Non-invasive detection of silicosis based on array sensing and pattern recognition

Wufan Xuan[1,2], Zhen Han[1,2], and Lina Zheng[1,2]

[1]School of Safety Engineering, China University of Mining and Technology, 221116 Xuzhou, P.R.China
[2]Xuzhou Engineering Research Center for Occupational Dust Control and Environmental Protection, 221116 Xuzhou, P.R.China

Abstract. Silicosis is a fibrotic lung disease caused by inhalation of silica dusts, early and accurate diagnosis of which remains a challenge. We aimed to assess the performance of a nanofiber sensor array and pattern recognition to promptly and noninvasively detect silicosis. A total of 210 silicosis cases and 430 non-silicosis controls were enrolled in a cross-sectional study. Exhaled breath was analysed by a portable analytical system incorporating an array of 16x organic nanofiber sensors. Models were established by Deep Neural Network and eXtreme Gradient Boosting. Linear Discriminant Analysis was used for dimensionality reduction and visualized data analysis. Receiver Operating Characteristic Curve, accuracy, sensitivity and specificity were used to evaluate models. Results: 99.3% AUC, 96.0% accuracy, 94.1% sensitivity, and 96.3% specificity were achieved in test set. Silicosis cases present different breath patterns from healthy controls, classification results using which were highly consistent with the experts' diagnosis. Breath analysis performed with the sensor array and pattern recognition is expected to provide a quick, stable recognition for silicosis. In this paper, different forms of features, different algorithms and data sets over long time periods were used, which provides a reference for silicosis expiratory diagnosis scheme.

1 Introduction

Silicosis is one of the most critical occupational diseases worldwide [**1, 2**], which is incurable and becomes less treatable in late stages [**3**]. Early detection of silicosis is critical for improving the life expectancy, as prompt treatment intervention significantly improves prognosis. At present, its diagnosis mainly relies on X-ray [**4**], which are challenging for early detection [**5, 6**], or invasive biopsies.

In recent years, respiratory analysis has shown the potential for detection of diseases (e.g., cancers) by monitoring the change in VOCs excreted from human breath or skin emanation [**7, 8, 9**]. However, most of the breath studies relied on the conventional bench-top analytical systems, such as GC-MS. These devices are complicated and cumbersome, which also need skilled workers to operate and are unsuitable as a point of-care tool. To address the challenge, electronic nose (e-nose) based on chemical sensor array has been developed and shown great potential in quick diagnosis of diseases [**10, 11**]. In 2018, prof. Yang et al. performed breath tests using Cyranose^{TM} to detect asbestosis, for which an accuracy rate of 70.0% was achieved in cross validation under a sensitivity of 66.7% [**12**]. Despite of the relatively small cohort group, this study showed the potential of developing a non-invasive tool for screening pneumoconiosis. In this context, we used a portable system to discriminate silicosis from healthy miners via sensors array and pattern recognition in a cohort of 640 subjects. Aimed to access and improve model effectiveness, we explored four forms of data features in cross combinations with two algorithms. Some exploratory analysis was also done to provide reference for potential clinical use.

2 Methods

2.1 Study design and subjects

This single-center, cross-sectional study with 640 miners obtain breath patterns by eNose. A total of 118 silicosis patients and 92 suspected silicosis patients were investigated in case group; 430 miners with no disease in lungs were healthy controls. Breath sample collection was done with a standardized process involving behaviour requirements, breath collection and pretreatment to reduce the interference. Extreme Gradient Boosting (XGBoost) and Deep Neural Network (DNN) were used to build classification models with four forms of features. Linear discriminant analysis (LDA) was used to visualize the classification performance. In model construction, 40% samples were set as the test set. The main research process is shown in Figure 1.

Behaviour requirements: subjects were asked not to eat, smoke, or take medicine for 10 hours prior to the breath collection. Subjects were asked not to eat onions, garlic or any other food with strong smells, or go to a

* Corresponding author: zhenglina@cumt.edu.cn

dusty environment for two days prior to 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.

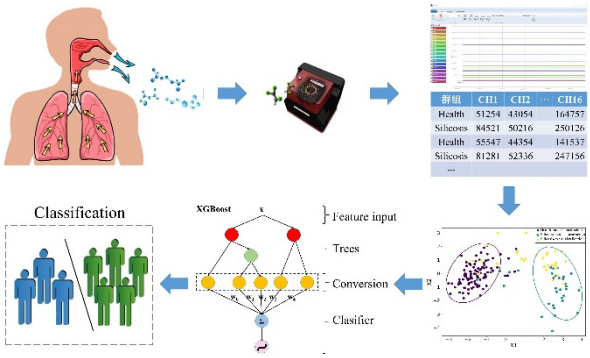

Fig 1. Schematic of study processes

2.2 Breath sampling and pretreatment

The breath collection site was well-ventilated over the whole analysis period to maintain clean ambient air as the reference. The collection system included mouthpiece, tubing, and a Tedlar^{TM} sample bag, all of which were made of polytetrafluoroethylene. All materials were disposable and the gas path was as streamlined and short as possible.

Human breath was stored in Teflon bags, which was pretreated by self-made system in Figure 2. The exhaled breath was analyzed along with the ambient air for 2 minutes.

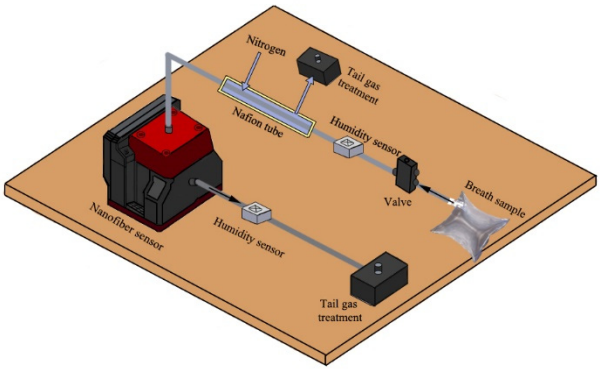

Fig 2. Pre-processing flow diagram

2.3 Data preprocessing

Min-Max method was used for normalization in Python 3.7.1 with Pycharm 2021.1 x64. Four features were extracted, which were the median value of breath exposure, median difference between the ambient air and breath data, the Pearson's correlation coefficient (PCC) between the ambient air and breath, and two well-behaved features combination.

3 Results

In total, 210 consecutive silicosis cases and 430 non-silicosis controls were enrolled in this cross-sectional study. Two-dimensional projection images using LDA with the median of breath data is shown Figure 3. In LDA diagram, points in the same group gathered into a mass apparently, and the points of "Confirmed silicosis" and "Suspected silicosis" couldn't be clearly distinguished by a straight line, both separated from those of "Healthy control" along axis X1. Point distribution of the "Healthy control" and "Confirmed silicosis" had little overlap, while some points of "Suspected silicosis" appeared in the point distribution region of "Healthy control".

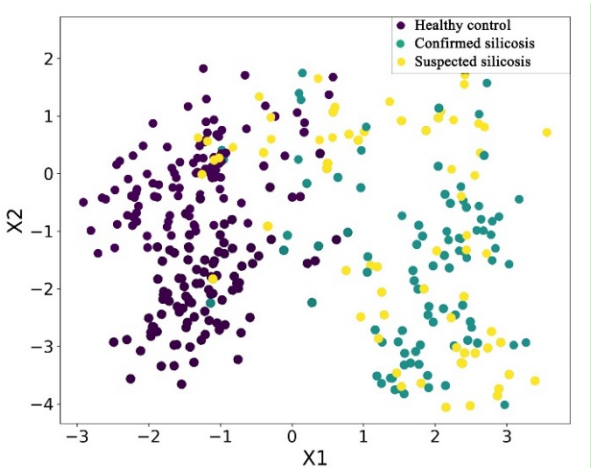

Fig 3. Two-dimensional projection images with the median of breath data using linear discriminant analysis

In model construction, 40% samples were set as the test set. Using DNN and XGBoost with the four forms of data feature extracted, classification results are shown in Table 1, including result of the train and test sets. The four evaluation index values of the test set result in the optimal classifier are all over 90%, among which "sensitivity" suggests apparent weakness, especially in models using DNN.

Receiver operating characteristic (ROC) analysis of models in test sets constructed with different algorithms and data features are shown in Figure 4, which is a visualized showing of AUC. ROC analysis is consistent with results shown in Table 1. In Table 1 and Figure 4, Models using the combined median and PCC in the four features and models using XGBoost in algorithm achieved superior performance to others.

4 Discussion

In train and test sets, silicosis cases were recognized with high accuracy. The best classifier used the XGBoost based on the combined feature of median and PCC, achieving the excellent performance with an AUC of 99.3%. DNN is a classic and powerful classification algorithm for deep learning but it did not perform well in this study, probably because that the number of samples or data dimensionality was well below the level for multilayers learning machine and big data. In terms of feature extraction, PCC not only contained exposure data about the exhaled air and air baseline, but it took their relation to deal with static characteristics of sensors as well, such as signal drifts. So the combined data of median and PCC can perform well

in datasets involving long time periods. And the appropriate balance of sensitivity and specificity is expected to be found in application cases.

In addition, attempts to identify subjects with suspected silicosis were shown to be possible because the cases were easy to be distinguished from healthy controls in breathprints as indicated by LDA plots. LDA is a dimension reduction technique of supervised learning. It projects data on the low dimension and selects the projection direction with the best classification performance. Its goal is to minimize intra-class variance and maximize inter-class variance, which is especially suitable for data sets containing samples of different categories.

Furthermore, it is interesting that subjects in this study covered large age ranges, various smoking habits and even some benign lesions of their bodies, and the classifiers still showed stable and outstanding performance in differentiating the two groups, suggesting that the these demographic variables may be less influential to silicosis stage on breathprints. This result shows breath analysis may be applied in screening of diverse subjects.

In this study, the population was not divided into three categories as we named above, which caused some limitations to the research on early screening and warning of this technology. And the reason we thought about it was that there would be a very large difference in quantities of samples of the three sets if we did so.

Table1. Classification results using DNN and XGBoost

Algorithm	Dataset	Feature	AUC	Accuracy	Sensitivity	Specificity
DNN	Train set	Median	0.787	0.919	0.550	0.976
		Median difference	0.693	0.872	0.333	0.946
		PCC	0.665	0.878	0.295	0.954
		Median+PCC	0.827	0.905	0.550	0.961
	Test set	Median	0.734	0.838	0.384	0.901
		Median difference	0.319	0.777	0.231	0.860
		PCC	0.562	0.727	0.4	0.764
		Median+PCC	0.877	0.858	0.571	0.905
XGBoost	Train set	Median	1.000	0.993	0.952	1.000
		Median difference	1.000	0.993	0.937	1.000
		PCC	1.000	0.993	0.944	0.565
		Median+PCC	1.000	0.993	0.933	0.856
	Test set	Median	0.945	0.980	0.867	1.000
		Median difference	0.891	0.818	0.500	0.846
		PCC	0.688	0.727	0.615	0.744
		Median+PCC	0.993	0.960	0.941	0.963

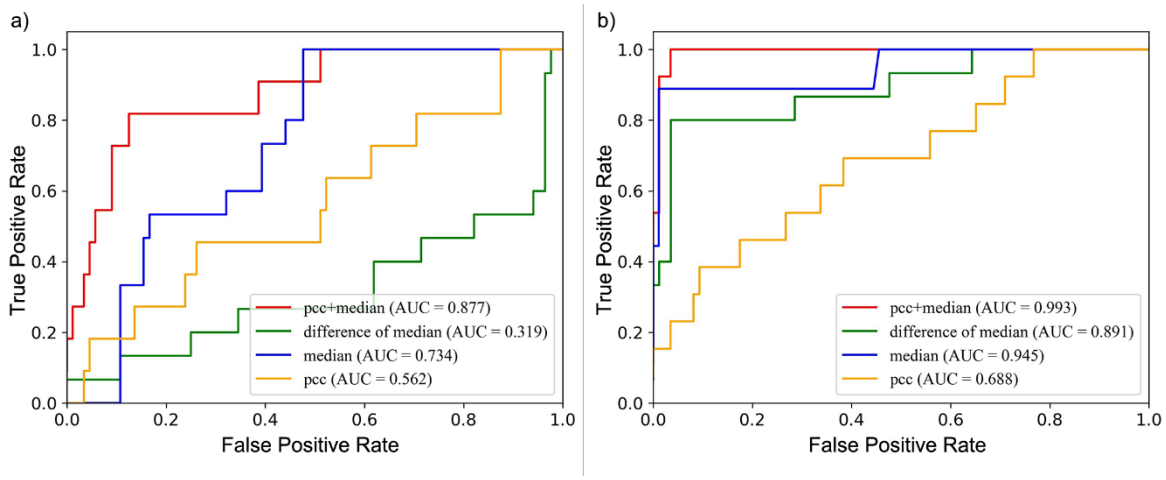

Fig 4. ROC Curves in test set. (a) Result using DNN; (b) Result using XGBoost

5 Conclusion

Breath analysis was investigated as a method for diagnosing silicosis with an array of chemical sensors. This study developed a breath testing system and showed excellent performance in a single-center study. Our investigation supports the hypothesis that similarities in silicosis group are expressed in breath patterns and the patterns are distinct from those of the healthy miners. This

study serves as a proof of concept that breath analysis can be adopted with cross-reactive sensor arrays and pattern recognition to enable silicosis screening.

We thank the participants for their volunteering in this study, and the doctors and nurses from local Institute of Occupational Disease Prevention for helping organize the sampling of the subjects.

References

1. P. Blanc and A. Seaton, Pneumoconiosis redux coal workers' pneumoconiosis and silicosis are still a problem, Am. J. Respir. Crit. Care Med. **193** (2016), no. 6, 603-605.
2. M. Greenberg, J. Waksman and J. Curtis, Silicosis: A review, DM, Dis.-Mon. **53** (2007), no. 8, 394-416.
3. J. Bell and J. Mazurek, Trends in pneumoconiosis deaths - united states, 1999-2018, Morb. Mortal. Wkly. Rep. **69** (2020), no. 23, 693-698.
4. K. Kim, C. Kim, M. Lee, K. Lee, C. Park, S. Choi and J. Kim, Imaging of occupational lung disease, Radiographics **21** (2001), no. 6, 1371-1391.
5. X. Baur, Diagnostic challenges of mixed dust silicosis (mixed dust pneumoconiosis) - 5 case reports, Pneumologie (Stuttgart, Germany) **74** (2020), no. 3, 159-172.
6. G. Guarnieri, R. Bizzotto, O. Gottardo, E. Velo, M. Cassaro, S. Vio, M.G. Putzu, F. Rossi, P. Zuliani, F. Liviero, P. Mason and P. Maestrelli, Multiorgan accelerated silicosis misdiagnosed as sarcoidosis in two workers exposed to quartz conglomerate dust, Occup. Environ. Med. **76** (2019), no. 3, 178-180.
7. T. Bruderer, T. Gaisl, M. Gaugg, N. Nowak, B. Streckenbach, S. Muller, A. Moeller, M. Kohler and R. Zenobi, On-line analysis of exhaled breath, Chem. Rev. **119** (2019), no. 19, 10803-10828.
8. M. Nakhleh, H. Amal, R. Jeries, Y. Broza, M. Aboud, A. Gharra, H. Ivgi, S. Khatib, S. Badarneh, L. Har-Shai, L. Glass-Marmor, I. Lejbkowicz, A. Miller, S. Badarny, R. Winer, J. Finberg, S. Cohen-Kaminsky, F. Perros, D. Montani, B. Girerd, G. Garcia, G. Simonneau, F. Nakhoul, S. Baram, R. Salim, M. Hakim, M. Gruber, O. Ronen, T. Marshak, I. Doweck, O. Nativ, Z. Bahouth, D.-y. Shi, W. Zhang, Q.-l. Hua, Y.-y. Pan, L. Tao, H. Liu, A. Karban, E. Koifman, T. Rainis, R. Skapars, A. Sivins, G. Ancans, I. Liepniece-Karele, I. Kikuste, I. Lasina, I. Tolmanis, D. Johnson, S.Z. Millstone, J. Fulton, J.W. Wells, L.H. Wilf, M. Humbert, M. Leja, N. Peled and H. Haick, Diagnosis and classification of 17 diseases from 1404 subjects via pattern analysis of exhaled molecules, Acs Nano **11** (2017), no. 1, 112-125.
9. C. Di Natale, A. Macagnano, E. Martinelli, R. Paolesse, G. D'Arcangelo, C. Roscioni, A. Finazzi-Agro and A. D'Amico, Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors, Biosens. Bioelectron. **18** (2003), no. 10, 1209-1218.
10. M. Farraia, J. Cavaleiro Rufo, I. Paciencia, F. Mendes, L. Delgado and A. Moreira, The electronic nose technology in clinical diagnosis: A systematic review, Porto biomedical journal **4** (2019), no. 4, e42-e42.
11. S. Erzurum, T. Burch, D. Laskowski, P.J. Mazzone, T. Mekhail, C. Jennings, J. Stoller, R. Machado, J. Pyle, O. Deffenderfer and R. Dweik, Can the electronic nose really sniff out lung cancer? Reply, Am. J. Respir. Crit. Care Med. **172** (2005), no. 8, 1060-1061.
12. H. Yang, H. Peng, C. Chang and P. Chen, Diagnostic accuracy of breath tests for pneumoconiosis using an electronic nose, J. Breath Res. **12** (2018), no. 1.