Artificial Intelligence Based Breath Analysis System for the Diagnosis of lung cancer

Binson V A¹ and Dr M Subramoniam²
¹Research Scholar, Sathyabama Institute of Science and Technology, Chennai, India
²Associate Professor, Sathyabama Institute of Science and Technology, Chennai, India
E-mail: binsonvsabraham@gmail.com

Abstract. Breath analysis has become a promising tool for the detection of pulmonary diseases in recent years. This paper describes the fabrication of an artificial intelligence (AI) based e-nose system for discriminating lung cancer patients from healthy controls. Breath volatile organic compounds (VOC) can be easily analyzed using an electronic nose (e-nose) made of various gas sensor arrays. Metal oxide semiconductor (MOS)-based e-noses are getting popular in the VOC analysis of exhaled breath from humans. Here we have developed the e-nose system with a sensor array system of five MOS gas sensors and incorporated the controller and machine learning algorithms to process the data. The sensor array was designed with MOS sensors developed by Figaro USA and the data acquisition was carried out with the help of the Arduino Uno developer board. 40 healthy control samples and 24 lung cancer patient samples were analyzed using the developed e-nose system. The data analysis was done by two supervised classification algorithms random forest and logistic regression. Among this, random forest with 5-fold cross-validation gave better results with 85.38 % classification accuracy and 0.87 of AUC. This system can be further extended to the diagnosis of various other pulmonary diseases.

Keywords: lung cancer, volatile organic compounds, electronic nose, random forest, logistic regression.

1. Introduction

Electronic nose devices are nowadays used in several fields such as agriculture, airline transportation, cosmetics, food and beverage, medical, and military applications [1-5]. The electronic nose system has specific features that mimic the features of the human olfactory system. The biological nose consists of olfactory receptors, olfactory bulb, and brain whereas the e-nose system consists of an array of sensors, preprocessing, database, and pattern recognition system. When we breathe through the nose, our brain recognizes the odor. The human olfactory system relies on the chemistry between volatile compounds in the nasal cavity and olfactory receptors. The signals generated are transmitted to the brain by secondary neurons and directed to the limbic system in the cortex. The odor is detected there using neural network pattern recognition. Primary neurons adapt to the electronic nose's chemical sensors with different sensitivity to different smells. A chemical reaction between volatile compounds and gas sensors changes the chemical state of the sensors and generates electrical signals. The device records these signals in a similar way to secondary neurons. Like the human olfactory system, the e-nose must recognize an odor. The e-nose creates a digital image of different smells, these images are easily remembered and the deviation of these images is easily observed. It should be remembered that the e-nose is much less sensitive and selective than the human nose [6].

The e-nose system consists of a sensor array, sensor data processing system, and pattern recognition system. When chemicals are applied to the electronic nose system, the electrical...
responses vary depending on the properties of the compounds applied to it. When the chemical compound is applied to the system; according to the type of sensors in the sensor chamber, the output pattern of the sensor array varies. The conductance of the sensors varies depending on the compounds/chemical made in contact with the sensor array [7].

Human exhaled breath consists of thousands of organic and inorganic volatile compounds, in addition to nitrogen, oxygen, and carbon dioxide. These VOCs are produced by both exogenous and endogenous sources [8]. These volatile organic compounds are biomarkers of certain illnesses since the volatile organic compound level in the breath of healthy people and patients with certain diseases vary. The major components present in the exhaled breath of humans are shown in figure 1 [9]. By examining the amount and variation in the VOCs present in the exhaled breath of humans certain diseases like lung cancer, cystic fibrosis, prostate cancer, chronic obstructive pulmonary disease, diabetes mellitus, etc can be easily diagnosed [10]. Breath analysis has become a promising tool for the detection of pulmonary diseases in recent years. Breath VOC analysis can be easily done by the use of electronic noses composed of various gas sensor arrays. Metal oxide semiconductor based e-noses are getting popular in the exhaled breath analysis of humans for the detection of VOCs.

Lung cancer, the subject of this study, while comparing with other cancer is the most common of all types and is difficult to detect at its early stage. Even the mortality rate of lung cancer is more. If lung cancer is detected in the early stages survival rate can be increased to 50%-70%. Presently the diagnosis happens in its late-stage where the curative treatment is impossible and the survival rate is just under 5% in India [11]. The volatile organic compounds which act as biomarkers for lung cancer are Isoprene, Styrene, Ethane, Decane, Pentane, Benzene, Ethanol, Acetone, Undecane, etc [12]. The most common methods for breath analysis are Gas Chromatography-Mass Spectrometry (GCMS), Ion-mobility spectrometry (IMS), and Gas chromatography-Flame ionization detector (GC-FID). All these methods are expensive, time-consuming, and require a trained operator to use the device. Breath analysis by an electronic nose is a simple, fast, and low-cost method that gives better accuracy, sensitivity, and specificity.

The main objective of this work was to develop an artificial intelligence (AI) based e-nose system for analyzing the exhaled breath from healthy participants and lung cancer patients and thereby classifying them in two groups. The sensor array was developed with metal oxide semiconductor (MOS) sensors developed by Figaro USA (TGS sensors). The data acquisition was carried out with
the help of the Arduino Uno developer board, and the data analysis was carried out in python using machine learning algorithms. The system was tested in 24 lung cancer patients and 40 healthy controls for determining the proper functioning of the e-nose device.

2. Literature Review
Evaluating the amount of volatile organic components emitted from biological materials, it is easy to classify them to various groups. Electronic nose systems found the opportunity to be used in various fields such as to determine the quality of the food, environmental control, and human health. An electronic nose system for determining different properties consists of sensors, signal acquisition unit, and software for pattern recognition. Table 1 gives an overview of the recent studies done on e-nose systems in the diagnosis of lung cancer using exhaled breath analysis method.

Table 1. Recent studies on e-nose based lung cancer diagnosis.

| Year | First Author & Reference | Methods/ Technology used | Advantages | Disadvantages |
|------|--------------------------|--------------------------|------------|---------------|
| 2017 | Li W [13]                | E-Nose with 14 gas sensors for breath detection | Sensitivity, specificity and accuracy of 91.58%, 91.72% and 91.59% | High cost | Only 24 lung cancer patients and 13 healthy non-smokers were included for the study. |
|      |                          | Feature extraction algorithms used- PCA, LDA, LE, LLE, tSNE |            |               |
|      |                          | Classifier used-Fuzzy k-NN, and Support vector machine |            |               |
| 2018 | Chang [14]               | E-Nose with 7 MOS gas sensors for breath detection | Low cost | Accuracy, sensitivity and specificity is only 75%, 79% and 72% |
|      |                          | Feature’s extraction algorithms and classifiers used linear discriminant analysis, LDA, support vector machine, SVM, and multilayer perception, MLP | 37 LC patients and 48 healthy controls were selected and also 3 types of samples were collected | More smokers were not selected. |
| 2018 | Wong DM [15]             | Sensor array system with 14 metal oxide gas sensors | Robust design with 14 MOS gas sensors, readout PCB, and a DAQ card with LabVIEW software. | Classification accuracy is not good. | Sensitivity and specificity is not discussed. |
|      |                          | Classifier used-kNN, support vector machine (linear, polynomial, and radial basis function) |            | Details on the e-nose system is not given Patient demographics is not mentioned in the paper |
| 2019 | Kononov A                | Sensor array system | The solid-state | The details |
with six metal oxide chemoresistance gas sensors
• Classifier used-kNN, logistic regression, random forest, linear discriminant analysis, and support vector machine.

2020 Wang [18]
• E-Nose with 10 gas sensors for breath detection
• Feature extraction algorithms used- PCA Classifier used- Logistic regression, random forest, and support vector machine.

2020 Chen [19]
• E-Nose was constructed by metal-ion induced assembly of graphene oxide
• Feature extraction algorithms used- PCA Classifier used- Linear discriminant analysis, artificial neural network

Although commercially e-noses are available, they are costly and difficult to use. Therefore, more economical and simple systems need to be developed. Most of the innovative developments recently identified in the area of lung cancer diagnosis are related to the use of electronic nose system. These devices show the chemical signature pattern of the exhaled breath of patients without predicting the staging of the disease. Despite real analytical potential, these systems are still confined to research laboratories today since their analysis performance can still be improved.

3. Materials and Methods

3.1. Experimental setup:
The e-nose developed for this study using MOS sensors is shown in figure 2. The features and specifications of the five MOS sensors used for the fabrication of the sensor array are shown in table 2. The data used for the analysis is obtained from the 5 MOS sensor-based sensor array. The sensors are placed in a sensor chamber and we have connected two tubes for air in and air out. Before the breath sample is delivered to the sensor array, it is cleaned using pure air after each sampling process. The participants were asked to exhale through the mouth to the Tedlar bag sensor array for about two to three exhalation.
We have reviewed the use of Figaro gas (TGS sensors) in the application of various disease diagnoses, focusing on lung cancer detection. TGS 26XX series and TGS 8XX series sensors have shown better accuracy in the detection of lung cancer [13, 14, 18].
3.2. Sampling Procedure:
Breath samples were collected from healthy controls and lung cancer patients after getting written informed consent from them. We have selected 24 lung cancer patients and 40 healthy controls for this study. Also, we have excluded the participants who have other respiratory diseases like asthma, COPD, chronic bronchitis, cystic fibrosis, etc. We have collected a total of 192 samples from the volunteers enrolled in the study. From that 120 samples were from healthy controls and 72 were from lung cancer patients. ie. We have collected three samples from each of the participants under study. The details of the study subjects are given in Table 3.

Table 3. Details of study subjects.

|                          | Lung Cancer | Healthy Control |
|--------------------------|-------------|-----------------|
| No. of Volunteers       | 24          | 40              |
| Age years                | 65±5        | 46±6            |
| Gender                   |             |                 |
| Male                     | 17          | 26              |
| Female                   | 7           | 14              |
| Smoking Status           |             |                 |
| Never smoker             | 7           | 14              |
| Current smoker           | 2           | 24              |
| Former Smoker            | 15          | 2               |
| Tumor Staging            |             |                 |
| IA                       | 1           |                 |
| IB                       | 3           |                 |
| IIA                      | 2           |                 |
| IIB                      | 4           |                 |
| IIIA                     | 5           |                 |
| IIIB                     | 3           |                 |
| IV                       | 6           |                 |

Breath samples are collected in a 1 L Tedlar gas sampling bag made of polyvinyl fluoride. While collecting the sample, volunteers under study are asked to exhale into the 1 L gas sampling Tedlar bag through a mouthpiece. After the bag is filled, it is connected to the sensor array system. The computer program controls the e-nose device to finish the cycle automatically. The experiments performed in the system implemented consist of 4 stages.

i. Baseline stage: For subsequent data preprocessing and analysis the baseline value of each sensor is recorded. This stage is only for the initial two seconds.

ii. Injection stage: In this phase, the pump is on and the collected breath in Tedlar bag is released to the gas chamber at a fixed velocity. The sensor responds to the VOCs comes out of the exhaled breath and this stage lasts for about 8 seconds.

iii. Reaction stage: In this phase, the pump is off and the sensor continues to respond to the elements present in the exhaled breath. The sensor response reaches its maximum value in this stage and this stage lasts for about 50 seconds.

iv. Purge stage: In this stage, the pump is on again and the pure air is applied to the sensor array system to clean the gas sensor chamber. This process takes almost 60 seconds. In this phase, the sensor output value is again reduced to the initial baseline value. The sensor array is ready to receive another sample of breath when it is steady at its baseline value. After completing the whole measurement cycle of a sample, we will get a digitized ‘breath signature’ or ‘breath print’
3.3. Data Preprocessing:
Sensor shifts in electronic nose studies are one of the major problems. To avoid this problem the most common method used is baseline manipulation [20]. This by applying the difference signal \( Y_S(t) \) that is obtained by subtracting the baseline signal \( X_S(0) \) from the sensor signal \( X_S(t) \) which is obtained by applying the sampled breath gas and is given by equation 1.

\[
Y_S(t) = X_S(t) - X_S(0)
\]  

The signal received from MOS sensors in e-nose circuits is voltage values on load resistors connected in series with sensors. When the studies in the literature are examined, it is seen that its conductance is examined rather than the voltage readings. In this study, raw signals obtained from MOS sensors are primarily baseline manipulated and converted into conductivity. The selection of baseline manipulation method of sensor drift compensation is greatly correlated to particular sensor technology and for specific applications. For metal oxide semiconductor-based sensors it is identified that the fragmentary variations in conductance give better pattern recognition performance.

3.4. Feature Extraction and Feature Selection:
In this study, both reducing the data size and attribute to find the most distinctive compounds belonging to the feature extraction and feature selection process has been carried out. The main objective of the feature extraction step is to extract features from the multidimensional sensor data. The features include the transient signals, steady-state signals, the parameters of various response curve-fitting models, and the coefficients of different transforms. The features extracted from the sensor response curves are as follows [21].

i. The steady-state value: This is indicated as the maximum or minimum value reached during the whole measurement process.

ii. The deviation of resistance/conductance from the baseline value to the minimum or maximum point reached in the transient response, which is expressed in both ration and difference.

iii. The transient’s derivative, calculated as difference and sampling points computed among various positions of the transient.

For dimensionality reduction, we have used Principal component analysis (PCA).

3.5. Classification Method:
The classification algorithms are powerful and the goal is to forecast the class of an unspecified input vector. The classification algorithm may either be supervised or unsupervised. The fundamental concept of the supervised learning algorithm is to train the classifier by giving it a train of similar examples and corresponding known classes. In unsupervised learning approaches, the classifier learns to distinguish the various classes from the response vectors regularly without processing some previous details on class membership. In this work, we have used two supervised classification methods random forest and logistic regression.

The classification algorithm random forest is developed in 2001 by Leo Breiman with the concept of ensemble learning. Even though the name is regression, logistic regression is a classification algorithm. This method generally attempts to plot samples by a sigmoid function.

4. Results and Discussion
In this study, we have collected 120 samples from healthy controls and 72 from lung cancer patients for analysis purposes. The study was a 5-fold cross-validation study and the mean classification accuracy is shown in table 4. The sensor array response to the exhaled breath of a lung cancer patient...
is shown in figure 3.

Table 4. Classification performance of various algorithms.

| Classification algorithm | Accuracy (%) |
|--------------------------|--------------|
| Random Forest            | 85.38        |
| Logistic regression      | 83.59        |

From figure 3 we can understand that the TGS 2610 which is highly sensitive to ethanol, isobutene and propane gave a maximum output voltage of 3.46 V for the gas sample which is from a lung cancer patient who still smokes. His daily medicines for diabetes mellitus, high cholesterol, and hypertension also affected the high output response. TGS 2600 and TGS 2620 are sensitive to ethanol and carbon monoxide thereby produces more output voltages for exhaled breaths of smokers and drinkers. It can be noted from the graph that TGS 2600 and TGS 2620 output was almost constant for the whole measurement period for all the study subjects. From figure we can measure that it is just above 600 mV for this particular measurement. TGS 826 gave high sensitivity to ammonia gas and for this patient the sensor gave a maximum value of 2.65 V. TGS 822 that highly sensitive to acetone and ethanol gave a maximum output voltage of 1.56 V. In all the sensor responses the baseline stage, injection stage, reaction stage and purge stage was around two seconds, eight seconds, fifty seconds, and sixty seconds respectively.

Figure 3. Sensor array response to a cancer patient.

Figure 4 (a) and 4 (b) show the receiver operating characteristic (ROC) curves of the classification algorithms logistic regression and random forest respectively. They depict the ROC curves classification of breath patterns from lung cancer patients and healthy controls using logistic regression and random forest respectively. The mean AUC (area under the curve) for differentiating lung cancer patients from healthy controls was around 0.87 for the random forest classification algorithm and 0.86 for logistic regression.
Since we have used various sensors for assembling the sensor array, we obtained different breath patterns for patients and healthy peoples as we have discussed earlier. It is also noted that there is a drastic change in the breath cycle in the purge stage of the measurement cycle. This is due to the presence of huge variations in the components of the gases that we inhale and exhale. From figure 3, we have also noticed that all the sensors were not able to provide better features for discriminating lung cancer patients from healthy controls whereas some of them performed well.

Lung cancer detection with the help of various sensor technologies gave meaningful and promising results. Exhaled breath VOC analysis for screening lung cancer is a rapidly developing area of research. Development of standardized and flexible breath sampling protocols, multicenter clinical research and deep understanding biochemical pathways associated with the development and the progression of lung cancer, will speed up the pace of development of a robust marker for the early diagnosis of the illness [5]. The presently available commercial e-noses are costly and difficult to use. Therefore, more economical and simple systems should be designed and developed. The present innovations in the area of breath research are based on the developments of e nose devices for early prediction of lung diseases. These systems show the VOC ‘breath pattern’ of the exhaled breath of patients without predicting the disease staging and also not calculating the exact amount of the VOCs in exhaled breath [13]. By further studies and analysis, this system can be extended to the diagnosis of other different pulmonary diseases.

5. Conclusion
The equipment developed to discriminate lung cancer patient breath print from healthy control breath print using 5 commercially available MOS sensors was able to perform the function with more than 80 % accuracy. The developed artificial intelligence (AI) based e-nose system was able to analyze the exhaled breath from healthy participants and lung cancer patients and thereby classify them in two groups. In this work, random forest with 5-fold cross-validation gave better accuracy than logistic regression and it is 85.38 %. This simple, low-cost electronic nose has given better results in the detection of lung cancer. This system can be further studied to use for the diagnosis of various other respiratory diseases like COPD, asthma, cystic fibrosis, etc.

In a brief, we designed and developed an electronic nose-based breath analysis system using five MOS sensors. The practical applicability of the system was evaluated with 192 human exhaled breath samples comprising of both healthy controls and lung cancer patients. The application of type different sensors may greatly improve the accomplishment of the electronic nose-based breath analysis systems, but still to what extent that may improve is not known. This simple, portable, low cost, non-invasive lung cancer diagnostic tool can be effectively utilized in the remote areas of the country.
References

[1] Vapnik, V.; Lerner, A. 1963, Pattern recognition using generalized portrait method. Autom. Remote Control, 24,774–780.

[2] S. Balasubramanian, S. Panigrahi, C. Logue, C. Doetkott, M. Marchello, J. Sherwood, 2008, Independent component analysis-processed electronic nose data for predicting Salmonella typhimurium populations in contaminated beef, Food Control, 19, 236–246.

[3] A.C. Romain, D. Godefroid, J. Nicolas, 2005, Monitoring of a compost pile with an e-nose and comparison with GC-MS data. Sens. Actuators B, 106, pp. 29-35.

[4] van Gelen WH, Bruins M, Kerstjens HA, 2016, Diagnosing viral and bacterial respiratory infections in acute COPD exacerbations by an electronic nose: a pilot study. J Breath Res, 10:036001

[5] Behera B, Joshi R, Vishnu GA, Blalerao S, Pandya HJ, 2019, Electronic nose: A non-invasive technology for breath analysis of diabetes and lung cancer patients. J Breath Res, 13(2):024001.

[6] Röck F, Barsan N, Weimar U, 2008, Electronic nose: current status and future trends, Chemical reviews, 13, 108(2):705-25.

[7] Pearce, T. C., Schiffman, S. S., Nagle, H. T., Gardner, J. W., 2003, Handbook of Machine Olfaction: Electronic Nose Technology, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

[8] Chnabel, R., Fijten, R., Smolinska, A. et al, 2015, Analysis of volatile organic compounds in exhaled breath to diagnose ventilator-associated pneumonia. Sci Rep 5, 17179.

[9] Das S, Pal M, 2020, Non-invasive monitoring of human health by exhaled breath analysis: A comprehensive review. Journal of The Electrochemical Society, 167(3):037562.

[10] Guo D, Zhang D, Li N, Zhang L, Yang J, 2010, A novel breath analysis system based on electronic olfaction. IEEE transactions on biomedical engineering, 57(11):2753-63.

[11] Binson VA, Subramoniam M, 2018, Advances in Early Lung Cancer Detection: A Systematic Review. In2018 International Conference on Circuits and Systems in Digital Enterprise Technology, pp. 1-5, IEE.

[12] Peng G, Tisch U, Adams O, Hakim M, 2009, Diagnosing lung cancer in exhaled breath using gold nanoparticles”, Nature Nanotechnology, 4(10):669-73, 2009.

[13] Li W, Liu H, Xie D, He Z, Pi X, 2017, Lung cancer screening based on type-different sensor arrays, Scientific reports, 7(1):1-2.

[14] Chang JE, Lee DS, Ban SW, Oh, 2018, Analysis of volatile organic compounds in exhaled breath for lung cancer diagnosis using a sensor system. Sensors and Actuators B: Chemical, 255:800-7.

[15] Wong DM, Fang CY, Chen LY, Chiu, 2018, Development of a breath detection method based on electronic nose for lung cancer identification. In2018 IEEE International Conference on Applied System Invention, pp. 1119-1120, IEEE.

[16] van de Goor R, van Hooren M, Dingemans AM, Kremer B, Kross K, 2018, Training and validating a portable electronic nose for lung cancer screening. Journal of Thoracic Oncology, 13(5):676-81.

[17] Kononov A, Korotetsky B, et al, 2019, Rassadina A. Online breath analysis using metal oxide semiconductor sensors (electronic nose) for diagnosis of lung cancer. Journal of breath research, 14(1):016004.

[18] Li W, Jia Z, Xie D, Chen K, Cui J, Liu H, 2020, Recognizing lung cancer using a homemade e-nose: A comprehensive study. Computers in Biology and Medicine, 103706.

[19] Chen Q, Chen Z, Liu D, He Z, Wu J, 2020, Constructing an E-Nose Using Metal-Ion-Induced Assembly of Graphene Oxide for Diagnosis of Lung Cancer via Exhaled Breath. ACS Applied Materials & Interfaces, 12(15):17713-24.

[20] Gutierrez-Osuna R, Nagle HT, Schiffman SS, 1999, Transient response analysis of an electronic nose using multi-exponential models. Sensors and Actuators B: Chemical, 61(1-3):170-82.

[21] Blatt R, Bonarini A, Matteucci M, 2010, Pattern classification techniques for lung cancer diagnosis by an electronic nose, Computational Intelligence in Healthcare, (pp. 397-423).