Lung Cancer 135 (2019) 123–129

Contents lists available at ScienceDirect

Lung Cancer

journal homepage: www.elsevier.com/locate/lungcan

Searching for selected VOCs in human breath samples as potential markers of lung cancer

Joanna Rudnicka[a,b], Tomasz Kowalkowski[a,b], Bogusław Buszewski[a,b,*]

[a] Department of Environmental Chemistry and Bioanalytics, Faculty of Chemistry, Nicolaus Copernicus University, 7 Gagarin St, 87-100, Toruń, Poland
[b] Centre for Modern Interdisciplinary Technologies, Nicolaus Copernicus University, 4 Wileńska St, 87-100 Toruń, Poland

ARTICLE INFO

Keywords:
Breath samples
Lung cancer
Volatile organic compounds
Statistical analysis
Discriminant function analysis
Artificial neural network

ABSTRACT

Objective: Evaluation of the potential of combined multivariate chemometric methods for seeking markers of lung cancer.

Methods: Statistical methods such as Mann-Whitney U test, discriminant function analysis (DFA), factor analysis (FA) and artificial neural network (ANN) were applied to evaluate the obtained data from GC/MS analysis of exhaled breath.

Results: The total number of compounds identified by GC/MS in human breath was equal to 88. The statistical analysis indicates seven analytes which have the highest discriminatory power. Cross validation of the obtained model shows that the sensitivity was 80% and the specificity was 91.23%, while for the test group the sensitivity and specificity were both 86.36%.

Conclusion: The application of combined statistical methods allowed to reduce the number of compounds to significant ones and indicates them as markers of lung cancer.

1. Introduction

Lung cancer, one of the most aggressive tumours, is the major cause of death both of men and women in industrialized countries [1]. One of the main causes of lung cancer is smoking (active either passive), as well as specific environmental factors. Another cause of cancer diseases incidence is exposure to carcinogenic substances such as: radon, cadmium, arsenic, nickel, beryllium or asbestos [2]. In recent years, there has been increased interest in searching for tools for early-stage lung cancer detection, especially those focusing on noninvasive medical methods such as analysis of human breath. The pioneer of exhaled air research was Linus Pauling, who in 1971 detected about 250 volatile organic compounds (VOCs) by gas chromatography [3]. Head space technique can be used for analysing volatile organic compounds as potential biomarkers for lung cancer in cancer cell, blood or exhaled air samples [4]. Till now, approximately 3000 compounds have been identified, of which approximately 250 volatile organic compounds present in the breath are detected at concentrations on the level from ppmv to pptv (parts per million by volume – parts per trillion by volume) [2]. It should be emphasized that most of the detected substances have exogenous origin. The endogenous compounds that could be identified as potential biomarkers in medical diagnostic include hydrocarbons (ethane, pentane, isoprene), oxygen-containing compounds (acetone, acetaldehyde, methanol, ethanol, 2-propanol), sulfur-containing compounds (dimethylsulfide, methylmercaptan, ethyl mercaptan, carbon disulfide) and nitrogen- containing compounds (ammonia, dimethylamine, trimethylamine) [5,6]. Because volatile organic compounds in human breath are present at the trace levels, two sampling techniques: solid phase microextraction [7,8] and thermal desorption (TD) [9,10] are mainly applied as pre-concentration and enrichment methods. After sampling, the identification of the components is performed by using: gas chromatography-mass spectrometry (GC/MS) [7,11,12], selected ion flow tube-mass spectrometry (SIFT-MS) [13,14], proton transfer reaction-mass spectrometry (PTR-MS) [15,16], ion mobility spectrometry [17] or electronic nose [18,19]. Another, unconventional method of early detection of cancers is the use of dogs, and especially their canine sense of smell. The first reports on the possibility of detecting cancer in humans by a dog appeared in the journal "The Lancet" in 1989 [20] and 2001 [21]. In the described cases, dogs without prior training contributed to the detection melanoma in their owners. Therefore, several research centres have undertaken study and tests to determine the effectiveness of trained dogs for detecting different types of cancers (prostate, bladder, ovarian, colon, lung and melanoma tumors) with the use of their well-developed olfactory sense, using various types of fragrant material, for example: tissue, urine or exhaled air [22].

[*] Corresponding author.
E-mail address: bbusz@chem.umk.pl (B. Buszewski).

https://doi.org/10.1016/j.lungcan.2019.02.012
Received 13 December 2018; Received in revised form 11 February 2019; Accepted 14 February 2019

All modern above instrumental analytical methods produce enormous flood of data. Such extensive and multidimensional datasets cannot be solely described by classical descriptive statistics as manual consideration of large data is practically impossible. Multivariate chemometric methods seem to be the only reasonable way for data classification, projection, modeling and final interpretation of the obtained results. The most commonly applied methods are: cluster analysis (CA) [23], principal component or factor analysis (PCA or FA) [23,24] and discriminant function analysis (DFA) [24]. Discriminant function analysis belongs to the so-called classification supervised methods, defining the relationship between variances within a group and between groups in a particular dataset and thus ensuring the greatest possible distinction between them. DFA allowed for the indication of the minimum number of parameters to classify data into specific groups of specified tolerance.

In this paper, the SPME-GC/MS technique was used for VOCs identification in the human breath sampled from patients with lung cancer and from healthy volunteers. The obtained data was classified by discriminant function analysis (DFA) which also allowed to select volatile organic compounds as potential markers of lung cancer. Additionally, factor analysis and artificial neural networks (ANN) were utilized for further evaluation of the results.

2. Experimental

2.1. Instrumentation

The GC/MS analysis was performed on 6890 N gas chromatograph (Agilent Technologies, Waldbronn, Germany) coupled with spectrometer mass Agilent 5975 Inert XL MSD equipped with CP-Porabond-Q (Varian Inc., Middelburg, The Netherlands) 25 m × 0.25 mm × 3 μm column. Oven temperature program was as follows: initial 40 °C maintained for 2 min, then ramped at 10 °C/min to 140 °C and next ramped at 5 °C/min to 270 °C and maintained for 5 min. The temperature of the split-splitless injector was 200 °C. The MS analyses were carried out in full-scan mode, with scan mass range of *m/z* 30 – 300. Spectra were collected at electron ionization (EI) of 70 eV, both ion source and line transfer temperatures were set to 200 °C. The acquisition of chromatographic data was performed by means of Chemstation software (Agilent). A manual SPME holder and carboxen/polydimethylsiloxane (CAR/PDMS) (75 μm) coated fiber (Supelco, Bellefonte, USA) were used for the solid phase microextraction method [12].

2.2. Chemical standards

Alkanes, alcohols, aldehydes and ketones were purchased from Sigma–Aldrich (Steinheim, Germany). Helium and argon, purity of 99.999%, were purchased from B.O.C. (Bydgoszcz, Poland) [12].

2.3. Solid phase microextraction

Before the first use, the fiber was conditioned in an injector at 200 °C for 5 h. During exposure, the SPME fiber was introduced into the bag containing sample of breath, through a silicone septum and was exposed for 10 min, at 25 °C. After extraction, the fiber was withdrawn into the needle, pulled out from the bag and injected into the GC. The compounds were desorbed in the hot GC injector port for 2 min at 200 °C [12].

2.4. Participants and breath samples collection

2.4.1. Basic information on volunteers

Exhaled breath samples were collected from: (i) 108 patients with lung cancer confirmed by histopathological examination, recruited from Department of Lung Diseases, Collegium Medicum, Nicolaus Copernicus University, Torun, Poland; (ii) 121 self-declared healthy volunteers.

For each participant the questionnaire was filled out. Samples of breath were collected from 76 males, age 39–80 (mean 60); number of smokers were 49, non-smokers 27 and 32 females, age 38–87 (mean 57); number of smokers were 20, non-smokers 12. These included males: 52 patients with non-small cell lung cancer, 10 patients with small cell lung cancer, and 14 patients with not specified histological type, whereas females: 22 patients with non-small cell lung cancer, 5 patients with small cell lung cancer and 5 patients with not specified histological type. Among the patients with non-small cell lung cancer 5 of them have stage I, 12 - stage II, 22 - stage III and 35 stage IV, while 15 patients have extensive disease (ED) of small cell lung cancer. Breath samples of healthy volunteers were collected from 121 persons, 31 males, age 21–47 (mean 40); number of smokers were 13, non-smokers 18 and 90 females, age 20–73 (mean 60); number of smokers were 37, non-smokers 53. All breath samples were collected from volunteers before eating and drinking. In order to follow the ethical requirements, all participants were informed about the aim of the study. The experimental procedure was approved by the Ethical Commission (Nicolaus Copernicus University in Torun, Ludwik Rydygier Collegium Medicum, Bydgoszcz).

2.4.2. Breath sample collection for GC/MS analysis

Breath samples were collected into 1 L Tedlar bags using a breath sampler (Medical University of Innsbruck, Austria). The device was operated in a CO_2-controlled manner and allowed to collect alveolar exhaled air samples. Before collection of breath, all bags were cleaned by flushing with argon gas and then filled with argon and heated at 60 °C for 12 h to remove any contaminants. Afterwards, a 200 mL sample from bags was transferred into another bag. Breath sample were analyzed within 3–4 hours to prevent loss of volatile organic compounds from bag. Ambient air samples were taken for blank measurement [12].

2.4.3. Calibration

Prior to the use, the glass bulb was cleaned with methanol and dried in an oven at 60 °C for at least 12 h. Afterwards, it was purged with pure argon for 15 min and pumped out using a vacuum pump within 30 min. Gaseous standards were prepared by injection using a microsyringe through the membrane of 1–2 μL of each compound into 1 L glass bulb and its evaporation. Afterwards, the mixture was moved using a gas syringe into 1 L Tedlar bag filled with 0.5 L of pure argon. During the sampling, SPME fiber was introduced into the bag through the septum to obtain concentrations in the range of 3–300 ppb, and exposed to the gas mixture [12].

2.5. Statistics

The chemometric analysis was performed using Statistica 7.1 Data Miner (Statsoft, Krakow, Poland). Firstly Mann-Whitney U test was adopted to select volatile organic compounds that significantly differ with peak area between healthy volunteers and those with lung cancer. After that, DFA was used to classify data, reduce number of variables and finally determine the selective group of analytes distinguishing between patients with cancer diagnosis and healthy control group. The factor analysis (FA) was further used to find relationships between compounds and health status and minimize the number of variables necessary for classification. The Varimax optimization of vectors was performed to maximize differences between variables needed to explain a given factor.

Moreover, variables identified as significant in the discriminant function analysis have been used to create and validate artificial neural network (ANN) as a classifier, which allowed the assignment of the examined object to a specific group. In the case of artificial neural networks, the multi-layer perceptron (MLP) concept was chosen. It was assumed that the neural network would have only one layer of hidden neurons (three-layer perceptron), while the number of neurons in the input and output layers was determined by the test criterion.

3. Results and discussion

3.1. Breath analysis

Total number of 86 volatile organic compounds has been identified in the exhaled air of healthy people and patients with lung cancer which belong to the following classes of compounds: ketones, aldehydes, alcohols, furans, nitriles, esters, aromatic hydrocarbons, sulfur compounds, saturated and unsaturated hydrocarbons. In the breath of patients with cancer, the content of alcohols (9.14%), aldehydes (12.83%) and ketones (12.64%) was higher compared to the breath of healthy people. The content of analytes belonging to the classes of saturated hydrocarbon compounds and aromatic hydrocarbons in the two groups was similar and ranging between of 25%–27% and 9.17%–9.64%, respectively. In the exhaled air of healthy persons 81 volatile organic compounds were determined while in case of lung cancer patients this number was 88. The major components of breath, acetone and isoprene were detected in all the analyzed samples. Acetonitrile, furan, 2-methylfuran, 3-methylfuran, 2,5-dimethylfuran, benzene and toluene were identified in exhaled samples from smokers. These compounds are found in tobacco smoke [25, 26]. Three of them, namely: acetonitrile, benzene and toluene were also identified in non-smoking air, which may be related to passive smoking. Also in the breath of healthy smokers and lung cancer patients unsaturated hydrocarbons such as: 2-methyl-2-butene, 2-pentene, 3-methyl-1-butene, 1,3- pentadiene, 1,4-pentadiene, 1-pentene-3-yn, 1,3-cyclopentadiene, 1,4-cyclohexadiene, 2,4-hexadiene, cyclohexene and 1-heptene were determined. Terpenes (limonene, pinene) present in human breath were probably of exogenous origin. Their source can be e.g. food, cosmetics, and household chemistry [27]. The exemplary GC/MS chromatogram of exhaled air for human with lung cancer and the frequency of detection of the compounds belonging to the specific groups is shown in Fig.1.

The mean concentrations of the main volatile organic compounds present in breath (acetone, isoprene) were significantly higher in the exhaled air of patients with cancer than in exhaled air from healthy persons. The mean concentrations of acetone and isoprene were 1000 ppb and 580 ppb, respectively. The concentration range was as follows: for alcohols (ethanol, 1-propanol, 2-propanol) 99–1203 ppb; for aldehydes (acetaldehyde, 2-propenal, propanal, pentanal, hexanal) 4–57 ppb; for ketones (2-butanone, 2-pentanone, methylvinylketone) 8–9 ppb; for esters (methyl acetate, ethyl acetate) 8–31 ppb; for compounds containing sulfur (dimethyl sulfide, carbon disulfide) 18–61 ppb. The 4-heptanone, benzaldehyde, heptanal, octanal, 2-methyl-1-propanol and 1-pentanol were not identified in the exhaled air taken from healthy volunteers [12].

3.2. Validation of the method

The precision of the method was determined by performing three replicates. The values of the relative standard deviation (RSD) were in the range from 3% to 10% for hydrocarbons, alcohols, aldehydes, ketones and aromatic compounds. The RSD values less than 10% show that the present method has good repeatability. The calibration curves were linear for aliphatic hydrocarbons in the range concentrations 3–106 ppb, for alcohols 4–419 ppb, for aldehydes 5–218 ppb, for ketones 4–333 ppb, for aromatic compounds 5–166 ppb, esters 6–154 ppb, compounds containing sulfur 8–203 ppb, acetonitrile 12–234 ppb, ethyl ether 6–118 ppb and terpenes 4–75 ppb. The linear correlation coefficients were higher than 0.991. The sensitivity of the method restricts detecting the limits of the SPME/GC–MS technique. The detection limit (LOD) and the quantification limit (LOQ) were defined as signal-to-noise ratio equal to three and ten, respectively. The lowest values of LOD were obtained for hydrocarbons and aromatic compounds ranging from 1 to 5 ppb and 2–5 ppb, respectively [12].

3.3. Statistical evaluation

The peak areas of the identified analytes have been used as input parameters for the calculation. At first stage of the statistical analysis, all 86 variables (compounds) were utilized. Due to the fact that for most variables lacking of normal distribution (checked by Shapiro-Wilks test), a non-parametric Mann-Whitney U test was used. From initial 88 compounds, the values of 34 variables indicated a significant difference ($p < 0.05$) between the healthy persons and the cancer patients. The results of the Mann Whitney U test for these compounds are presented in Table 1.

In the next step, the discriminant function analysis (DFA) was performed on the dataset containing selected 34 compounds. The group of 98 healthy volunteers and 86 patients with lung cancer was randomly chosen to build classification model. The stepwise forward mode of discriminant function analysis was chosen. It allowed further reduction of variables to 21 parameters. The results of the discriminant function analysis for these variables are presented in Table 2 including coefficients to build discriminant function. The marked compounds have a significant F-remove value ($p < 0.05$) higher than the threshold value of F (1.119). The discriminant ability of particular component is determined by Wilks' partial Lambda. It has also been shown that isoprene has the greatest discriminatory power. Other highly discriminating compounds were as follows: cyclohexane, cyclohexanone, acetone, 2-methylheptane, methyl acetate and methyl vinyl ketone. Discriminative power of model expressed by Lambda Wilks value was high (0.4939). The canonical analysis of the discriminant function was performed for 21 selected variables. It allowed to reduce parameter's space to single canonical root. Root's values were statistically significantly different between the group of patients and healthy individuals ($p = 0.0000$). However, the total separation of the studied groups was not achieved (Fig.1S).

At the next stage, the obtained discriminant model was validated. The canonical coefficients were calculated for the test group containing patients not included in the model building phase (23 healthy volunteers and 22 lung cancer patients) (Fig.2). As a result of cross validation, the model correctly classified 80% of lung cancer samples (18 out of 90 cases were misclassified) and 91.23% of healthy samples (8 out of 91 cases were misclassified). Statistically significant differences were found for all the investigated cases (model and test group) ($p = 0.0000$) (Fig. 3).

Additional analyses were also performed. Variables with a significant value F-remove from Table 2 were used to perform factor analysis (FA) with Varimax normalized rotation of vectors. This method replaces the original variables with so-called factors correlated with individual variables and allows to determine the impact of a given group of compounds on the classification of samples. Factor 1 is related to cyclohexane and 2-methylheptane, while factor 2 is negatively correlated to acetone and isoprene. In the second group, compounds that have endogenic origin [21] are identified as opposed to the first group (Table 1S).

Artificial neural network (ANN) was applied to improve the classification. ANN model was constructed using parameters that were indicated as significant in discriminant function analysis (Table 2). A database containing 7 parameters (compounds) and 226 cases (healthy volunteers and patients) was used. 50% of cases were selected for network learning, 25% for testing, and 25% for validation. 1000 different artificial neural networks topologies were tested. The best model had 7 neurons in the input layer, 6 neurons in the hidden layer and 2 neurons in the output layer. Exponential functions were used both in the hidden and the first layer. The use of selected network gave a 27.7% of incorrectly classified cases. For the selected model, an ROC curve was

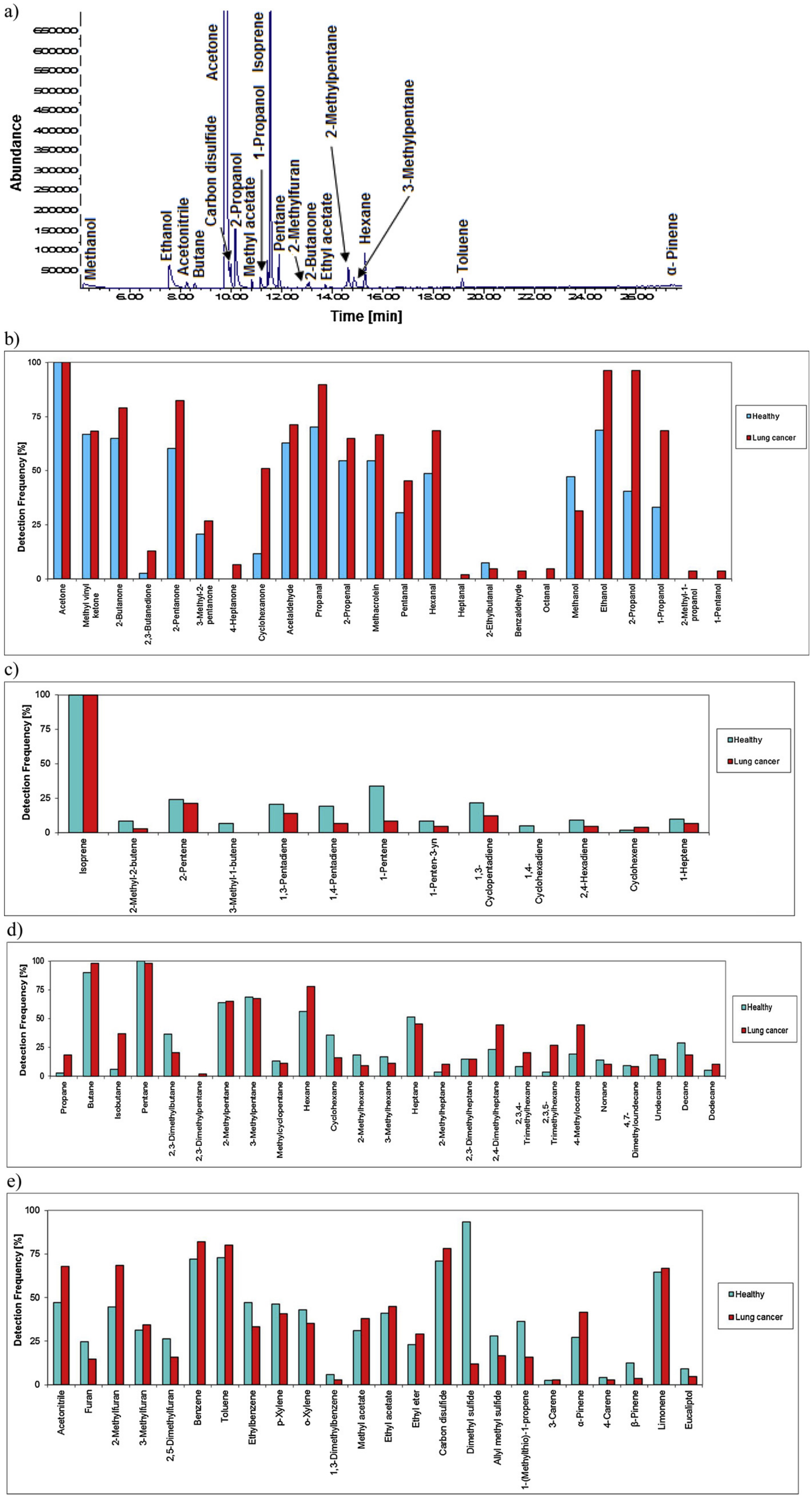

Fig. 1. Detection frequency of VOCs in exhaled air of healthy and lung cancer volunteers belonging to particular classes compounds: a) the GC /MS chromatogram of breath of patient with lung cancer b) aldehydes, ketones, alcohols; c) saturated hydrocarbons; d) unsaturated hydrocarbons; e) compounds containing nitrogen, furans, aromatics, sulfur-containing compounds and terpenes.

Table 1
The summary of Mann–Whitney's U test.

Compound	CAS number	Rank sum lung cancer	Rank sum healthy	U	p-Level
Propane	74-98-6	8772.5	7698.5	3512.5	0.0011
Isobutane	75-28-5	8717.0	7754.0	3568.0	0.0098
Acetonitrile	75-05-8	8775.0	7696.0	3510.0	0.0498
Acetone	67-64-1	9163.0	7308.0	3122.0	0.0058
Carbon disulfide	75-15-0	9089.0	7382.0	3196.0	0.0099
2-Propanol	67-63-0	8822.5	7648.5	3462.5	0.0143
Dimethyl sulfide	75-18-3	4915.0	11556.0	820.0	0.0000
Methyl acetate	79-20-9	8911.0	7560.0	3374.0	0.0348
Isoprene	78-79-5	9427.0	7044.0	2858.0	0.0004
2-Methyl-2-butene	513-35-9	7872.5	8598.5	3777.5	0.0168
1,4-Pentadiene	591-93-5	7647.0	8824.0	3552.0	0.0089
1-Pentene	109-67-1	7682.5	8788.5	3587.5	0.0096
Methyl vinyl ketone	78-94-4	6951.0	9520.0	2856.0	0.0002
2,3-Butanedione	431-03-8	8651.0	7820.0	3634.0	0.0024
2-Methylfuran	534-22-5	7517.5	8953.5	3422.5	0.0283
Ethyl acetate	141-78-6	8752.5	7718.5	3532.5	0.0078
Hexane	110-54-3	7546.0	8925.0	3451.0	0.0402
Cyclohexane	110-82-7	7066.0	9405.0	2971.0	0.0000
2-Pentanone	107-87-9	10093.0	6378.0	2192.0	0.0000
Hexanal	66-25-1	7225.0	9246.0	3130.0	0.0001
2-Methylheptane	592-27-8	8504.0	7967.0	3781.0	0.0466
Cyclohexanone	108-94-1	9664.5	6806.5	2620.5	0.0000
Ethylbenzene	100-41-4	7340.0	9131.0	3245.0	0.0029
p-Xylene	106-42-3	7411.5	9059.5	3316.5	0.0073
o-Xylene	95-47-6	7656.0	8815.0	3561.0	0.0279
4-Heptanone	123-19-3	8372.0	8099.0	3913.0	0.0426
2,4-Dimethylheptane	2213-23-2	9534.5	6936.5	2750.5	0.0000
2,3,4-Trimethylhexane	921-47-1	8909.0	7562.0	3376.0	0.0028
4-Methyloctane	2216-34-4	9581.5	6889.5	2703.5	0.0000
Nonane	111-84-2	8835.5	7635.5	3449.5	0.0079
ß-Pinene	127-91-3	7830.0	8641.0	3735.0	0.0274
4,7-Dimethylundecane	17301-32-5	8463.0	8008.0	3822.0	0.0125
Limonene	138-86-3	6969.5	9501.5	2874.5	0.0003
Dodecane	112-40-3	8516.5	7954.5	3768.5	0.0385

Table 2
The summary of discriminant function analysis.

Compound	Wilks' Lambda	Partial Wilks' Lambda	F-remove (1.119)	p-Level	Standardized Coefficients for Canonical Root
2,4-Dimethylheptane	0.4997	0.9885	1.8511	0.1756	−0.219
Isoprene	0.5634	0.8767	22.3593	0.0000	−0.590
Cyclohexanone	0.5202	0.9495	8.4539	0.0042	−0.346
Cyclohexane	0.5239	0.9429	9.6368	0.0023	0.427
2-Methylheptane	0.5099	0.9687	5.1308	0.0249	−0.313
2-Methylfuran	0.5004	0.9870	2.0959	0.1497	0.174
Acetone	0.5167	0.9559	7.3308	0.0075	−0.320
Dodecane	0.5045	0.9790	3.4026	0.0670	−0.208
4-Methyloctane	0.5002	0.9875	2.0165	0.1576	−0.226
Acetonitrile	0.5033	0.9813	3.0292	0.0837	0.198
Methyl vinyl ketone	0.5078	0.9726	4.4731	0.0360	0.241
Dimethyl sulfide	0.5048	0.9784	3.5109	0.0628	0.233
Hexane	0.5041	0.9798	3.2754	0.0722	0.211
1,4-Pentadiene	0.4961	0.9957	0.6872	0.4084	0.119
Methyl acetate	0.5126	0.9636	6.0100	0.0153	−0.332
4-Heptanone	0.5017	0.9846	2.4949	0.1162	−0.177
Ethyl acetate	0.5015	0.9848	2.4467	0.1198	0.241
Isobutane	0.5025	0.9829	2.7675	0.0982	−0.214
p-Xylene	0.4979	0.9919	1.2946	0.2569	0.129
Propane	0.5015	0.9850	2.4261	0.1213	0.216
2-Pentanone	0.5013	0.9853	2.3729	0.1254	−0.190

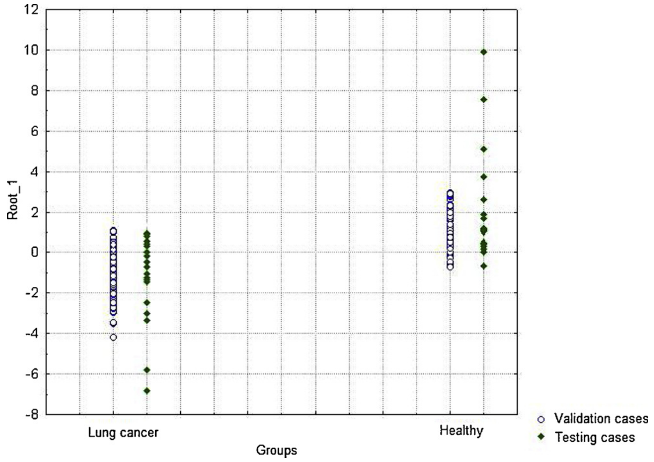

Fig. 2. Canonical scores for investigated cases.

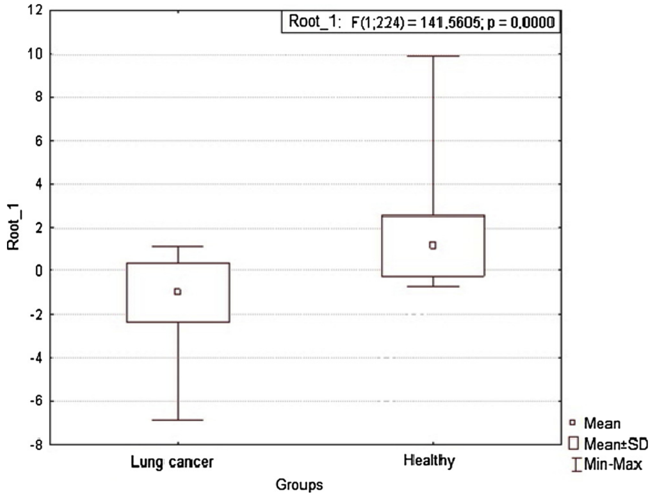

Fig. 3. The box-and-whisker graph for all cases studied.

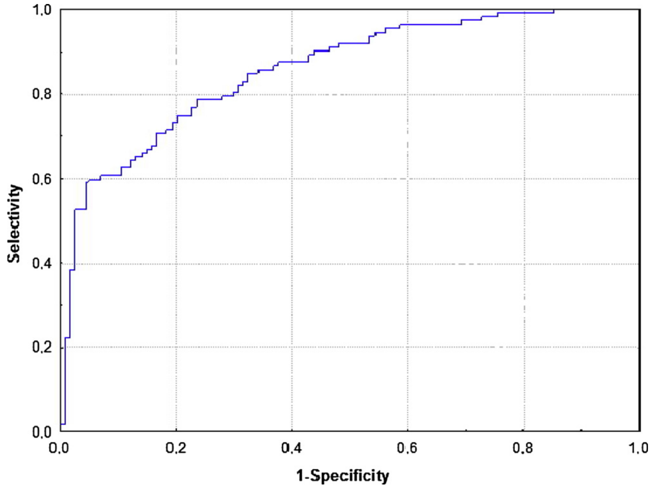

Fig. 4. The ROC curve for selected ANN model.

constructed (Fig. 4), where the predictive value of lung cancer was 72%, specificity was 79% and AUC was 0.86.

4. Discussion

Compounds such as alkanes, ketones, aldehydes are produced in the human body by oxidation of fatty acids. Their increase in concentration occurs during inflammation and oxidative stress. There are several sources of aldehydes in the human body. The major one is associated with the metabolism of ethanol. Aldehydes that are formed in the body are oxidized by aldehyde dehydrogenase (ALDH) to carboxylic acids. The presence of aldehydes in the body is also associated with smoking.

Both saturated (formaldehyde, ethanal, propanal, butanal) and unsaturated compounds (acrolein, croton aldehyde) are present in tobacco smoke. It is supposed that sugars are the main source of formaldehyde in cigarette smoke, while cellulose may be the precursor of acetaldehyde in the main stream of cigarette smoke. Another source of aldehydes in the body is associated with their formation as a by-product of tobacco metabolism catalyzed by cytochrome P450, which is part of the detoxification process. Alcohols are also derived from hydrocarbon metabolism. It is likely that alcohols are metabolized to aldehydes through the action of such enzymes as alcohol dehydrogenase (ADH) and cytochrome P450 (CYP2E1), which function mainly in the liver. Alcohol dehydrogenase belongs to a general group of enzymes called oxidoreductases, which contribute to the formation of aldehydes or ketones from alcohol by the reduction of nicotinamide adenine dinucleotide (NAD + to NADH). ADH can catalyze the oxidation of primary alcohols in the human body [4]. Phillips et al. [28] have suggested that branched chain hydrocarbons (C4-C20) may also be potential markers of oxidative stress. This hypothesis, however, is difficult to confirm or exclude, as a result of inflammation, methylated alkanes are produced and they may be present in human breath. These compounds can also be of exogenous origin. No increase in pentane content (potential oxidative stress) in exhaled air was observed in patients with lung cancer during the present study. Also, alkanes are produced during the oxidation of polyunsaturated fatty acids by reactive oxygen forms (ethane, pentane). Other saturated hydrocarbons, e.g. C3-C11, may also be formed as a result of the lipid oxidation process. However, the presence of branched hydrocarbons in the body is not possible due to this mechanism, because there are no branched polyunsaturated fatty acids in it [4]. Alkanes such as propane or butane can be produced as a result of protein oxidation or come from the intestinal flora [29]. Acetone is produced by the decarboxylation of acetylacetate and acetyl CoA, while isoprene is formed along the mevalonic pathway of cholesterol synthesis [29]. Analytes such as furan, 2-methylfuran, 3-methylfuran, 2,5-dimethylfuran, acetonitrile, benzene, toluene, cyclohexanone and methyl acetate are found in smokers' exhaled air.

Rudnicka et.al [30] used statistical methods such as artificial neural network and chi-squared automatic interaction detector (CHAID) without any data reduction. For obtained model, the ROC was construed to predict lung cancer with sensitivity 74%, specificity 73% and AUC = 0.97. Phillips et al. [31] examined the exhaled air samples from 178 patients with lung cancer and 41 healthy persons by gas chromatography- mass spectrometry and identified 80 volatile organic compounds (C4-C20). 87 out of 178 samples from patients were used to build the model. As a result of forward stepwise discriminant function analysis, 9 analytes were selected which had the biggest discriminatory power between lung cancer patients and healthy individuals. The sensitivity for this method was 89.6%, while the specificity was 82.9%. Cross validation correctly classified 85.1% of patients with lung cancer and 80.5% without cancer.

In this paper a different approach was taken to analyze the received data. At first stage the Mann- Whitney U test was applied to select VOCs which significant differentiate healthy volunteers and patients with lung cancer. Next, DFA was applied to reduce number of variables and determine group of analytes which distinguishing two research groups. Than the variables identified as significant in the DFA were used to create artificial neural network. As a result of cross validation for the obtained model, the sensitivity was 80% and the specificity was 91.23%, whereas for the test group the sensitivity and specificity of the examined groups was 86.36%.

5. Conclusions

The statistical analysis allowed selecting compounds which significantly differentiate between groups of healthy volunteers and patients with lung cancer. The combined statistical methods allowed to reduce the number of compounds to seven: acetone, methyl acetate, isoprene, methyl vinyl ketone, cyclohexane, 2-methylheptane, cyclohexanone, which separated human breath samples of two research groups containing lung cancer sufferers and healthy persons. The application of non-parametric test and classification method could be a useful tool which can support screening of lung cancer.

Conflict of interest

The authors have no conflict of interest in relationship with the content of the manuscript to be disclosed.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.lungcan.2019.02.012.

References

[1] A. Kishor Ganti, J.L. Mulshine, Lung cancer screening, Oncologist 11 (2006), https://doi.org/10.1634/theoncologist.11-5-481 481-48. https://doi: 10.1634/theoncologist.11-5-481.

[2] B. Buszewski, J. Rudnicka, T. Ligor, M. Walczak, T. Jezierski, A. Amann, Analytical and unconventional methods of cancer detection using odor, TrAC 38 (2012) 1–12, https://doi.org/10.1016/j.trac.2012.03.019.

[3] L. Pauling, A.B. Robinson, R. Teranishi, P. Cary, Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography, Proc. Nat. Acad. Sci. 68 (1971) 2374–2376, https://doi.org/10.1073/pnas.68.10.2374.

[4] M. Hakim, Y.Y. Broza, O. Barash, N. Peled, M. Phillips, A. Amann, H. Haick, Volatile organic compounds of lung cancer and possible biochemical pathways, Chem. Rev. 112 (2012) 5949–5966, https://doi.org/10.1021/cr300174a.

[5] W. Miekisch, J.K. Schubert, G.F.E. Noeldge-Schomburg, Diagnostic potential of breath analysis-focus on volatile organic compounds, Clin. Chim. Acta 347 (2004) 25–39, https://doi.org/10.1016/j.cccn.2004.04.023.

[6] B. Buszewski, M. Kęsy, T. Ligor, A. Amann, Human exhaled air analytics: biomarkers of diseases, Biomed. Chromatogr. 21 (2007) 553–566, https://doi.org/10.1002/bmc.835.

[7] M. Kosmo, E. Mendez, K.G. Furton, Development of headspace SPME method for analysis of volatile organic compounds present in human biological specimens, Anal. Bioanal. Chem. 400 (2011) 1817–1826, https://doi.org/10.1007/s00216-011-4950-2.

[8] F. Di Francesco, R. Fuoco, M.G. Trivella, A. Ceccarini, Breath analysis: trends in techniques and clinical applications, Microchem. J. 79 (2005) 405–410, https://doi.org/10.1016/j.microc.2004.10.008.

[9] A. Hryniuk, B.M. Ross, Detection of acetone and isoprene in human breath using a combination of thermal desorption and selected ion flow tube mass spectrometry, Int. J. Mass Spectrom. 285 (2009) 26–30, https://doi.org/10.1016/j.ijms.2009.02.027.

[10] A.W. Jones, V. Lagesson, C. Tagesson, Determination of isoprene in human breath by thermal desorption gas chromatography with ultraviolet detection, J. Chromatogr. B 672 (1995) 1–6, https://doi.org/10.1016/0378-4347(95)00207-Y.

[11] R. Hyspler, S. Crhova, J. Gasparic, Z. Zadak, M. Cizkova, V. Balasova, Determination of isoprene in human expired breath using solid-phase microextraction and gas chromatography-mass spectrometry, J. Chromatogr. B 739 (2000) 183–190, https://doi.org/10.1016/S0378-4347(99)00423-5.

[12] J. Rudnicka, T. Kowalkowski, T. Ligor, B. Buszewski, Determination of volatile organic compounds as biomarkers of lung cancer by SPME-GC-TOF/MS and chemometrics, J. Chromatogr. B 879 (2009) 3360–3366, https://doi.org/10.1016/j.jchromb.2011.09.001.

[13] M.S. Abbott, B. Elder, P. Spanel, D. Smith, Quantification of acetonitrile in exhaled breath and urinary headspace using selected ion flow tube mass spectrometry, Int. J. Mass Spectrom. 228 (2003) 655–665, https://doi.org/10.1016/S1387-3806(03)00212-4.

[14] M. Storer, J. Salmond, K.N. Dirks, S. Kingham, M. Epton, Mobile selected ion flow tube mass spectrometry (SIFT-MS) devices and their use for pollution exposure monitoring in breath and ambient air–pilot study, J. Breath Res. 8 (2014) 037106, , https://doi.org/10.1088/1752-7155/8/3/037106 (7pp).

[15] J. King, P. Mochalski, A. Kupferthaler, K. Unterkofler, H. Koc, W. Filipiak, S. Teschl, H. Hinterhuber, A. Amann, Dynamic profiles of volatile organic compounds in exhaled breath as determined by a coupled PTR-MS/GC-MS study, Physiol. Meas. 31 (2010) 1169–1184, https://doi.org/10.1088/0967-3334/31/9/008.

[16] X. Zhan, J. Duan, Y. Duan, Recent developments of proton-transfer reaction mass spectrometry (PTR-MS) and its applications in medical research, Mass Spectrom. Rev. 32 (2013) 143–165, https://doi.org/10.1002/mas.21357.

[17] P. Mochalski, J. Rudnicka, A. Agapiou, M. Statheropoulos, A. Amann, B. Buszewski, Near real-time VOCs analysis using an aspiration ion mobility spectrometer, J. Breath Res. 7 (2013) 026002, , https://doi.org/10.1088/1752-7155/7/2/026002 (11pp).

[18] A. D'Amico, G. Pennazza, M. Santonico, E. Martinelli, C. Roscioni, G. Galluccio, R. Paolesse, C. Di Natale, An investigation on electronic nose diagnosis of lung cancer, Lung Cancer 68 (2010) 170–176, https://doi.org/10.1016/j.lungcan.2009.

11.003.

[19] M. Tirzīte, M. Bukovskis, G. Strazda, N. Jurka, I. Taivans, Detection of lung cancer in exhaled breath with an electronic nose using support vector machine analysis, J. Breath Res. 11 (2017) 50–57, https://doi.org/10.1088/1752-7163/aa7799.

[20] H. Williams, A. Pembroke, Sniffer dogs in the melanoma clinic? Lancet 1 (1989) 73, https://doi.org/10.1016/S0140-6736(89)92257-5.

[21] J. Church, H. Williams, Another sniffer dog for the clinic? Lancet 358 (2001) 930, https://doi.org/10.1016/S0140-6736(01)06065-2.

[22] A. Amann, W. Miekisch, J. Schubert, B. Buszewski, T. Ligor, T. Jezierski, J. Pleil, T. Risby, Analysis of exhaled breath for disease detection, Annu. Rev. Anal. Chem. Palo Alto Calif (Palo Alto Calif) 7 (2014) 455–482, https://doi.org/10.1146/annurev-anchem-071213-020043.

[23] J. Bartel, J. Krumsiek, F.J. Theis, Statistical methods for the analysis of high-throughput metabolomics data, Comput. Struct. Biotechnol. J. 4 (2013) 1–9, https://doi.org/10.5936/csbj.201301009.

[24] J. Pereira, P. Porto-Figueira, C. Cavaco, K. Taunk, R.D. Rapole, H. Nagarajaram, J.S. Câmara, Breath analysis as a potential and non-invasive frontier in disease diagnosis: an overview, Metabolites 5 (2014) 3–55, https://doi.org/10.3390/metabo5010003.

[25] M.M. Sampson, D.M. Chambers, D.Y. Pazo, F. Moliere, B.C. Blount, C.H. Watson, Simultaneous analysis of 22 volatile organic compounds in cigarette smoke using gas sampling bags for high-throughput solid-phase microextraction, Anal. Chem. 86 (2014) 7088–7095, https://doi.org/10.1021/ac5015518.

[26] G.M. Polzin, R.E. Kosa-Maines, D.L. Ashley, C.H. Watson, Analysis of volatile organic compounds in mainstream cigarette smoke, Environ. Sci. Technol. 41 (2007) 1297–1302, https://doi.org/10.1021/es060609l.

[27] Z. Xie, Q. Liu, Z. Liang, M. Zhao, X. Yu, D. Yang, X. Xu, The GC/MS analysis of volatile components extracted by different methods from Exocarpium Citri Grandis, J. Anal. Methods Chem. 2013 (2013) 1–8, https://doi.org/10.1021/es060609l.

[28] M. Philips, R.N. Cataneo, J. Greenberg, R. Gunawardena, A. Naidu, F. Rahbari-Oskoui, Effect of age on the breath methylated alkane contour, a display of apparent new markers of oxidative stress, J. Lab. Clin. Med. 136 (2000) 243–248, https://doi.org/10.1067/mlc.2000.108943.

[29] W. Miekisch, J.K. Schubert, G.F.E. Noeldge-Schomburg, Diagnostic potential of breath analysis-focus on volatile organic compounds, Clin. Chim. Acta 347 (2004) 25–39, https://doi.org/10.1016/j.cccn.2004.04.023.

[30] J. Rudnicka, M. Walczak, T. Kowalkowski, T. Jezierski, B. Buszewski, Determination of volatile organic compounds as potential markers of lung cancer by gas chromatography – mass spectrometry versus trained dogs, Sens. Actuat. B: Chemical 202 (2014) 615–621, https://doi.org/10.1016/j.snb.2014.06.006.

[31] M. Phillips, R.N. Cataneo, A.R.C. Cumin, A.J. Gagliardi, J. Greenberg, R.A. Maxfield, W.N. Rom, Detection of lung cancer with volatile markers in the breath, Chest 123 (2003) 2115–2123, https://doi.org/10.1378/chest.123.6.2115.