MEDICAL APPLICATIONS of Laser Molecular Imaging and Machine Learning

MEDICAL APPLICATIONS of Laser Molecular Imaging and Machine Learning

Yury V. Kistenev
Alexey V. Borisov
Denis A. Vrazhnov

SPIE PRESS
Bellingham, Washington USA

Library of Congress Control Number: 2021938748

Published by
SPIE
P.O. Box 10
Bellingham, Washington 98227-0010 USA
Phone: +1 360.676.3290
Fax: +1 360.647.1445
Email: books@spie.org
Web: http://spie.org

Printed in the United States of America.
First Printing.
For updates to this book, visit http://spie.org and type “PM333” in the search field.

Table of Contents

Preface

This book examines various biophotonics applications associated with modern machine learning techniques and laser molecular imaging and spectroscopy. Most of the existing books focus on either a specific instrumental method, such as terahertz and IR spectroscopy or Raman scattering, or a limited number of mathematical tools for raw data analysis. We describe a thorough review of molecular imaging technologies and current machine learning approaches to perform data analysis of gaseous, liquid samples of biological origin and biotissues. Much of the material highlights applications of machine learning to develop non-invasive medical diagnostics tools.

Here, we present the basics of machine learning methods, which consider the specificity of laser molecular imaging and spectroscopy medical data features, such as the high dimensionality of raw data and a low number of samples leading to a lack of representation. Modern trends such as deep learning are not applied broadly in similar tasks because of the small volume of available samples. There are two main reasons for this. The first is the high variability of biological systems, which makes biophysical relations difficult to discover. The second is ethical restrictions on studies with living beings. These reasons require new methods to deal with high-dimensional but low-numbered data (contrary to big data, which operates with low-dimensional yet outnumbered data).

Speaking of the development of both biophotonics hardware and software, we try to make future forecasts based on current trends in these fields. We also discuss available hardware platforms: home and self, medical screening, and specialized devices for end-level diagnosis. General trends include personalized medicine and bringing high-tech diagnostics from hospitals directly to individuals.

This book focuses on the most suitable approaches for medical screening and monitoring. Some ideas can be used in personal diagnosis tool design and production. Machine learning pipeline algorithms can be useful for high-accuracy multi-modal diagnosis.

This book is intended for specialists in the fields of biomedical optics, laser spectroscopy, bioengineering, and medical engineering. To provide practical

help to readers who plan to use the machine learning methods in their research, the Supplemental Materials include sample datasets and the Python modules for the most useful algorithms described in the book. The link to the Supplemental Materials website is

https://github.com/biophotonics-lab-tsu/monograph

For convenience, we use Roman superscript numbers to link key terms and specific methods with the chapter that defines them. Also, the first mention of a method is italicized.

Acknowledgments

The authors would like to thank their colleagues from the Laboratory of Biophotonics of Tomsk State University, especially Anatasia Knyazkova and Olga Zakharova, for their help with the technical work in this book.

The materials have been collected under partial support of the Russian Foundation for Basic Research (Grant No. 7-00-00186), RFBR, and Administration of Tomsk Region (grant No. 18-42-703012), FCPIR, contract No. 14.578.21.0082.

This work was supported by the Government of the Russian Federation (Agreement No. 075-15-2021-615 of 04 June 2021) to support scientific research projects implemented under the supervision of leading scientists at Russian institutions (Russian institutions of higher education).

Yury V. Kistenev
Alexey V. Borisov
Denis A. Vrazhnov
July 2021

Chapter 1
Fundamental Concepts Related to Laser Molecular Imaging

Introduction

Laser molecular imaging deals with analyzing the spatial distribution and temporal variation of biomolecules in a human body and samples of a biological origin. Similar studies are associated with the discovery and analysis of biomarkers. Suitable biomarkers are vital for monitoring a person's current metabolism and disease detection, but the dependence of a disease, a shift in metabolism, and registered spectral data are latent and complicated. Accordingly, specific methods of spectral data analysis are necessary. Currently, artificial intelligence is the most promising approach in this field. This chapter gives general information about the biomarker conception, molecular laser imaging, and artificial intelligence, including machine learning. The basic concepts introduced here are described in detail in Chapters 2–4.

1.1 Molecular Biomarkers

1.1.1 Biomarker conception

Biomarkers were understood initially as "cellular, biochemical, or molecular alterations that are measurable in biological media such as human tissues, cells, or fluids."[1] The concept of a biomarker proposed in 2001 described a physical, functional, or biochemical characteristic that can be measured quantitatively and can serve as an indicator of physiological or pathological processes or pharmacological responses to therapeutic intervention. This definition implies the dependence of biomarker values on physiological or pathological changes.

Biomarkers are becoming a mandatory part of clinical studies, allowing investigation of biologically active substances' mechanisms, forming groups of risk associated with a disease.

1.2.1.1 Molecular vibrational absorption spectra

Vibrations of a diatomic molecule

A molecule's stability is connected with

- a repulsive force among nuclei and electron clouds of different atoms, which form a molecule;
- the attractive force between an atom nucleus and electrons.

The equilibrium distance R_0 of atoms' nuclei in a molecule is determined by the equality of the repulsive and attractive forces. Accordingly, modeling the interatomic interaction by the force of elasticity, the molecule can be represented as a harmonic oscillator.

The Schrödinger equation for a diatomic molecule in the approximation of a 1D harmonic oscillator has the following solution:

$$E_{vibr} = \hbar\omega\left(\nu + \frac{1}{2}\right), \tag{1.8}$$

where $\nu = 1,2, \ldots, n$ is the vibrational quantum number. Thus, in the framework of this model, the spectrum of vibrational transitions is equidistant. An anharmonicity accounting is usually described by the Morse potential:[79,80]

$$V(R) = E_d[1 - \exp\{a(R_0 - R)\}]^2, \tag{1.9}$$

where a is the constant, specific for a molecule, R is the distance between atoms' nuclei, and E_d is the energy of dissociation.

The vibrational energy levels are described by the expression

$$E_{vibr} = \hbar\omega_e\left(\nu + \frac{1}{2}\right) - \hbar\omega_e \cdot X_e\left(\nu + \frac{1}{2}\right)^2 + \ldots, \tag{1.10}$$

where $X_e > 0$ is the constant characterizing anharmonicity, i.e., a measure of the deviation of the actual potential function $V(R)$ from the potential function of the anharmonic oscillator, $\omega_e = a\sqrt{2E_d/\mu}$, where $\mu = \frac{m_1 m_2}{m_1 + m_2}$ is the reduced molecule mass, and m_1 and m_2 are the atom's masses.

The selection rules for transitions in the anharmonic oscillator are

$$\Delta\nu = \pm 1;\ \pm 2;\ \pm 3,\ \ldots.$$

Vibrations of polyatomic molecules

A polyatomic molecule containing N atoms has $3N$ degrees of freedom, including 3 degrees of freedom for translational motion (three spatial coordinates) and 3 degrees of freedom for rotation of a nonlinear molecule

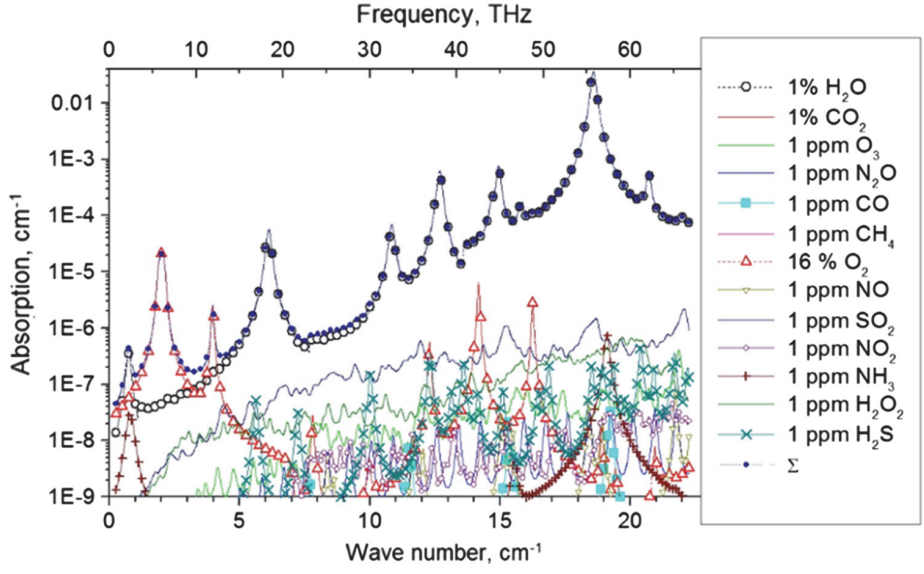

Figure 1.12 Absorption of light molecules of atmospheric air in the THz spectral range.

processes. The external optical field induces a dipole moment μ, which varies with the frequency of the optical field v_o:

$$\mu = \alpha \cdot F_0 \cos(\nu_o t). \tag{1.30}$$

Here, F_0 is the amplitude of the external optical field, and α is the polarizability of the molecule. The polarizability varies during the vibrational motions of the molecule. Let the fundamental vibration frequencies of the molecule be ν_k, $k = 1,2, \ldots, M$. Then, a variation of the polarizability with vibrations of the molecule can be expressed by expanding components of the α in a Taylor series concerning the normal coordinates of vibration Q_k, as follows:[92]

$$\alpha = \alpha_0 + \sum_k \left(\frac{\partial \beta}{\partial Q_k}\right) Q_k + \ldots \tag{1.31}$$

where α_0 is the polarizability at the zero displacements. Thus, during harmonic vibrational motions of the molecule,

$$\alpha = \alpha_0 + \sum_{k=1}^{M} \alpha_k \cos(\nu_k t + \phi_k) + \ldots. \tag{1.32}$$

The induced dipole moment is

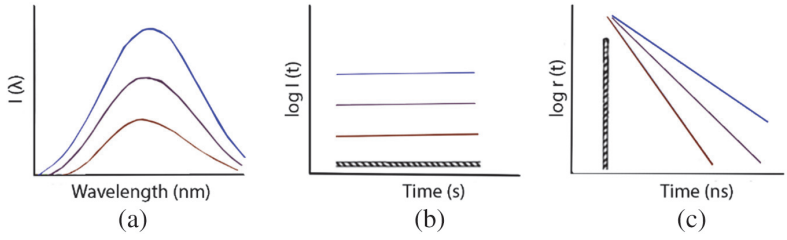

Figure 1.19 Comparison of steady-state and time-resolved fluorescence spectroscopy: (a) fluorescence spectra of various analytes, (b) their fluorescence irradiation intensity for steady-state pumping (marked by the horizontal black shaded box), and (c) time-resolved fluorescence irradiation intensity for pulse pumping (marked by the vertical black shaded box)[96] (reprinted under license of Springer Nature, License Number 4701911058954).

1.2.4 Molecular imaging

Molecular imaging provides a spatial-temporal visualization and characterization of biomolecules in tissue.[99]

In vivo molecular imaging is associated with identifying and quantifying the molecular marker profile in tissue without a surgical biopsy. While anatomical imaging is focused on diagnosis, surgical guidance, and monitoring of treatment, molecular imaging provides the ability for screening and early diagnosis, personalization of therapy, and earlier treatment follow-up.[100]

There are labeled and label-free variants of molecular imaging. Most non-optical molecular imaging modalities use labels or contrast agents (see Fig. 1.20). They include molecular targeting substances such as antibodies, peptides, nucleic acids, and labels for readout by an imaging modality.[100]

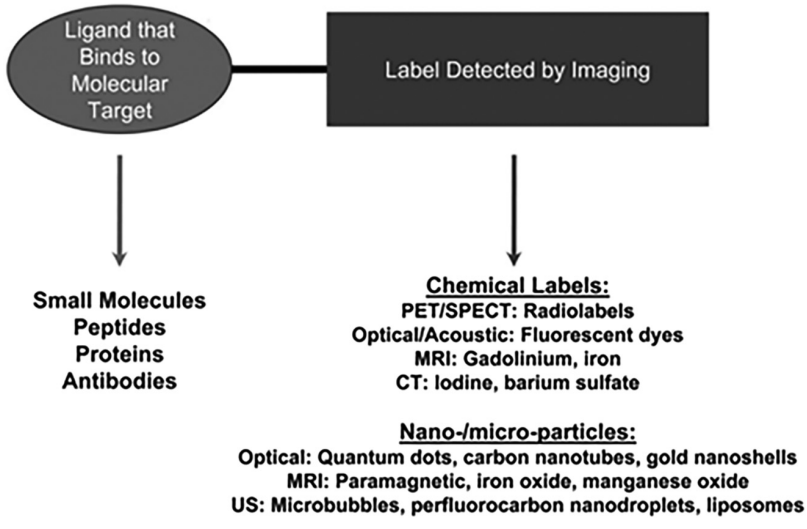

Figure 1.20 Contrast agents used for molecular imaging[100] (reprinted under Elsevier's license, License Number 4694920914851).

1.3 Basics of Machine Learning

The first step in machine learning (ML) was made by Alan Turing in 1950 by proposing three different strategies, which might achieve a "thinking machine" or artificial intelligence (AI).[102]

The first strategy is to develop AI by manually programming a computer.[103] This idea was transformed into modern expert or rule-based systems. The main drawback of such systems is the need for experts to develop and maintain them. They are not flexible, and there is a problem with contractions in the rules.

Another strategy is *ab initio* ML.[103] This approach teaches a computer, which consists of analyzing positive and negative data samples with rewards and punishments. If an expert provides links among samples and possible classes, then datasets are called labeled. A part of the latter forms a training set. The rest of the labeled samples, which were not used in teaching, form a testing set used to validate results. This idea leads to a supervised learning approach.

It started with the invention of a human brain cell prototype called "perceptron" in 1957 by Frank Rosenblatt.[104] Perceptron was initially designed as a tool for image recognition and was a kind of associative link between input stimuli (x_i) and the necessary response (y) at the output (Fig 1.22).[105] Here, ω_i are weight coefficients. The teaching consisted of adjustments to the weight coefficients to make the proper conclusion depending on input signals. This procedure can also be considered "highlighting" an object feature vector's particular coordinates to increase the distance between feature vectors belonging to various states. This branch of machine learning has recently been developed in the form of deep learning (DL).

There are four major branches of ML (see Fig. 1.23). In supervised ML, a trained algorithm can classify new data.[64] Unsupervised ML algorithms allow us to make conclusions about the presence of possible classes with unknown origins in unlabeled data. Semi-supervised learning combines supervised and unsupervised learning, which uses relatively small amounts of labeled data and unlabeled data for training.[106] It requires a lot of data to process, so a way to apply semi-supervised learning in biomedicine is still under discussion. The

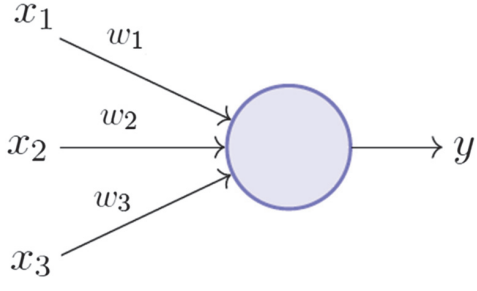

Figure 1.22 Perceptron model by Minsky and Papert (1969).

Chapter 2
Laser-based Molecular Data-Acquisition Technologies

Introduction

A "gold standard" for the verification of many diseases is histopathology analysis of a biopsy sample. A biopsy is the extraction of cells or tissues for examination.[1] The latter's disadvantages are that it is time consuming and invasive. In cancer detection, there is a high risk of metastasis due to the cancer cells possible dissimilation through blood or lymph vessels from the region of surgery.

The term "optical biopsy" has entered into common usage in the field of biomedical optics. This term has internal inconsistency because "biopsy" refers specifically to tissue removal, whereas the implication of "optical" is that tissue is not removed. Regardless, "optical biopsy" is commonly understood as optical measurements, often a kind of spectroscopy, to noninvasively (or minimally invasively) perform *in vivo* and real-time diagnosis.[2] Depending on an analyzed diagnostic agent, the optical biopsy is often divided into breath biopsy, liquid biopsy, and tissue biopsy.

Optical biopsy can be used as a diagnostic tool or to reveal specific (patho-) physiological mechanisms. The latter is connected with the chemical-based identification of particular compounds. But an individual molecular compound hardly serves as a biomarker of a specific disease due to low specificity. Reliable diagnostics is possible through the control of a group (profile) of molecular biomarkers. Probabilistic discrimination of biomarker profiles can be conducted by a pattern-recognition approach, which forms the basis for assessing acceptable diagnostic accuracy. The chemical analytical-based identification of individual molecular biomarkers is not strictly necessary in a clinical setting; also, note that the biochemical origin of most molecular biomarkers is unknown.[4]

$$\Delta N \geq \frac{a}{\sigma L(\frac{S}{N})}, \tag{2.5}$$

where α is the total multipass absorbance

$$\alpha = -\ln\left[\frac{I}{I_0}\right], \tag{2.6}$$

L is the path length, S/N is the signal-to-noise ratio, and σ is the absorption cross-section.

A multipass cell based on confocal mirrors was created.[34] A path length of about 300 m was achieved in a cell with the mirrors spaced 0.5 m apart. The multipass cell's sensitivity was tested by measuring the absorption spectra of CH_4, CO, and CO_2. The LOD for CH_4 was 6 ppmv. Similarly, for CO_2 and CO, the LOD was found as 640 and 320 ppmv, respectively.

The drawback of MPAS that the individual passes of light must be spatially separated in the absorption cell. It leads to the need for relatively large mirrors. Also, an overlap of light beams that make different numbers of passes through the cell causes interference noise in the transmitted light, limiting sensitivity.[37]

2.1.4.3 Cavity ring-down spectroscopy

Cavity ring-down spectroscopy (CRDS) tends to provide the highest sensitivity in terms of absorption based instruments allowing the detection of molecular components in a gaseous sample with ppb concentrations.[37–39]

A typical CRDS setup consists of a laser and a high-quality optical resonator (two or more mirrors with reflectivity $R > 99.9\%$). The principle of CRDS is illustrated in Fig. 2.6.[40] Here, L is the cavity length, and R is the mirror reflectivity.

The laser beam is reflected many times in the high-quality resonator, providing an effective path length of several kilometers. Simultaneously, the laser beam intensity released from the resonator after each pass in the cavity is decreased in time following the expression

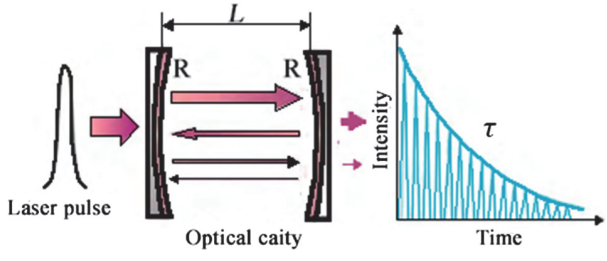

Figure 2.6 The principle of CRDS.[40]

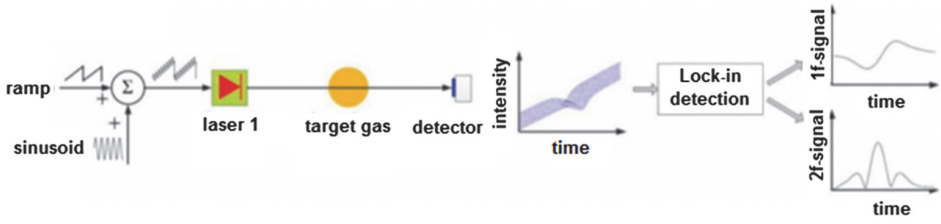

Figure 2.8 The idea of WMS[68] (reprinted under Taylor & Francis's license, License Number 4700831059288).

first and second harmonics are preferable (Fig. 2.8). The lock-in amplifiers act by multiplying the detector signal by a reference sinusoid at the frequency of interest ($1f$ or $2f$). Taking into account that

$$\cos(\alpha)\cos(\beta) = \frac{1}{2}\cos(\alpha - \beta) + \frac{1}{2}\cos(\alpha + \beta), \tag{2.13}$$

a low-pass filter is applied to isolate this current and eliminate all components outside of the filter bandwidth.[68,69]

WMS was mainly used with diode lasers., It can also be performed by using external modulators in the near-IR.[70,71] A tunable diode laser-based WMS sensor for *in situ* temperature and water measurements in flames was presented in Ref. 72. The sensor enables measurements without calibration or knowledge of the mixture collisional-broadening coefficient.

Modulation spectroscopy is potentially more sensitive than DAS. The modulation technique provides two main advantages: it measures a difference signal, which is directly proportional to the species concentration; and it allows for shifting a measured signal to higher frequencies, thereby offering a larger signal-to-noise ratio and higher sensitivity.

2.1.4.6 Fourier transform spectroscopy

Fourier transform IR (FTIR) spectroscopy is based on two beams' interference (Fig. 2.9).

If a moving mirror is shifted with a constant speed, then the signal on a photodetector is modulated by a sinusoid; each maximum corresponds to the beam's optical path difference δ equal to $k\lambda$ ($k = 0, \pm1, \pm2, \ldots$). The detected signal intensity for monochromatic optical radiation with a frequency ν is determined by the expression

$$I_0(\delta) = 0.5I(v)\left(1 + \cos 2\pi\frac{\delta}{\lambda}\right) = 0.5I(v)(1 + \cos 2\pi v\delta), \tag{2.14}$$

where $I(\nu)$ is the optical wave intensity. In most FTIR commercial spectrometers based on a Michelson interferometer, the mirror moves at a constant speed V and $\delta = 2Vt$. Varied in time, only the term

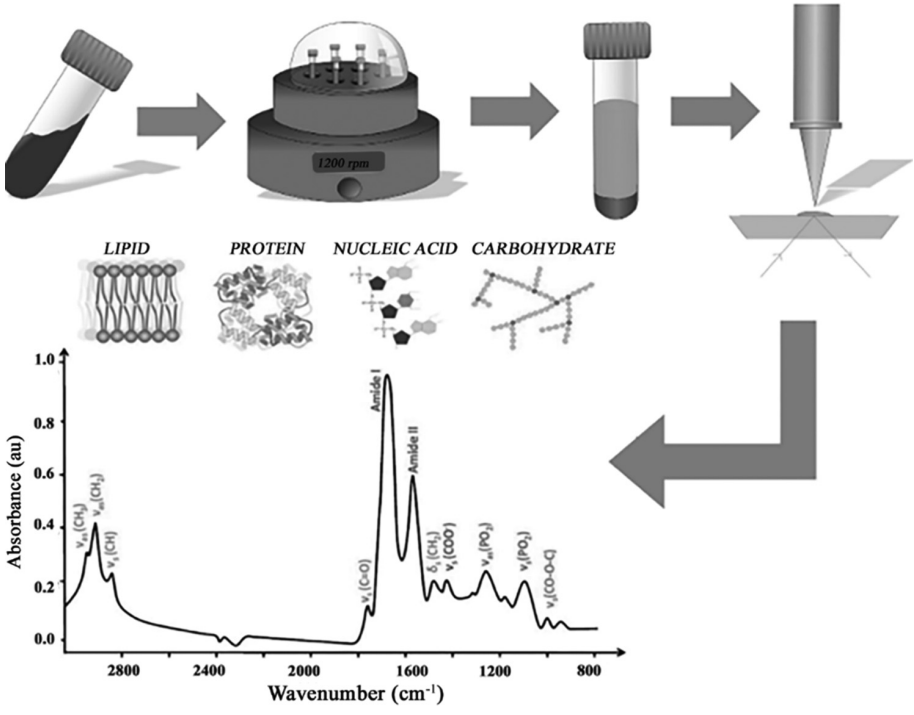

Figure 2.18 The typical blood sample analysis pipeline, from the blood sampling, serum separation step to the spectrum acquisition.[100]

to a medium dielectric permittivity. These are typically Lorentzian resonances so that a full dielectric response can be modeled as

$$\varepsilon(\omega) = \varepsilon_\infty + \sum_i \frac{\Delta\varepsilon_i}{1 + i\omega\tau_i} + \sum_j \frac{f(\omega_j)}{(\omega_j^2 - \omega^2) - i\gamma\omega}, \tag{2.20}$$

where ε_∞ is the medium permittivity high-frequency limit, $\Delta\varepsilon = \varepsilon_s - \varepsilon_\infty$ is the amplitude of the relaxation component of the total permittivity, ε_s is the medium permittivity stationary value, τ is a characteristic re-orientation time for a single dipole in a material, $f(\omega_j)$ is the amplitude of an oscillation component, and γ is its relaxation time.[98,99]

Time-domain spectroscopy (TDS) is the most frequently used instrumental realization of spectral analysis of biomedical samples in the THz range. TDS uses the generation of THz probe pulses by a GaAs antenna pumped by a repetitive train of femtosecond laser pulses and their synchronous registration (after interaction with the investigated sample) by the other GaAs antenna. The GaAs detector is sensitive to the THz wave's electric field only at its illumination by the laser pulse. Thus, laser pulses provide the generation of THz probe pulses and the GaAs detector's strobing using an

2.3.4 Possible experimental laser molecular imaging methods for *in vivo* tissue optical biopsy

2.3.4.1 Multiphoton microscopy

Multiphoton microscopy is based on laser radiation nonlinear multiphoton absorption by biomolecules. For example, two photons can excite a molecular transition jointly when the transition energy equals the double energy of incident photons (i.e., half the wavelength). The nonlinear excitation requires high power photon flux, typically 10^{20}–10^{30} photons/(cm^2s). To fulfill this condition, femtosecond pulsed lasers in combination with high-numerical-aperture focusing lenses are used.

Second-harmonic generation (SHG) is a second-order nonlinear effect, which appears in noncentrosymmetric structures.[127] Collagen molecules form the triple helix, consisting of three α-chains forming the triple helix (Fig. 2.30). The amino acid sequence of the triple helix α-chain consists of the repeated Gly-X-Y motif, where X is, as a rule, proline (Pro), and Y can be any amino acid, usually, hydroxyproline (Hyp) or hydroxylysine. Therefore, collagen has no symmetry center, which makes possible its visualization based on SHG

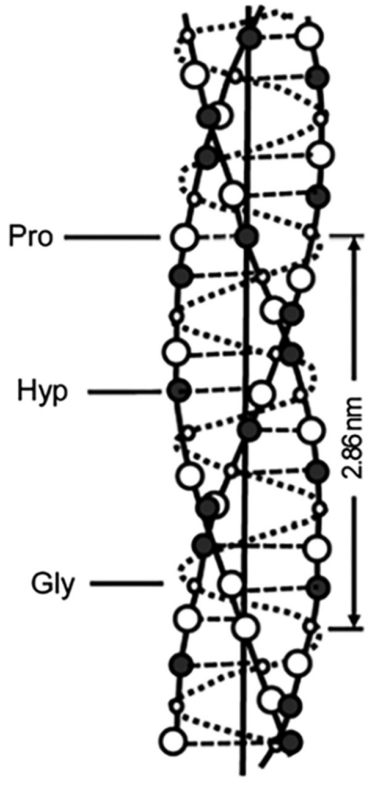

Figure 2.30 Collagen molecule spatial structure.

Chapter 3
Informative Feature Extraction

Introduction

Laser molecular imaging produces high-dimension data with the structure dependent on the optical modality, laser type, detection method, kind of sample, etc.[1] Generally, data's high dimension corresponds to a situation where the number of initial parameters exceeds by orders of magnitude the number of hidden independent variables, e.g., when the number of measured absorption coefficients of a complex gas mixture exceeds by an order or more the quantity of pure components in the mixture.

The high-dimension data are hard to use for predictive data model construction[I] due to the "curse of dimensionality" problem formulated by R. Bellman.[2] Essentially, when the feature vector's dimension increases, the volume of data needed for classifier training grows exponentially. This is because the difference between two random vectors tends to zero as their dimension increases according to the central limit theorem.

One of the main goals of feature extraction is to overcome this problem. The universal approach for this is in decreasing the data dimension. Concrete ways depend on the data origin. In particular, 2D-3D images can be decomposed into small geometrical parts with similar properties named textures.[3,4] The texture approach allows one to find a compact description of the initial image.

Molecular spectra can be considered as a degenerate case of molecular imaging data in a case of a homogeneous medium when we can study only one "point" to describe the whole sample.

Feature vector dimension reduction includes feature selection and feature extraction.[I] The difference between them is only in the ways used to get the result. This chapter describes these methods in details sufficient for practical applications. The Python codes for the most useful analytical methods described in the chapter are presented in the Supplemental Materials.

3.1 Feature Selection

Feature selection is realized in three steps; some of them are optional. The first step is the feature subset selection (generation), based on either classical

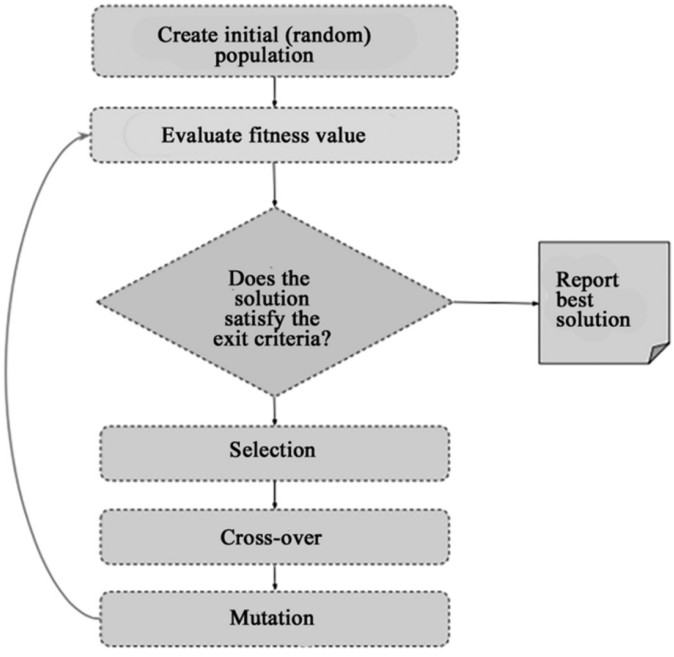

Figure 3.6 A block scheme of the genetic algorithm.[22]

for feature selection. (The Python code modules #FFS, #BFS, #BP, #GA, and test examples are available in the Supplemental Materials).

The **embedded methods**' main idea is to combine the advantages of filter and wrapper methods and find the best subset of informative features automatically through their built-in feature selection methods (see Fig. 3.7).

Examples of embedded methods include *least absolute shrinkage selector operator* (LASSO) regularization in linear regression[23] and *gradient boosting machine* (GBM) over a decision tree.[24]

Regression models suffer from complexity, which increases overfitting risk. The solution is a reduction of the magnitude of linear regression coefficients.

A penalty term proportional to a linear regression coefficients' absolute values are added in the LASSO method, so uninformative features automatically become zeros.

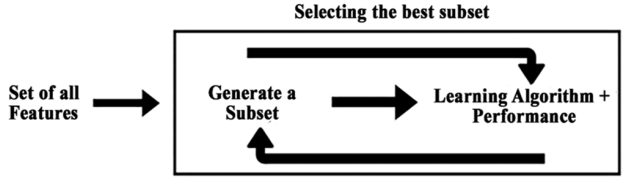

Figure 3.7 Embedded methods pipeline.

Chapter 4
Clusterization and Predictive Model Construction

Introduction

The most crucial step in the machine learning pipeline is related to experimental data content and semantic analysis to predict new data's meaning.

Let's consider a set of studied objects, which are described by corresponding feature vectors. The set of these vectors can be presented in the form of a matrix:

$$\hat{X} = \left\| \begin{array}{c} x_{11}x_{12}\ldots x_{1M} \\ \ldots \\ x_{N1}x_{N2}\ldots x_{NM} \end{array} \right\|,$$

where a line contains an object's feature vector, and a column $f_j = (x_{1j}x_{2j}\ldots x_{Nj})^T$ contains a component of the feature vector for all objects.

Possible states (classes) y_l to which studied objects belong are described by a vector of labels:

$$\boldsymbol{y} = \left\| \begin{array}{c} y_1 \\ y_2 \\ \ldots \\ y_N \end{array} \right\|.$$

The main assumption of machine learning approach is a compact distribution of feature vectors of different classes in the feature space.

Function $\alpha : \hat{X} \to \boldsymbol{y}$ is called a *predictive model*. Estimation of α is a fundamental task of machine learning algorithms. The predictive model construction is based on information about data structure and similarity measure or distance metric between feature vectors. If vector y is preliminarily

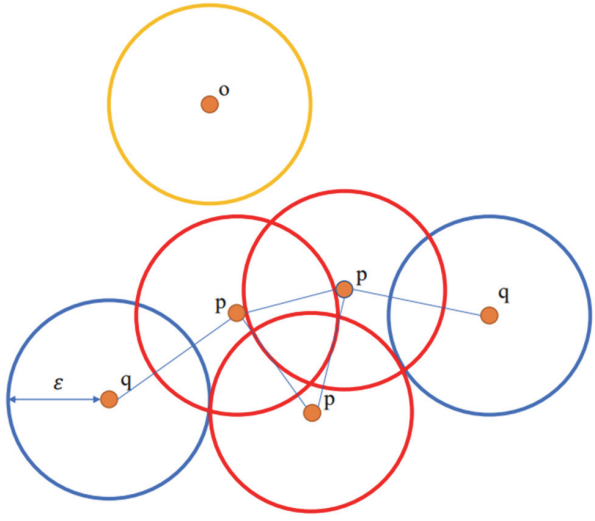

Figure 4.2 DBSCAN illustration. Dots p are core points, q are reachable points, o is an outlier, and $NumP = 3$.

- Point q is directly reachable from point p if p is a core point and the distance between p and q is less than ε.
- Point q is reachable from point p if there is a route from p and each point is directly reachable from the previous point of the route (all route points should be core, except q).
- Outliers o are the points that are not reachable from the core.
- The cluster is formed by the core point, and all points reachable from the core point. Each cluster contains at least one core point.
- Point p from the cluster is called a dense point if ε is the neighborhood that is also a part of the same cluster.

DBSCAN algorithm:

1. Choose a random point that is not visited yet.
2. The ε is the chosen neighborhood, and if it contains at least $NumP$ points, the cluster is started. Otherwise, it is marked as an outlier. This outlier can be included in a cluster later. The point is a dense one if all points in the ε neighborhood are in this cluster. Dense points of a specific cluster form its dense part.
3. This process continues until a dense part of the cluster is found.
4. Another random, unvisited point is selected.

DBSCAN has the following advantages. First, there is no need for knowledge of the number of clusters. Second, DBSCAN can find clusters of arbitrary shapes.[4,5,6] It can determine noise and outliers.

1.2.1.1 Molecular vibrational absorption spectra

Vibrations of a diatomic molecule

A molecule's stability is connected with

- a repulsive force among nuclei and electron clouds of different atoms, which form a molecule;
- the attractive force between an atom nucleus and electrons.

The equilibrium distance R_0 of atoms' nuclei in a molecule is determined by the equality of the repulsive and attractive forces. Accordingly, modeling the interatomic interaction by the force of elasticity, the molecule can be represented as a harmonic oscillator.

The Schrödinger equation for a diatomic molecule in the approximation of a 1D harmonic oscillator has the following solution:

$$E_{vibr} = \hbar\omega\left(\nu + \frac{1}{2}\right), \tag{1.8}$$

where $\nu = 1,2, \ldots, n$ is the vibrational quantum number. Thus, in the framework of this model, the spectrum of vibrational transitions is equidistant. An anharmonicity accounting is usually described by the Morse potential:[79,80]

$$V(R) = E_d[1 - \exp\{a(R_0 - R)\}]^2, \tag{1.9}$$

where a is the constant, specific for a molecule, R is the distance between atoms' nuclei, and E_d is the energy of dissociation.

The vibrational energy levels are described by the expression

$$E_{vibr} = \hbar\omega_e\left(\nu + \frac{1}{2}\right) - \hbar\omega_e \cdot X_e\left(\nu + \frac{1}{2}\right)^2 + \ldots, \tag{1.10}$$

where $X_e > 0$ is the constant characterizing anharmonicity, i.e., a measure of the deviation of the actual potential function $V(R)$ from the potential function of the anharmonic oscillator, $\omega_e = a\sqrt{2E_d/\mu}$, where $\mu = \frac{m_1 m_2}{m_1 + m_2}$ is the reduced molecule mass, and m_1 and m_2 are the atom's masses.

The selection rules for transitions in the anharmonic oscillator are

$$\Delta\nu = \pm 1;\ \pm 2;\ \pm 3,\ \ldots.$$

Vibrations of polyatomic molecules

A polyatomic molecule containing N atoms has $3N$ degrees of freedom, including 3 degrees of freedom for translational motion (three spatial coordinates) and 3 degrees of freedom for rotation of a nonlinear molecule

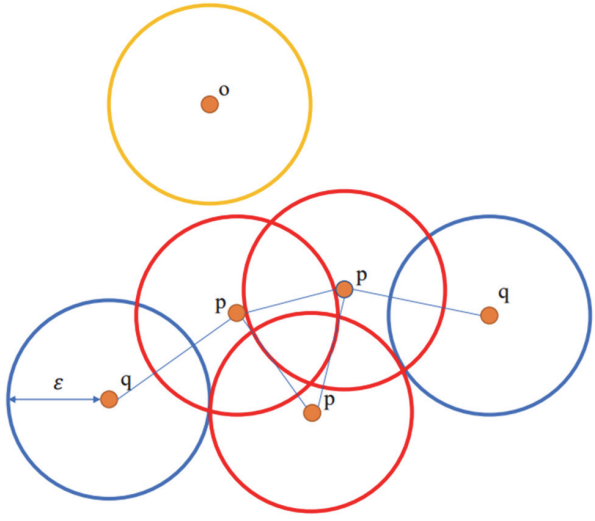

Figure 4.2 DBSCAN illustration. Dots p are core points, q are reachable points, o is an outlier, and $NumP = 3$.

- Point q is directly reachable from point p if p is a core point and the distance between p and q is less than ε.
- Point q is reachable from point p if there is a route from p and each point is directly reachable from the previous point of the route (all route points should be core, except q).
- Outliers o are the points that are not reachable from the core.
- The cluster is formed by the core point, and all points reachable from the core point. Each cluster contains at least one core point.
- Point p from the cluster is called a dense point if ε is the neighborhood that is also a part of the same cluster.

DBSCAN algorithm:

1. Choose a random point that is not visited yet.
2. The ε is the chosen neighborhood, and if it contains at least $NumP$ points, the cluster is started. Otherwise, it is marked as an outlier. This outlier can be included in a cluster later. The point is a dense one if all points in the ε neighborhood are in this cluster. Dense points of a specific cluster form its dense part.
3. This process continues until a dense part of the cluster is found.
4. Another random, unvisited point is selected.

DBSCAN has the following advantages. First, there is no need for knowledge of the number of clusters. Second, DBSCAN can find clusters of arbitrary shapes.[4,5,6] It can determine noise and outliers.

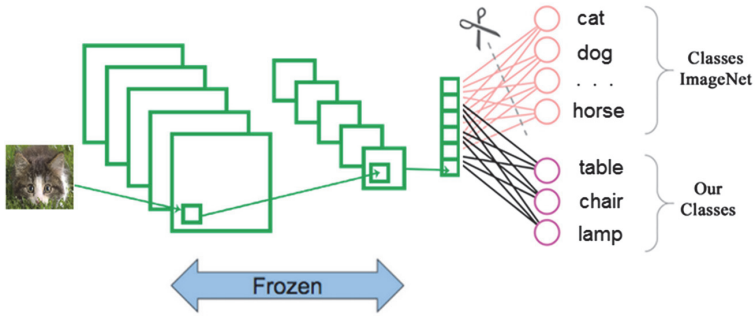

Figure 4.18 Transfer learning idea.

3. Inference. A mask that provides the best noise reduction on the output is selected.

4.2.12 Improving prediction models; ensemble learning

The main challenge for machine learning is to achieve more accuracy, more robustness, and generalization. There are several approaches to decrease bias, variance and improve predictions. You can work with preprocessed data by removing outliers, adding more positive and negative samples to reduce unseen data.

The ensemble learning approach is based on combining several machine learning algorithms into one meta-algorithm. The latter can be done sequentially and in parallel. Sequential ensemble learning is also known as boosting machine learning. It assumes that you have several "weak" classifiers with accuracy better than random guessing. The goal is to construct a "strong" classifier, which has accuracy significantly better than any "weak" classifier does. The tactics are as follows: the mislabeled data from the previous "weak" classifier gets bigger weights, and the subsequent "weak" classifier is focused on classifying correctly more data. This process repeats until the desired accuracy is obtained. The SVM, ANN, random forests, and other techniques can be used as "weak" classifiers in this approach.[89–92] An example of boosting ensemble learning is presented in Fig. 4.19.

The so-called stacking ensemble is a further development of the boosting approach. Using the same input dataset, you should train many classifiers (CNNs or else), choose several with the best accuracy, and use them as the input for another ANN called a meta-learner.[93] The Python code modules #BEL and test examples are available in the Supplemental Materials.

A parallel ensemble learning is presented by bagging and random forests.[94] In this approach, one should generate random samples with replacement subsets from initial data and train classifiers for each such subset. A final decision will be carried out using the majority vote procedure or averaging (see Fig. 4.20).

43. Hsu, C.W. and C.J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Trans Neural Netw* **13**(2), 415–25 (2002). [doi:10.1109/72.991427]
44. Gasteiger, J. and J. Zupan, *Neural Networks for Chemists: An Introduction*, 1st — s.l.: VCH Publishers, New York, NY (1993).
45. Platt, J.C., N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multi-class classification," *In Advances in Neural Information Processing Systems, MIT Press* **12**, 547–553 (2000).
46. Vapnik, V., *Statistical Learning Theory*, Wiley, New York, NY (1998).
47. Weston, J. and C. Watkins, "Multi-class support vector machines," *In M. Verleysen, editor, Proceedings of ESANN99 Brussels, D. Facto Press*, 219–224 (1999).
48. Crammer, K. and Y. Singer, "On the learnability and design of output codes for multi-class problems," *Machine learning* **47**(2), 35–46 (2002).
49. Breiman, L., "Random Forests," *Machine Learning* **45**(1), 5–32 (2001).
50. Breiman, L., *Random Forests*, Statistics Department University of California Berkeley, CA 94720 Technical Report 567 (1999).
51. Lin, Z., et al., "Exploring metabolic syndrome serum profiling based on gas chromatography mass spectrometry and random forest models," *Analytica Chimica Acta* **827**, 22–27 (2014).
52. Lanz, C., et al. "Radiation metabolomics. 3. Biomarker discovery in the urine of gamma-irradiated rats using a simplified metabolomics protocol of gas chromatography-mass spectrometry combined with random forests machine learning algorithm," *Radiation Research* **172**(2), 198–212 (2009).
53. Qi, Y., "Random forest for bioinformatics," *In: Ensemble machine learning*, C. Zhang and Y. Ma (eds), Springer, Boston, MA (2012).
54. Ball, G., et al., "An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers," *Bioinformatics* **18**(3), 395–404 (2002).
55. Aljović, A., A. Badnjević, and L. Gurbeta, "Artificial neural networks in the discrimination of Alzheimer's disease using biomarkers data," *In 2016 5th Mediterranean Conference on Embedded Computing (MECO), IEEE*, 286–289 (2016).
56. Wang, L., et al., "Identifying biomarkers of endometriosis using serum protein fingerprinting and artificial neural networks," *International Journal of Gynecology & Obstetrics* **101**(3), 253–258 (2008).
57. Kohonen, T., Self-Organization and Associative Memory, *8 — s.l.: Springer Berlin Heidelberg* (1989).
58. Huang, G.B., et al., "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **42**(2), 513–529 (2011).

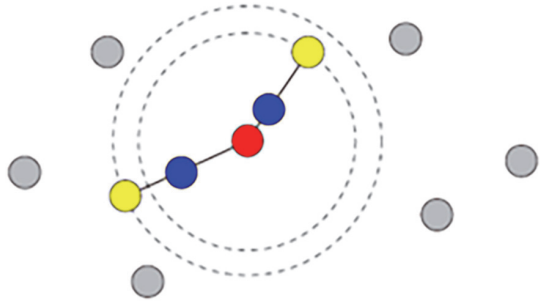

Figure 5.8 Oversampling procedure. Two virtual samples were interpolated at a random point between a chosen random sample and some two samples that are nearest the chosen one.[13]

A target group of 107 patients with LC and a control group of 29 healthy volunteers was studied.[13] The authors pointed out that the imbalanced sample numbers in the VOC data set is likely to cause an inappropriate classification. To complete the sample numbers, they used a synthetic minority oversampling technique. An original sample was chosen randomly. Two virtual samples were interpolated at a random point between the chosen sample and some two samples nearest to the chosen one (Fig. 5.8).

The authors used a nonlinear SVM with a Gaussian RBF kernel[IV] and the leave-one-out CV.[IV] The maximum accuracy was achieved using 9–10 VOCs. Still, the lowest number of the corresponding support vectors of the true positive rate classification was performed using 5 VOCs (CHN, methanol, CH_3CN, isoprene, 1-propanol) in a training set, and that of the true negative rate reached almost the bottom for 4 VOCs. These results suggested that 5 VOCs are sufficient for 89.0% diagnostic accuracy without overfitting and that the 95% and 89% correct negative rate-based diagnoses are possible when using 5 and 4 VOCs, respectively.

Malignant pleural mesothelioma (MPM) is a rare type of pulmonary cancer mainly caused by asbestos exposure. Thirty-nine patients with clinically verified MPM were recruited for breath air analysis.[36] Breath air samples were analyzed by the GC-MS tool. Three machine learning algorithms in combination with a leave-one-out CV[IV] were applied to create and evaluate a predictive model (Fig. 5.9).

Ten VOCs, including ketones, alkanes, methylate derivates, hydrocarbons, allowed distinguishing MPM patients and healthy controls with the accuracy presented in the form of the ROC curve[I] in Fig. 5.10. The naïve Bayes[IV] expectedly demonstrated the worst accuracy.

19 VOCs were studied in the breath of lung cancer patients with a confirmed diagnosis ($N = 51$) and controls ($N = 53$) using GC-MS and machine learning.[37] Isoprene, acetone, 2-propanol, and nonanal were detected in the breath of all participants. Thiophene (1.4%), 1-butanol (40.1%), and

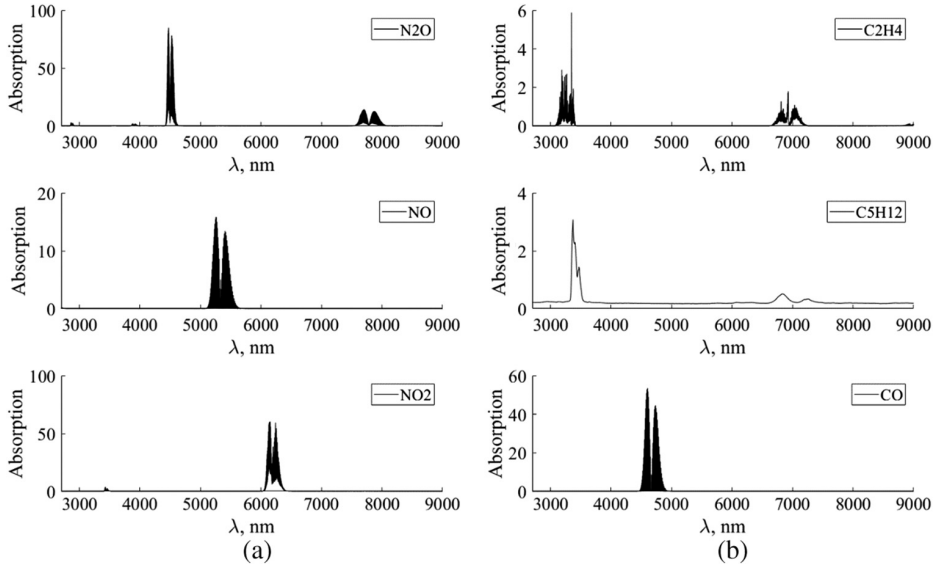

Figure 5.17 Absorption spectra of several VOCs, which are typical for broncho-pulmonary diseases.

including propionic acid, ethanol, triethylamine, hexane, toluene, and dimethylsulfide.[66] In any case, the sensors' signals were partially overlapped and, strictly speaking, could not be associated with a specific component. The latter is typical for a pattern recognition approach.

The Mann-Whitney U and Kruskal Wallis tests were used to compare the medians of groups under study. The statistical significance threshold was $p < 0.05$. PCA[IV] and k-NN[IV] classification methods were applied to study the sensors' signals. The created predictive model was validated using a leave-one-out CV.[IV] The mean value of sensitivity was 94.1% (95% confidence interval [CI], 83.8-98.8%) and the same for specificity was 90.0% (95% CI, 68.3-98.8%).

Differential diagnosis of pulmonary diseases

The IR LPAS and the pattern-recognition-based analysis of the patient breath air samples' absorption spectra also can be used for non-invasive differential express diagnostics of pulmonary diseases.[3,4,67]

The study involved four groups: patients with broncho-pulmonary diseases including LC patients ($N = 30$), COPD patients ($N = 40$), patients with pneumonia ($N = 40$), and a control group of healthy volunteers ($N = 130$). All patients had been diagnosed preliminarily by clinical methods.

The sampling procedure is described in Section 5.1.1.2. The samples were analyzed using the LaserBreeeze gas analyzer.[12] All measurements were carried out at room temperature (20–25°C) and humidity (50–60%). The