GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification

A.-C. Hauschild[1,2], J.I. Baumbach[1] and J. Baumbach[2,3]

[1]Department Microfluidics and Clinical Diagnostics, KIST Europe, Saarbrücken, Germany
[2]Computational Systems Biology Group, Max Planck Institute for Informatics, Saarbrücken, Germany
[3]Cluster of Excellence for Multimodel Computing and Interaction, Saarland University

Corresponding author: J. Baumbach
E-mail: jan.baumbach@mpi-inf.mpg.de

Genet. Mol. Res. 11 (3): 2733-2744 (2012)
Received March 27, 2012
Accepted May 2, 2012
Published July 10, 2012
DOI http://dx.doi.org/10.4238/2012.July.10.17

ABSTRACT. Exhaled air carries information on human health status. Ion mobility spectrometers combined with a multi-capillary column (MCC/IMS) is a well-known technology for detecting volatile organic compounds (VOCs) within human breath. This technique is relatively inexpensive, robust and easy to use in every day practice. However, the potential of this methodology depends on successful application of computational approaches for finding relevant VOCs and classification of patients into disease-specific profile groups based on the detected VOCs. We developed an integrated state-of-the-art system using sophisticated statistical learning techniques for VOC-based feature selection and supervised classification into patient groups. We analyzed breath data from 84 volunteers, each of them either suffering from chronic obstructive pulmonary disease (COPD), or both COPD and bronchial carcinoma (COPD + BC), as well as from 35 healthy volunteers, comprising a control group (CG). We standardized and

integrated several statistical learning methods to provide a broad overview of their potential for distinguishing the patient groups. We found that there is strong potential for separating MCC/IMS chromatograms of healthy controls and COPD patients (best accuracy COPD *vs* CG: 94%). However, further examination of the impact of bronchial carcinoma on COPD/no-COPD classification performance is necessary (best accuracy CG *vs* COPD *vs* COPD + BC: 79%). We also extracted 20 high-scoring VOCs that allowed differentiating COPD patients from healthy controls. We conclude that these statistical learning methods have a generally high accuracy when applied to well-structured, medical MCC/IMS data.

Key words: Ion mobility spectrometry; Machine learning; Chronic obstructive pulmonary disease; Bronchial carcinoma; Feature selection

INTRODUCTION

Multi-capillary column-ion mobility spectrometry (MCC/IMS) is a comparatively inexpensive, sensitive high through-put method to analyze human exhaled air carrying information about health status. The resulting MCC/IMS chromatograms, contain this information. Sophisticated computational approaches can be utilized for classifying patients into disease-specific profile groups and identifying the volatile organic compounds (VOCs) that are important.

First, a brief introduction to chronic obstructive pulmonary disease (COPD) is provided, followed by an overview of the MCC/IMS technique. Various preprocessing steps for the analysis of MCC/IMS data and the different machine learning methods applied in this study are also elucidated. Finally, results obtained from various machine learning methods are presented and followed by a discussion and comparison with the state-of-the-art techniques.

COPD

COPD is an inflammatory lung disease characterized by a permanent blockage of airflow from the lungs. The primary cause of COPD is tobacco smoke (through smoking or second-hand smoke). The disease is widely under-diagnosed, although it is a life-threatening lung disease, which is not fully reversible. The World Health Organization (WHO) reported it to be one of the most frequent causes of death. It is in fourth place after ischemic heart disease, cerebrovascular disease, and lower respiratory infections (World Health Organization, 2008). According to a WHO report in 2008, an estimated 64 million people worldwide suffered from COPD in 2004, and more than 3 million people died of COPD in 2005. This number will most likely increase by ≥30% (http://www.who.int/healthinfo/global_burden_disease/en/). Young et al. reported in 2009 that COPD is both a common and important independent risk factor for lung cancer. Hence, our study includes samples of both patients suffering from COPD in combination with bronchial carcinoma and patients with only COPD.

In clinical practice, the diagnosis of COPD is based on three different parts, the symp-

toms, the assessment of lung function and the evaluation of the responses to inhaled pharmacological agents (Rabe et al., 2007). Although these tests are generally considered informative, they are time-consuming and strongly dependent on the personnel's experience.

MCC/IMS

Human exhaled air contains a combination of VOCs reflecting the state of health of the body. Therefore, it is a potential information carrier for novel diagnostic techniques. The ion mobility spectrometer combined with a multi-capillary column as a pre-separation (MCC/IMS) is a well known technology for detecting VOCs in human breath. There are several advantages of MCC/IMS: it is very sensitive (detection limit at the nanogram and picogram per liter levels).

It can handle the moisture that comes with exhaled air, and polar molecules are detectable. Also, most measurements are inexpensive and fast (≈5 min). In the study presented here, a BioScout device (B&S Analytik, Dortmund, Germany) was utilized for collecting the metabolomic data. It consists of a SpiroScout (Ganshorn Medizin Electronic, Niederlauer, Germany) as sample inlet unit and the MCC/IMS. The end-tidal breath gathered through the SpiroScout is collected in a sample loop and transferred to the multi-capillary column for the first chromatographic separation. Entering the IMS, the compounds are ionized by a ^{63}Ni β-radiation source and intermittently injected into the drift tube and detected by the Faraday plate. For further details, see Baumbach (2009). Finally, we obtained a three-dimensional data file for each of the measurements. The first dimension is defined by the retention time (RT), the time the molecule needs to pass through the multi-capillary column. The second dimension corresponds to the drift time K_0, which is the time the compound flies through the ion mobility spectrometer. Finally, a Faraday detector measures the third dimension, the electric charge h. The combination of the three measures can be visualized as a heat map (see Figure 1).

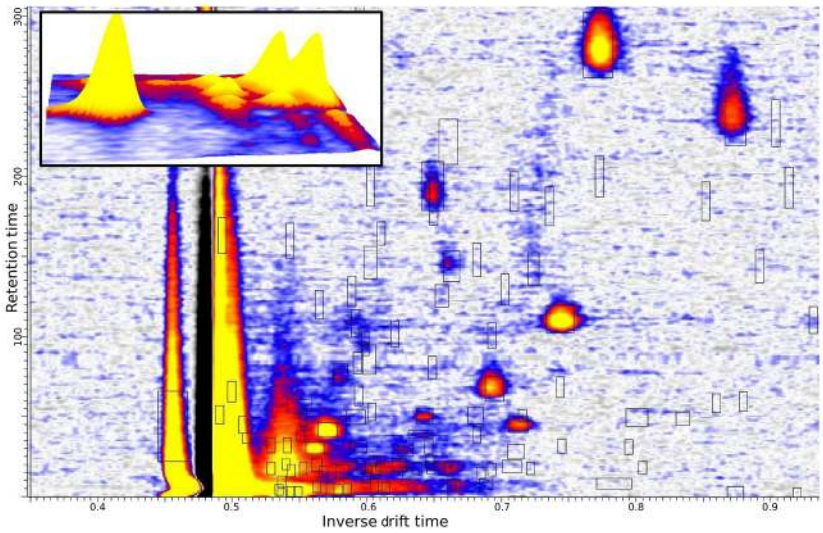

Figure 1. Example of a multi-capillary column-ion mobility spectrometry chromatogram (human exhaled air). Each volatile organic compound is marked by a black rectangle.

Our contribution

In this study, we analyzed the metabolomic profiles of volunteers suffering from COPD and healthy controls. First, the measurements were captured by the BioScout and the data set stored (see Figure 2, step 1). Afterwards, the three-dimensional data files were preprocessed for noise reduction and VOC identification (Figure 2, step 2). The data set consists of three different groups of volunteers: healthy controls, COPD patients, and COPD with bronchial carcinoma (BC) patients (COPD+BC).

We addressed the following two major questions (Figure 2, step 3):

Can we generally distinguish between healthy and diseased volunteers (Healthy *vs* COPD and COPD+BC)?

Can we find differences between all three groups, healthy, COPD, and COPD+BC?

To tackle these problems, we applied and evaluated several statistical learning techniques (Figure 2, steps 4 and 5). Furthermore, we aimed at finding those VOCs that are most relevant for classification power.

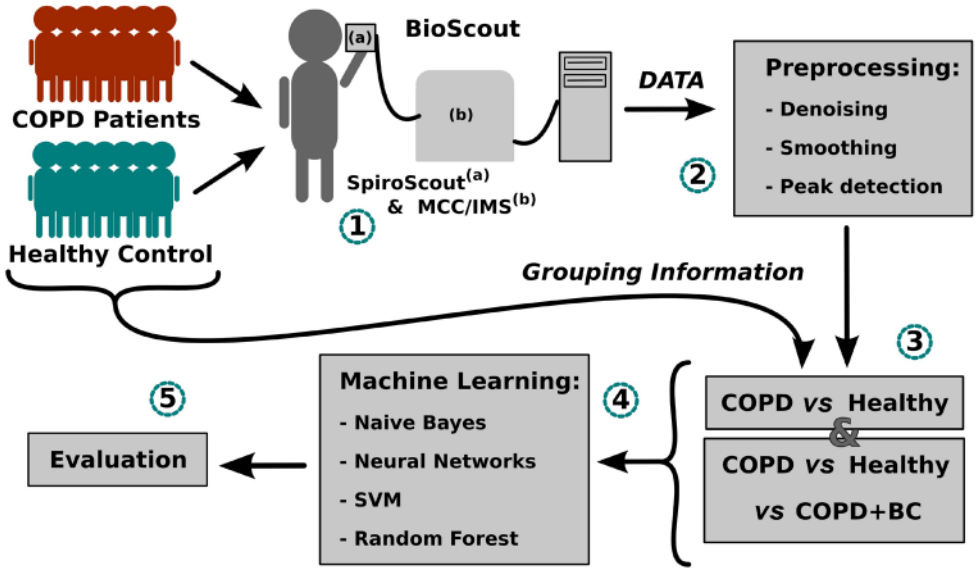

Figure 2. Overview of the integrated statistical approach to evaluate chronic obstructive pulmonary disease (COPD)-related metabolic multi-capillary column-ion mobility spectrometry (MCC/IMS) profiles. SVM = Support vector machine; BC = bronchial carcinoma. The numbering indicates the different steps of the approach.

MATERIAL AND METHODS

Data and preprocessing

In total, we analyzed the data of 119 volunteers: 35 healthy controls and 84 patients suffering from COPD. Note that 54 of the COPD patients also suffered from bronchial carcinoma. If not stated otherwise, we used the following abbreviations as class descriptors: HC =

healthy controls, COPD = COPD patients, BC = bronchial carcinoma patients, and COPD+BC = only those COPD patients also suffering from BC. The patients were recruited from cooperating German hospitals.

Preprocessing

Our MCC/IMS measurements contained 1 million entries, 500 for the retention time axes and 2000 for the drift time axes. We applied data reduction methods for denoising, smoothing and peak detection, such as log-normal detailing and wavelet transformation (Bader et al., 2008). The most intuitive way of feature reduction is to define regions within the heat map where VOCs are detected in at least one of the measurements analyzed. The maximal height within a specific region of an IMS chromatogram indicates whether the compound is present or absent. It represents the quantity of the compound and is stored as a matrix of maxima, for all regions and chromatograms.

This was done by utilizing the VisualNow software (provided with BioScout by B&S Analytik), which allows for manual expert-based peak picking or execution of fully automatic peak finding algorithms. In our case, the data were preprocessed using the standard settings of VisualNow. Using the manual peak picking procedure, we determined 120 VOC positions. The final preprocessed data table consisted of 119 measurements and a continuous value of the quantity of each of the 120 components.

Statistical methods

We used several standardized statistical learning methods in this study, to get a broad overview of the potential of the data and the different classification techniques. Thus, we included some rather simple methods, decision tree, naive Bayes, and linear support vector machine (SVM), for instance. On the other hand, we used more recent and sophisticated techniques, such as neural net, random forest and radial SVM. The R language package (Version 2.13.1) was used (R Development Core Team, 2011) to implement the statistical analysis and feature selection. A brief overview of the different techniques follows.

Decision tree

This method is a fundamental machine learning tool based on recursive partitioning. The features span the space of the classification task. This space is cut into one dimension/feature in such way that the accuracy is maximized. The method repeats this procedure for the resulting subspaces recursively until the the accuracy is optimal. Each split corresponds to one internal node. Each leaf-node is labeled by the majority class of the samples in the corresponding subspace. The decision tree was implemented using the R package rpart (Therneau and Atkinson, 2011), using the standard parameters.

Naive Bayes

The most widely tested and straightforward method for statistical induction is known as the naive Bayesian classifier (Langley and Sage, 1994). In this method, each class is repre-

sented by a single probabilistic summary. To minimize the probability of error in the classification assignment, the state of action that maximizes the posterior probability is chosen each time. This is calculated by the Bayesian formula simply from the prior probabilities and the conditional densities. Despite its simplicity, the naive Bayes works quite well in practice and outperforms far more sophisticated techniques. The reason for this is that although bias can hurt the individual class density estimates, this might not influence the posterior probabilities as much, especially near the decision (Hastie et al., 2009). The naive Bayes method was implemented using the R package NaiveBayes (Weihs et al., 2005), using the standard parameters with activated usekernel parameter.

Neural networks

A neural network is a two-stage regression or classification model, which is typically represented as a network. The smallest model contains three layers, the input layer, the output layer, and at least one hidden layer, whereas the complexity of the model increases by the number of hidden layers. For a *k*-class classification problem the model contains *k* output units, each showing the probability of the associated class. The result of an output or hidden node is created from a linear combination of the nodes of the previous layer, respectively. Applying an activation function, which is usually chosen to be the sigmoid $1/1+e^{-v}$ within each hidden node leads to a non-linear model. The classification by neural networks was implemented using the nnet package (Venables and Ripley, 2002). In our case, we set the number of hidden layers to 2. The weight decay and the maximal number of iterations were kept at the standard settings 0 and 100, respectively.

Random forest

Random forest is based on the strategy of bagging. Bagging is a sampling technique applying a method with low-bias and high-variance on subsets of the data. To reduce the high-variance of the unbiased method, the outcome is averaged. Random forest builds a large collection of de-correlated trees and averages the results (regression) or uses a majority vote (classification). Since trees can capture complex interactions in the data and are unbiased if they are grown sufficiently deep, they are the perfect candidates for bagging.

A specific bootstrapping approach is used to reduce the correlation between the trees without increasing the variance of the whole classifier too much. This means that each decision tree is grown on a new set of samples, randomly drawn from the original data set and of equal size. Finally, during the assembly of the decision tree, a random set of variables is drawn out of the bootstrapping sample for each of the nodes (Hastie et al., 2009).

The random forest method provides two measures of importance, both dependent on the accuracy of the trees. The first, called Gini index, accumulates the improvement in the split-criterion, while growing the trees for each variable and corresponding splits. The second uses the left out samples (called out of the box samples, OOB) of each tree to measure the prediction strength of each variable; it is further referred to as OOB randomization. See Hastie et al. (2009) for more details. The random forest classification and feature selection were done using the randomForest R package, by Liaw and Wiener (2002). Again, standard parameters were used, the number of trees was equal to 500, and no limits in the number of nodes for a single tree were set.

Support vector machine

SVM is one of the most widely used statistical learning methods. This technique is based on the maximization of the margin, defining the region surrounding the hyperplane that best splits the different classes. The original optimal hyperplane method was a linear classifier, which is also used in this study. Additionally, the radial SVM was used. In 1992, Boser et al. suggested the application of the kernel trick as a solution to create non-linear classifiers, for example by using the Gaussian radial basis function. SVM was implemented using the e1071 package (Dimitriadou et al., 2011), with the cost and tolerance parameters of the linear SVM set to 100 and 0.01 and the cost and gamma parameters of the radial SVM fixed to 1000 and 0.1, respectively.

Integrative approach

As mentioned earlier, the data set, generated in cooperation with the associated physicians, consisted of MCC/IMS chromatograms that cover the health status of three groups of volunteers, i.e., HC, COPD and COPD+BC. These data were preprocessed utilizing the VisualNow software, followed by an expert-driven component detection. Based on the three classes, two different classification tasks were considered: COPD *vs* HC and COPD *vs* COPD+BC *vs* HC.

The accuracy of the different statistical learning techniques was evaluated in a 10-fold cross validation environment. In general, the set of samples was split into training and test sets. The test set is used to evaluate the models created by the training set. In settings where the data set is small, in our case restricted to 119 samples, this leads to relatively noisy estimates of predictive performance. Therefore, cross validation is used to give an estimate for the actual accuracy of the predictive model. The data set is split into k preferably equal-sized subsets (≈ #samples/k). Each of the subsets W_j, is evaluated by a model trained on the opposite set of samples excluding subset $W_{1,..,i-1,i+1,..,k}$ (Hastie et al., 2009). To ensure that each subset covers the variety of all classes, the classes are balanced within each subset, for the two-class as well as the three-class-problem.

To assess the information content within the breath data and to avoid overtraining the statistical learning methods, simpler methods as well as more sophisticated methods were applied without further tuning of the parameters. The R package pROC was used to compute various measures of prediction quality (Robin et al., 2011): sensitivity, specificity and the AUC, which is the area under the receiver operating characteristics (ROC) curve. In contrast to the other methods used in this study, random forest and linear support vector machine allow for judging the features' importance to the trained models' performance. In this study, we evaluated the result of the importance-measure of random forest (Gini index) and the weights fitted by the linear SVM model. Therefore, the vector of importances of each of the ten models resulting from the cross validation was extracted and the average value for each of the features was determined. The best ten features for each of the two methods will be discussed.

RESULTS AND DISCUSSION

COPD classification

The results of the COPD *vs* HC performance comparison of the six different meth-

ods is depicted in Table 1. The more simplistic methods, i.e., decision tree and naive Bayes, achieved an accuracy between 82 and 85% and an AUC of around 80%. The linear SVM performed slightly better with an AUC of 83% and an accuracy of around 87%. While the more sophisticated methods, i.e., neural net and radial SVM, gave an accuracy of 89% and AUCs of 86 and 87%, respectively. The best performing method of the classification, distinguishing between COPD patients, was random forest, which had the best prediction accuracy of 94% as well as high values for AUC (92%), sensitivity (98%) and specificity (86%). As expected, the more sophisticated methods performed best, having a relatively low bias, which means they do infer less than the simpler methods. On the other hand, the basic methods performed surprisingly well in terms of AUC and accuracy, which indicates that the data provided some information to distinguish the two classes. However, due to the unbalanced data set (COPD ≈70% *vs* HC ≈30%), one has to take a closer look at the sensitivity and specificity. While the sensitivity of both types of methods was good (between 87 and 98%), the specificity of the enhanced methods (80 to 85%) was in general higher than the specificity of the simplistic methods (71 to 74%).

Table 1. Results of the two-class-classification problem, evaluating the differences between chronic obstructive pulmonary disease and the healthy controls.

Method	AUC	Accuracy	Sensitivity	Specificity
Decision tree	81	85	91	71
Linear SVM	83	87	92	74
Naive Bayes	79	82	87	71
Neural net	86	89	93	80
Radial SVM	87	89	92	83
Random forest	92	94	98	86

AUC = area under the curve. SVM = support vector machine.

To the best of our knowledge, we present the first comprehensive study about the performance of state-of-the-art statistical learning tools for IMS-based metabolic profiling of COPD and bronchial carcinoma. A recent study presented by Westhoff et al. (2011) solely concentrated on rank sum tests and decision trees in VOC marker detection for COPD *vs* HC. However, the major issue of this study (also briefly discussed in Westhoff et al., 2011) was a lack of cross validation to avoid data overfitting. The best solution with the only classification method directly implemented in the VisualNow software, however, is based on the best splitting VOC (peak 98), given by the rank sum test, resulting in a comparatively good training accuracy of 91%.

Feature selection

In our classification, setting features directly corresponded to molecules in human exhaled air. Hence, we are interested in finding those VOCs/molecules/peaks/biomarkers/features that contribute the most to the classification performance. Therefore, the weights of the linear SVM and the Gini index were taken as a measure of importance.

Figure 3 shows the ten best features provided by the two methods linear SVM and random forest. The linear SVM feature importance depends on the weights of the variables according to their influence on the final linear model. The Gini index of the random forest

measures to what extend the variable improves the accuracy of the fit. Tables 2 and 3 show the two resulting subsets of features.

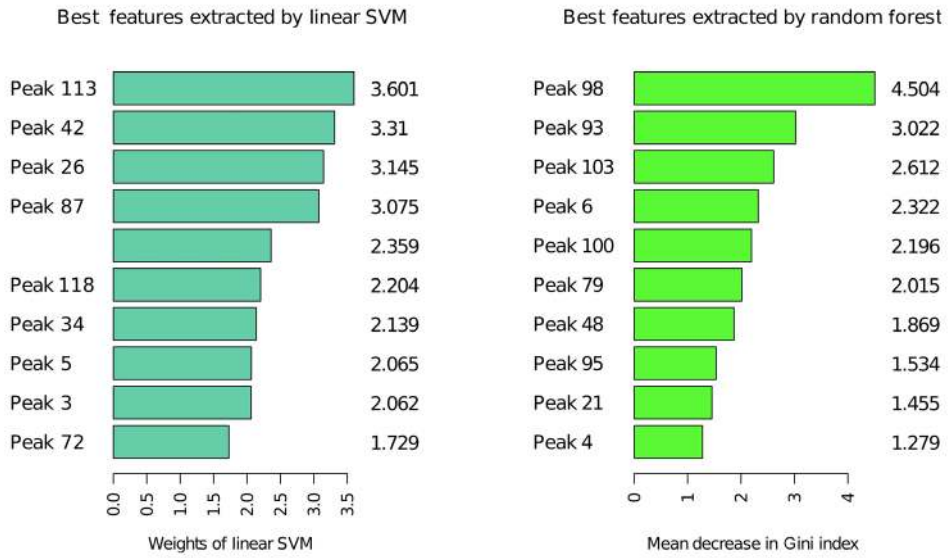

Figure 3. Result of the feature selection of the linear support vector machine (SVM) and the random forest, on the chronic obstructive pulmonary disease *vs* healthy controls classification. Depicted are the 10 best features according to their weights generated by the linear model (left) and the Gini index provided by the random forest (right). The right side of each Figure lists the names of peaks/volatile organic compounds.

Table 2. The 10 best features selected by linear support vector machine (SVM).

	Linear SVM selected peaks									
Peak No.	113	42	26	87	37	118	34	5	3	72
$1/K_0$	1.119	0.844	0.514	0.55	0.693	0.932	0.563	0.544	0.535	0.904
RT	189.8	501.1	248.1	56.5	100.7	110.4	6.2	3.4	4.3	233.4

The peaks are ordered by the rank of their weight in the linear model. Additionally, the coordinates in the multi-capillary column-ion mobility spectrometry chromatogram, the inverse drift time ($1/K_0$) and the retention time (RT) are shown.

Table 3. Comparison of the 10 best features selected by random forest to the peaks identified by the study of Westhoff et al. (2011) using rank sum test.

	Random forest selected peaks									
Peak No.	98	93	103	6	100	79	48	95	21	4
$1/K_0$	0.605	0.607	0.61	0.581	0.553	0.563	0.553	0.648	0.647	0.6
RT	22.3	16.9	18.9	78.5	49.6	20	17.9	17.6	21.9	11.9
Westhoff et al., 2011	X		X		X	X	X			

The peaks are ordered by the rank of the Gini index generated during the training of the random forest model. Additionally, the coordinates in the multi-capillary column-ion mobility spectrometry chromatogram, the inverse drift time ($1/K_0$) and the retention time (RT) are shown.

Both best feature subsets did not overlap. This resulted from the unequal underlying mathematical approaches, in one case a linear model, in the other a non-linear. Additionally, we compared both subsets to the peaks identified by the study of Westhoff et al. (2011) using the rank sum test. Interestingly, five of the compounds found in this study have also been found to be important by random forest. This indicates that further analysis of the compounds found by the variable selection performed is promising, especially for the peaks 48, 79, 98, 100, and 103. Another study by Bessa et al. (2011) analyzed a data set of 13 COPDs and 33 HCs, and reported a peak at position $1/K_0 = 0.50$ and RT = 26 to be the best discriminant for that classification purpose with 100% accuracy. However considering 120 VOCs and only 46 training and evaluation data sets, does not allow for any save conclusions at this point.

COPD+BC classification

Table 4 depicts the classification results of this three-class problem. Prediction quality was low (accuracy ≈ 70%) for each of the applied machine learning techniques, except random forest (accuracy ≈ 79%). The AUC dropped by at least ten percent for all of the methods except naive Bayes, which may be due to its simplicity and its robustness. Therefore, it remains unclear whether the data's information content is high enough for distinguishing the three groups of volunteers.

Table 4. Results of the three-class-classification problem, evaluating the differences between COPD patients, COPD patients suffering from bronchial carcinoma, and the control group.

Method	AUC	Accuracy	COPD		COPD+BC	
			Sensitivity	Specificity	Sensitivity	Specificity
Decision tree	70	60	23	82	69	65
Linear SVM	71	59	0	93	80	48
Naive Bayes	75	62	43	79	61	72
Neural net	73	61	20	82	69	62
Radial SVM	73	62	0	91	78	57
Random forest	79	67	6	99	85	55

The class-specific sensitivity and specificity assessed for class COPD, as well as COPD+BC, is based on the equations discussed in the Methods section. AUC = area under the curve. COPD = chronic obstructive pulmonary disease. COPD+BC = COPD + bronchial carcinoma. SVM = support vector machine.

In fact, all of the methods showed a very low sensitivity for the COPD class, which indicates that the differentiation between class COPD and COPD+BC is a difficult problem using all of the methods. While the methods were still able to identify the HCs in a quite robust manner, most of the measurements of COPD patients were falsely predicted to suffer from both COPD and bronchial carcinoma, i.e. class COPD+BC. This was due to the fact that the number of BC patients was almost double the number of patients solely suffering from COPD. This is not surprising, since the cause of both diseases is highly dependent on the smoking behavior of the patients, and both are reported to be strongly related to each other. This is also supported by Young et al. (2009), where they identified COPD as a common and important independent risk factor for lung cancer. The prevalence of the different groups of COPD within lung cancer goes up to 60% (Young et al., 2009). Consequently, we can assume that the probability for each of the COPD patients to get a bronchial carcinoma is comparatively high.

Hence, we cannot exclude the possibility that an early stage bronchial carcinoma might have been undetected, particularly since most types of lung cancer are detected in late stages in the majority of cases.

Although there has been no study using modern machine learning methods on COPD MCC/IMS chromatograms, there were two studies for MCC/IMS data about BC patients. One study applied naive Bayes, multi layer perceptron, and SVM to a set of MCC/IMS chromatograms and achieved an outstanding performance (accuracy and AUC both 99%) (Baumbach et al., 2007). Despite the good results, one has to consider that 1) the prediction was done on a comparatively large feature set, where each chromatogram was separated by a grid, while each feature was calculated as the average intensity of the corresponding grid element, and 2) the accuracy and AUC were evaluated on the training set, without cross validation, as in the study of Westhoff et al. (2011). Another study used relational probabilistic learning for the extraction of Markov logic network formulas and achieved a cross validation accuracy of up to 90% during classification (Finthammer et al., 2010). Unfortunately, a comparison with the results of the peak selection of those two methods is difficult, as the authors did not make the peak or grid positions publicly available. Besides, no further information on additional medical conditions of the patients, such as COPD, was known.

Another problem for the analysis of COPD measurements is the medical COPD categorization itself. The patients were categorized with respect to the severity of their COPD disease, which is done according to a defined set of rules and tests. In addition, clinical practice and the subjective impression of the physician play an important role, as well as other factors that can influence the compounds in human breath, such as diet, smoking, and other secondary diseases, that have not necessarily been tested.

CONCLUSION

Ion mobility spectrometry data of human breath can generally be utilized for distinguishing between lung diseases if used properly with statistical learning environments. To demonstrate this, we evaluated sophisticated machine learning techniques on MCC/IMS chromatograms regarding their classification performance and ability to identify the most important molecular compounds. Therefore, the breath of 84 patients either suffering from COPD or both COPD and bronchial carcinoma was processed and compared with 35 healthy volunteers. The by far best test error estimates (AUC 91%, ACC 94% for COPD *vs* HC; AUC 79%, ACC 67% for COPD *vs* COPD+BC *vs* HC) were achieved with the random forest method. These results pinpoint a strong potential to separate healthy from COPD, but also suggest that a further examination of the differences between COPD and COPD+BC is needed. Linear SVM and random forest extracted 20 important VOCs, while five of them were confirmed to have discriminative properties regarding COPD and healthy IMS chromatograms in other studies.

In the future, we plan to determine whether these 20 molecules are biologically related to COPD and to eliminate those that are related to diet or other environmental influences. A further objective is to optimize and enhance the standardized learning methods for enhancing prediction performance, on the one hand, and the identification of the smallest discriminating set of biomarkers, on the other hand. This will be a tremendous progress in the field of COPD and cancer diagnostics. However, larger COPD+BC data sets are necessary here.

ACKNOWLEDGMENTS

We thank the anonymous data sets of IMS chromatograms related to the three groups as obtained by Dr. Michael Westhoff, Dr. Patrick Litterst and Barbara Obertrifter, all from the Lung Hospital Hemer. J. Baumbach is grateful for the financial support from the Cluster of Excellence for Multimodel Computing and Interaction of the German Research Foundation, Saarbrücken, Germany.

REFERENCES

Bader S, Urfer W and Baumbach JI (2008). Preprocessing of ion mobility spectra by lognormal detailing and wavelet transform. *Int. J. Ion Mobility Spectrom.* 11: 43-49.

Baumbach J, Bunkowski A, Lange S, Oberwahrenbrock T, et al. (2007). IMS2 - An integrated medical software system for early lung cancer detection using ion mobility spectrometry data of human breath. *J. Integr. Bioinform.* 4: 75.

Baumbach JI (2009). Ion mobility spectrometry coupled with multi-capillary columns for metabolic profiling of human breath. *J. Breath Res.* 3: 1-16.

Bessa V, Darwiche K, Teschler H, Sommerwerck U, et al. (2011). Detection of volatile organic compounds (VOCs) in exhaled breath of patients with chronic obstructive pulmonary disease (COPD) by ion mobility spectrometry. *Int. J. Ion Mobility Spectrom.* 14: 7-13.

Boser BE, Guyon IM and Vapnik VN (1992). A Training Algorithm for Optimal Margin Classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM Press, New York, 144-152.

Dimitriadou E, Hornik K, Leisch F, Meyer D, et al. (2011). Misc Functions of the Department of Statistics (e1071), TU Wien Version 1.5-27. Available at [http://cran.r-project.org/web/packages/e1071/index.html]. Accessed September, 2011

Finthammer M, Beierle C, Fisseler J, Kern-Isberner G, et al. (2010). Probabilistic relational learning for medical diagnosis based on ion mobility spectrometry. *Int. J. Ion Mobility Spectrom.* 80: 365-375.

Hastie T, Tibshirani R and Friedman JH (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd end. Springer, New York.

Langley P and Sage S (1994). Induction of Selective Bayesian Classifiers. Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, Seattle, 399-406.

Liaw A and Wiener M (2002). Classification and regression by random forest. *R News* 2/3: 18-22.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Available at [http://www.r-project.org]. Accessed April, 2011.

Rabe KF, Hurd S, Anzueto A, Barnes PJ, et al. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am. J. Respir. Crit. Care Med.* 176: 532-555.

Robin X, Turck N, Hainard A, Tiberti N, et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12: 77.

Therneau TM and Atkinson B (2011). rpart: Recursive Partitioning. R package version 3.1-50.

Venables WN and Ripley BD (2002). Modern Applied Statistics with S. 4th edn. Springer.

Weihs C, Ligges U, Luebke K and Raabe N (2005). klaR Analyzing German Business Cycles. In: Data Analysis and Decision Support (Baier D, Decker R and Schmidt-Thieme L, eds.). Springer-Verlag, Berlin, 335-343.

Westhoff M, Litterst P, Maddula S, Bödeker B, et al. (2011). Statistical and bioinformatical methods to differentiate chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control by breath analysis using ion mobility spectrometry. *Int. J. Ion Mobility Spectrom.* 14: 139-149.

World Health Organization (2008). The Global Burden of Disease, 2004 Update. Available at [http://www.who.int/healthinfo/global_burden_disease/en/]. Accessed December 2011.

Young RP, Hopkins RJ, Christmas T, Black PN, et al. (2009). COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. *Eur. Respir. J.* 34: 380-386.