PAPER

Data analysis of electronic nose technology in lung cancer: generating prediction models by means of Aethena

To cite this article: Sharina Kort *et al* 2017 *J. Breath Res.* **11** 026006

Journal of Breath Research

PAPER

Data analysis of electronic nose technology in lung cancer: generating prediction models by means of Aethena

RECEIVED
24 October 2016

REVISED
27 March 2017

ACCEPTED FOR PUBLICATION
4 April 2017

PUBLISHED
1 June 2017

Sharina Kort[1], **Marjolein Brusse-Keizer**[2], **Jan-Willem Gerritsen**[3] **and Job van der Palen**[2,4]

[1] Department of Pulmonary Medicine, Medisch Spectrum Twente, Enschede, The Netherlands
[2] Medical School Twente, Medisch Spectrum Twente, Enschede, The Netherlands
[3] The eNose Company, Zutphen, The Netherlands
[4] Department of Research Methodology, Measurement, and Data Analysis, University of Twente, Enschede, The Netherlands

E-mail: s.kort@mst.nl

Keywords: lung cancer, electronic nose, exhaled breath, aeonose, prediction models, data analysis

Abstract

Introduction. Only 15% of lung cancer cases present with potentially curable disease. Therefore, there is much interest in a fast, non-invasive tool to detect lung cancer earlier. Exhaled breath analysis using electronic nose technology measures volatile organic compounds (VOCs) in exhaled breath that are associated with lung cancer. *Methods*. The diagnostic accuracy of the Aeonose™ is currently being studied in a multi-centre, prospective study in 210 subjects suspected for lung cancer, where approximately half will have a confirmed diagnosis and the other half will have a rejected diagnosis of lung cancer. We will also include 100–150 healthy control subjects. The eNose Company (provider of the Aeonose™) uses a software program, called Aethena, comprising pre-processing, data compression and neural networks to handle big data analyses. Each individual exhaled breath measurement comprises a data matrix with thousands of conductivity values. This is followed by data compression using a Tucker3-like algorithm, resulting in a vector. Subsequently, model selection takes place after entering vectors with different presets in an artificial neural network to train and evaluate the results. Next, a 'judge model' is formed, which is a combination of models for optimizing performance. Finally, two types of cross-validation, being 'leave-10%-out' cross-validation and 'bagging', are used when recalculating the judge models. These judge models are subsequently used to classify new, blind measurements. *Discussion*. Data analysis in eNose technology is principally based on generating prediction models that need to be validated internally and externally for eventual use in clinical practice. This paper describes the analysis of big data, captured by eNose technology in lung cancer. This is done by means of generating prediction models with Aethena, a data analysis program specifically developed for analysing VOC data.

Introduction

Lung cancer is the leading cause of cancer death among males and females worldwide, accounting for approximately 5% of total mortality in many countries [1]. Lung cancer is not a well-defined single entity. It is a heterogeneous disease, arising in many different clinical pathological patterns. The World Health Organization classification recognizes 20 different types of malignant lung neoplasms [2]. The main types of lung cancer are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) where the latter can be subdivided into three major histological types: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Chest radiography and computed tomography (CT), which are considered non-invasive diagnostic techniques, are the first steps in the diagnostic work-up to detect and stage lung cancer. Histopathological diagnosis following an invasive bronchoscopic intervention still remains the gold standard to prove or rule out the diagnosis of lung cancer. However, this investigation is accompanied by associated risks and substantial costs, which makes it unsuitable for population-based screening.

The diagnosis of early stage lung cancer is essential for curative therapy by means of surgery, and

substantially determines life expectancy [3]. Five-year survival for those with pathological stage IA NSCLC is 73%, whereas metastatic disease has a miserable prognosis with a five-year survival of merely 13% [4, 5]. Unfortunately, only 15% of the lung cancer cases present with localized, potentially curable disease, which means that the majority of the cases is diagnosed in an advanced stage with consequently poor survival rates.

There has been a lot of interest in secondary prevention involving screening tests for the detection of early stage lung cancer. Screening tests using sputum cytology and chest radiography have been attempted with unfortunately limited success [6]. Although low-dose computed tomography (LDCT) is able to detect early stage lung cancers [3], in practice it does not sufficiently demonstrate a survival benefit, reduce the incidence of advanced stage cancers or reduce lung cancer mortality [7, 8]. The observed increased survival time with screening can be overestimated due to lead time bias, when survival time is measured from the time of diagnosis. Length bias can also give an overestimation of survival duration among screening detected cases by the relative excess of slowly progressing cases. These cases are disproportionately identified by screening because the probability of detection is directly proportional to the length of time during which they are detectable (and thereby inversely proportional to the rate of progression). Furthermore, maybe less important in lung cancer, overdiagnosis bias can play a role in screening research, which could lead to overestimation of survival duration among screen-detected cases caused by inclusion of pseudo-disease—subclinical disease that would not become overt before the patient dies of other causes [9, 10]. However, there are several ongoing lung cancer screening trials by means of CT scanning, with some optimistic results [11–15], but these results still are insufficient for screening to be incorporated in clinical practice since the high numbers needing to be screened, and a large number of false positives, continue to question the cost-effectiveness, especially concerning determining the definition of the screening population and the screening frequency [16–19]. Hence, there is much interest in a fast, simple, cost-effective and non-invasive tool for detecting lung cancer at an early stage, preferably during a visit at the general practitioner.

This has led to the introduction of exhaled breath analysis by means of electronic nose technology. This diagnostic approach seems very promising in the lung cancer field, though it is yet far from being incorporated in clinical practice [20–23].

The concept of an electronic nose is based on the availability of powerful personal computing making it possible to apply pattern recognition techniques to complex measurement data. The desire is to have a general, broadly responsive sensor system that generates complex multidimensional measurement data and uses pattern recognition techniques to match measured response patterns to previously observed response patterns in order to identify specific scents present within complex mixtures. This is analogous to the physiology of the human smell, where the brain combines received signals and determines what characteristic scent pattern is smelled, but does not distinguish specific components. Hence, the name 'electronic nose'.

Electronic nose technology is based on the usability of volatile organic compounds (VOCs) in exhaled breath. Exhaled breath is mainly composed of inorganic compounds, such as nitrogen, oxygen, carbon dioxide, water vapour and inert gases. In addition, it contains thousands of VOCs, which are exhaled in very low concentrations, but reflect pathological processes such as inflammation, oxidation, infection and neoplasms, where they can serve as non-invasive biomarkers for certain diseases [24]. The perspective is that metabolic and biochemical processes that occur in different pathological situations cause different endogenous VOCs to arise, which can be detected with different chemical sensors and can therefore be promising disease biomarkers. All these methods are directed at measurable changes in physical properties of the sensors when being exposed to a gas mixture.

However, the use of VOCs in electronic nose technology is only one method. There are several other methods utilized for breath sampling, such as multi-capillary column-ion mobility spectrometry or gas chromatography mass spectrometry that look for specific compounds in exhaled air [25–28]. By contrast to determining VOCs in exhaled breath, these techniques do not apply pattern recognition techniques, since they are aimed at identifying individual molecules in exhaled breath instead of a unique composite breath signal. Recently, Schallschmidt *et al* published results of an observational study on the profiles of volatile organic compounds where they showed that the use of solid phase micro-extraction gas chromatography mass spectrometry is not reliable enough to discriminate between cancer patients and healthy controls [29]. An important remark they make relates to the limited capability of current analytical procedures to detect unstable marker candidates.

The use of human breath as a diagnostic tool is not completely innovative. The use of smell as a diagnostic aid has been known since ancient times when Hippocrates mentioned the diagnostic value of smell in his work 'Aphorisms' which was written in 400 BC [30]. However, it was only when Pauling described in 1971 the presence of VOCs in exhaled breath that this method became of great scientific interest [31]. Over the last few decades, several electronic nose devices have been developed, which contain different sensors to detect the VOCs and generate a quantifying measure for these VOCs. A lot of research has been performed using the Cyranose 320, and analyses performed by Machado *et al* and Dragonieri *et al* provided some promising results in the lung cancer field [26, 32]. Also promising was the gold particle

nanosensor developed by Peng *et al* [33]. Peled *et al* showed an accuracy of the nanoarray in discriminating between malignant and benign pulmonary disease of 88% with an area under the curve (AUC) of 0.986 [34]. However, these results are based on a small study population ($n = 69$) without external validation being performed. In this paper, the Aeonose™, developed by The eNose Company (Zutphen, The Netherlands) will be discussed. The Aeonose™ differs from other electronic nose devices in that it offers the opportunity for transferring calibration models and therefore enables large-scale application [35].

An important aspect of the electronic nose concept is that a substance, or a mixture of substances (VOCs), can only be recognized after a calibration phase, i.e. the pattern must be known beforehand ('seen' beforehand). This is why the electronic nose must be trained and a database of patterns, called breath prints, must be developed. This searchable, digital database systematically stores previous measurements with characteristic scent patterns. In this way, new scent patterns can be matched with an existing scent profile through comparative pattern recognition analysis.

When comparing breath patterns between subjects diagnosed with and without a certain disease, the eNose can be trained to distinguish between these two groups. In this way, a new diagnostic device can be developed for screening or diagnosing diseases based on people's exhaled breath.

The aim of this manuscript is to describe our study concerning the detection of lung cancer with the Aeonose™, where we will focus on the statistical analysis in the 'black box' of the Aeonose™ measurements for classifying whether lung cancer is present or not.

Objectives

The main objective of this study is training the Aeonose™ to build a database to detect lung cancer. This study aims to investigate the diagnostic accuracy of exhaled breath analysis with the Aeonose™ to distinguish the breath of subjects suspected for lung cancer, who are truly diagnosed with lung cancer, from subjects suspected for lung cancer in which this diagnosis is rejected after histopathological diagnosis following a bronchoscopic intervention. The obtained patterns will also be compared with breath patterns of healthy subjects who are not suspected for lung cancer. Additionally, we will investigate whether the Aeonose™ recognizes patterns between different types of lung cancer (NSCLC versus SCLC) and between different lung cancer stages.

Material and methods

Design

This concerns a multi-centre, prospective, non-invasive study in subjects suspected for lung cancer, who are referred for a histological biopsy through bronchoscopy. Subjects who are suspected for lung cancer will be compared in a cross-sectional design, where breath patterns from those who are truly diagnosed with lung cancer are compared to those where this diagnosis is rejected. Also, breath patterns of healthy subjects will be compared with confirmed and rejected lung cancer cases. This concerns a single measurement in the pulmonology departments of Medisch Spectrum Twente Enschede, Ziekenhuis Bernhoven Uden, Medisch Centrum Leeuwarden, and Deventer Ziekenhuis, all in the Netherlands.

Study population

Adult subjects who have a scheduled visit at the outpatient clinic of the pulmonology departments of the participating hospitals due to suspicion of lung cancer will be asked to participate. Suspected subjects will be divided in a group with a confirmed diagnosis of lung cancer and a group with a rejected diagnosis of lung cancer based on histopathology following a bronchoscopic intervention. Healthy subjects will be recruited from partners, relatives or friends of eligible subjects. They will be frequency matched on age and gender distribution to the subjects suspected for lung cancer. When we calculated a sample size to ensure a study with a reasonable power we took into account a desired sensitivity of 90% with a two-sided confidence interval of 82.5%–95%. In this way, we need approximately 105 subjects diagnosed with lung cancer. When we presume a realistic 1:1 ratio of a confirmed versus a rejected diagnosis of lung cancer in suspected subjects, we also need approximately 105 subjects with a rejected diagnosis, which gives a total of 210 suspected subjects. Given the possibility of observing a bigger contrast between suspected subjects with a confirmed diagnosis of lung cancer and subjects not suspected for lung cancer at all, we will also include 100–150 'healthy' subjects without any suspicion for lung cancer. This should be sufficient for training the Aeonose™ and determining whether it can reliably detect differences in breathing substances.

Inclusion criteria

Recruitment of these subjects started in June 2015 and is expected to conclude in the summer of 2017. We aim to include a total of 210 patients where the number of patients per hospital depends on the catchment population of each hospital. Suspected subjects need to meet the following criteria to be eligible.

(1) Referred for a histological biopsy due to suspicion for lung cancer.

(2) Age $\geqslant$ 18 years.

Eligible healthy subjects need to meet the following criterion

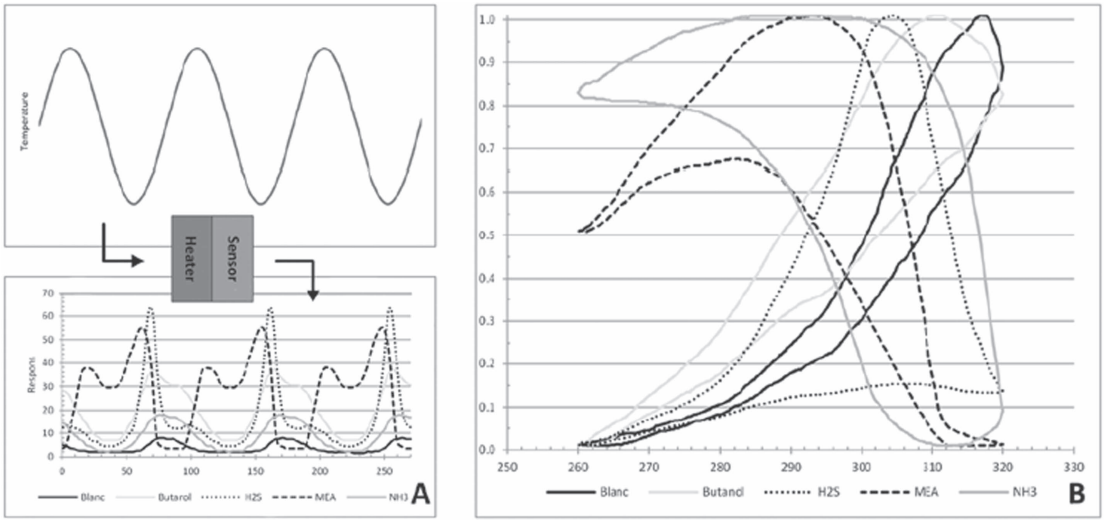

Figure 1. (A) Measuring principle of the Aeonose™: a continuous sinusoidal temperature cycle in which the heater is thermally cycled (top) and the conductivity of the sensor (bottom) is recorded as a function of momentary temperature. The temperature profile is applied to the heater while the response is recorded at the sensor. Substances can show temperature-dependent responses for the same chemical sensor type. (Blanc = clean air, H2S = hydrogen sulphide, MEA = methylamine, NH3 = ammonia). (B) Thermal response loops resulting from the normalized conductivity of the sensor plotted as a function of the heater temperature during a full period. (Blanc = clean air, H2S = hydrogen sulphide, MEA = methylamine, NH3 = ammonia). (Reproduced from [48] with permission from the author.)

(1) Age $\geqslant$ 18 years.

The only exclusion criterion for all subjects is

(1) Known to have an active malignancy.

In setting up the study protocol, we tried to exclude correlated features between cases and controls as much as possible. In an exploratory analysis, however, we noticed an (unexpected) decrease in AUC when we used supposedly healthy partner controls. This might be due to correlated features, such as similar diet and smoking behaviour, or at least residing in the same indoor atmosphere. In the case of suspicion of correlated features, cluster analysis could be helpful using, e.g., a software package like Carotta [36].

Aeonose™ technology

The Aeonose™ consists of three micro hotplate metal-oxide sensors (MOS) that are rigid, mass producible, and offer the opportunity for transferring calibration models. This means that once a calibration model has been developed, it can easily be transferred to other Aeonose™ devices. Several metal oxides behave as semi-conductors at higher temperatures. The sensors vary in terms of metal-oxide type and catalysing agent. Redox reactions occurring at the sensor surface result in changes in conductivity that can be measured and quantified, resulting in a unique breath signal. These redox reactions depend on the type of metal oxide and catalyst, the reacting gas(es), and the temperature. A broad range of VOCs in exhaled breath will give a redox reaction.

Thermal cycling

Redox reactions are temperature dependent, and by using thermal cycling this temperature dependency can be determined as a function of time. Different VOCs show different responses at varying temperatures for the same chemical sensor type (figure 1(A)). The breath patterns are obtained by taking the response of a complete cycle and can be presented as a function of the temperature (figure 1(B)). In this way the temperature dependency of the redox reactions is acquired on a single sensor. The patterns obtained by thermal cycling do not only depend on the applied temperatures, but also on the dynamics of the temperature, because intermediate products created at the sensor surface have limited lifetimes.

Statistical analysis

Predictive models are important tools to provide estimations of diagnostic outcomes. There are various resampling methods to estimate the performance of a model in a new sample of independent subjects (the test set) after a training set of observations has been created, i.e. these methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model. The resampling methods provide estimates of the test-set prediction error (test error) and error of the parameter estimates for future observations (prediction error). First, it is important to know whether a data set is either low-dimensional or high-dimensional. Low-dimensional implies that there are more subjects present than parameters in a data set ($n > p$). By contrast, high-dimensional implies having more

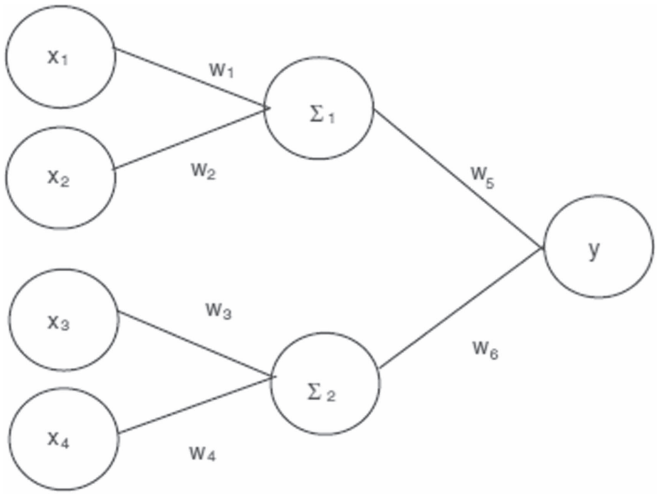

Figure 2. General structure of an ANN. A feed-forward neural network with one input layer consisting of four nodes (x), one hidden layer with two nodes (Σ), and one output layer (y). The connections between the layers have associated connection strengths or weights (w_i), which can be varied. (Reproduced from [49] © 2008 Wiley-Liss, Inc.)

parameters than subjects in a data set ($p > n$). A high-dimensional data set, as obtained with the Aeonose™, poses statistical challenges where too many predictors will overfit the data and result in a model that looks appropriate on the training data used to develop it, but will poorly perform on future observations from the test data. This problem of overfitting can be avoided by using a combination of analytical techniques such as data compression, cross-validation and bootstrapping. In statistics, cross-validation is a model validation technique for assessing how the results of a prediction model will generalize to a new independent data set [37, 38]. Bootstrapping is a useful technique for getting an idea of the variability or standard deviation of an estimate and its bias [39, 40].

The eNose Company uses a proprietary software for data analysis, called 'Aethena'. This package retrieves raw data from a database and takes care of data compression, data analysis and data reporting. In this section we will illustrate the methods used to obtain the best prediction models. During an exhaled breath measurement, for each sensor, 64 × 36 data points are being recorded. In this way, each individual patient measurement comprises of a data matrix with thousands of records. In the course of the data analysis and pattern recognition, the following steps can be distinguished:

Pre-processing

As mentioned before, the sensor's temperature control enables accurate reproducibility of the results. However, slight variations between sensors among Aeonoses™ can be seen. In order to cope with these variations, the data are being standardized in several ways, creating multiple representations of the same data set. 1. Data of a measurement are scaled between 0 and 1 per measurement cycle. 2. Data of the full measurement are scaled between 0 and 1.

Data compression

As the matrix sizes are too large for classification, the data are compressed using a Tucker3 solution [41]. This needs to be done to avoid so-called spurious correlations. Spurious correlations become of greater importance since modern eNoses collect increasing amounts of data. The compression results into a vector for one of the seven sensor combinations of the three metal-oxide sensors (A, B, C, AB, AC, BC and ABC). In the case of lung cancer this results in 11 components per patient in which redundant information and noise is removed, but in which information concerning the distinction between healthy and sick subjects is maintained.

We start with all subjects, called data set A. When classifying subjects, we set aside 20% of the data in a blinded fashion in order to create a test set for external validation (data set C, also called the test set). Of the remaining 80% (data set B, also called the training set) the true lung cancer status based on pathology is known. The vectors generated in the study will be entered into an artificial neural network (ANN). Figure 2 describes the principle of an artificial neural network. There is one input layer consisting of the obtained vector in the compression phase. By means of algorithms based on trial and error the components of the input layer and hidden layer will be given different weights to determine the best output. Several statistical learning methods could be applied for data classification. For optimal results all of them require fine-tuning. Up until now we have been focusing on applying neural networks. However, also other methods such as random forest and support vector machine could be applied. Actually, in another study (submitted for publication), the neural network results were compared to results obtained from random forest, support vector machine, and the Gaussian process showing comparable AUC values. Hauschild

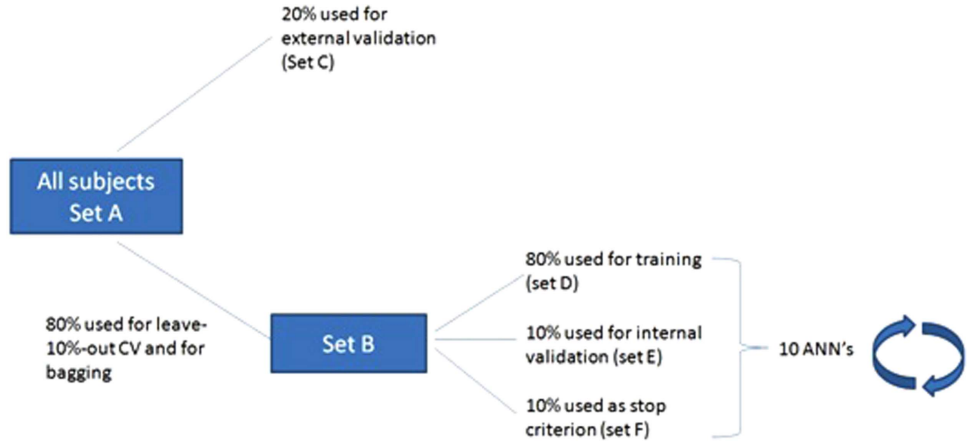

Figure 3. First round cross-validation.

et al have also described different classification methods [42]. Up until now we have no compelling evidence that other classification techniques will show better results than neural networks. However, for specific diseases, it could be favourable to use other classification techniques (e.g. random forest). Therefore, we intend extending our software package to other classification techniques in the near future.

Ten-fold cross-validation or leave-10%-out cross-validation

Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used when one wants to estimate how accurately a predictive model will perform in practice.

Ten-fold cross-validation comprises ten rounds of validation (figure 3). One round of cross-validation involves partitioning the training set (data set B) into complementary subsets (80% (data set D), 10% (data set E), 10% (data set F)), performing the analysis on data set D, and validating the analysis on the 10% in subset E. Data set F is used as a stop criterion in order to decide how long the model needs to be trained. To reduce variability, ten rounds of cross-validation are performed using different partitions in such a way that after ten rounds all data have been used once in datasets D, E, and F, and all patients are predicted once. The validation results are averaged over the ten rounds, resulting in one combined AUC.

Model selection

The process described is executed for all seven sensor combinations and for different pre-processing techniques. In this way, a large amount of possible ANN-models are being generated, each with its specific performance measures. The output consists of a list including ranked receiver operating characteristics curves (ROCs) with performance calculated by means of the AUC, sensitivity and specificity. Higher AUCs usually indicate better performance.

Subsequently, based on ranked AUCs, various models will be selected for optimizing diagnostic performance. First, a model is selected that is able to properly separate positive and negative subjects. Subsequently, for both negative and positive subjects, two different complementary *combined* models are constructed, based on single models, which minimize the number of false positives and false negatives. For positive subjects we use models that accurately predict positives and for negative subjects models are used that accurately predict negatives.

A *combination* of the best models showing the smallest error is called a 'judge' model (figure 4). The smallest error is defined as the sum of false positives and false negatives. Every judge model gives one AUC and all models are independent from each other. The next step will be to fine-tune and validate these selected models. Two sets of ROC plots can be constructed: at first, the neural network is being trained using samples with known classifications, and applying leave-10%-out cross-validation. The results can be represented in a ROC plot that should also be representative for blind samples because of the cross-validation process. Secondly, blind samples are classified using the trained neural network. When these classifications are compared with the gold standard results, threshold dependent confusion matrices can be constructed followed by a corresponding ROC plot. If the blind samples have similar characteristics as the training set, the ROC curves of training set and blind samples can be expected being almost identical.

Subsequent cross-validation

In our analysis, two types of cross-validation techniques are used: 'leave-10%-out' and 'bagging' (bootstrap aggregation) [43, 44]. When using the 'leave-10%-out' method, the selected single models and obtained judge models are recalculated as previously

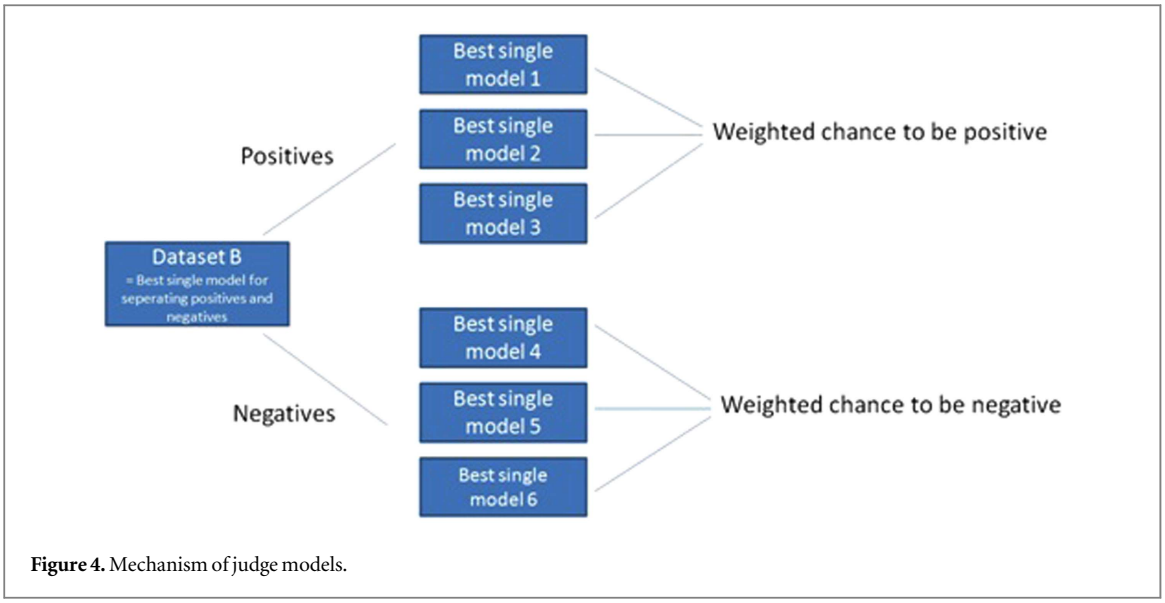

Figure 4. Mechanism of judge models.

described. However, fine-tuning of the ANNs is being applied for optimal performance where new weights for every model are calculated, which means that the ANNs are generated a few more times from scratch to determine whether the ANNs are stable, i.e. whether comparable ROC curves are derived. However, the input is not the full data set. Only the positives from the first separation are entered in the upper models and only the negatives from the first separation are entered in the lower models, which eventually lead to one AUC. Bagging is an alternative cross-validation technique to provide stable networks. From data set B, a random sample of measurements is chosen, used for training an ANN, and this sample is subsequently replaced, contrary to ten-fold cross-validation. This procedure is repeated many times (i.e. >1000). Per person, a large number of calculated risks for lung cancer are derived and are averaged to one chance, which is used to calculate the AUC. This obtained AUC will be compared with the obtained AUC from the ten-fold cross-validation. Finally, the best ANNs generated by bagging will be used to classify the blind measurements from data set C, the test set.

The bagging technique is mainly used to check whether the leave-10%-out procedure succeeded and gives a smoother model fit with a better balance between potential bias and variance. An important difference compared with ten-fold cross-validation is that in bagging models are not further adapted and no judge models are constructed. The calculated weight remains constant.

Example

Figure 5(A) shows a separation plot, based on training data, showing predicted values for 50 patients with lung cancer (pos) and 60 healthy controls (neg), according to the statistical procedures as described in this manuscript, representing a demonstration of the principle. Figure 5(B) shows the corresponding ROC curve, again based on preliminary training data.

Discussion

Despite modest advances in the treatment of lung cancer, it remains a fatal disease with overall five-year survival rates not having increased over a few decades [45, 46]. Therefore, it is of great importance to detect lung cancer at an early, potentially curable state. Screening programmes concerning lung cancer have proven evidence of reducing lung cancer-specific mortality, but results must be implemented carefully. There should be a clear balance between maximizing benefits and minimizing harm with acceptable costs. As seen with lung cancer screening, the high number of false positives involves substantial costs and therefore drives the cost-effectiveness of lung cancer screening downward. A positive CT scan triggers additional diagnostics ranging from rather easily repeating the CT scan to invasive diagnostics such as biopsy and surgical resection. These interventions, however, also involve associated risks, such as morbidity and mortality from complications and high emotional stress. Therefore, the lung cancer screening field can be extended with alternative forms of diagnostics instead of just focusing on imaging techniques. Exhaled breath analysis by means of electronic nose technology is a young field of research, but has been of great scientific interest over the last few years and is a rapidly emerging field of medical diagnostics. However, it has not yet been implemented in clinical practice. Several electronic noses with varying underlying technologies have been tried with some promising results, but the limited amount of external validation studies have not yet given sufficient trust in these methods. Recently, Leopold *et al* published an article concerning external validation in studies using various methods of electronic nose technology in lung

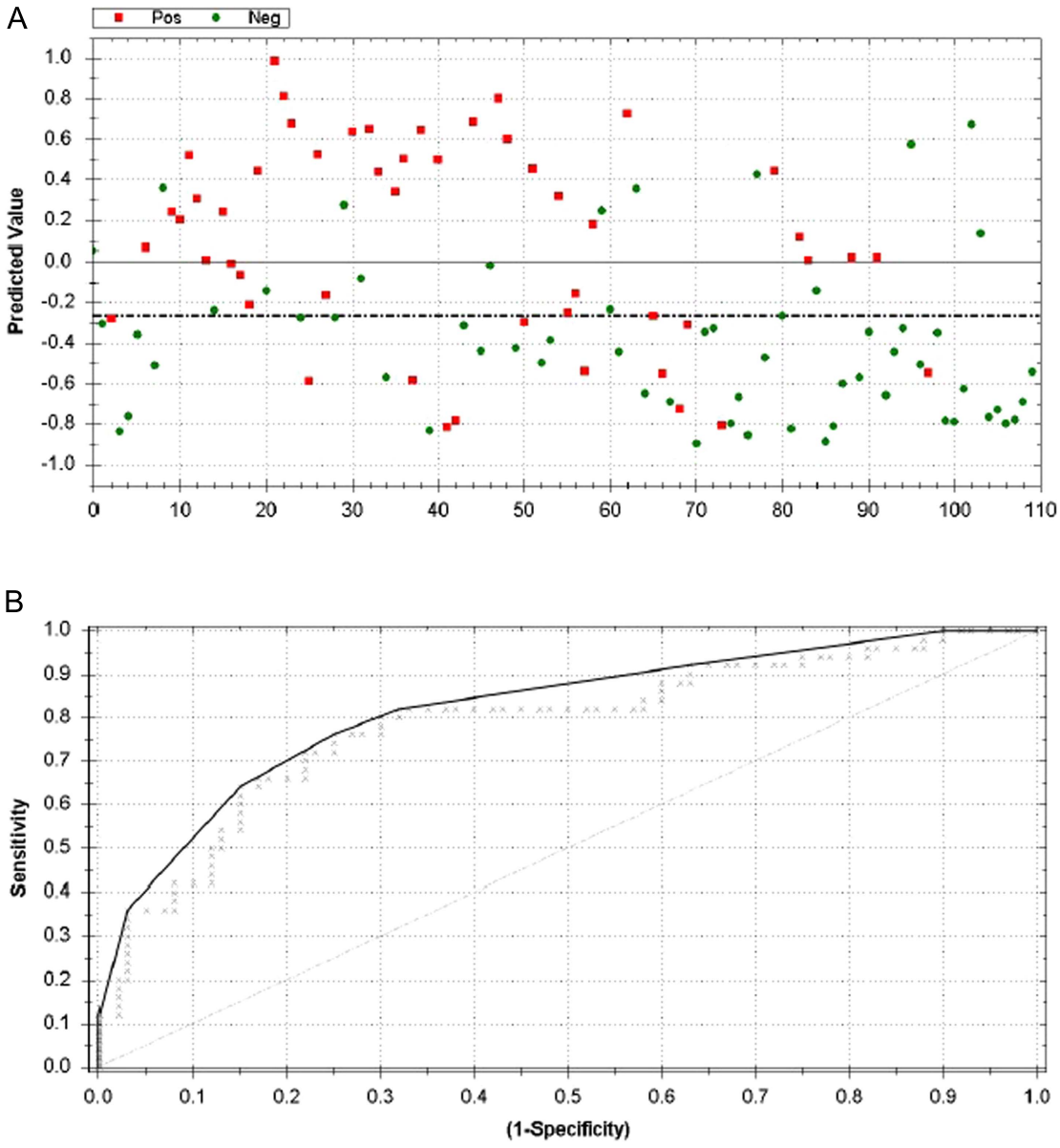

Figure 5. (A) Separation plot. (B) ROC curve.

cancer [47]. They evaluated 46 studies regarding different approaches to dimension reduction, classification and validation in electronic nose technology. Only seven studies had performed external validation on an independent data set with four datasets available for re-analysis. External validation resulted in a lower area under the receiver operating characteristics curve (ROC-AUC) compared to the internal validation in two out of four datasets. The other two datasets did not show decreased ROC-AUCs when applying external validation. However, no single combination of dimension reduction and classification methods gave consistent results between internal and external validation sets in these four datasets. Therefore, to show accurate diagnostic performance, it is important to estimate diagnostic performance on an independent data set (external validation). Robustness of the models is important, especially when one plans on classifying blind samples. Next to high overall AUC, we therefore also require models to show a small AUC standard deviation between the ten consecutive steps during the ten-fold leave-10%-out cross-validation.

The ideal diagnostic test should be both sensitive (a high percentage of sick subjects who are correctly identified as having the condition) and specific (a high percentage of healthy subjects who are correctly identified as not having the condition). This overall percentage of correctly diagnosed subjects determines the test accuracy. The results of the new diagnostic test are compared to the results of the reference test called the gold standard.

The Aeonose™ used in our study is a hand-held electronic nose device, which is convenient to use, includes non-invasiveness and gives fast results with consistent copy-and-paste between different Aeonoses™. However, possible disadvantages that need

to be taken into account are the inability to differentiate between endogenous and exogenous compounds and the influence of many exogenous factors, such as smoking, diet and other scents. In this study, we investigate whether exhaled breath patterns from patients with lung cancer can be distinguished from healthy subjects. After completing the training phase with approximately 350 subjects, we should have an idea whether the Aeonose™ can reliably detect differences in breath patterns of patients with lung cancer and subjects without lung cancer. After the training phase, an external validation phase must follow with an independent group of sick and healthy subjects in a different setting.

Conflicts of interest and source of funding

Professor Dr J van der Palen and Dr M Brusse report no conflict of interest. Miss S Kort was partly financed by an unrestricted research grant from The eNose Company. Dr J W Gerritsen is employed by the company producing the e-nose devices used.

References

[1] Torre L A, Bray F, Siegel R L, Ferlay J, Lortet-Tieulent J and Jemal A 2015 Global cancer statistics, 2012 *CA Cancer J. Clin.* **65** 87–108
[2] Beasley M B, Brambilla E and Travis W D 2005 The 2004 World Health Organization classification of lung tumors *Semin. Roentgenol.* **40** 90–7
[3] Henschke C I, Yankelevitz D F, Libby D M, Pasmantier M W, Smith J P and Miettinen O S 2006 Survival of patients with stage I lung cancer detected on CT screening *New Engl. J. Med.* **355** 1763–71
[4] Goldstraw P *et al* 2007 The IASLC lung cancer staging project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM classification of malignant tumours *J. Thorac. Oncol.* **2** 706–14
[5] Quadrelli S, Lyons G, Colt H, Chimondeguy D and Buero A 2015 Clinical characteristics and prognosis of incidentally detected lung cancers *Int. J. Surg. Oncol.* **2015** 287604
[6] Manser R L, Irving L B, Byrnes G, Abramson M J, Stone C A and Campbell D A 2003 Screening for lung cancer: a systematic review and meta-analysis of controlled trials *Thorax.* **58** 784–9
[7] Bach P B, Jett J R, Pastorino U, Tockman M S, Swensen S J and Begg C B 2007 Computed tomography screening and lung cancer outcomes *J. Am. Med. Assoc.* **297** 953–61
[8] Infante M *et al* 2009 A randomized study of lung cancer screening with spiral computed tomography: three-year results from the DANTE trial *Am J. Respir. Crit. Care Med.* **180** 445–53
[9] Gates T J 2014 Screening for cancer: concepts and controversies *Am. Fam. Physician* **90** 625–31
[10] Gill R R, Jaklitsch M T and Jacobson F L 2016 Controversies in lung cancer screening *J. Am. Coll. Radiol.* **13** R2–7
[11] Aberle D R *et al* 2011 Reduced lung-cancer mortality with low-dose computed tomographic screening *New Engl. J. Med.* **365** 395–409
[12] Baldwin D R, Duffy S W, Wald N J, Page R, Hansell D M and Field J K 2011 UK lung screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer *Thorax* **66** 308–13
[13] Becker N *et al* 2012 Randomized study on early detection of lung cancer with MSCT in Germany: study design and results of the first screening round *J. Cancer Res. Clin. Oncol.* **138** 1475–86
[14] Lopes P A *et al* 2009 Design, recruitment and baseline results of the ITALUNG trial for lung cancer screening with low-dose CT *Lung Cancer* **64** 34–40
[15] van Klaveren R J *et al* 2009 Management of lung nodules detected by volume CT scanning *New Engl. J. Med.* **361** 2221–9
[16] Chirikos T N, Hazelton T, Tockman M and Clark R 2003 Cost-effectiveness of screening for lung cancer *J. Am. Med. Assoc.* **289** 2358–9
[17] Mahadevia P J, Fleisher L A, Frick K D, Eng J, Goodman S N and Powe N R 2003 Lung cancer screening with helical computed tomography in older adult smokers: a decision and cost-effectiveness analysis *J. Am. Med. Assoc.* **289** 313–22
[18] Marshall D, Simpson K N, Earle C C and Chu C W 2001 Economic decision analysis model of screening for lung cancer *Eur. J. Cancer* **37** 1759–67
[19] Marshall D, Simpson K N, Earle C C and Chu C 2001 Potential cost-effectiveness of one-time screening for lung cancer (LC) in a high risk cohort *Lung Cancer* **32** 227–36
[20] Bajtarevic A *et al* 2009 Noninvasive detection of lung cancer by analysis of exhaled breath *BMC Cancer* **9** 348
[21] Chen X *et al* 2007 A study of the volatile organic compounds exhaled by lung cancer cells *in vitro* for breath diagnosis *Cancer* **110** 835–44
[22] D'Amico A *et al* 2010 An investigation on electronic nose diagnosis of lung cancer *Lung Cancer* **68** 170–6
[23] Hubers A J *et al* 2014 Combined sputum hypermethylation and eNose analysis for lung cancer diagnosis *J. Clin. Pathol.* **67** 707–11
[24] Boots A W, Bos L D, van der Schee M P, van Schooten F J and Sterk P J 2015 Exhaled molecular fingerprinting in diagnosis and monitoring: validating volatile promises *Trends. Mol. Med.* **21** 633–44
[25] Gordon S M, Szidon J P, Krotoszynski B K, Gibbons R D and O'Neill H J 1985 Volatile organic compounds in exhaled air from patients with lung cancer *Clin. Chem.* **31** 1278–82
[26] Machado R F *et al* 2005 Detection of lung cancer by sensor array analyses of exhaled breath *Am J. Respir. Crit. Care. Med.* **171** 1286–91
[27] Phillips M *et al* 2007 Prediction of lung cancer using volatile biomarkers in breath *Cancer Biomark* **3** 95–109
[28] Phillips M *et al* 2008 Detection of lung cancer using weighted digital analysis of breath biomarkers *Clin. Chim. Acta.* **393** 76–84
[29] Schallschmidt K *et al* 2016 Comparison of volatile organic compounds from lung cancer patients and healthy controls-challenges and limitations of an observational study *J. Breath Res.* **10** 046007
[30] Soled M 1991 Aphorisms of Hippocrates *New J. Med.* **88** 33
[31] Pauling L, Robinson A B, Teranishi R and Cary P 1971 Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography *Proc. Natl Acad. Sci. USA* **68** 2374–6
[32] Dragonieri S *et al* 2009 An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD *Lung Cancer* **64** 166–70
[33] Peng G *et al* 2009 Diagnosing lung cancer in exhaled breath using gold nanoparticles *Nat. Nanotechnol.* **4** 669–73
[34] Peled N *et al* 2012 Non-invasive breath analysis of pulmonary nodules *J. Thorac. Oncol.* **7** 1528–33
[35] Bruins M, Gerritsen J W, van de Sande W W J, van Belkum A and Bos A 2013 Enabling a transferable calibration model for metal-oxide type electronic noses *Sensors Actuators B* **188** 1187–95
[36] Hauschild A C, Frisch T, Baumbach J I and Baumbach J 2015 Carotta: revealing hidden confounder markers in metabolic breath profiles *Metabolites* **5** 344–63
[37] Schumacher M, Hollander N and Sauerbrei W 1997 Resampling and cross-validation techniques: a tool to reduce bias caused by model building? *Stat. Med.* **16** 2813–27
[38] Steyerberg E W, Harrell F E Jr, Borsboom G J, Eijkemans M J, Vergouwe Y and Habbema J D 2001 Internal validation of predictive models: efficiency of some procedures for logistic regression analysis *J. Clin. Epidemiol.* **54** 774–81

[39] de la Cruz R, Fuentes C, Meza C and Nunez-Anton V 2016 Error-rate estimation in discriminant analysis of non-linear longitudinal data: a comparison of resampling methods *Stat. Methods Med. Res.* (https://doi.org/10.1177/0962280216656246)
[40] Hesterberg T C 2015 What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum *Am. Stat.* **69** 371–86
[41] Sidiropoulos N D and Kyrillidis A 2012 Multi-way compressed sensing for sparse low-rank tensors *IEEE Signal Process. Lett.* **19** 757–60
[42] Hauschild A C, Baumbach J I and Baumbach J 2012 Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification *Genet. Mol. Res.* **11** 2733–44
[43] Hughey J J, Hastie T and Butte A J 2016 ZeitZeiger: supervised learning for high-dimensional data from an oscillatory system *Nucleic Acids Res.* **44** e80
[44] Jaffe A E, Storey J D, Ji H and Leek J T 2013 Gene set bagging for estimating the probability a statistically significant result will replicate *BMC Bioinform.* **14** 360
[45] Jemal A, Bray F, Center M M, Ferlay J, Ward E and Forman D 2011 Global cancer statistics *CA Cancer J. Clin.* **61** 69–90
[46] Rafiemanesh H *et al* 2016 Epidemiology, incidence and mortality of lung cancer and their relationship with the development index in the world *J. Thorac. Dis.* **8** 1094–102
[47] Leopold J H *et al* 2015 Comparison of classification methods in breath analysis by electronic nose *J. Breath Res.* **9** 046002
[48] Bruins M G 2014 Transferable odor differentiation models for infectious disease diagnostics *Thesis*
[49] Motsinger-Reif A A, Dudek S M, Hahn L W and Ritchie M D 2008 Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiololgy *Genet. Epidemiol.* **32** 325–40