An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction

Xianghao Zhan[a,c], Zhan Wang[a], Meng Yang[b], Zhiyuan Luo[d], You Wang[a], Guang Li[a,*]

[a] *State Key Laboratory of Industrial Control Technology, Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, Zhejiang, China*
[b] *Department of Computer Science and Technology, School of Mechanical Electronic & Information Engineering, China University of Mining & Technology, Beijing, 100083, China*
[c] *Department of Bioengineering, Stanford University, Stanford, 94305, CA, USA*
[d] *Computer Learning Research Center, Royal Holloway, University of London, Egham Hill, Egham, Surrey TW20 0EX, UK*

Abstract

Lung cancer leads to high mortalities in various countries while the reliability of cancer diagnosis has not been paid enough attention. In this work, a novel application of conformal prediction in lung cancer diagnosis with electronic nose is introduced. The nonconformity measurement is based on k-nearest neighbors. In offline prediction, accuracies of 87.5% and 83.33% have been achieved by conformal predictors based on 1NN and 3NN respectively, outperforming those of simple k-nearest neighbor predictors. Additionally, conformal predictors provides confidence and credibility information of each prediction that could inform the patients of diagnostic risks. In online prediction, with increasing number of samples, the frequency of errors given by conformal predictions can gradually be limited by the significance level set by users. This project manifests that electronic nose promises to be an applicable cheaper analytic tool in assisting lung cancer diagnosis and conformal prediction provides a promising method to ensure reliability.

Keywords: Conformal prediction, electronic nose, lung cancer, online

*Corresponding author
Email addresses: xzhan96@stanford.edu (Xianghao Zhan), 11732003@zju.edu.cn (Zhan Wang), m.yang@cumtb.edu.cn (Meng Yang), zhiyuan@cs.rhul.ac.uk (Zhiyuan Luo), king_wy@zju.edu.cn (You Wang), guangli@zju.edu.cn (Guang Li)

prediction, reliability

1. Introduction

Lung cancer has gradually become one of the most fatal diseases which contributes to huge mortalities in various nations [1, 2, 3, 4, 5]. The diagnosis of lung cancer is usually in its terminal stages and the five-year survival rate is quite low. Earlier diagnosis can improve the survival rate to a great extent [6, 7].

Scientists has already applied various methods to diagnose lung cancer in an earlier state [8, 9, 10, 11] based on biopsy. For instance, Tan used support vector machine to analyze the gene expression statistics, and reached a classification accuracy of 96.61% [11]. However, biopsy, as the state-of-art and most reliable diagnostic technique, may do harm to patients and could not be applied frequently in a short period of time. Computed Tomography(CT) offers a frequently used non-invasive method that could help early detection but with the disadavantages of high cost, ionizing radiation and false positive results. Therefore, a cheaper, more convenient, radiationless and non-invasive diagnostic method for lung cancer is much in need.

Breath air, as a product of metabolism containing volatile organic compound(VOC), serves as an indicator of human health. Using breath air to diagnose such disease as lung cancer has been a research spot [12, 13, 14]. For instance, some scholars applied gas chromatography(GC) and mass spectroscopy(MS) to find out patterns associated with lung cancer[15, 16]. Nevertheless, GC-MS is quite expensive, complicated and time-consuming for widespread applications. Electronic nose is an artificial olfaction system capable of analyzing volatile gas mixture with sensors sensitive to different VOCs [17, 18, 19, 20]. As a low-cost, relatively compact analytic tool, electronic nose has been applied to many domains such as air environment quality evaluations [21, 22, 23], assistant medical diagnosis [24, 25, 26, 27] and food and beverage quality tests [28, 29, 30, 31, 32, 33, 34]. The application of electronic nose in lung cancer

diagnosis promises a cheaper and more convenient assistant diagnostic method without invasion and radiation.

In addition, what has been frequently ignored in lung cancer diagnosis is the prediction reliability and overall accuracy. Diagnostic misinterpretation and erroneous prediction may exert tremendous psychological and financial burdens on patients and their families. Therefore, lung cancer diagnosis should not only provide prediction results but also confidence associated with each prediction, to help doctors to make the best decisions and provide patients with effective information about risks. To solve the problem of reliability of each prediction and the accuracy of overall predictions, such methods as probably approximately correct learning(PAC), Bayesian learning, hold-out validation and cross validation have been put forwards by scholars. However, PAC generally requires a large number of samples and does not offer specific information on the reliability of individual prediction [35]. Bayesian learning, logistic regression [36] and Platts method [37] provides additional probabilistic information about reliability for each prediction. For instance, before outputting a label to an observation, logistic regression and Bayesian methods give out conditional probability given the features. Some of these methods rely on distribution assumption or model assumption. For instance, linear discriminant analysis and quadratic discriminant analysis, two Bayesian methods, rely on normal distribution with the same or different covariance in different classes and employ maximum likelihood to evaluate distribution parameters. Others rely on the data to calculate priori to infer the true conditional distribution of each class. Due to sensor drifts, measurement noises and collinearity among features, the distribution assumptions and model assumptions may not be proper. Meanwhile, the small research training set in biomedical researches are usually biased in classes. The frequency of the disease cases may not be an accurate estimate of the prior probability. Therefore, the models and reliability may not be robust enough with the process of seeking for prior information. In addition, such methods as hold-out validation and cross-validation often predict overoptimistically in reality in that responses given by electronic noses may be influenced by sensor drifts, sensor aging or

sensor poisoning.

Conformal prediction, which is based on the assumption that all samples and their associate labels are generated from an independent and identical distribution (IID), was firstly put forward by Vladimir Vovk. and his co-workers [35, 36, 38, 39]. Based on a weaker assumption when compared with methods mentioned above, when applied to the actual data from electronic nose, it can provide a promising method for the evaluation of prediction reliability. Conformal prediction not only gives individual prediction, but also reveals the conformity of each new prediction when compared to the group of observations already examined, able to offer additional yet important information about the confidence and credibility for each prediction and avoid overestimating the overall accuracy of prediction [40].

In this work, samples of breath air from lung cancer patients and controls are gathered and analyzed with an electronic nose, which is followed by the usage of conformal prediction based on k-nearest neighbors for lung cancer sample classification. In Section 2, the definition of conformal prediction and its applications in online prediction and offline prediction are illustrated respectively. Then, the sampling processes, the electronic nose system and data processing methods are illustrated in Section 3. Results and implications are discussed and analyzed in Section 4. Finally, the conclusion is drawn in Section 5.

2. Methods

2.1. Conformal Prediction

2.1.1. Definition

In the problem of machine learning classification, there are usually a set of samples as the training sets.

Each observation contains one object $x_i \in X$ and one label $y_i \in Y$, in which X denotes object space and Y denotes label space. As long as a new sample x_n is given, the task of classification is to predict the label to which this sample

belongs.

$$((x_1, y_1), ..., (x_{n-1}, y_{n-1})) \tag{1}$$

$$z_i = (x_i, y_i), i = 1, 2, ..., (n-1) \tag{2}$$

With the samples and labels combined, example space Z can be set.

What simple predictor does is to find one mapping F which helps the prediction of new samples x_n based on present example space Z^*.

$$F : Z^* \times X \longrightarrow Y \tag{3}$$

When compared with simple predictors, conformal predictors have another parameter $\epsilon \in (0, 1)$, which is defined as the significance level. Meanwhile, 1-ϵ is defined as the confidence level, which reflects the confidence for individual prediction. A conformal predictor output a subset of label space Y based on a given significance level:

$$\Gamma^{\epsilon}((x_1, y_1), ..., (x_{n-1}, y_{n-1}), x_n) \tag{4}$$

The subsets output by conformal predictors are nested, as is shown below:

$$\Gamma^{\epsilon_1}(z_1, ..., z_{n-1}), x_n) \subset \Gamma^{\epsilon_2}(z_1, ..., z_{n-1}), x_n)(\forall \epsilon_1 \geq \epsilon_2) \tag{5}$$

In order to output the prediction sets, it is necessary to introduce a measurable function A mapping each observation $z_i (i = 1, 2, \ldots, n) \in Z$ to a real number $\alpha_i (i = 1, 2, \ldots, n) \in R$. The process is defined as nonconformity measurement, which reflects how conformed each observation is when placed in a group of other observations:

$$\alpha_i = A(z_1, ...z_{i-1}, z_{i+1}, ..., z_n), i = 1, ..., n \tag{6}$$

Additionally, A should be based on a specific statistical learning algorithm and satisfy the following property of exchangeability assumption: for any n and any permutation π,

$$(\alpha_1, ..., \alpha_n) = A(z_1, ..., z_n) \longrightarrow (\alpha_{\pi(1)}, ..., \alpha_{\pi(n)}) = A(z_{\pi(1)}, ..., z_{\pi(n)}) \tag{7}$$

Then, the conformal predictor dependent on A is defined as:

$$\Gamma^{\epsilon}(z_1, z_2, ...z_{n-2}, z_{n-1}, z_n) = \{y|p^{\epsilon} > \epsilon\} \tag{8}$$

For a new sample x_n, conformal predictor outputs all those possible labels in the set Γ^{ϵ} based on a given significance level $\epsilon \in (0, 1)$ and the training set. For each possible label $y \in Y$, the p-value associated with it is defined as:

$$p^y = \frac{|\{i = 1, ..., n|\alpha_i^y > \alpha_n^y\}|}{n} \tag{9}$$

where p^y represents that when the label of x_n is y, how well the unseen observation conforms to other observations.

The corresponding sequence of nonconformity scores for one predicted label y is defined by:

$$(\alpha_1^y, \alpha_2^y, ..., \alpha_n^y) = A(z_1, z_2, ..., z_{n-1}, (x_n, y)) \tag{10}$$

Based on α_n^y, it is recognized that the lower α_n^y is, the higher confidence we have for the prediction, since it means this combination of feature and label conforms better with other observations.

For conformal prediction, the validity of the prediction means:

$$P(y_n \notin \Gamma^{\epsilon}(x_1, y_1, ..., x_{n-1}, y_{n-1}, x_n)) \leq \epsilon \tag{11}$$

2.1.2. Nonconformity Measurement

Theoretically speaking, any algorithm can be modified to be the fundamental algorithm on which nonconformity measure can be based. Vladmir Vovk. originally based the calculation of nonconformity of conformal prediction on k-nearest neighbor. The typical method based on k-nearest neighbors(KNN) is also applied in this work. To make it simple and clear, CP-1NN and CP-3NN are used to denote conformal predictors based on 1NN and 3NN respectively and the corresponding simple predictors are denoted as 1NN and 3NN. The nonconformity measure algorithm is illustrated as below:

For a specific observation (x_i, y_i), firstly, the distance between this observation and any other observation in the training set is calculated, as is denoted

as:

$$d(x_i, x_j), j = 1, 2, ..., i-1, i+1, ..., n \tag{12}$$

Then, k nearest observations with the same label as that of(x_i, y_i) are found and denoted as (x_{is}, y_{is}),s=1,...,k. Meanwhile, k nearest observations with different labels from that of (x_i, y_i) are found and denoted as (x_{js}, y_{js}),s=1,...,k. The nonconformity score is given by the sum of distances between the test observation and k nearest homogeneous observations (with the same label) divided by the sum of distances between the test observation and k nearest heterogeneous observations (with different labels). The k is a number set for both the heterogeneous observations and homogenous observations.

$$\alpha_i = \frac{\sum_{s=1}^{k} d(x_{is}, y_{is})}{\sum_{s=1}^{k} d(x_{js}, y_{js})} \tag{13}$$

Based on the method listed above, the closer to observations with the same label a new observation is, the better conformed this new observation is and the lower α is.

2.1.3. Prediction in Online Mode

As one of the most prevailing protocols for machine learning problems, online prediction observes an ongoing and step-by-step principle which is practical and useful in real-world application. In this mode, after a fixed number of observations are listed in the training set, the label y_{n+1} of the new sample x_{n+1} is predicted based on the existing training set:

$$((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)) \tag{14}$$

Then, the effectiveness of the prediction is evaluated, from the perspective of erroneous prediction err_n^ϵ, multiple prediction $mult_n^\epsilon$ or empty prediction emp_n^ϵ, which will be illustrated later. Afterwards, this new observation with its correct label, is added to the training set to enlarge the number of observations available for further prediction. The process is conducted repeatedly, which means the training set is continuously updated during the whole process of

online prediction. The steps showing how online mode works are listed as follows [40]:

ONLINE PREDICTION PROTOCOL:

$Err_0^\epsilon := 0, \epsilon \in (0,1)$;

$Mult_0^\epsilon := 0, \epsilon \in (0,1)$;

$Emp_0^\epsilon := 0, \epsilon \in (0,1)$;

Training set= $\{(x_1, y_1), (x_2, y_2), ..., (x_i, y_i)\}$;

FOR n=1,2,...:

Reality inputs $x_n \in X$;

Predictor outputs $\Gamma^\epsilon \subset Y$ for all $\epsilon \in (0,1)$;

Reality outputs $y_n \in Y$;

$err_n^\epsilon = \begin{cases} 1 & if\ y_n \notin \Gamma^\epsilon \\ 0 & otherwise \end{cases}$

$Err_n^\epsilon = Err_{n-1}^\epsilon + err_n^\epsilon$

$mult_n^\epsilon = \begin{cases} 1 & if |\Gamma_n^\epsilon| > 1 \\ 0 & otherwise \end{cases}$

$Mult_n^\epsilon = Mult_{n-1}^\epsilon + mult_n^\epsilon$

$emp_n^\epsilon = \begin{cases} 1 & if |\Gamma_n^\epsilon| = 0 \\ 0 & otherwise \end{cases}$

$Emp_n^\epsilon = Emp_{n-1}^\epsilon + emp_n^\epsilon$

Training set=Training set,(x_n, y_n)

END FOR

When formula (11) and strong law of large number are taken into consideration, the validity for online conformal predictors is shown that:

$$\lim_{n \to \infty} sup \frac{Err_n^\epsilon}{n} \leq \epsilon \tag{15}$$

2.1.4. Prediction in Offline Mode

Offline prediction is the counterpart of online prediction and is characterized by prediction based on a fixed training set and fixed model. In this project, the offline prediction is done in a leave-one-out cross validation mode. The prediction in offline mode rests on certain rules gained from static training set, and the validity of conformal prediction in offline mode is not strictly proved. Conformal prediction in offline mode can provide users with additional information respecting confidence for each prediction, instead of just giving out the

predictions. The reliability information provides users with risk information for better decisions. For the prediction of each individual, conformal predictor can output the label with the highest p-value, which is defined as the forced prediction. Along with it, conformal predictor offers reliability information with two features, confidence and credibility:

$$confidence = sup\{1 - \epsilon :\mid \Gamma^{\epsilon} \mid \leq 1\} \tag{16}$$

$$credibility = inf\{\epsilon :\mid \Gamma^{\epsilon} \mid = 0\} \tag{17}$$

In the problem of classification, confidence equals to 1 minus the second largest p-value, which represents the confidence of prediction when the predicted label with the largest p-value is chosen while others are rejected. Credibility equals to the largest p-value, indicating how well conformed this selected choice is.

A prediction is judged to be reliable as long as confidence approximates 1 while credibility does not approximate 0 [38], meaning there are no better choices when compared to this prediction. If credibility is close to 0, the prediction tends to be nonconformal in the group and may probably be an unreliable outlier.

2.2. Experiments and Data Processing

2.2.1. Sample Collection

Samples were gathered from patients aging from 30-80 who were firstly diagnosed with lung cancer from the Second Affiliated Hospital of Zhejiang University. Meanwhile, the patients selected had not received chemotherapy or radiotherapy for 6 months before the tests and had abstained from smoking for at least 6 months. In addition, the samples for controls came from teachers and officials with a similar gender ratio and age distribution as the patients. They came from Zhejiang University and were not diagnosed with any respiratory disease, diabetes or any other health problem that may negatively influence the experiments. Meanwhile, they had not received any surgical operation in one

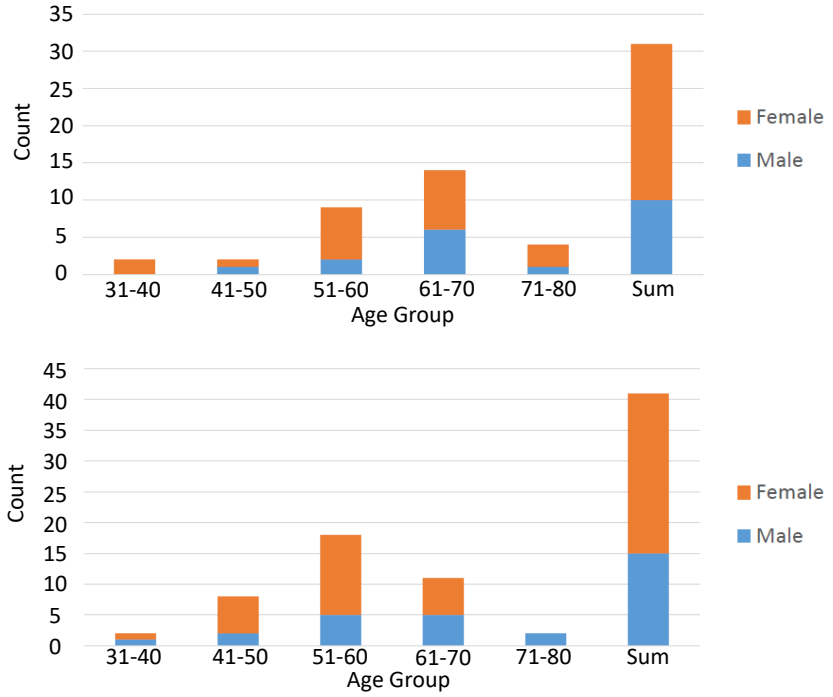

Figure 1: The distribution of volunteers' ages and gender in the patient group (up) and control group (down).

year. All the volunteers tested offered written approvals. The distribution of volunteers' ages were shown in Fig.1.

Among the patients, 26 were adenocarcinoma cases, 1 was small cell carcinoma case, 2 were squamous-cell carcinoma cases, 2 were large cell carcinoma cases. 13 cases were in patients' left lungs while 18 cases were in the right lungs.

All volunteers were banned from eating after 22:00 the day before sample collection. Then, samples were collected from 6:00 to 7:00 in the morning on empty stomachs. Volunteers could wash their mouths with water but couldn't brush their teeth. Meanwhile, they had been banned from vigorous activities such as running for 2 hours before the tests. All of the samples were collected under the instructions of researchers.

The M3014-4 offline air collection equipment (ECO MEDICS AG, Duern-

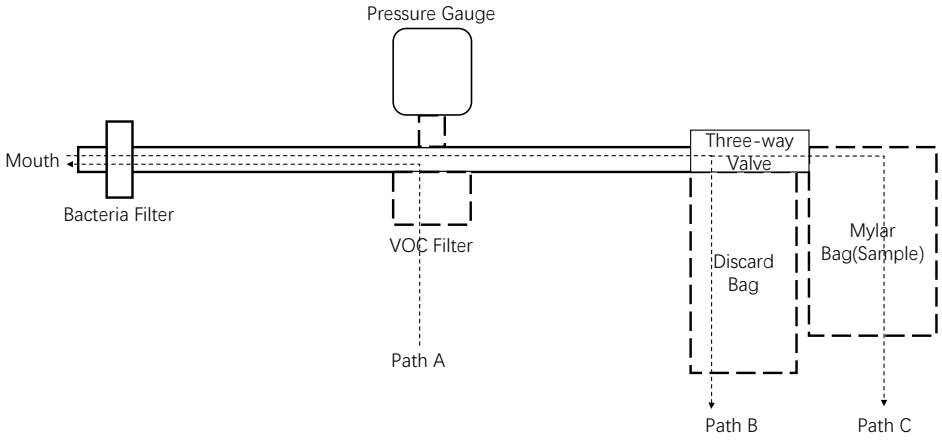

Figure 2: The illustration of the sampling equipment and process

ten, Switzerland) was used for the sampling process. The air bags were made from aluminized polyethylene terephthalate. The illustration of the sampling equipment was shown in Fig 2. Firstly, a volunteer breathed in through the mouth. As was indicated by Path A, the air went through the VOC filter and bacteria filter and entered the mouth. The process lasted for 3-5 min. As the air was filtered, this process meant to reduce environmental effect. Secondly, a volunteer took a deep breath and exhaled into the mouth. The requirement was that the pressure should be maintained within the range of $10 - 15cmH_2O$ to reduce the effect of flow rate. The first part of breath air was collected by Discard Bag through Path B and as the Discard Bag was filled, breath air followed Path C into Mylar Bag. This process ensured the breath air sample represented the air deep in lungs and reduced the influences of air in trachea and oral cavity. Finally, the samples were stored in a thermostat at a temperature of 37 degrees Celsius, transported to the lab and analyzed with electronic nose within an hour. A total of 72 breath air samples (including 31 samples from patients and 41 samples from controls) were collected with air collectors.

2.2.2. Electronic Nose System and Experiments

All the experiments and analysis were based on an electronic nose made in State Key Laboratory of Industrial Control Technology in Zhejiang University [40, 41, 42, 43], with 16 metal-oxide semi-conductive (MOS) sensors, of TGS type and purchased from Figaro Engineering Inc. (Osaka, Japan). This type of sensors performed well in multicle classification tasks[44, 45]. The 16 sensors had overlapped specificity[42, 40]. And the reliability and accuracy of this electronic nose system has been verified in previous studies [42, 40]. For example, we[46] have successfully used this system to classify 12 different categories of alternative herbal medicines with a leave-one-out cross validation(LOOCV) accuracy of 98.94% over 600 observations. What's more, we have also validated the effectiveness of online learning with the electronic nose[47]. In another work, we[48] applied the electronic nose to discriminate between different origins of the same type of alternative herbal medicine and reached the LOOCV accuracy ranging from 85.63% to 99.78% for each classification task.

The schematic description of the e-nose system could be represented in Fig. 3, which included the sensor array in a 200 ml gas chamber, a three-way valve changing from the flow of sample and standard clean air, air pumps pumping air at a rate of 1 L/min, power supply system and data-acquisition units. The three-way valve was the major controller of the test gas and standard clean air. This was an generalized electronic nose used for multifunctional applications[40, 41]. The typical sensitive components of each sensor were shown in table 1 but the sensors were not only sensitive to components listed. The sensors were able to react to the volatile organic compounds of breath air with diverse components including phenols and aldehydes. Meanwhile, they were not excessively sensitive to one specific type of gas, appropriate for breath air analysis which involves complex volatile organic compounds.

The whole process of measurement experiment of each collected sample was illustrated in Fig. 4 [40, 46, 47, 48]: 400 seconds were set for each individual test sample experiment. The sampling frequency of the electronic nose system was

Table 1: Information of the sensors used in 16-sensor array in the electroninc nose

Sensor Type	Count	Most sensitive to:
TGS800	1	Carbon monoxide, ethanol, methane, hydrogen, ammonia
TGS813	2	Carbon monoxide, ethanol, methane, hydrogen, isobutane
TGS816	1	Carbon monoxide, ethanol, methane, hydrogen, isobutane
TGS821	1	Carbon monoxide, ethanol, methane, hydrogen
TGS822	2	Carbon monoxide, ethanol, methane, acetone, n-hexane, benzene, isobutane
TGS826	1	Ammonia, trimethyl amine
TGS830	1	Ethanol, R-12, R-11, R-22, R-113
TGS832	1	R-134a, R-12, R-22, ethanol
TGS880	1	Carbon monoxide, ethanol, methane, hydrogen, isobutene
TGS2620	1	Carbon monoxide, methane, isobutene, hydrogen
TGS2600	1	Carbon monoxide, hydrogen
TGS2602	1	Hydrogen, ammonia ethanol, hydrogen sulfide, toluene
TGS2610	1	Ethanol, hydrogen, methane, isobutene, propane
TGS2611	1	Ethanol, hydrogen, isobutene, methane

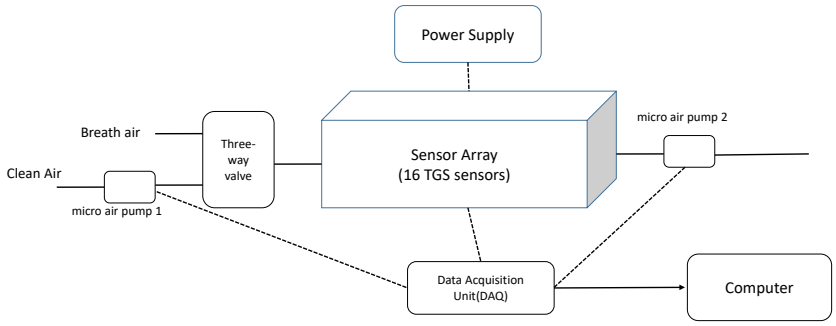

Figure 3: The brief illustration of the electronic nose system used in this project

100Hz. To begin with, standard clean air was pumped into the sensor panel at a rate of 1 L/min for 20 seconds to let the sensor responses return to baseline. Then, upon the time when the flow of standard clean air was stopped, the test gas sample was injected to the chamber as soon as possible. Medical injectors were used to extract 10ml gas mixtures. Afterwards, we set 180 seconds for the reaction period and recorded the sensor responses in terms of voltage change. At t=200s, the standard clean air was pumped into the system again and the test gas sample was pumped out. From t=200s to t=340s, we recorded the declining patterns of sensor responses. Finally, another 60 seconds were set for the sensors to further stabilize and return to baseline.

After extracting breath air samples, all of the tests with e-nose were conducted in the same laboratory with the same electronic nose at environment temperatures varying from 22 to 25 degrees Celsius and with relative humidity records ranging from 50-65%.

2.2.3. Data Processing and Feature Extraction

The data processing of this work was done on a personal computer without GPU acceleration. Typical sensor responses were shown in Fig. 5, which showed

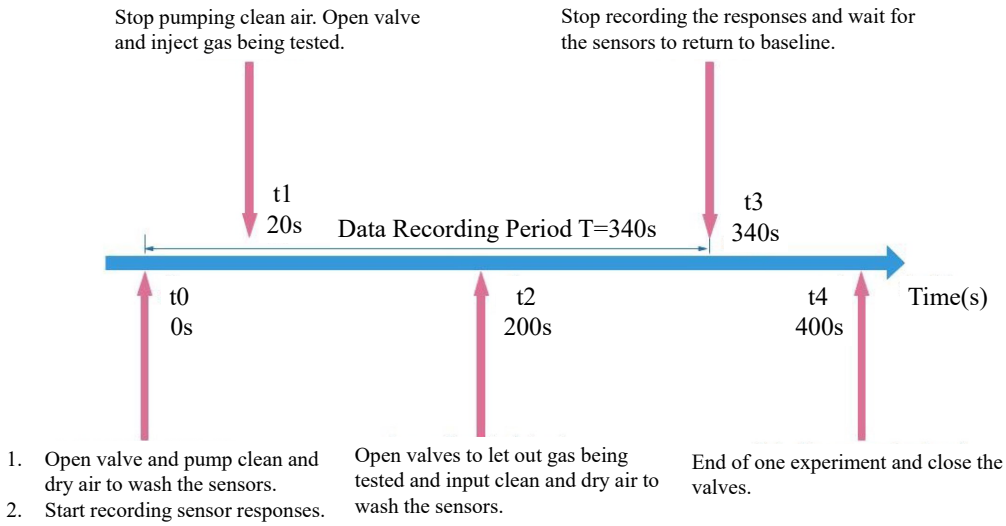

Figure 4: The whole process of one single experiment

the responses to the first control sample with each curve representing one of the 16 sensors. The sampling time was 0.01s. Discrete Wavelet Transform(db2) was used to decompose the signals and the fifth level of decomposed signals were selected to deal with noises. Then, all the recorded signals were calibrated by subtracting the baseline values to minimize influences of sensor drifts:

$$V = V_S - V_0 \tag{18}$$

V_s denoted the response of a sensor after filtering and V_0 denoted the baseline value after filtering.

Afterwards, based on previous time-series data-mining performances [40, 41, 46, 47, 48], 9 commonly used features for e-nose (a total of 9*16=144 features) were extracted:

1.Maximum Value

$$V_{max} = max(\mid V \mid) \tag{19}$$

2.Integral Value

$$V_{int} = \int_0^T V(t)dt \tag{20}$$

T denoted the whole time for measurement, T=340s.

3.Phase Information

$$V_{phase} = \int_0^{|V_{max}|} (\frac{dV(t)}{dt})^2 dt \tag{21}$$

4-9.Exponential moving average of the derivative of V

$$E_a(V) = [min(y(k)), max(y(k))], 2000 < k < 34000 \tag{22}$$

The discrete sampling exponential moving average $y(k)$ and smoothing factor a were defined as:

$$y(k) = (1-a)y(k-1) + a(V(k) - V(k-1)) \tag{23}$$

$$a = \frac{1}{100 * SR}, \frac{1}{10 * SR}, \frac{1}{SR} \tag{24}$$

$$y(1) = aV(1) \tag{25}$$

$SR = 100$ denoted the sampling rate. $E_a(V)$ denoted the vector containing the largest value and the smallest value in the period of time after the injection of a gas sample. The feature extraction is guided by previous researches but feature significance evaluation, sensor optimization and more effective information could be done, which is open to further researches in the future.

Afterwards, these features were pieced together to a 72*144 matrix with each column representing one feature. Finally, each column of features were normalized by subtraction of the minimum and division of the range of each feature. After normalization, the features were in the range from 0 to 1. All the data preprocessing and model training were done with MATLAB R2017b.

3. Results and Discussion

3.1. Offline Prediction with Conformal Prediction

In the offline prediction mode, leave-one-out cross validation protocol was applied. The results of forced prediction given by conformal predictors with

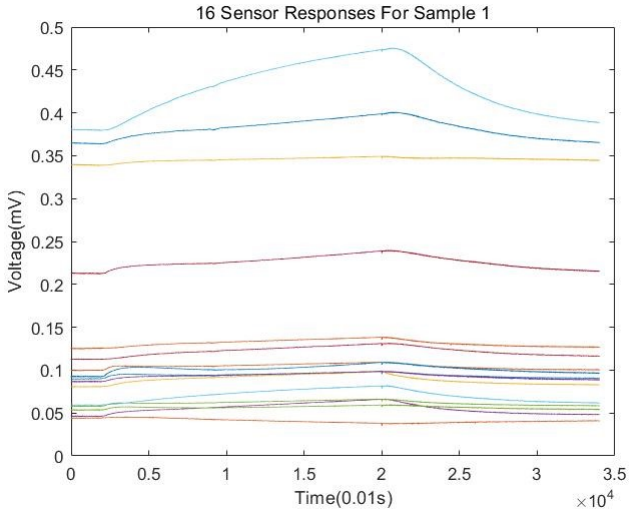

Figure 5: A typical sensor response curve given by the e-nose system

Table 2: Conformal Prediction Accuracy In Offline Mode

Prediction Method	1NN	3NN
Forced Conformal Prediction	87.50%	83.33%
Simple Prediction	87.50%	81.94%

1NN (CP-1NN) and 3NN (CP-3NN), in which the label with the highest p-value was the output, and the results of simple prediction given by 1NN and 3NN were presented in table 2.

According to the results, with regard to the accuracy of prediction, CP-1NN performs as well as 1NN while CP-3NN outperforms 3NN. Without sacrifice of accuracy, conformal prediction provides additional information about the confidence and credibility which underlies each prediction. To be concrete, five typical predictions are selected randomly and the results are shown in table 3 (CP-1NN). Label 0 denotes control while label 1 denotes patient. According to the results of CP-1NN, given by the forced predictor, the first sample is predicted to be of label 0 with confidence 0.9167, which indicates that the predictor is confident to reject the other label. Meanwhile, this prediction has a credibility

Table 3: Five typical individual predictions with CP-1NN

Sample Index	True Label	Forced Prediction	Confidence	Credibility	Simple Prediction
1	0	0	0.9167	0.1667	0
25	0	0	0.9722	0.3333	0
37	0	0	0.9861	0.7083	0
59	1	1	0.9861	0.9306	1
64	1	1	0.9583	0.3194	1

of 0.1667, which shows that although the other label is inappropriate compared with the chosen label, this prediction itself does not conform to the training set very well and the reliability of this prediction is not high. The 59^{th} sample is predicted to be of label 1, which is correct, with confidence 0.9861 and credibility 0.9306, which indicates that this prediction is reliable that it not only conforms well to the training set but also conforms better than its counterpart label.

For sample 1 and sample 64, lower credibility means the results are not quite reliable. Under these circumstances when the reliability of prediction is not high enough, though specific diagnosis can be given, doctors should inform customers of the risks of misleading diagnosis and recommend these customers to keep a closer look at their health and take further medical analysis once low confidence or low credibility occurs, as an erroneous diagnosis may bring huge burden to customers.

From the results in offline mode, in addition to giving the predicted labels, conformal predictors provide additional and significant reliability information about individual prediction with confidence and credibility, which may play an important role in diagnosis.

3.2. Online Prediction with Conformal Prediction

According to the protocol illustrated in Section 2, CP-1NN and CP-3NN were used for online prediction. Samples were reshuffled randomly and the first

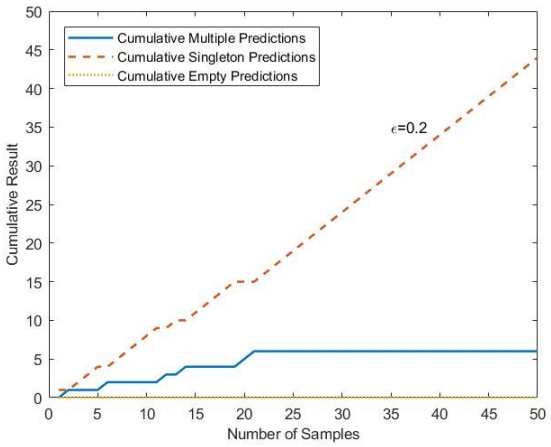

Figure 6: Results of online conformal prediction with 1NN,$\epsilon = 0.2$

20 samples were used as the initial training set. Afterwards, each new sample from the reshuffled sample sets underwent prediction.

The results of cumulative singleton prediction ($Sing_n^\epsilon$), multiple predictions ($Mult_n^\epsilon$) and empty predictions (Emp_n^ϵ) under the significance levels 0.2 and 0.1 are illustrated in Fig. 6 and 7 respectively. Initially, with limited training samples, the conformal predictors tend to output multiple predictions. Gradually, as more observations are added, the model become more robust. Multiple predictions stop arising while singleton predictions prevail, which indicates that with more samples, the sample can be classified into one group while rejecting the counterpart choice with higher confidence. Possible reason may be that more samples expand the space covered by a specific type of sample in the sample space which helps improve the predictor's knowledge and generalization ability.

In addition to the tendency of predictions discussed above, by giving such prediction forms as multiple, empty or singleton, conformal predictions provide users with information about the distribution of samples in the sample space. For instance, the 20^{th} sample is given multiple prediction, which indicates that this sample may locate near the borderline between two labels. Meanwhile, if one sample is given an empty prediction, it means that neither of the two

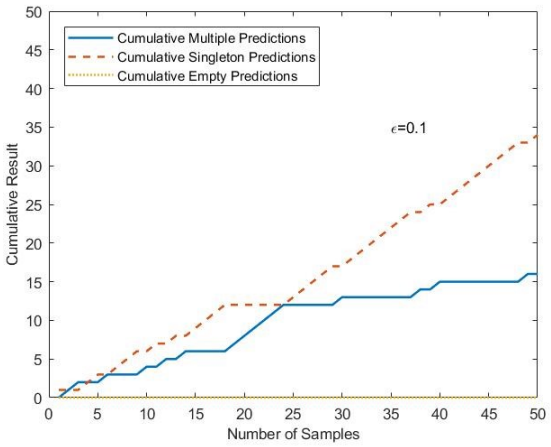

Figure 7: Results of online conformal prediction with 1NN, $\epsilon = 0.1$

labels provides reliable prediction for this sample, and this sample may serve as an outlier not conforming to both groups. As for the implications in lung cancer diagnosis, once multiple prediction or empty prediction arises, doctors should think twice before diagnosing since it is not definitely clear whether the prediction is accurate and reliable.

Two specific features are generally used to describe the performances of online conformal predictions. Firstly, validity of conformal prediction refers to the frequency of error predictions: the total number of erroneous predictions divided by number of samples. The tendencies of error rates under significance levels 0.2 and 0.1 are shown in Fig. 8. According to the results, the validity of online conformal prediction can be manifested that with the accumulation of observations, the error rate tends not to exceed the upper bound set by the significance level.

Based on the validity illustrated, it seems that with lower significance level ϵ and higher confidence level $1 - \epsilon$, the error rate is guaranteed to be lower with enough samples. Nevertheless, it does not necessarily mean the higher confidence level $1 - \epsilon$, the better the conformal predictor will be, when the efficiency of conformal prediction is taken into consideration.

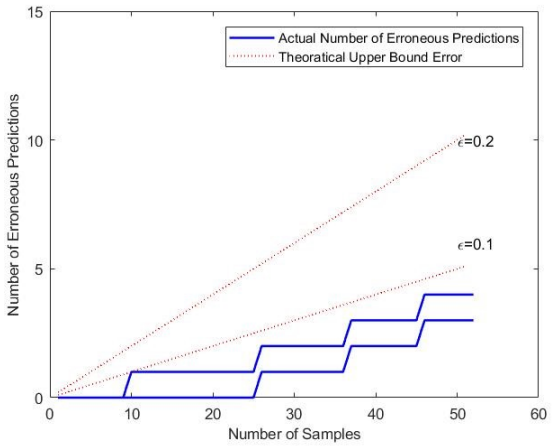

Figure 8: Tendency of erroneous prediction with CP-1NN ($\epsilon = 0.2, 0.1$)

To fulfill the reliability requirements, conformal predictors tend to output multiple predictions (double predictions in this problem) to avoid misclassifying samples and abide by the confidence level. This leads the predictor to be less efficient. Therefore, as long as the nonconformity measurement method is fixed, it is necessary for users to strike a balance between confidence and efficiency. Generally, there are two major criteria specifying the efficiency of conformal predictors [49]: the percentage of multiple predictions in all tested sample $Mult_n^\epsilon$/n, denoted as M criterion ('M' for 'Multiple'), and the average number of predicted labels labels in the predict region of multiple prediction, $W\_Mult_n^\epsilon$/n which is denoted as E criterion ('E' for 'Excess'). Based on these two criteria, the results are listed in table 4 and the tendency of multiple predictions is shown in Fig. 9 and 10. According to the results, since this is a binary classification problem, the average number of multiple predictions is 2. Choosing both labels indicates the efficiency of conformal predictor declines, which calls for further diagnostic analysis for this sample to avoid misclassification under the current reliability requirement. Meanwhile, it is evident that higher confidence level $1 - \epsilon$ leads to higher rate of multiple predictions. Therefore, it is necessary for users to set an appropriate significance level to balance efficiency and reliability.

Table 4: Efficiency of online conformal prediction

Predictor	$Mult_n^\epsilon$/n		$W\_Mult_n^\epsilon$/n	
	$\epsilon = 0.2$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.1$
CP-1NN	0.1154	0.3077	2	2
CP-3NN	0.0962	0.2115	2	2

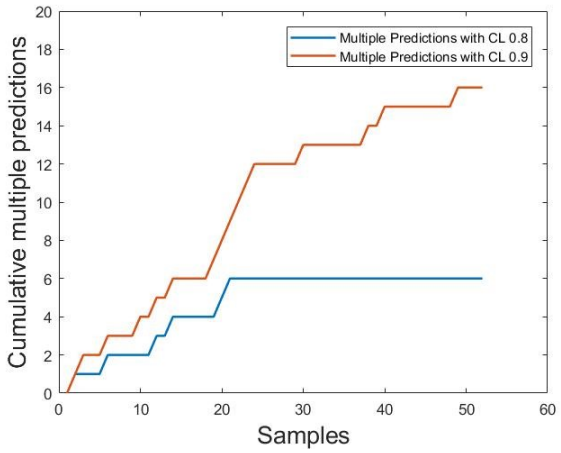

Figure 9: Tendency of multiple predictions with CP-1NN

What is more, when comparisons are made between CP-1NN and CP-3NN, it is clear from the results that in this case of classifying lung cancer samples, CP-3NN tends to output fewer multiple predictions, indicating that it is more efficient when compared with CP-1NN.

3.3. Comparison with Previous Studies

Previous studies have been done related to lung cancer screening with electronic nose data[50, 51]. For instance, Rens van de Goor et al[50] applied electronic nose data and artificial neural network(ANN) and successfully differentiated patients with controls with a sensitivity of 88%, specificity of 86% and diagnostic accuracy of 86%. Madara Tirzïte et al[46] employed logistic regression analysis on detection of lung cancer with electronic nose and gained an

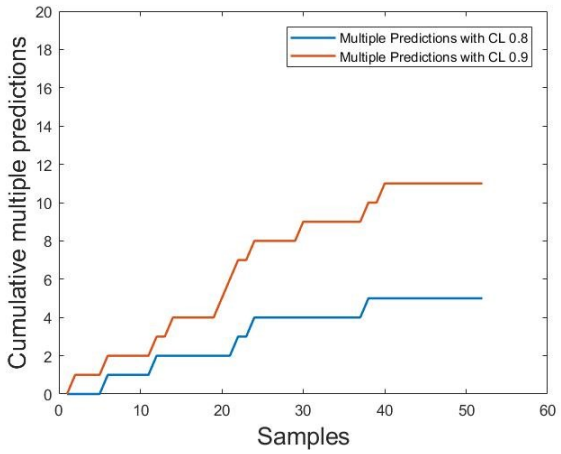

Figure 10: Tendency of multiple predictions with CP-3NN

overall sensitivity of 95.8% for smokers and 96.2% for non-smokers.

Admittedly, with larger dataset, the metrics of prediction in these previous studies tend to be better. When compared to these two studies, the specific contributions of this paper are listed as follows:

Firstly, with regard to generalization ability from one dataset to another, this study offers better risk information and uncertainty for future classification tasks for unseen datasets. The datasets and models are pretty consistent in previous studies, which means the researchers base their evaluation on the same overall cohort in e-nose experiment. What is more important is how well the classifiers perform when dealing with inhomogeneous datasets, such as e-nose system with manufacturing differences, different temperature, humidity and different conditions of sensor aging. Therefore, high sensitivities, specificities are not a guarantee of overall performances across from different datasets. As no one could examine all the real-world electronic nose datasets, in this sense, prediction itself and the metrics related to predictions in the specific retrospective tests may not be enough without considering reliability, the uncertainty information and the confidence of prediction which could be expanded for more clinical values. The current study lay more emphasis on the reliability of prediction

with CP-KNN. This could bring more value in addition to merely giving out the prediction itself. Moreover, the mathematic foundation of the reliability of Conformal Prediction is strictly justified by Vladimir Vovk. and colleagues. In this study, we also experimentally validated the reliability in an online manner. The confidence in the reliability evaluation can be ensured, which could not be interpretable in ANN and LR. The ANN and LR may work really well under specific circumstances with some systems, the generalization ability, interpretability of their reliability is not beyond doubt.

Secondly, both of these two previous studies are typical parametric methods and rely on specific model assumption. For artificial neural network, the architecture, weights, bias of activation function trained from training set may be incompatible to real-world cases if transferred to a different dataset acquired from a different electronic nose. This could take a lot more time in the training process for each individual electronic-nose system. In logistic regression, a linear model assumption and maximum likelihood estimate of the parameters may not be appropriate for electronic nose data. The data is both influenced by system noises, sensor drifts and influenced by collinearity, which could lead to extremely not robust classifiers. For instance, the collinearity between features for one sensor and across sensors could lead the coefficients of the logistic regression model to variate over a large range and be numerically unstable. Therefore, when dealing with electronic nose, a non-parametric method such as CP-KNN may be a wiser choice.

4. Conclusion

In this work, a novel application of conformal prediction in lung cancer prediction with an electronic nose system is introduced. Breath air samples from lung cancer patients and controls are collected and analyzed in e-nose system. Afterwards, the data are processed with conformal prediction in both offline mode and online mode. Nonconformity measurement for conformal predictions are based on 1NN and 3NN in this work. In offline mode, the accuracies of con-

formal prediction based on 1NN (CP-1NN) and 3NN (CP-3NN) are 87.50% and 83.33% respectively, which are slightly better than those results gained from simple predictors 1NN and 3NN. In addition to predicted results, conformal predictors enable users to know the reliability of individual prediction by giving confidence and credibility of each prediction, which is important in cancer diagnosis. In online mode, validity of conformal prediction is manifested that with growing number of samples, the erroneous prediction rate gradually lies below the significance level set by users, meaning the predictor can not only effectively perform online lung cancer diagnosis but also become more accurate and robust with more samples. Additionally, the potential of conformal prediction to indicate the distribution of samples has been discussed. Meanwhile, it is necessary for users to balance confidence and efficiency while considering confidence level for conformal prediction, since higher confidence level usually leads to more frequent multiple predictions, a symbol of lower efficiency. The developed software provides an additional analytical solution. With the hardware of electronic nose, the data can be obtained and analyzed with the software in labs and clinics.

In future work, different methods of nonconformity measurement can be applied in conformal predictions for data gathered from electronic nose systems to improve validity and efficiency as this project only proposes the framework based on CP-KNN. Additionally, artificial sensor selection and optimization could also be done on a supervised way based on the data to find out the sensors that contribute most to the classification problem. Finally, when it comes to the dealing with low-reliability diagnosis, more solutions such as data fusion with different analytic methods and diagnostic imaging such as GC-MS, CT to deal with conditions with low credibility could also be studied in order to provide more diagnostic values for patients.

Acknowledgements

The work is supported by the Natural Science Foundation of China (Grant No. 61773342) and the Autonomous Research Project of the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT1914).

References

[1] E. Office, World cancer day 2017: Fact sheet and promotion 3 (1).

[2] R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2017, CA: A Cancer Journal for Clinicians 67 (1) (2017) 5.

[3] R. Lozano, M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, J. Abraham, T. Adair, R. Aggarwal, S. Y. Ahn, et al., Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010, The lancet 380 (9859) (2012) 2095–2128.

[4] P. Ong, D. Ost, Lung cancer epidemiologic changes: Implications in diagnosis and therapy.

[5] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, A. Jemal, Global cancer statistics, 2012, Ca A Cancer Journal for Clinicians 65 (2) (2015) 87–108.

[6] A. J. Linz, J. J. Picken, Early detection of lung cancer: a method for improving survival rates, Journal of the American Osteopathic Association 79 (6) (1980) 364.

[7] F. Taher, H. Al-Ahmad, N. Werghi, Early detection of lung cancer based on sputum color image analysis, in: IEEE International Conference on Electronics, 2014.

[8] S. Peng, Q. Xu, X. B. Ling, X. Peng, W. Du, L. Chen, Molecular classification of cancer types from microarray data using the combination of genetic

algorithms and support vector machines, Febs Letters 555 (2) (2003) 358–362.

[9] B. Yu, Y. Zhang, L. Zhao, Cancer classification by a hybrid method using microarray gene expression data, Journal of Computational and Theoretical Nanoscience 12 (10) (2015) 3194–3200.

[10] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine 7 (6) (2001) 673.

[11] B. Yu, Q. Wang, X. Wang, S. Li, L. Lou, W. Qiu, On extraction of cancer informative genes and gene expression data mining, Journal of Bionanoscience 10 (4) (2016) 293–299(7).

[12] A. Bikov, Z. Lázár, I. Horvath, Established methodological issues in electronic nose research: how far are we from using these instruments in clinical settings of breath analysis? 9 (3) (2015) 034001.

[13] F. D. Francesco, R. Fuoco, M. G. Trivella, A. Ceccarini, Breath analysis: trends in techniques and clinical applications, Microchemical Journal 79 (1) (2005) 405–410.

[14] B. Schmekel, F. Winquist, A. Vikström, Analysis of breath samples for lung cancer survival, Analytica chimica acta 840 (2014) 82–86.

[15] D. Poli, M. Goldoni, M. Corradi, O. Acampa, P. Carbognani, E. Internullo, A. Casalini, A. Mutti, Determination of aldehydes in exhaled breath of patients with lung cancer by means of on-fiber-derivatisation spme–gc/ms, Journal of Chromatography B Analytical Technologies in the Biomedical & Life Sciences 878 (27) (2010) 2643–2651.

[16] F. Wojciech, S. Andreas, F. Anna, A. Clemens, S. Jochen, M. Wolfram, A. Anton, T. Jakob, Td-gc-ms analysis of volatile metabolites of human

lung cancer and normal cells in vitro, Cancer Epidemiol Biomarkers Prev 19 (1) (2010) 182–195.

[17] E. Gobbi, M. Falasconi, G. Zambotti, V. Sberveglieri, A. Pulvirenti, G. Sberveglieri, Rapid diagnosis of enterobacteriaceae in vegetable soups by a metal oxide sensor based electronic nose, Sensors & Actuators B Chemical 207 (2015) 1104–1113.

[18] C. Olguín, N. Laguarda-Miró, L. Pascual, E. García-Breijo, R. Martínez-Mañez, J. Soto, An electronic nose for the detection of sarin, soman and tabun mimics and interfering agents, Sensors & Actuators B Chemical 202 (10) (2014) 31–37.

[19] S. Sankaran, L. R. Khot, S. Panigrahi, Biology and applications of olfactory sensing system: A review, Sensors & Actuators B Chemical 171-172 (8) (2012) 1–17.

[20] G. C. Green, A. D. C. Chan, H. Dan, M. Lin, Using a metal oxide sensor (mos)-based electronic nose for discrimination of bacteria based on individual colonies in suspension, Sensors & Actuators B Chemical 152 (1) (2011) 21–28.

[21] S. De Vito, M. Piga, L. Martinotto, G. Di Francia, Co, no2 and nox urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization, Sensors and Actuators B: Chemical 143 (1) (2009) 182–191.

[22] L. Zhang, F. Tian, H. Nie, L. Dang, G. Li, Q. Ye, C. Kadri, Classification of multiple indoor air contaminants by an electronic nose and a hybrid support vector machine, Sensors and Actuators B: Chemical 174 (2012) 114–125.

[23] R. Munoz, E. C. Sivret, G. Parcsi, R. Lebrero, X. Wang, I. M. Suffet, R. M. Stuetz, Monitoring techniques for odour abatement assessment, Water Research 44 (18) (2010) 5129–5149.

[24] T. Saidi, O. Zaim, M. Moufid, N. El Bari, R. Ionescu, B. Bouchikhi, Exhaled breath analysis using electronic nose and gas chromatography–mass spectrometry for non-invasive diagnosis of chronic kidney disease, diabetes mellitus and healthy subjects, Sensors and Actuators B: Chemical 257 (2018) 178–188.

[25] P. Montuschi, N. Mores, A. Trové, C. Mondino, P. J. Barnes, The electronic nose in respiratory medicine, Respiration 85 (1) (2013) 72–84.

[26] A. K. Pavlou, N. Magan, C. McNulty, J. M. Jones, D. Sharp, J. Brown, A. P. Turner, Use of an electronic nose system for diagnoses of urinary tract infections, Biosensors and Bioelectronics 17 (10) (2002) 893–899.

[27] V. S. Kodogiannis, J. N. Lygouras, A. Tarczynski, H. S. Chowdrey, Artificial odor discrimination system using electronic nose and neural networks for the identification of urinary tract infection, IEEE Transactions on information technology in biomedicine 12 (6) (2008) 707–713.

[28] V. Y. Musatov, V. Sysoev, M. Sommer, I. Kiselev, Assessment of meat freshness with metal oxide sensor microarray electronic nose: A practical approach, Sensors and Actuators B: Chemical 144 (1) (2010) 99–103.

[29] J. Ragazzo-Sanchez, P. Chalier, D. Chevalier-Lucia, M. Calderon-Santoyo, C. Ghommidh, Off-flavours detection in alcoholic beverages by electronic nose coupled to gc, Sensors and Actuators B: Chemical 140 (1) (2009) 29–34.

[30] A. Loutfi, S. Coradeschi, G. K. Mani, P. Shankar, J. B. B. Rayappan, Electronic noses for food quality: A review, Journal of Food Engineering 144 (2015) 103–111.

[31] M. Tohidi, M. Ghasemi-Varnamkhasti, V. Ghafarinia, S. S. Mohtasebi, M. Bonyadian, Identification of trace amounts of detergent powder in raw milk using a customized low-cost artificial olfactory system: A novel method, Measurement 124 (2018) 120–129.

[32] S. Faal, M. Loghavi, S. Kamgar, Physicochemical properties of iranian ziziphus honey and emerging approach for predicting them using electronic nose, Measurement 148 (2019) 106936.

[33] M. Ezhilan, N. Nesakumar, K. J. Babu, C. Srinandan, J. B. B. Rayappan, Freshness assessment of broccoli using electronic nose, Measurement.

[34] A. Sanaeifar, S. S. Mohtasebi, M. Ghasemi-Varnamkhasti, H. Ahmadi, Application of mos based electronic nose for the prediction of banana quality properties, Measurement 82 (2016) 105–114.

[35] A. Gammerman, V. Vovk, Hedging predictions in machine learning the second computer journal lecture, Computer Journal 10 (2) (2007) 151–163.

[36] I. Nouretdinov, D. Devetyarov, V. Vovk, B. Burford, S. Camuzeaux, A. Gentry-Maharaj, A. Tiss, C. Smith, Z. Luo, A. Chervonenkis, Multi-probabilistic prediction in early medical diagnoses, Annals of Mathematics and Artificial Intelligence 74 (1-2) (2015) 1–20.

[37] C. Zhou, I. Nouretdinov, Z. Luo, D. Adamskiy, L. Randell, N. Coldham, A. Gammerman, A Comparison of Venn Machine with Platt's Method in Probabilistic Outputs, Springer Berlin Heidelberg, 2011.

[38] V. Vovk, Conditional validity of inductive conformal predictors, Machine Learning 92 (2-3) (2013) 349–376.

[39] V. Vovk, A. Gammerman, G. Shafer, Algorithmic learning in a random world (2005) xvi.

[40] Z. Wang, X. Sun, J. Miao, Y. Wang, Z. Luo, G. Li, Conformal prediction based on k-nearest neighbors for discrimination of ginsengs by a home-made electronic nose, Sensors 17 (8) (2017) 1869.

[41] X. Sun, L. Liu, Z. Wang, J. Miao, Y. Wang, Z. Luo, G. Li, An optimized multi-classifiers ensemble learning for identification of ginsengs based on electronic nose, Sensors & Actuators A Physical 266.

[42] J. Miao, Z. Luo, Y. Wang, G. Li, Comparison and data fusion of an electronic nose and near-infrared reflectance spectroscopy for the discrimination of ginsengs, Analytical Methods 8 (6) (2016) 1265–1273.

[43] Y. Wang, J. Miao, X. Lyu, L. Liu, Z. Luo, G. Li, Valid probabilistic predictions for ginseng with venn machines using electronic nose, Sensors 16 (7) (2016) 1088.

[44] Z. Haddi, M. Boughrini, S. Ihlou, A. Amari, Geographical classification of virgin olive oils by combining the electronic nose and tongue, in: Sensors, 2012, pp. 1–4.

[45] K. Timsorn, C. Wongchoosuk, P. Wattuya, S. Promdaen, S. Sittichat, Discrimination of chicken freshness using electronic nose combined with pca and ann, in: International Conference on Electrical Engineering/electronics, Computer, Telecommunications and Information Technology, 2014, pp. 1–4.

[46] X. Zhan, X. Guan, R. Wu, Z. Wang, Y. Wang, G. Li, Discrimination between alternative herbal medicines from different categories with the electronic nose, Sensors 18 (9) (2018) 2936.

[47] X. Zhan, X. Guan, R. Wu, Z. Wang, Y. Wang, Z. Luo, G. Li, Online conformal prediction for classifying different types of herbal medicines with electronic nose.

[48] X. Zhan, X. Guan, R. Wu, Z. Wang, Y. Wang, G. Li, Feature engineering in discrimination of herbal medicines from different geographical origins with electronic nose, in: 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB), IEEE, 2019, pp. 56–62.

[49] V. Vovk, V. Fedorova, I. Nouretdinov, A. Gammerman, Criteria of efficiency for conformal prediction (2016) 23–39.

[50] R. van de Goor, M. van Hooren, A.-M. Dingemans, B. Kremer, K. Kross, Training and validating a portable electronic nose for lung cancer screening, Journal of Thoracic Oncology 13 (5) (2018) 676–681.

[51] M. Tirzïte, M. Bukovskis, G. Strazda, N. Jurka, I. Taivans, Detection of lung cancer with electronic nose and logistic regression analysis, Journal of breath research 13 (1) (2018) 016006.