascular disease, Appl. Med. Inf. 36 (1) (2015) 23–32.
[31] T. Takenouchi, O. Komori, S. Eguchi, An extension of the receiver operating characteristic curve and AUC-optimal classification, Neural Comput. 24 (10) (2012) 2789–2824.
[32] A. Fawzi, O. Fawzi, P. Frossard, Analysis of classifiers' robustness to adversarial perturbations, Mach. Learn. 107 (3) (2017) 481–508.
[33] V. Paliwal, N.R. Babu, Prediction of stability boundaries in milling by considering the variation of dynamic parameters and specific cutting force coefficients, Procedia CIRP 99 (2021) 183–188.
[34] X. Chen, et al., Calculated indices of volatile organic compounds (VOCs) in exhalation for lung cancer screening and early detection, Lung Cancer 154 (2021) 197–205.
[35] H. Lu, et al., A hybrid ensemble algorithm combining AdaBoost and genetic algorithm for cancer classification with gene expression data, IEEE ACM Trans. Comput. Biol. Bioinf 18 (3) (2021) 863–870.

[36] S. Wu, H. Nagahashi, Penalized AdaBoost: improving the generalization error of gentle AdaBoost through a margin distribution, IEICE Trans. Info Syst. E98.D (11) (2015) 1906–1915.
[37] A. Mahabub, A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers, SN Appl. Sci. 2 (4) (2020).
[38] R. de Vries, et al., Prediction of response to anti-PD-1 therapy in patients with non-small-cell lung cancer by electronic nose analysis of exhaled breath, Ann. Oncol. 30 (10) (2019) 1660–1666.

Hao Lijun, a doctoral candidate at University of Shanghai for Science and Technology and is also working at Shanghai University of Medicine & Health Science. She graduated the Institute of biomedical instruments of Shanghai Jiaotong University in 2007. She is mainly engaged in teaching and research in biomedical engineering. In terms of teaching, she participated in the construction of Shanghai, high-quality courses, Shanghai public education training base, etc.; participated in the preparation and completion of several textbooks. And in terms of scientific research, she has presided over the Chenguang project of the Shanghai Municipal Education Commission and several school-level scientific research startup projects and participated in several science and Technology Commission and natural science fund projects. In the past five years, she has published more than ten papers as the first author, and applied for and approved many patents.

Huang Gang, M.D., second-grade Professor, and doctoral supervisor, concurrently serves as the director of the Institute of clinical nuclear medicine of Shanghai Jiaotong University, the president of the Asian Nuclear Medicine College, the chairman of Shanghai Medical Education Association, and the leader of the national key clinical specialty of imaging medicine, key disciplines of Shanghai and first-class disciplines of Shanghai. He has won the young and middle-aged experts with outstanding contributions from the Ministry of Health, Shanghai's leading talents Shanghai Medical's leading talent, Shanghai's 100-person plan, and other titles. He has undertaken more than 30 projects such as the National Natural Science Foundation and key projects, and the "973" project. So far, He has published more than 200 papers in domestic and foreign journals, including more than 80 papers included in SCI or EI; More than 10 authorized patents; He edited more than 10 planned textbooks and monographs for medical colleges. In addition, he has won more than 10 awards, including the second prize of National Science and Technology Progress Award and the first prize of Huaxia Medical Science and Technology Award. His main research fields are molecular probes, tumor imaging omics, biomedical engineering.