# IMS$^2$ – An integrated medical software system for early lung cancer detection using ion mobility spectrometry data of human breath

**Jan Baumbach[1,2], Alexander Bunkowski[1], Sita Lange[1,3], Timm Oberwahrenbrock[1], Nils Kleinbölting[1,3], Sven Rahmann[1], Jörg Ingo Baumbach[4]**

[1] Computational Methods for Emerging Technologies, Genome Informatics, Technische Fakultät, Bielefeld University, 33594 Bielefeld, Germany
[2] International NRW Graduate School in Bioinformatics and Genome Research, Bielefeld University
[3] Bioinformatics Resource Facility, Bielefeld University
[4] Department of Metabolomics, ISAS - Institute for Analytical Sciences, Bunsen-Kirchhoff-Str. 11, 44139 Dortmund, Germany

**Abstract**

IMS$^2$ is an Integrated Medical Software system for the analysis of Ion Mobility Spectrometry (IMS) data. It assists medical staff with the following IMS data processing steps: acquisition, visualization, classification, and annotation. IMS$^2$ provides data analysis and interpretation features on the one hand, and also helps to improve the classification by increasing the number of the pre-classified datasets on the other hand. It is designed to facilitate early detection of lung cancer, one of the most common cancer types with one million deaths each year around the world.
After reviewing the IMS technology, we first describe the software architecture of IMS$^2$ and then the integrated classification module, including necessary pre-processing steps and different classification methods. The Lung Hospital Hemer (Germany) provided IMS data of 35 patients suffering from lung cancer and 72 samples of healthy persons. IMS$^2$ correctly classifies 99% of the samples, evaluated using 10-fold cross-validation.

## 1 Introduction

Lung cancer is the most common cancer type in men (fourth in women), with ca. 200 000 new cases and ca. 140 000 deaths each year in the European Union. The 5-year survival rate is approx. 10% for both sexes [10]. The National Cancer Institute of the United States estimates ca. 213,000 new cases and ca. 160,000 deaths solely for 2007 in the USA. It is further estimated that approximately 9.6 billion US dollars are spent in the United States on the treatment of lung cancer. Nowadays, the screening of blood and urine as invasive standard methods are applied to potentially diseased patients. Especially for the identification of lung cancer, chest X-ray, sputum cytology, and spiral computer tomography scans are used. The patient's chance to survive is relatively low compared to other types of cancer, partly because of the usually very late detection of the disease. An early identification of lung cancer would considerably increase the chance for recovery.

It is well known in the medical community that human exhaled air contains volatile metabolites that potentially carry information on the health status of the human organism. Hence, a sensitive

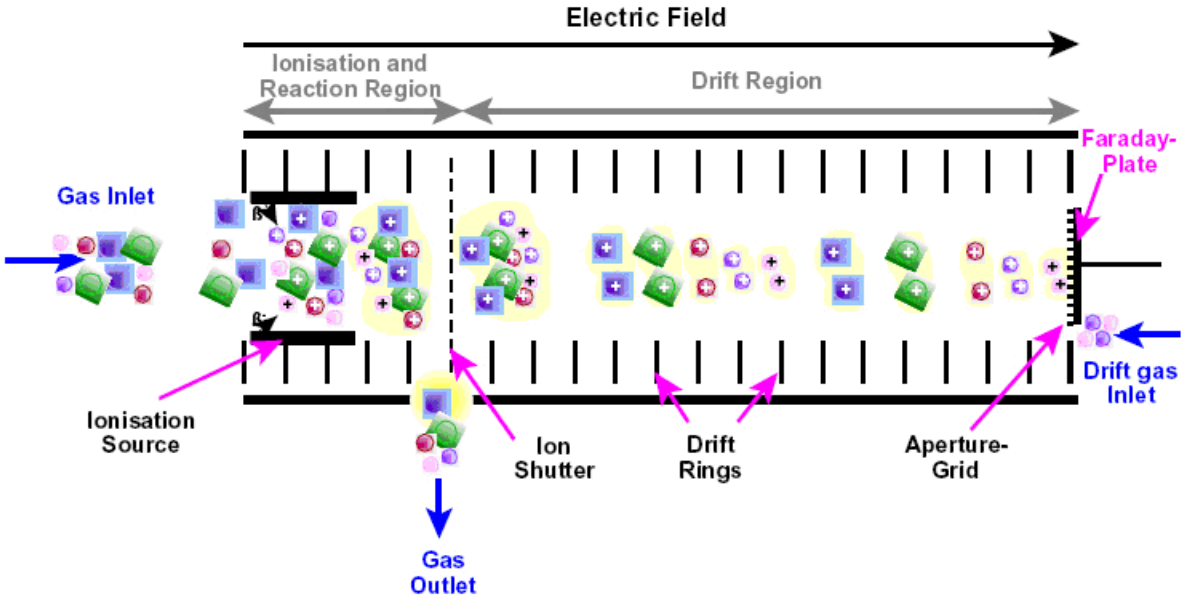


**Figure 1: Schematic overview of the working principle of an ion mobility spectrometer.**

metabolic profiling of these molecules can provide essential data for the early classification of lung diseases. The application of spectrometric methods, such as mass spectrometry (MS) and ion mobility spectrometry (IMS), allows the identification and quantification of molecules in gases. To detect very small concentrations of volatile metabolites, the detection limit has to be very low (down to the $\text{ppt}_V$, pg/L ranges). Nowadays, the most common approaches utilize MS techniques [3], but these instruments are large and very expensive. If spectrometric methods are to become widely established beside other clinical tests in hospitals and point-of-care centers, the instruments have to be small, easy to use, and the price has to be moderate. For these reasons, the application of miniaturized ion mobility spectrometers is an appropriate, fast, low-cost, and non-invasive method, which has recently been tested in several clinical studies [1, 9, 8, 11].

**IMS Technology.** The working principle of IMS is described in detail in [2]. Thus, we just give a brief summary here.

IMS is based on an appropriate ionization of gaseous analytes and a subsequent separation of the emerging positive and negative ions at ambient air temperature and pressure. Figure 1 illustrates the main working principle. Ion swarms formed within the ionization region enter the drift tube for very short shutter opening times (a few microseconds) and are separated in an electrical field. A drift gas flows towards the ions. The drift velocity $v$ of the ions is related to the electrical field strength $E$ and the mobility $k$ by $v = kE$. Hence, the drift time $t$, which is measured at a fixed drift length $l$ is inverse proportional to the mobility $k$. The mobility depends on the collision rate of the swarm ions with the drift gas molecules, the temperature, the ion structure, and the collision integral. The collision integral is related to the ions' size, structure, and polarisibilities. The measurement of $t$ is performed by means of a Faraday plate, whose charge variation over time is called the ion mobility spectrum. In contrast to mass spectrometry, the drift tube has ambient pressure. Hence, beside the ions masses, also the collisions with neutral molecules influence the drift time. Compared to MS, IMS cannot

be used for the identification of unknown molecules, but the method is much more sensitive (ng to pg, $ppm_V$ to $ppt_V$), especially when using humid air (as in human breath) and when the sample is handled directly without any pre-enrichment. Both MS and IMS instruments are often coupled with gas chromatographic (GC) columns for fast pre-separation: GC/MS or MCC/IMS (MCC = multi capillary column).

Many other spectroscopic methods are cost intensive, time consuming, error prone, and need well qualified staff. In contrast, recently the development of IMS technology has provided a sensitive screening in spite of comparatively low costs. IMS works with ambient pressure, ambient air, and within milliseconds (ca. 50 ms). A miniaturized IMS instrument is available for less than 30,000 US dollars, which is inexpensive in comparison to similar technologies.

**Our contributions.** We contribute an **I**ntegrated **M**edical **S**oftware system for the analysis of **IMS** data (**IMS²**). IMS² is developed to assist the medical practitioner with the following data processing steps:

1. Acquisition,
2. Visualization,
3. Annotation,
4. Classification,
5. Automatic improvement of classification results.

The last point refers to the aim of an automatic improvement of pre-classified samples, resulting in a better classification over time.

We present the software architecture along with the used libraries and data processing pipelines. Afterwards, we introduce and discuss the integrated classification module and evaluate the system on a dataset provided by the Lung Hospital Hemer (Germany). A discussion concludes the paper.

# 2 Methods

## 2.1 Software architecture

The IMS² software has two main goals:

- The visualization and enhancement of the measurement
- An automatic improvement of the classification

The first is to present a clear and interpretable view on the measurement taken, which is necessary to control the quality and to ensure that no problems occurred during the data generation.

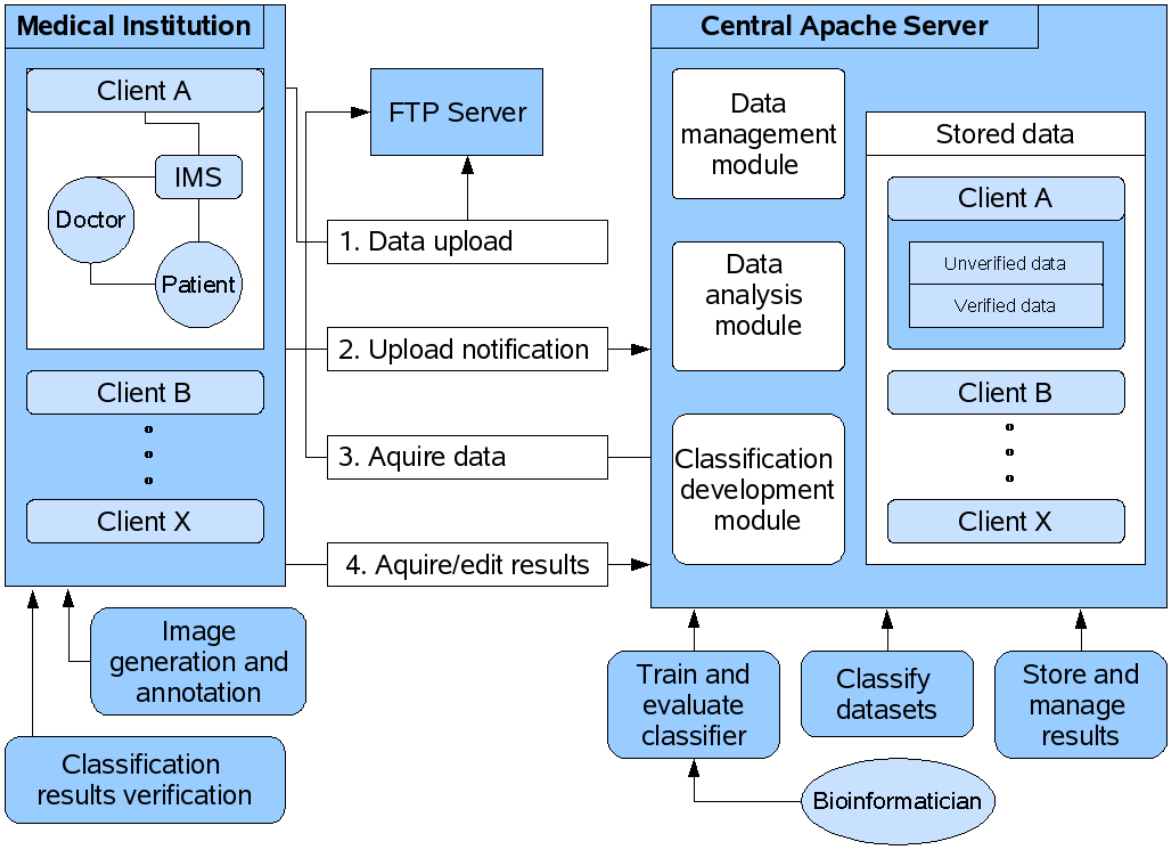


**Figure 2: Schematic overview of the IMS$^2$ software architecture and the data flow.**

This is achieved by applying normalization, histogram spreading, and filtering on the visualized image. After this process, there are possibilities to add previous knowledge based on the experience of the medical personnel. This could be a general comment or an annotation of a particular region of interest. This information could later help to determine why a measurement was wrongly classified. After the visualizing and commenting step, the dataset is sent to a central server which processes and classifies it (see Figure 2).

Multiple clients can upload their data and classification requests at any time. The data is first sent to the FTP server, after which the client calls a PHP script on the Apache server. This notifies the data management module that new data is available. Thereupon the client data is transferred to the central server and stored as being unverified. The connection between FTP and the Apache server is fastest when running on the same hardware.

The classification of the data is done by the data analysis module. For this purpose, a previously built classifier, which can be generated by the classification development module, is used. The classification development module provides methods for a competent user to train and evaluate classifiers using the verified data.

When the classification is complete, the client can access the classification results, which remain on the central server. As soon as new information about a measurement, or more precisely about the associated patient is available, the doctor is asked to comment and verify the classification result. This verified data is added to the training set, which can be used to generate a

new and perhaps improved classifier. This leads to the second main goal of IMS$^2$, an automatic enhancement of the classification by extending the size of pre-classified sample.

All data transfer is done encrypted and no patient information is sent over the internet. A unique ID is generated, which can not be traced back to the patient's identity, because the IMS$^2$ system is physically unconnected to the hospital's patient administration network.

We use PHP version 4.3.2, Apache 2.0.49, Java 1.6.0, and WEKA 3.

## 2.2 Classification of MCC/IMS data

MCC/IMS data can be seen as an image where the peak intensities correspond to different color values. Hence, methods successfully applied to image processing tasks can also be used for the classification of this data with respect to some specialities. Therefore we apply several pre-processing steps to the data, followed by feature selection and classification. Since there is no need to reinvent the wheel, we use methods available in WEKA [12], a public data mining library in Java for feature selection and classification.

### 2.2.1 Pre-processing

First we use a standard two-dimensional Gaussian filter implemented in WEKA to reduce the effect of background noise. We choose a standard deviation of $\sigma$=1.5 and a kernel size of 5.

Secondly, we detect the position of the RIP (Rest Ion Peak), which exists in every MCC/IMS image as a continuous band from top to bottom. To this end, we use the fact that the RIP intensities exceed every other peak. For each fixed y-coordinate, we calculate from right to left the slope between each pair of consecutive points. If the slope exceeds an empirically determined value, the right point is assigned to the set of points that define the border between the RIP and the right (important) part of the image. The mean value of all x-coordinates in this set is used as a cutoff to exclude the left part of the image, which contains only irrelevant information for the classification. Finally, we linearly normalize all remaining values to the interval $[0, 1]$.

### 2.2.2 Feature selection

We compress the data by laying a grid over the relevant part of the image and calculate the average intensity value within one grid element. For each classification process, we iterate over the grid size to determine which one is optimal. In this case, each grid element with its respective intensity value is treated as one feature. By means of these attributes we attempt to classify the given data. Since using the entire feature set results in very large problems, we select significant features using the built-in methods in WEKA and compare both approaches for different classifiers. For a successful feature selection, we need to combine an attribute evaluator with a search method. Most methods for attribute selection search for the subset that makes the best predictions as to which class the instance belongs to. We choose two methods for this purpose which empirically achieve satisfying results on our data. Further information on the process of feature selection and the methods we use here can be found in [12, pages 288 and 420].

**Best-First Search Method.** The best-first method performs a greedy hill climbing with backtracking. We use the class 'weka.attributeSelection.BestFirst' with the parameters '-D 1 -N 5'. Parameter 'D' indicates the direction in which is searched, for our case we search in the forward direction and start with the empty set. It is also possible to scan backwards from the full set, or start at an intermediate point and look in both directions. Parameter 'N' denotes how many consecutive non-improving nodes must be encountered before the system backtracks.

**CFS Subset Evaluator.** This evaluator individually assesses the predictive ability of each attribute and also evaluates the degree of redundancy between them. It assigns a high significance to attributes with a high correlation with the class and with a low intercorrelation. We use the WEKA class 'weka.attributeSelection.CfsSubsetEval'.

### 2.2.3 Classification

We compare the classification performance of the Naive Bayes (NB) classifier, MultiLayer Perceptrons (MLPs), and the Support Vector Machine (SVM). All of them have been previously used for data mining with GC/MS data. Hence, they may be applicable to MCC/IMS data as well. In the following, we very briefly describe the WEKA methods we use, along with the chosen parameters.

**Naive Bayes classifier.** The NB classifier is a simple probabilistic method based on Bayes' theorem. It assumes independent variables, which usually does not reflect reality. However, it requires just a small amount of data to estimate the necessary means and variances of the variables; only the variances of the variables for each class need to be determined and not the entire covariance matrix [7].

We use the class 'weka.classifiers.bayes.NaiveBayes'. Further information on the implementation of the NB classifier in WEKA can be found in [12, p.403].

**Multi Layer Perceptron.** A MLP is an interconnected group of artificial neurons (neural network) consisting of multiple layers of interconnected computational units. Each neuron in one layer has directed connections to the neurons of the subsequent layer with adjustible weights. The sum of input weights of a neuron is usually transformed by a sigmoid activation function and then passed on to the next layer. Using the training data, the algorithm iteratively readjusts the weights of the connections between the neurons (using back-propagation) in order to minimize the prediction error. The network usually converges to some state where the error is small; thus the networks learns a target function.

We use the class 'weka.classifiers.functions.MultilayerPerceptron' with the default options '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a'. Further information on the implementation of MLPs in WEKA can be found in [12, p.223].

**Support Vector Machine.** We use the SMO feature of WEKA: an implementation of the sequential minimal optimization algorithm for training a support vector classifier [6, 4]. The basic idea is to find a function that approximates the training data by minimizing the prediction error. The main difference compared to linear regression methods is that all deviations up to a user-defined threshold are discarded, so that the threshold defines a tube around the function. The risk of overfitting is reduced by trying to maximize the flatness of the regression function simultaneously.

We use the class ‘weka.classifiers.functions.SMO‘ with the default options ‘-C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0‘. Further Information on the implementation of support vector regression in WEKA can be found in [12, p.219].

# 3 Evaluation Results

## 3.1 The Dataset

We obtained data on exhaled air of lung cancer patients provided by the Lung Hospital in Hemer (Germany) that specializes in lung diseases. The data is expertly pre-classified and split into two groups: 35 MCC/IMS sets (called IMS-chromatograms) of patients suffering lung cancer and 72 samples of healthy patients as the control group. Furthermore, all test patients have not been allowed to drink, eat, or smoke within two hours before the experiments.

All measurements of lung cancer patients exhaled breath and of the control group are performed with an ISAS home made $^{63}$Ni $\beta$-ionization source IMS. Table 1 summarizes the main parameters of the IMS used in this study.

One IMS chromatogram of the exhaled breath of a lung cancer patient is exemplarily shown in Figure 3 and visualized using IMS$^2$. The colors refer to different peak heights. To give a clear and comparable view on the measurement, the medical practitioner can use image processing methods, such as normalization to values between 0 and 1, and inverting in the case of reverse measurements. To visualize and define regions of special interest, the user can additionally calculate a histogram using user-defined minimal and maximal values for spreading. These values can also be determined automatically but for a visual comparison of different measurements it could be of advantage to use fixed values. Moreover, it is possible to tag regions of interest manually and to use a peak detection algorithm as well as a Gaussian filter to decrease background noise. The ’Received Results’ tab panel shows the results gathered from the central (classification) server.

| Parameter | Value |
|---|---|
| Ionization source | $^{63}$Ni $\beta$-radiation source, 510 MBq |
| Drift region length | 12 cm |
| Electrical field strength | 330 V/cm |
| Drift voltage | 4 kV |
| Shutter opening time | 300 $\mu$s |
| Drift gas | synthetic air |
| Drift gas flow | 100 mL/min |
| Sample gas flow | 150 ml/min |
| Temperature | ca. 25°C (ambient) |
| Pressure | 100 kPa (ambient) |

**Table 1: Main parameters of $^{63}$Ni Ion mobility spectrometer.**

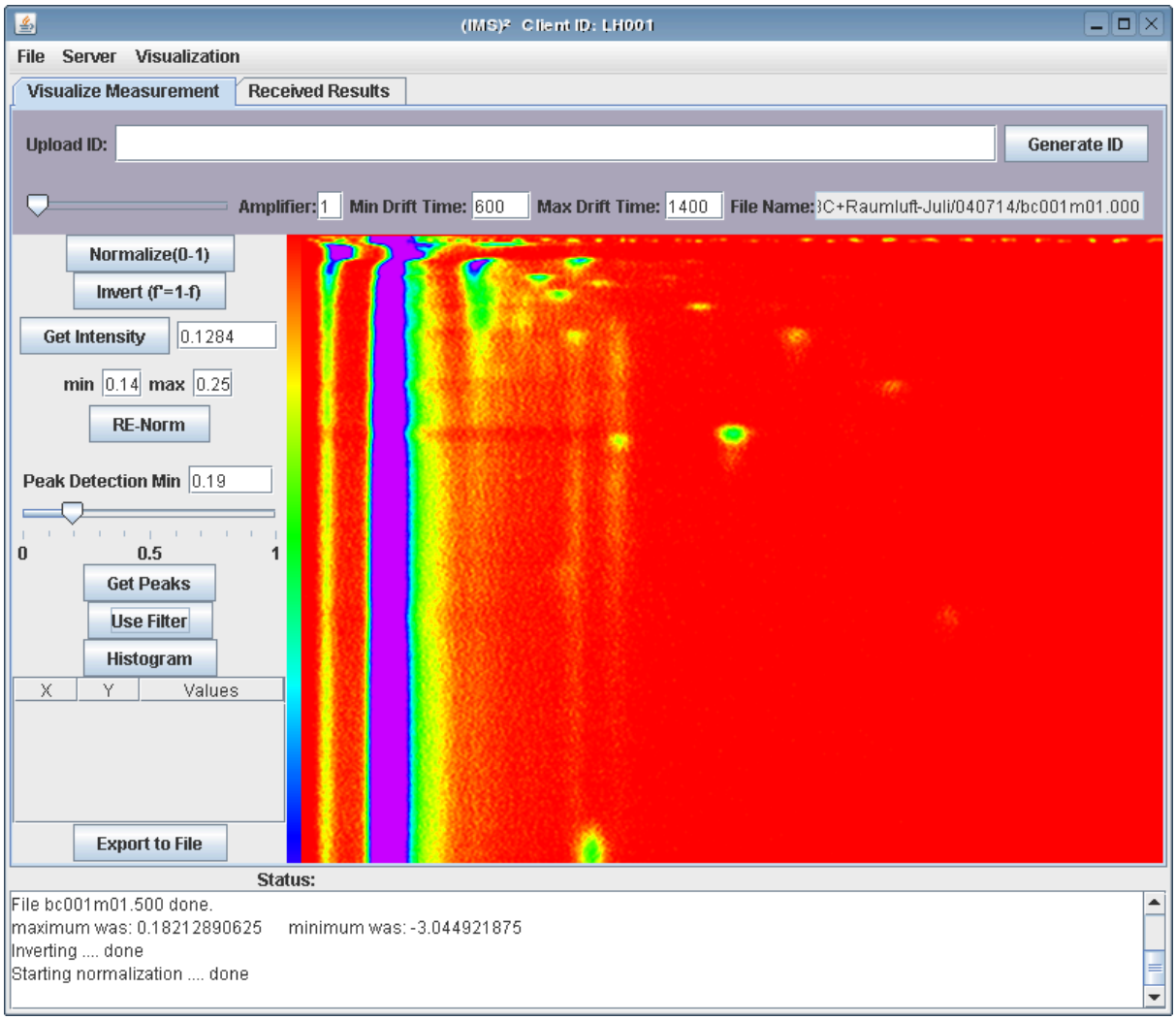


**Figure 3: Screenshot of the IMS$^2$ client software with a sample lung cancer MCC/IMS result file visualized. The continuous band on the left side of the visualization is the so called RIP (Rest Ion Peak), which is present in every MCC/IMS image.**

## 3.2 Classification Results

The classification performance of the classifiers is evaluated by means of 10-fold cross-validation [5], as implemented in WEKA. We report the percentage of the correctly classified MCC/IMS spectra against the size of the grids used for data reduction and filtering. We test all classifiers with and without prior feature selection (FS) for grid sizes between 3 and 150 pixels (px). Due to the amount of features in the MCC/IMS spectra, we could not perform training on MLPs without prior data reduction using FS for grid sizes $\leq$ 25 px.

Figure 4 shows the classification performances for grid sizes between 3 and 40 px in increments of one and between 40 and 60 px in increments of ten. For grid sizes between 60 and 150 px significantly worse results are achieved. Table 2 shows the best classifiers. Both MLPs and SVM provide 99.07% accuracy with very low error rates for both classes. Interestingly, the classification results do not depend too much on the chosen grid size but more on the used method. Hence, we suggest to use MLPs or SVM for the classification of lung cancer from MCC/IMS data using a grid size of approximately 23 px.

| Classifier | FS | Grid size | Correct (%) | ROC | ErrLC (%) | ErrHl (%) |
|---|---|---|---|---|---|---|
| MLP | yes | 28 | 99.07 | 0.986 | 0 | 2.9 |
| MLP | yes | 27 | 99.07 | 0.986 | 0 | 2.9 |
| MLP | yes | 18 | 99.07 | 0.986 | 1.4 | 0 |
| SVM | no | 27 | 99.07 | 0.986 | 0 | 2.9 |
| SVM | no | 24 | 99.07 | 0.986 | 0 | 2.9 |
| SVM | no | 23 | 99.07 | 0.986 | 0 | 2.9 |
| SVM | no | 21 | 99.07 | 0.986 | 0 | 2.9 |
| SVM | no | 18 | 99.07 | 0.986 | 0 | 2.9 |
| NB | yes | 24 | 98.13 | 0.997 | 1.4 | 2.9 |
| MLP | no | 70 | 98.13 | 0.943 | 0 | 5.7 |
| SVM | yes | 28 | 97.2 | 0.914 | 0 | 8.6 |
| NB | no | 40 | 93.46 | 0.857 | 2.8 | 14.3 |

**Table 2: The best classifiers achieve a correct classification of 99.07 % into lung cancer or healthy patients, which is measured using 10-fold cross-validation. The last 4 lines correspond to the best results for the other classifier/feature selection combinations. MLPs and SVM perform equally and in the same grid size range. FS = feature selection, Correct = percentage of correctly classified samples, ROC = area under the ROC curve, ErrLC = healthy patients falsely classified as lung cancer, ErrHl = lung cancer patients falsely classified as healthy.**

## 4 Discussion and Conclusion

This paper introduced IMS$^2$, a semi-automatic medical assistance system, which can easily be used by medical staff for the detection of lung cancer using human breath air measured with MCC/IMS instruments. A software client provides visualization and annotation features. A central server is used for all space- and time-consuming calculations: data collection, training, and classification.

Using the Java-based data mining library WEKA we demonstrated that MCC/IMS data can be used for the automatic separation of healthy and lung cancer patients. The success of the method depends (1) on the existence of small metabolites present in breath samples that allow to distinguish between lung cancer and healthy patients, and (2) on the amount of available pre-classified sample data for training.

In order not to raise false medical expectations, a few words of caution are in order. First, the machine learning techniques implemented in IMS$^2$ at present do not allow the direct identification of the distinguishing metabolites, even if one or two responsible peaks can be clearly identified. Additional wet lab experiments would be required to isolate and identify the molecules. Second, it should be noted that the sample used in this study was collected under well-controlled conditions. While the classification results are very encouraging, this study does not imply that its results are directly transferable to clinical practice.

In fact, we want to stress that the classification procedure used is just a module of the IMS$^2$ system, which by itself is an important step towards streamlined data collection and integration: IMS$^2$ provides an easy-to-use front-end, which assists the medical staff with the data analysis and interpretation on the one hand but also motivates them to help with the enhancement of the pre-classified datasets on the other hand. The additional time investment is low when compared

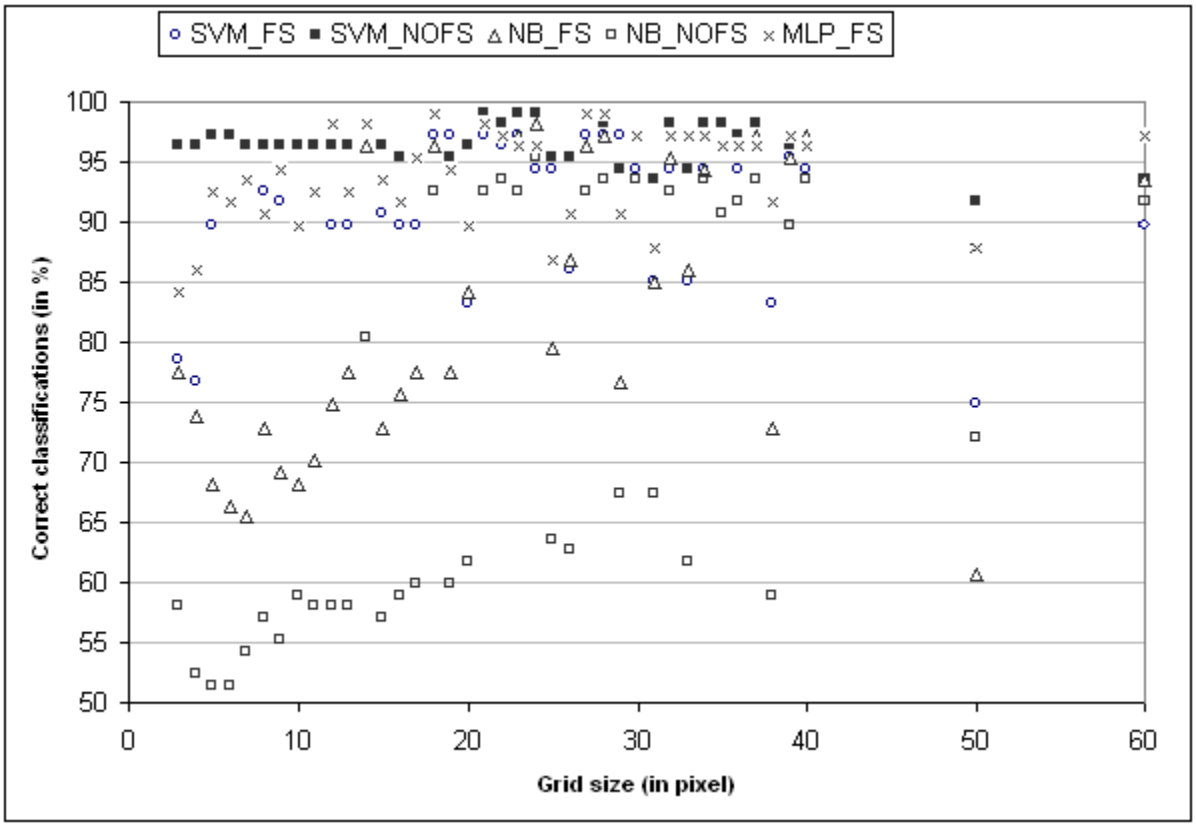


**Figure 4: Comparison of classification results for different grid sizes for the following methods: Support Vector Machine (SVM), Naive Bayes Classifier (NB) and Multi Layer Perceptron (MLP). SVM and NB have been evaluated with and without prior feature selection (FS and NOFS respectively). For runtime reasons, MLPs without FS have only been tested for grid sizes $\geq$ 25 px. The picture shows that the classification performance is best on grids sizes between 18 and 28 px using MLPs or SVM.**

to that of manually interpreting the results. The more data that has been analyzed with IMS$^2$, the better the results will be in future.

The described techniques may also be used to screen metabolites for the detection of other (lung) diseases and airway infections [8], as long as enough data is available for training.

In the future, we plan to establish and extend the system with more data and additional analysis methods. Furthermore, we plan to extend IMS$^2$ to be applicable to other pulmonary diseases and infections.

## 5 Availability and Acknowledgments

The software package IMS$^2$ along with installation instructions and the used evaluation dataset are available upon request (`Jan.Baumbach@CeBiTec.Uni-Bielefeld.DE`).

We thank Dr. Michael Westhoff, Dr. Patric Literst, Dr. Lutz Freitag and Barbara Oberdrifter of the Lung Hospital Hemer (Germany) for providing the lung cancer test datasets and their continuous support of the experiments and studies, Sabine Bader from ISAS for helpful discussions and for her introduction into the field of data handling using IMS, Stefanie Güssgen,

Lucia Seifert, and Dr. Wolfgang Vautz (all ISAS) for their major contributions to experimental and laboratory work, and Dr. Alexander Goesmann and Ralf Nolte (both Center for Biotechnology, Bielefeld University) for expert technical support. In particular we express our thanks to Dr. Vera Ruzsanyi at ISAS for providing data and results obtained during the work of her PhD thesis. We thank Tobias Wittkop from CeBiTec, Bielefeld University for proof reading and helpful discussions. The financial support by the Ministerium für Innovation, Wissenschaft, Forschung und Technologies des Landes Nordrhein-Westfalen and the Bundesministerium für Bildung und Forschung is gratefully acknowledged. JIBB wish to thank the European Union for funding the project BAMOD: Breath-gas analysis for molecular-oriented detection of minimal diseases (Contract: LSHC-CT-2005-019031).

## References

[1] J. I. Baumbach. Process analysis using ion mobility spectrometry. *Anal Bioanal Chem*, 384(5):1059–1070, Mar 2006.

[2] J. I. Baumbach and G. A. Eiceman. Ion mobility spectrometry: arriving on site and moving beyond a low profile. *Appl Spectrosc*, 53(9):338A–355A, Sep 1999.

[3] C. Baumgartner and A. Graber. Data mining and knowledge discovery in metabolomics. In F. Masseglia, P. Poncelet, and M. Teisseire, editors, *Successes and new directions in data mining*. 2007.

[4] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. Improvements to platt's SMO algorithm for SVM classifier design. *Neural Comp.*, 13(3):637–649, 2001.

[5] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2 (12):1137–1143, 1995.

[6] J. Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods: Support vector learning. Cambridge, MA: MIT Press*, 1998.

[7] I. Rish. An empirical study of the naive bayes classifier. *Proceddings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 2001.

[8] V. Ruzsanyi, J. I. Baumbach, S. Sielemann, P. Litterst, M. Westhoff, and L. Freitag. Detection of human metabolites using multi-capillary columns coupled to ion mobility spectrometers. *J Chromatogr A*, 1084(1-2):145–151, Aug 2005.

[9] V. Ruzsanyi, S. Sielemann, and J. Baumbach. Determination of vocs in human breath using IMS. *Int. J. Ion Mobility Spectrom.*, 5:45–48, 2002.

[10] M. Sant, T. Aareleid, F. Berrino, M. B. Lasota, P. M. Carli, J. Faivre, P. Grosclaude, G. Hedelin, T. Matsuda, H. Moeller, T. Moeller, A. Verdecchia, R. Capocaccia, G. Gatta, A. Micheli, M. Santaquilani, P. Roazzi, D. Lisi, and E. U. R. O. C. A. R. E. W. Group. Eurocare-3: survival of cancer patients diagnosed 1990-94–results and commentary. *Ann Oncol*, 14 Suppl 5:v61–118, 2003.

[11] M. Westhoff, P. Litterst, L. Freitag, V. Ruzsanyi, S. Bader, W. Urfer, and J. I. Baumbach. Ion Mobility Spectrometry: A new method for the detection of lung cancer and airway infection in exhaled air? First results of a pilot study. *Chest*, 128(4):155S, 2005.

[12] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition edition, 2005.