

PAPER

Engineering solutions to breath tests based on an e-nose system for silicosis screening and early detection in miners

To cite this article: Wufan Xuan *et al* 2022 *J. Breath Res.* **16** 036001

View the [article online](#) for updates and enhancements.

You may also like

- [Online breath analysis using metal oxide semiconductor sensors \(electronic nose\) for diagnosis of lung cancer](#)
Aleksandr Kononov, Boris Korotetsky, Igor Jahatspanian *et al.*
- [Detection of lung cancer with electronic nose using a novel ensemble learning framework](#)
Lei Liu, Wang Li, ZiChun He *et al.*
- [Infrared cavity ring-down spectroscopy for detecting non-small cell lung cancer in exhaled breath](#)
Robyn Larracy, Angkoon Phinyomark and Erik Scheme



Breath Biopsy[®] OMNI

The most advanced, complete solution for global breath biomarker analysis



SEE WHAT OMNI CAN DO FOR YOU





PAPER

Engineering solutions to breath tests based on an e-nose system for silicosis screening and early detection in miners

RECEIVED
9 February 2022ACCEPTED FOR PUBLICATION
18 March 2022PUBLISHED
7 April 2022Wufan Xuan^{1,2}, Lina Zheng^{1,2}, Benjamin R Bunes³, Nichole Crane³, Fubao Zhou^{1,*}  and Ling Zang^{4,*} ¹ School of Safety Engineering, China University of Mining and Technology, Xuzhou, People's Republic of China² Xuzhou Engineering Research Center for Occupational Dust Control and Environmental Protection, Xuzhou, People's Republic of China³ Vaporsens, Inc., Salt Lake City, Utah, United States of America⁴ Nano Institute of Utah and Department of Materials Science and Engineering, University of Utah, Salt Lake City, Utah, United States of America

* Authors to whom any correspondence should be addressed.

E-mail: f.zhou@cumt.edu.cn and lzang@eng.utah.edu**Keywords:** breath test, silicosis, electronic nose, nanofiber sensor, ensemble learningSupplementary material for this article is available [online](#)**Abstract**

This study aims to develop an engineering solution to breath tests using an electronic nose (e-nose), and evaluate its diagnosis accuracy for silicosis. Influencing factors of this technique were explored. 398 non-silicosis miners and 221 silicosis miners were enrolled in this cross-sectional study. Exhaled breath was analyzed by an array of 16 organic nanofiber sensors along with a customized sample processing system. Principal component analysis was used to visualize the breath data, and classifiers were trained by two improved cost-sensitive ensemble algorithms (random forest and extreme gradient boosting) and two classical algorithms (K-nearest neighbor and support vector machine). All subjects were included to train the screening model, and an early detection model was run with silicosis cases in stage I. Both 5-fold cross-validation and external validation were adopted. Difference in classifiers caused by algorithms and subjects was quantified using a two-factor analysis of variance. The association between personal smoking habits and classification was investigated by the chi-square test. Classifiers of ensemble learning performed well in both screening and early detection model, with an accuracy range of 0.817–0.987. Classical classifiers showed relatively worse performance. Besides, the ensemble algorithm type and silicosis cases inclusion had no significant effect on classification ($p > 0.05$). There was no connection between personal smoking habits and classification accuracy. Breath tests based on an e-nose consisted of $16 \times$ sensor array performed well in silicosis screening and early detection. Raw data input showed a more significant effect on classification compared with the algorithm. Personal smoking habits had little impact on models, supporting the applicability of models in large-scale silicosis screening. The e-nose technique and the breath analysis methods reported are expected to provide a quick and accurate screening for silicosis, and extensible for other diseases.

1. Introduction

Pneumoconiosis is one of the most common occupational diseases in China and other developing countries [1]. Though prevention methods have been applied for many decades, it remains a critical problem worldwide, especially in construction and mining [2, 3]. The total number of reported occupational cases in China was 97 500 by 2018, and 90% of the cases were pneumoconiosis [4]. This kind of disease

affects drastically the patients' quality of life, causing breathing difficulties and even death. A study in the United States showed that during 1999–2018, a total of 43 366 decedents aged ≥ 15 years had pneumoconiosis listed on their death certificates [5]. And beyond that, pneumoconiosis is incurable, latent, and has a delayed progression [6], leading to difficulty in intervention when detected at a late stage. Hence, early detection is critical to suppress pneumoconiosis, which can be achieved through regular and

targeted screening [7]. Silicosis is a typical case of pneumoconiosis, characterized by nodular fibrosis of the lungs [8]. At present, the diagnosis of silicosis mainly relies on pulmonary function tests along with medical imaging technologies like x-ray imaging and computerized tomography scanning. However, early manifestations of silicosis are not apparent in medical images; diagnosis of early stage silicosis requires multidisciplinary expertise of specialized pulmonologists, radiologists, and pathologists [9, 10]. Moreover, the heterogeneity of dust exposures and intrapulmonary depositions results in pulmonary structural variations as well as difficulties in image recognition [2, 11]. Biopsies with tissue analysis are capable of yielding a clear diagnosis and disclosing potential lesions, but this is an invasive method that may lead to complicated procedures and severe complications in some cases [12]. Clearly, current diagnostic technologies are not suitable for convenient, high volume screening of silicosis, especially for early cases. Therefore, it is imperative to develop a new technology that is capable of silicosis screening and early detection for massive population.

In recent years, respiratory analysis has shown the potential for minimally (or non-) invasive disease detection by monitoring the volatile organic compounds (VOCs) excreted from human breath or skin emanation [13–15]. VOCs in the expired breath are believed to give information about general metabolic conditions and, particularly those of the lung [16]. In the past two decades, with the continuous improvement of modern analytical technologies, electronic nose (e-nose) has been developed and shown great potential in quick diagnosis of diseases via breathprints [17–19]. E-noses based on varying sensor materials have been studied for breath tests with the aim to detect different types of disease such as cancers [20, 21], chronic obstructive pulmonary disease [22], breast cancer [21], etc. On one hand, sensors that respond to specific biomarkers were developed, like graphene-based sensors designed for known concentrations of acetone for diabetes diagnosis [23]. On the other hand, the sensor array that makes a multidimensional response to the target gas can distinguish the breath pattern of different populations using supervised classification algorithms [22]. However, few studies have been performed on pneumoconiosis thus far. In 2018, Yang *et al* performed breath tests using an array of polymer sensors to detect asbestosis and achieved an accuracy of 70.0% in the validation set [24], which suggested the feasibility of developing a respiratory diagnostic method for pneumoconiosis. This study was theoretically supported by Yang's previous research on the breath biomarker exploration of pneumoconiosis [25]. Despite this technique's promising application, there are some challenges and pitfalls in the clinical application of e-nose on breath analysis, like limited number of test participants, lack of standardization, environmental

interference, and technical issues of the e-nose regarding noise, reproducibility etc [26]. It is recognized that these factors hinder its application in medical surveillance or screening, and need to be addressed by exploring new e-nose with alternative sensors. A number of studies have been conducted on breath tests proposing a series of standardized requirements, but few of them have been verified in large cohorts or practically use in pneumoconiosis [19, 27].

In this paper, an exploratory study was conducted on an e-nose consisted of 16 nanofiber sensors for silicosis diagnosis through breath tests. Objectives of this study are to (a) verify the e-nose system along with a homemade breath sample processing device, and apply it to silicosis screening in miners, especially to early stage detection; (b) investigate the performance difference between screening models and early detection models by including silicosis cases in various disease stages; (c) evaluate the impact of algorithms and personal smoking habits on classification, and explore other influencing factors.

2. Materials and methods

2.1. Subject recruitment

Participants were local gold miners in the region of Yantai City of Shandong Province, China. A total of 619 subjects, including 398 non-silicosis miners and 221 silicosis miners, were recruited, and their breath samples were collected and tested during three onsite visits made in October, November and December 2019. All the participants took health examinations at Shandong Gold occupational disease hospital, and their demographic information was collected simultaneously through structured questionnaires. Exclusion Criteria were: (a) Involving with other occupational exposure, such as welding; (b) Having had a lung lavage or any other form of lung surgery; (c) Having autoimmune diseases, cancers, chronic inflammation, or endocrine, metabolic diseases. Clinical diagnosis of silicosis and classification of health status of the participants were performed by a panel of three experienced physicians according to the National Diagnosis of Occupational Pneumoconiosis (Criteria Code: GBZ 70-2015).

Ethical review: The study conforms to the Declaration of Helsinki and has little harm or health risk to the participants. Involvement in the study imposes no additional burden on the participants in clinical practice. The sampling and testing protocol we adopted was approved by the Research Ethics Committee of Xuzhou No. 1 Peoples Hospital (No. xyyl [2020]69). All participants were informed of the collection process and signed written informed consent. One potential benefit to subjects participating the study is that their breathprints generated from the testing may provide them with some health information regarding risk of silicosis or progression of the disease. For the general public, the outcome of the

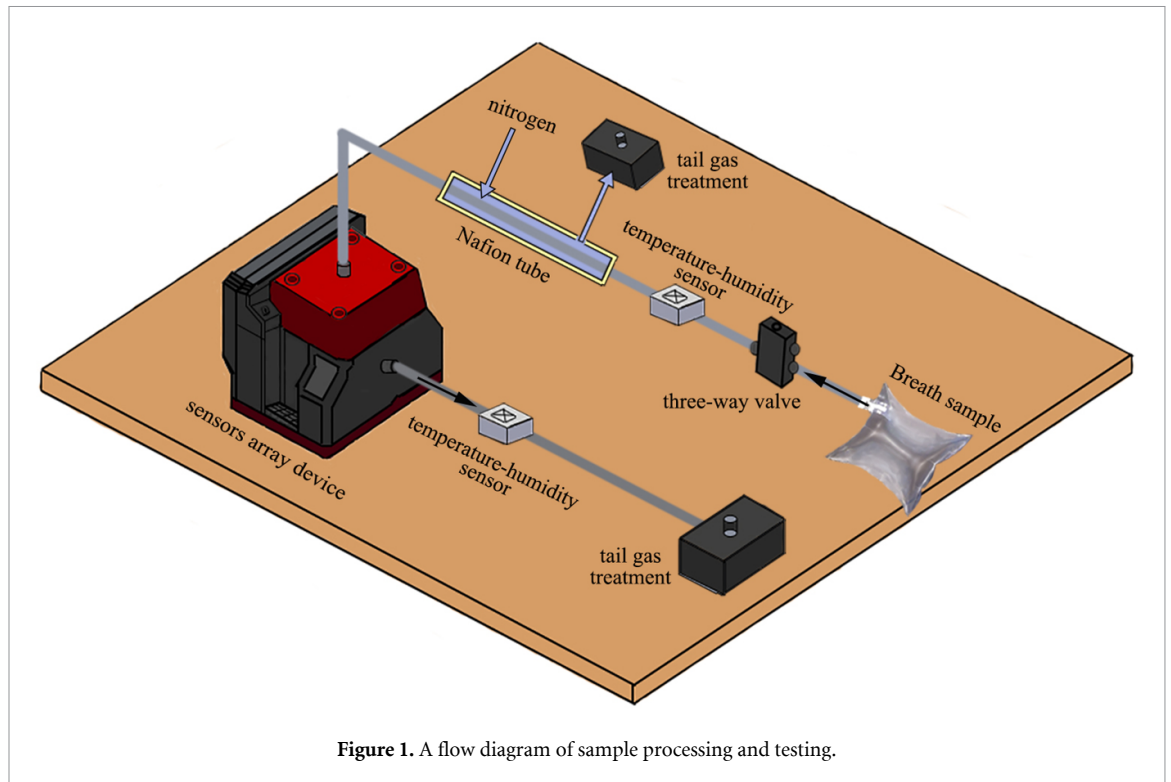


Figure 1. A flow diagram of sample processing and testing.

study may help eventually develop a point-of-care and non-invasive tool for quick screening of silicosis, especially at the early stage, which remains challenging to be diagnosed at the moment.

The sample size was evaluated based on equations (1) and (2), among which the larger size number would be adopted. According to similar diagnostic models reported in the literature [24, 28], 95% sensitivity, 85% specificity, statistical significance level of 0.05 and 5% allowable error were substituted into equations. The calculations showed that a sample size of at least 196 subjects would suffice in this study.

$$n_{se} = \frac{Z_{1-\alpha/2}^2 * p_{se} * (1 - p_{se})}{\delta^2} \quad (1)$$

$$n_{sp} = \frac{Z_{1-\alpha/2}^2 * p_{sp} * (1 - p_{sp})}{\delta^2} \quad (2)$$

where n is the estimated sample size, α the significance level, $Z_{1-\alpha/2}$ the rejection region cut-off point in a two-sided test, p_{se} the reference value of sensitivity, p_{sp} the specificity, and δ the allowable error.

2.2. Study design

Two types of diagnosis models were trained in this single-center, cross-sectional study. The first one was the screening model, including 398 non-silicosis miners as healthy controls and 221 silicosis patients as cases. To break silicosis's continuous and irreversible progression early [29], the early detection model was further studied with 85 patients in stage I as cases and 398 non-silicosis as healthy controls.

The good clinical practice (GCP) regards quality control (QC) as an essential part in tests. QC is the process by which samples are tested and measured in the scope of the standard set in advance to minimize errors and inconsistencies and ensure the ongoing implementation of appropriate data entry [30]. And, QC itself does not have a uniform, universal process or protocol. In our work, GCP guidelines were adopted as a basis. QC involves consistent subject recruitment, device operation specifications, sample measurement and recording, and data processing [26]. Highly standardized conditions and model assessment for testing were also taken into account.

2.3. Breath collection and customized system for sample processing

The breath testing in this study was performed on an e-nose system based on a 16-nanofiber sensor array (Pilot™, Vaporsens, Salt Lake City, USA), which was equipped with a customized sample processing system (figures 1 and 2). The data were analyzed by pattern recognition algorithms. Detailed information of the sensor array and operation mechanism were provided in supporting information (available online at stacks.iop.org/JBR/16/036001/mmedia). Considering the breathing velocity variation among subjects, testing by breathing directly into the Pilot™ instrument may cause significant fluctuation in sensor signals. To mitigate this problem, sampling bags were used for collecting and storing the exhaled breath, which allowed all the samples to be tested under exactly the same conditions including humidity level, temperature and flow rate. Briefly, subjects tidally

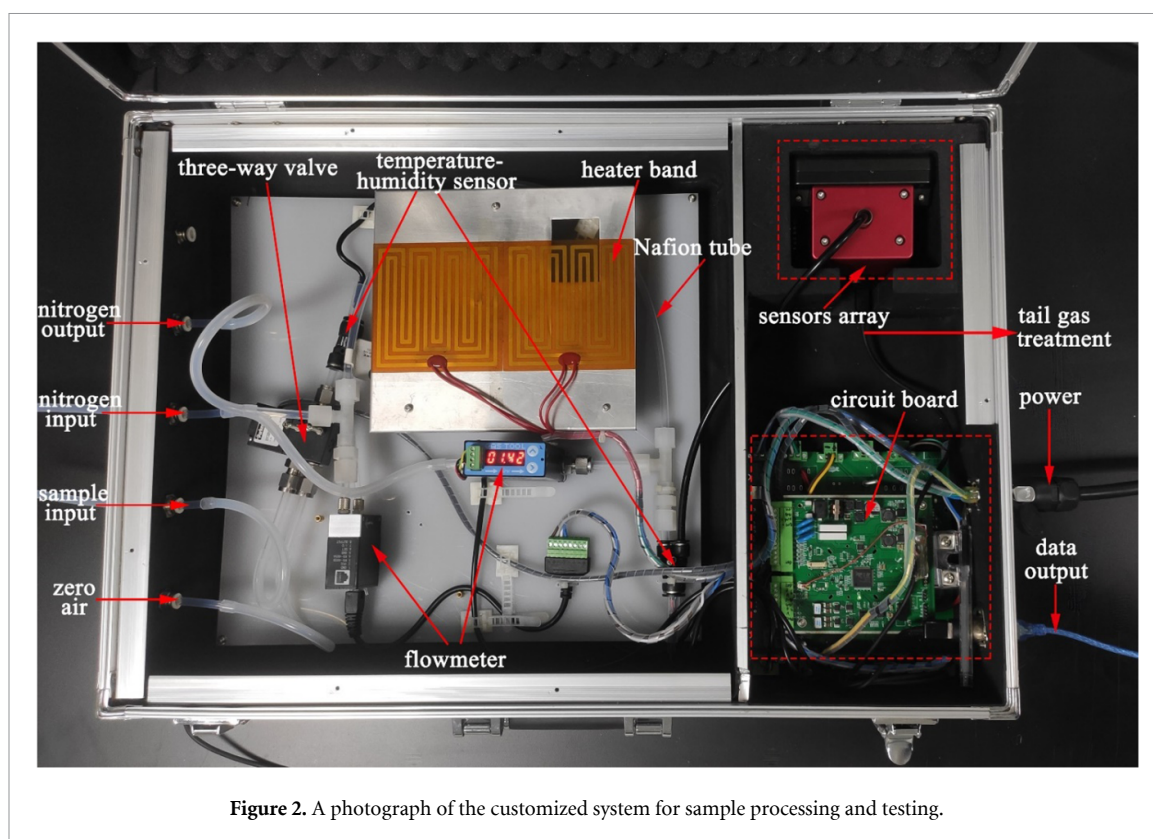


Figure 2. A photograph of the customized system for sample processing and testing.

blew into a 1-litre sampling bag through the mouth-piece with a standardized and validated breathing process [31]. The breath collection site was well-ventilated over the whole period to maintain clean and stable ambient air. Items used for sample collection and storage included mouthpieces, tubes, and sampling bags were all new and made of polytetrafluoroethylene. Prior to being introduced to the sensor array for testing, the samples were processed for dehumidification in a customized system at a flow rate of around 400 sccm. Dehumidification was realized by flowing the sample through a Nafion® tube (MD070-24F, Perma Pure LLC., Lakewood, NJ 08701, USA), which adjusted the relative humidity to $\sim 10\%$ for the breath samples. The dehumidified sample was then introduced into Pilot™ to be exposed to the sensor array. Processing and testing of each sample was finished within half an hour after being collected. During the whole period of sensor testing, the response signal was recorded in real time before and after the exposure to the breath sample, named period data. Period data is the continuous monitoring data in the duration of breath exposure, which represents a set of contiguous time granules within the exposure period.

Pre-collection behavior requirements: Sampling of breath was performed in the morning, and subjects were asked not to eat, smoke, or take medicine for 10 h before the breath collection. Subjects were asked not to eat onions, garlic, or any other food with a strong odor or go to a dusty environment for two days before the collection. Subjects stayed in a

naturally ventilated environment and did not exercise within an hour before sampling. Subjects rinsed their mouths with purified saline and then with distilled water before sampling.

2.4. Sensor materials and array

The sensor array used in this work was consisted of 16 organic nanofiber materials, which were fabricated from solution phase self-assembly of different building blocks molecules based on perylene tetracarboxylic diimide (PTCDI) [32–34]. By changing the side-binding group of PTCDI molecules, the sensor selectivity of nanofibers can be tuned and optimized toward different VOCs present in the exhaled breath. In this study, the side groups were selected to target the common VOCs found in exhaled breath of humans that are redox-active for initiating interfacial charge transfer interaction with the nanofiber and thus resulting in a change in resistance [32, 34]. Such VOCs include aldehydes, ketones, ammonia, hydrogen peroxide, nitric oxide, and the concentrations of these compounds in the breath may change (increase or decrease) when people get a disease like pneumoconiosis. Depending on the disease type and stage, the concentration change could be dramatic, ranging from multiple times to orders of magnitude. For example, formaldehyde, 2-pentanone, and propionaldehyde are detected in the breath of lung cancer patients, and the concentration increases by two to more than ten times compared to the healthy people [35]. The same or similar VOCs were taken into account in our work. Indeed, as recently reported

[25], the concentration of 3-methyl-butanol increases 2.5 times for people diagnosed with pneumoconiosis. More biomarker VOCs such as hydrogen peroxide, nitric oxide, acetoin have been identified for idiopathic pulmonary fibrosis, a disease having similar symptoms as pneumoconiosis.

To target the above-mentioned VOCs, the side-binding groups of PTCDI building blocks were designed to contain amines, amide, hydroxyl and other heterocycles, which provide strong and specific binding towards the VOCs via hydrogen bonding, charge transfer and electrostatic interactions. Typical molecular structures of the binding groups can be found in one of our patents [36]. It should be noted that rather than separating and detecting the individual VOCs (as usually done with the benchtop instrumentations like gas chromatograph-mass spectrometry), the sensor array was exposed directly to the breath mixture, and the composite profile (breathprints) generated from the responses of all the 16 nanofibers will enable discrimination between disease and healthy controls, reflecting the metabolic change in breath composition [17, 37, 38]. Such an array-based detection is analogous to the mammalian olfactory system [17, 39], wherein a large number of olfactory receptors (sensors) work as a cooperative array to generate specific patterns for different odors or mixtures, but without knowing the details of the individual components. This is also how a dog can be trained to sniff out certain diseases by differentiating the odor pattern of breath or sweat between diseased and healthy people [40].

2.5. Statistical analysis and modeling

The electronic nose made a response to breath samples with multiple signals, which were called breathprints. Representative diagrams of sensor response signals of a healthy control and a silicosis patient in stage III are provided in figure S2, in supporting information. Measurement time or batches were not included as covariant in data modeling in this study because the period of measurement caused only slight data drifts (see figure S2), which could be neglected by adopting appropriate feature extraction methods, and stricter and more detailed QC requirements. Therefore, 16-dimensional signals obtained from the sensor array with demographic information were input as the raw data.

Feature extraction is a technique that selects variables and combines them into features, effectively reducing the data amount to be processed, while still accurately and fully describing the original data set [41]. The reduction of the data and the effort in building combinations of features facilitate the learning and generalization steps in the machine learning process. To extract as much information as possible from the original data, we have conducted linear regression on period data of the breath exposure, which was about fitting a linear model ($y = k*x + b$) to data

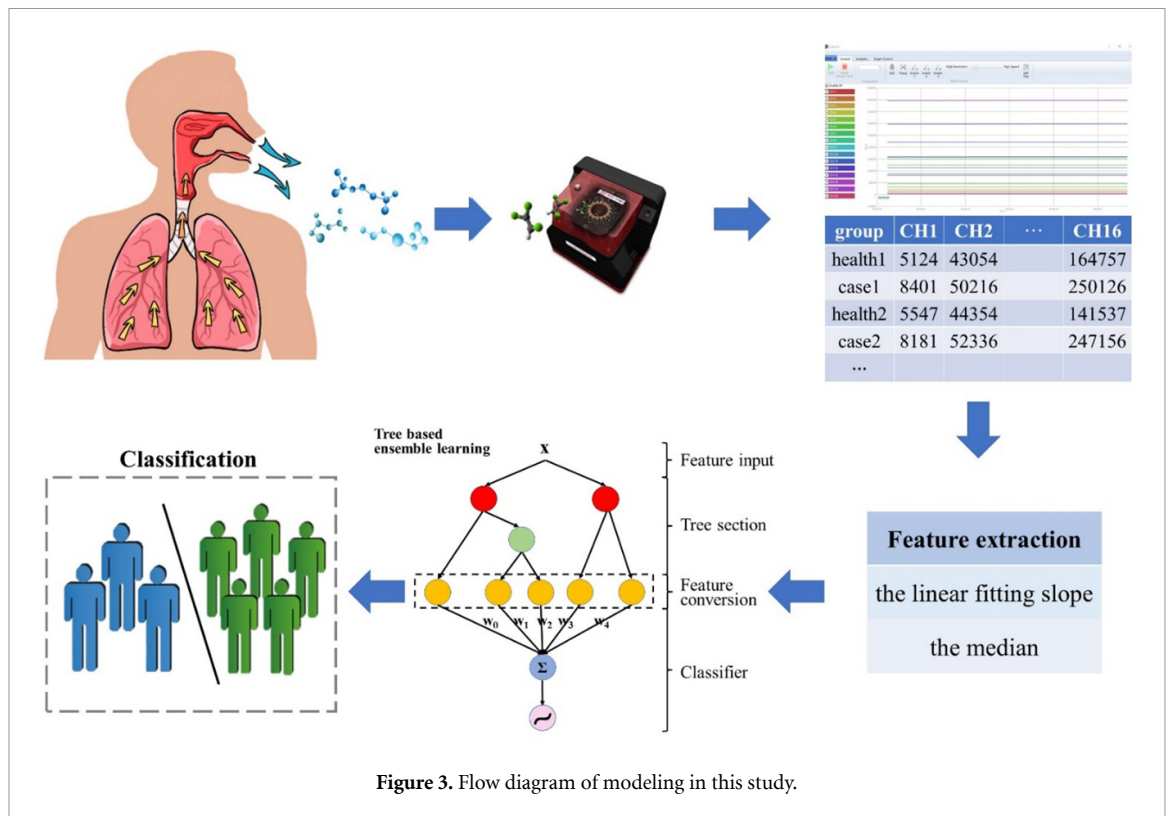
points. Hence, data were reduced and generalized into a few parameters in a regression: k in the function is the slope, which quantifies the steepness of the line. It equals the change in Y for each unit change in X . b in the function is intercept term. The intercept term and the median of the period data indicates the data range of samples.

As a result, the linear fitting slope and the median of the period data were extracted as two data features, forming a 32-dimensional dataset with 16 sensor signals. Discrete information like personal smoking habits was coded by one-hot encoding. All the data processing and analysis were done in Python version-3.7.1 with PyCharm 2021.1 \times 64 (Python Software Foundation, Delaware, USA).

We reduced the dimension of sensors' response by principal component analysis (PCA), aiming to visualize points aggregation between groups. Three-dimensional PCA were adopted, getting three principal components extracted after dimensionality reduction. To identify representative sensors that contribute significantly to the classification, partial least square variable importance in projection (PLS-VIP) analysis of the two models were conducted, respectively. PLS-VIP is often used to identify the importance of each indicator based on variance calculation [42]. The variance of each PLS dimension could measure the variable's data explanatory ability and its influence on the model classification.

For data modeling, raw data were divided into a training dataset (data obtained in October and November tests) and a validation dataset (data obtained in December tests). Classifiers were constructed by 5-fold cross-validation with the training dataset, which was divided into a training set and test set further. Blinded identification was conducted via unknown data in the validation dataset, namely external validation. To ensure the inaccessibility of the validation dataset and prevent information leaks, the training dataset and validation dataset were processed independently. Classifiers to discriminate between the controls and cases were established with two improved ensemble learning algorithms, eXtreme gradient boosting (XGBoost) and random forest (RF). The two algorithms were popular ensemble learning algorithms, which were modified to be cost-sensitive algorithms in this study. The core of cost-sensitive algorithms was that we assigned different weights to false negatives (FNs) and false positives (FPs). Consequently, cost-sensitive algorithms can flexibly adjust 'the cost of misdiagnosing a patient as a healthy person' and 'the cost of misdiagnosing a healthy person as a patient' to reduce FNs in medical diagnosis. The modeling flow diagram of this study is shown in figure 3.

We also used two classical classification algorithms, support vector machine (SVM) and K-nearest neighbor (KNN), to contrast with our improved ensemble learning algorithm. SVM is



a kind of generalized linear classifier by supervised learning method. Its decision boundary is the maximum-margin hyperplane solved for the learning samples. KNN quantifies the distance of samples based on the Euclidean distance. This method determines the classification of the samples according to the category of the nearest one or several samples in the feature space. Normalization and dimensionality reduction were combined to provide input data for KNN models.

In KNN, normalization method based on equation (3) changes numbers to decimals in the range of 0–1.

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}. \quad (3)$$

These four algorithms were all optimized for parameter tuning during modeling.

Since there were two types of diagnostic models in this study due to different silicosis stages included, eight classifiers were run based on the combination of algorithms and diagnostic model types. Statistical analyses were done using receiver operating characteristic (ROC) analysis. Classifiers were evaluated using the area under the ROC curve (AUC), accuracy, sensitivity, and specificity. The eight classifiers were compared further, and two-factor analysis of variance was conducted to assess whether the algorithm and case stages could make a significant difference to classification. Moreover, the association between personal smoking habits and classifier performance was investigated using a chi-square test with the significance level α as 0.05. Personal

smoking habits were quantified into four types: current smoker, former smoker, never smoker, and second-hand smoker. The number of the true positive (TP), the true negative (TN), the FP, and the FN were employed as data input to the chi-square test.

3. Results

3.1. Subject characteristics

A total of 619 subjects were enrolled in this study. The screening model included 398 healthy controls and 221 silicosis cases. The early detection model had 398 healthy controls and 85 silicosis cases in stage I. Age, personal smoking habits, and silicosis stages were labeled individually. The demographic characteristics of all subjects and silicosis phase statistics are provided in table 1. The subjects were all males since the female was not allowed to do the hard physical labor in underground mines in China.

3.2. Data visualization by PCA

The score plots for PCA analysis are shown in figure 4. In figures 4(a) and (b), almost 97% of the variance was explained by the first three principal components. Figure 4(a) presented the observation of all subjects in the screening model, in which scattered regions of different groups separated to a great extent. Points of the control group and case group were apart along the approximate direction of the PC1 axis in figure 4(a). Figure 4(b) presented the observation of subjects in the early detection model, points of which aggregated. However, some points of early silicosis cases were

Table 1. Subject characteristics in this study.

Characteristics	Controls (<i>n</i> = 398)	Cases in screening model (<i>n</i> = 221)	Cases in early detection model (<i>n</i> = 85)
Age (year), mean (SD)	48.0 ± 7.2	50.3 ± 8.3	50.0 ± 7.3
Sex, <i>n</i> (%)			
Male	398.0(100.0)	221.0(100.0)	85.0(100.0)
Female	0(0)	0(0)	0(0)
Smoking habit, <i>n</i> (%)			
Current smoker	244(61.3)	132(59.7)	52(61.2)
Former smoker	25(6.3)	26(11.8)	11(12.9)
Never smoker	100(25.1)	52(23.5)	16(18.8)
Second-hand smoker	29(7.3)	11(5.0)	6(7.1)
Clinical stage, <i>n</i> (%)			
Stage I		85(38.5)	85(100.0)
Stage II		83(37.6)	
Stage III		53(24.0)	

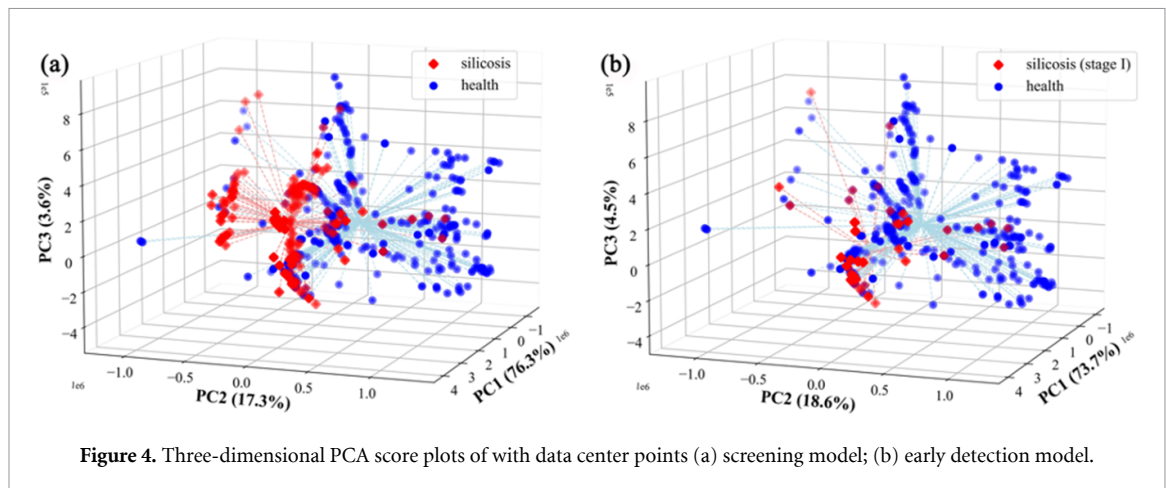


Figure 4. Three-dimensional PCA score plots of with data center points (a) screening model; (b) early detection model.

Table 2. The variable importance of all the 16 sensors (positive correlation).

Sensor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Scores	0.34	0.01	1.08	0.62	0.29	0.34	0.26	0.26	0.37	0.47	0.73	1.31	1.63	0.26	0.50	0.32

close to those of healthy controls with region overlap. On the other hand, from the comparison of models, the aggregation of case points in the early detection model was not as tight as those in the screening model. This may be due to the stage difference of silicosis cases in the two case groups.

PLS-VIP analysis of the 16 sensor signals were done. The VIP values were scored, as shown in table 2. The scores helped to reveal sensors' contribution to classification results. Sensor CH13, CH12, and CH3 are the top three responsive sensors; meanwhile, CH2, CH14, CH7, and CH8 were the four sensors that had the least contribution to classification.

3.3. Classifier performance

Eight classifiers were established: screening model using RF, XGBoost, KNN and SVM; early detection model using RF, XGBoost, KNN and SVM. The

screening model allocated 197 healthy controls and 128 silicosis cases to 5-fold validation to construct classifiers, 201 healthy controls, and 93 silicosis cases for external validation to test classifiers. The early detection model allocated 197 healthy controls and 43 silicosis cases to 5-fold validation, 201 healthy controls, and 42 silicosis cases for external validation.

ROC analysis investigated the classifier performance of the test set and external validation, as shown in figure 5. All the ROCs reached a relatively high value, and it was clear to find ROCs of RF and XGBoost were close in all four pictures. ROCs of figures 5(a) and (b) seem similar in terms of trend and AUCs, while the ROC of figure 5(d) presented an evident decline in comparison with figure 5(c). ROC analysis of KNN and SVM is in supporting information.

Results of the classification are presented in table 3. Classifiers of ensemble learning were further

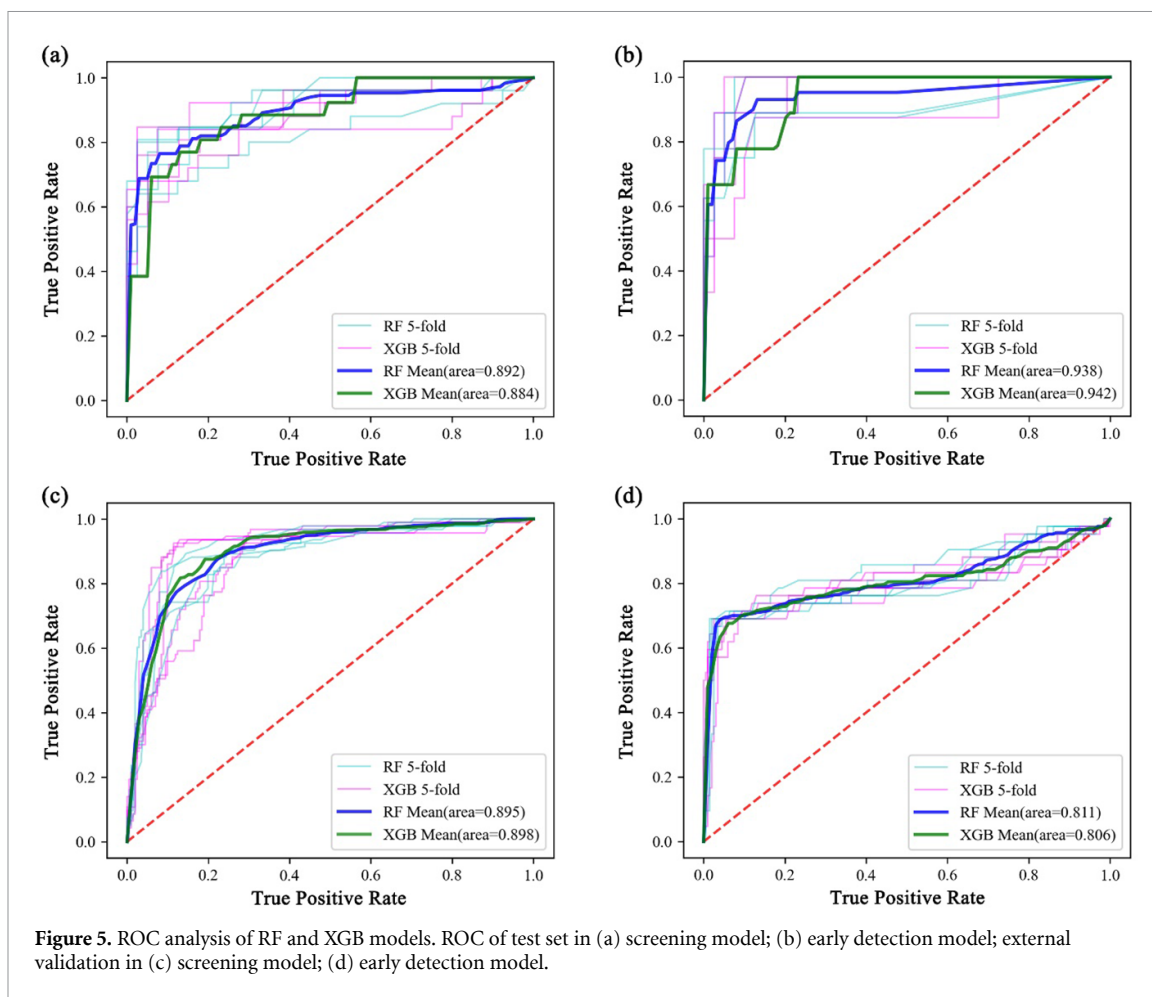


Figure 5. ROC analysis of RF and XGB models. ROC of test set in (a) screening model; (b) early detection model; external validation in (c) screening model; (d) early detection model.

Table 3. Classification results using different diagnostic model types and algorithms.

Model type	Algorithm	Data set	AUC	Accuracy	Sensitivity	Specificity
Screening model	RF	Training	0.999	0.987	0.984	0.992
		Test	0.892	0.843	0.765	0.893
		External validation	0.895	0.839	0.776	0.869
	XGB	Training	0.989	0.945	0.879	0.987
		Test	0.884	0.843	0.719	0.923
		External validation	0.898	0.858	0.806	0.882
	KNN	Training	0.985	0.987	0.971	1.000
		Test	0.769	0.785	0.692	0.846
		External validation	0.488	0.490	0.484	0.493
SVM	Training	0.815	0.800	0.882	0.747	
	Test	0.782	0.785	0.769	0.795	
	External validation	0.585	0.673	0.344	0.826	
Early detection model	RF	Training	0.999	0.982	0.907	0.999
		Test	0.938	0.917	0.811	0.939
		External validation	0.811	0.817	0.729	0.836
	XGB	Training	0.998	0.975	0.959	0.978
		Test	0.942	0.9	0.811	0.919
		External validation	0.806	0.867	0.705	0.901
	KNN	Training	0.926	0.974	0.853	1.000
		Test	0.479	0.708	0.111	0.846
		External validation	0.509	0.811	0.048	0.970
	SVM	Training	0.851	0.849	0.853	0.848
		Test	0.744	0.792	0.667	0.821
		External validation	0.788	0.852	0.690	0.886

Table 4. P value calculated with two-factor analysis of variance.

Data set	Factor	AUC	Accuracy	Sensitivity	Specificity
Training	Model	0.500	0.605	0.981	0.844
	Algorithm	0.437	0.395	0.775	0.590
Test	Model	0.073	0.082	0.205	0.555
	Algorithm	0.795	0.500	0.500	0.874
External validation	Model	0.029*	0.747	0.223	0.833
	Algorithm	0.844	0.269	0.930	0.374

Note: * means $p < 0.05$.

Table 5. Chi-square tests for smoking habits and classification in models using RF.

	Screening model				Early detection model			
	Never smoker	Former smoker	Second-hand smoker	Current smoker	Never smoker	Former smoker	Second-hand smoker	Current smoker
TP	14	3	8	47	3	2	5	21
TN	42	12	12	109	40	12	11	105
FP	6	2	2	16	8	2	3	20
FN	4	0	3	14	2	0	2	7
χ^2	4.745				7.552			
p	0.856				0.580			

Table 6. Chi-square tests for smoking habits and classification in models using XGBoost.

	Screening model				Early detection model			
	Never smoker	Former smoker	Second-hand smoker	Current smoker	Never smoker	Former smoker	Second-hand smoker	Current smoker
TP	15	3	9	48	4	2	5	19
TN	42	12	12	111	43	13	12	113
FP	6	2	2	14	5	1	2	12
FN	3	0	2	13	1	0	2	9
χ^2	4.900				7.694			
p	0.850				0.565			

compared using a two-factor analysis of variance, and p values calculated are shown in table 4. Pictures in figure 5 supported discoveries in tables 3 and 4.

In general, four classifiers based on ensemble learning had achieved good performance. In these models, all AUCs obtained with the screening model were larger than 0.88, indicating excellent stability. Results of the early detection model were slightly inferior to the screening model, but it was still good, with the AUC ranging from 0.806 to 0.999. For four evaluation indexes, sensitivity was generally lower than the other three in the validation, with differences around 0.1. Since it is not unusual for cases detection models to classify subjects as healthy controls by mistake, the sensitivity range between 0.705 and 0.984 is acceptable.

Models using KNN and SVM showed a good performance in model training and a rapidly decline in testing set and external validation set, which meant poor generalization capacity in internal and external validation.

From table 4, all p values of algorithms in variance analysis were more than 0.05, indicating no significant difference between the two algorithms.

P values of model types were more than 0.05 in most cases, indicating that the difference between the two models had no statistical significance, except in the AUC of external validation. This result conformed to table 3, as the AUC of the early detection model was obviously poorer than that of the screening model in external validation. Given the study design, this phenomenon meant that identifying early silicosis cases in this study was harder than running a screening model.

3.4. Chi-square tests for smoking habits and classification

After simple statistics on the predicted results of classifiers, the amount of the TP, the TN, the FP, and the FN were calculated in each classifier. Statistics generated acted as data input to investigate the association between personal smoking habits and classifier performance by chi-square test (significance level $\alpha = 0.05$). Results of the chi-square tests conducted on RF and XGBoost models were summarized in tables 5 and 6.

It can be concluded that the TP rate, the TN rate, the FP rate, and the FN rate were different in people

with various smoking habits, but chi-square tests demonstrated a nonsignificant impact of smoking habits on both models as $p > 0.05$.

4. Discussion

As some studies have demonstrated the feasibility of respiratory diagnosis for several diseases [43], we focused on developing a technical scheme for silicosis diagnosis and exploring the influencing factors of this approach. Experimentally, breath samples collected from 619 participants (221 in the case group and 398 in the control group) were measured with the Pilot™ platform incorporated in a customized e-nose system. Classifiers with different case stages and algorithms all achieved good performance.

In data visualization, points of silicosis cases aggregated and were clearly separated from points of healthy controls in screening model. This indicated that silicosis cases had a similar breath pattern, which was different from healthy controls. In figure 4(a), points of the control group and case group were apart along the approximate direction of the axis of PC1, which was the first principal component containing the most information of the original data. This also revealed the separability of the control group and case group. In figure 4(b), region of early silicosis cases overlapped a part of that of healthy controls, showing a weaker separability. Thus, efforts should be made to optimize classifiers and elevate the classification capability.

In both the classifier training and validation, silicosis cases were recognized accurately in models using ensemble learning, taking recent literature as a [17, 24]. Five-fold cross-validation and external validation gave a robust evaluation of classification. The consistency of performance in the screening model and early detection model demonstrated that results were reproducible. It is probably worth noting that sensitivity in nearly all the classifications was slightly lower than the other three evaluation indexes. It meant that classifiers were inclined to classify samples as healthy controls. This problem has also been found in previous studies, presenting a wider gap between sensitivity and specificity [24, 44–46]. This may be attributed to the imbalance of sample sizes in the two groups, as healthy controls are often much more than cases. Due to this reason, we introduced the cost-sensitive algorithms in advance to adjust the potential bias. Thus, other evaluation indexes did not hold an overwhelming advantage over the sensitivity, with differences around 0.1 in the validation.

For models using KNN and SVM, classification performance showed a huge decline in internal and external validation, compared with the model training. This result suggested the two algorithms' weak generalization ability in large-scale practice. And all the model performance of the two classical

algorithms were weaker than the improved ensemble learning methods. The difference between the classical classifiers and ensemble learning models suggested the importance of appropriate feature extraction and model selection. Furthermore, the sensitivity of KNN and SVM models was obviously low, while the sensitivity of RF and XGBoost models was close to other indexes. This supported the utility of the cost-sensitive algorithms in diagnosis modeling.

From the two-factor analysis of variance, RF and XGBoost algorithms exerted little influence on models in this paper. On the other hand, different model types showed a significant difference in the AUC of the external validation, in which the early detection model's AUC was inferior to that of the screening model. Since the difference between the two models lied in silicosis stages included, we believed that the reason was the uneven quality of the raw data input. Data size or distribution bias were underlying reasons, e.g. the amount of category data in the early detection model was more unbalanced. In summary, our results suggested that the data input and improved algorithms exerted a significant impact on classification, while algorithms had little effect on classifiers if already improved. This study leads us to emphasize both the algorithmic optimization and the enhancement of raw data quality. Therefore, standardized clinical care, behavior requirements, appropriate data processing and modeling are important in future guidelines.

Concerns have also been raised in the previous literature on individual smoking habits, supposing that this variable would mask pathologically induced differences in breath compounds [47–49]. Our results of Chi-square tests indicated there was little difference in identifying people with different smoking habits. This phenomenon was accordant with some relevant studies using electronic noses [46, 50], showing its great potential to be applied in a diverse population. We speculated that one reason was the effective smoking ban and gargling before breath tests. Another reason might be the sensor array was highly selective and little sensitive to smoking-related components. Therefore, whether this conclusion is extensive and universal needs further verification.

The novel contribution of our study was that a customized e-nose system was applied to explore two types of diagnostic models with independent external validation. Although some articles did not do external validation, many scholars confirmed the necessity of this vital step to check the reproducibility of tests [51–53]. In addition to reducing the environmental interference, some actions were adopted to elevate the test reproducibility in this study. On the one hand, we regulated individual behaviors before sampling and customized the sample pre-processing system to reduce system error. On the other hand, cost-sensitive algorithms was adopted to improve the

classifiers. The linear fitting slope of each signal was extracted along with the median also, covering the data information of the whole breath exposure. Consequently, the static characteristics of sensors like drift and linearity were naturally incorporated into the integrated data features. Classification results of the validation and training datasets, or between sensitivity and specificity, did not show a wide gap in this work, indicating that our frame of algorithms was technically reasonable and applicable. Successful external validation of models is an important step forward in silicosis detection.

5. Limitations

It should be noted that this study has examined only the population from one place and there were some limitations. Ideally, data from another location to form a multicenter test could better verify the adaptability of a diagnostic model. At present, technical noise and lack of multicenter trials still hamper the application of this technique to screen larger cohorts of miners.

6. Conclusion

Our investigation proposed a technical scheme for silicosis detection based on breath tests and electronic nose. A customized system, standard operating procedure, and improved modeling played a huge role in exploring this technique. In comparison with other diagnostic techniques, this method suggested a great potential of identifying early cases. The classification results in validation have proved the potential and robustness of the improved diagnostic models using ensemble learning algorithms. Overall, breath tests using the cross-reactive e-nose is worth developing to provide a quick and accurate detection method for silicosis.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

The article was financially supported by the 2nd Xplorer Prize sponsored by the Tencent Foundation, National Natural Science Foundation of China (Grant No. 52074274), and Changjiang Scholar Program of Chinese Ministry of Education (Grant No. IRT_17R103).

Conflict of interest

None declared

ORCID iDs

Fubao Zhou  <https://orcid.org/0000-0002-9464-2252>

Ling Zang  <https://orcid.org/0000-0002-4299-0992>

References

- [1] Wang T, Sun W, Wu H, Cheng Y, Li Y, Meng F and Ni C 2021 Respiratory traits and coal workers' pneumoconiosis: mendelian randomisation and association analysis *Occup. Environ. Med.* **78** 137–41
- [2] Blanc P D and Seaton A 2016 Pneumoconiosis redux coal workers' pneumoconiosis and silicosis are still a problem *Am. J. Respir. Crit. Care Med.* **193** 603–5
- [3] Greenberg M I, Waksman J and Curtis J 2007 Silicosis: a review *Dis.-A-Mon.* **53** 394–416
- [4] 2019 Improving occupational health in China *Lancet* **394** 443
- [5] Bell J L and Mazurek J M 2020 Trends in pneumoconiosis deaths—United States, 1999–2018 *Morb. Mortal. Wkly. Rep.* **69** 693–8
- [6] Wagner G R 1997 Asbestosis and silicosis *Lancet* **349** 1311–5
- [7] Hoy R F, Glass D C, Dimitriadis C, Hansen J, Hore-Lacy F and Sim M R 2021 Identification of early-stage silicosis through health screening of stone benchtop industry workers in Victoria, Australia *Occup. Environ. Med.* **78** 296–302
- [8] Mossman B T and Churg A 1998 Mechanisms in the pathogenesis of asbestosis and silicosis *Am. J. Respir. Crit. Care Med.* **157** 1666–80
- [9] Kim K I, Kim C W, Lee M K, Lee K S, Park C K, Choi S J and Kim J G 2001 Imaging of occupational lung disease *Radiographics* **21** 1371–91
- [10] Guarnieri G et al 2019 Multiorgan accelerated silicosis misdiagnosed as sarcoidosis in two workers exposed to quartz conglomerate dust *Occup. Environ. Med.* **76** 178–80
- [11] Baur X 2020 Diagnostic challenges of mixed dust silicosis (mixed dust pneumoconiosis)—5 case reports *Pneumologie* **74** 159–72
- [12] Lerman Y, Ribak J and Selikoff I J 1986 Hazards of lung biopsy in asbestos workers *Br. J. Ind. Med.* **43** 165–9
- [13] Brodrick E, Davies A, Neill P, Hanna L and Williams E 2015 Detection of lung cancer using expired breath analysis by ion mobility spectrometry *Eur. Respir. J.* **46** PA2999
- [14] Moor C C et al 2020 Exhaled breath analysis by use of eNose technology: a novel diagnostic tool for interstitial lung disease *Eur. Respir. J.* **57** 2002042
- [15] Manney S et al 2012 Association between exhaled breath condensate nitrate plus nitrite levels with ambient coarse particle exposure in subjects with airways disease *Occup. Environ. Med.* **69** 663–9
- [16] Di Natale C, Macagnano A, Martinelli E, Paolesse R, D'Arcangelo G, Roscioni C, Finazzi-Agro A and D'Amico A 2003 Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors *Biosens. Bioelectron.* **18** 1209–18
- [17] Farraia M V, Cavaleiro Rufo J, Paciencia I, Mendes F, Delgado L and Moreira A 2019 The electronic nose technology in clinical diagnosis: a systematic review *Porto Biomed. J.* **4** e42
- [18] Erzurum S C et al 2005 Can the electronic nose really sniff out lung cancer? Reply *Am. J. Respir. Crit. Care Med.* **172** 1060–1
- [19] Roeck F, Barsan N and Weimar U 2008 Electronic nose: current status and future trends *Chem. Rev.* **108** 705–25
- [20] Brooks S W, Moore D R, Marzouk E B, Glenn F R and Hallock R M 2015 Canine olfaction and electronic nose detection of volatile organic compounds in the detection of cancer: a review *Cancer Invest.* **33** 411–9

- [21] de Leon-Martinez D L et al 2020 Identification of profiles of volatile organic compounds in exhaled breath by means of an electronic nose as a proposal for a screening method for breast cancer: a case-control study *J. Breath Res.* **14** 046009
- [22] Binson V A, Subramoniam M and Mathew L 2021 Discrimination of COPD and lung cancer from controls through breath analysis using a self-developed e-nose *J. Breath Res.* **15** 046003
- [23] Kalidoss R, Umapathy S, Kothalam R and Sakthivelu U 2021 Adsorption kinetics feature extraction from breathprint obtained by graphene based sensors for diabetes diagnosis *J. Breath Res.* **15** 016005
- [24] Yang H-Y, Peng H-Y, Chang C-J and Chen P-C 2018 Diagnostic accuracy of breath tests for pneumoconiosis using an electronic nose *J. Breath Res.* **12** 016001
- [25] Yang H-Y, Shie R-H, Chang C-J and Chen P-C 2017 Development of breath test for pneumoconiosis: a case-control study *Respir. Res.* **18** 178
- [26] Bruderer T, Gaisl T, Gaugg M T, Nowak N, Streckenbach B, Muller S, Moeller A, Kohler M and Zenobi R 2019 On-line analysis of exhaled breath *Chem. Rev.* **119** 10803–28
- [27] Horvath I et al 2017 A European respiratory society technical standard: exhaled biomarkers in lung disease *Eur. Respir. J.* **49** 1600965
- [28] Zou Y, Wang Y, Jiang Z, Zhou Y, Chen Y, Hu Y, Jiang G and Xie D 2021 Breath profile as composite biomarkers for lung cancer diagnosis *Lung Cancer* **154** 206–13
- [29] Lee H S, Phoon W H and Ng T P 2001 Radiological progression and its predictive risk factors in silicosis *Occup. Environ. Med.* **58** 467–71
- [30] Spitzer S, Kiesewetter H, Bach R, Gutensohn J, Jung F, Schieffer H and Kleinsorge H 1993 Good clinical practice. Reorientation in clinical research. European society for good clinical practice *Dtsch. Med. Wochenschr* **118** 838–43
- [31] Dragonieri S, Annema J T, Schot R, van der Schee M P C, Spanevello A, Carratu P, Resta O, Rabe K F and Sterk P J 2009 An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD *Lung Cancer* **64** 166–70
- [32] Zang L, Che Y and Moore J S 2008 One-dimensional self-assembly of planar π -conjugated molecules: adaptable building blocks for organic nanodevices *Acc. Chem. Res.* **41** 1596–608
- [33] Zang L 2015 Interfacial donor-acceptor engineering of nanofiber materials to achieve photoconductivity and applications *Acc. Chem. Res.* **48** 2705–14
- [34] Chen S, Slattum P, Wang C and Zang L 2015 Self-assembly of perylene imide Molecules into 1D nanostructures: methods, morphologies, and applications *Chem. Rev.* **115** 11967–98
- [35] Konvalina G and Haick H 2014 Sensors for breath testing: from nanomaterials to comprehensive disease detection *Acc. Chem. Res.* **47** 66–76
- [36] Zang L and Slattum P 2017 Sensor compounds and associated methods and devices, PCT application #WO2017008074, US application # US20200354356A1, Europe application #EP3320329A4, China application # CN107850538A (available at: <https://patents.google.com/patent/US20200354356A1/en>)
- [37] Broza Y Y, Vishinkin R, Barash O, Nakhleh M K and Haick H 2018 Synergy between nanomaterials and volatile organic compounds for non-invasive medical evaluation *Chem. Soc. Rev.* **47** 4781–859
- [38] Anton A et al 2014 The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva *J. Breath Res.* **8** 034001
- [39] Lewis N S 2004 Comparisons between mammalian and artificial olfaction based on arrays of carbon black–polymer composite vapor detectors *Acc. Chem. Res.* **37** 663–72
- [40] Koushlesh R and Rajeev S 2018 Dog nose to E-nose in disease diagnosis *J. Adv. Biol.* **11** 2294–2306
- [41] Choi S-I, Oh J, Choi C-H and Kim C 2012 Input variable selection for feature extraction in classification problems *Signal Process.* **92** 636–48
- [42] Mukherjee R, Sengupta D and Sikdar S K 2015 *Computer Aided Chemical Engineering* ed F You (Amsterdam: Elsevier) pp 311–29
- [43] Nakhleh M K 2017 Diagnosis and classification of 17 diseases from 1404 subjects via pattern analysis of exhaled molecules *ACS Nano* **11** 112–25
- [44] Talaei-Khoei A and Wilson J M 2018 Identifying people at risk of developing type 2 diabetes: a comparison of predictive analytics techniques and predictor variables *Int. J. Med. Inform.* **119** 22–38
- [45] Jiang H, Mao H, Lu H, Lin P, Garry W, Lu H, Yang G, Rainer T H and Chen X 2021 Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease *Int. J. Med. Inform.* **145** 104326
- [46] Chen C-Y, Lin W-C and Yang H-Y 2020 Diagnosis of ventilator-associated pneumonia using electronic nose sensor array signals: solutions to improve the application of machine learning in respiratory research *Respir. Res.* **21** 45
- [47] Shehada N, Bronstrup G, Funke K, Christiansen S, Leja M and Haick H 2015 Ultrasensitive silicon nanowire for real-world gas sensing: noninvasive diagnosis of cancer from breath volatilome *Nano Lett.* **15** 1288–95
- [48] Filipiak W et al 2012 Dependence of exhaled breath composition on exogenous factors, smoking habits and exposure to air pollutants *J. Breath Res.* **6** 036008
- [49] Yang H-Y, Wang Y-C, Peng H-Y and Huang C-H 2021 Breath biopsy of breast cancer using sensor array signals and machine learning analysis *Sci. Rep.* **11** 103
- [50] Bofan M et al 2013 Within-day and between-day repeatability of measurements with an electronic nose in patients with COPD *J. Breath Res.* **7** 017103
- [51] Marco S 2014 The need for external validation in machine olfaction: emphasis on health-related applications *Anal. Bioanal. Chem.* **406** 3941–56
- [52] Capuano R, Catini A, Paolesse R and Di Natale C 2019 Sensors for lung cancer diagnosis *J. Clin. Med.* **8** 235
- [53] Covington J A, Marco S, Persaud K C, Schiffman S S and Nagle H T 2021 Artificial olfaction in the 21st century *IEEE Sens. J.* **21** 12969–90