Nonlinear Analysis: Modelling and Control, Vol. 23, No. 3, 285–302
https://doi.org/10.15388/NA.2018.3.1
ISSN 1392-5113

The ROC-based analysis of spectroscopic signals from medical specimens

Vilmantas Gėgžna[a,b,1], **Olga Kurasova**[c], **Gintautas Dzemyda**[c], **Rūta Kurtinaitienė**[d], **Ignas Čiplys**[a,b], **Juozas Vidmantis Vaitkus**[a], **Aurelija Vaitkuvienė**[a]

[a]Institute of Photonics and Nanotechnology, Vilnius University,
Sautėtekio ave. 3, LT-10257 Vilnius, Lithuania
vilmantas.gegzna@tmi.vu.lt

[b]Institute of Biosciences, Life Sciences Center, Vilnius University,
Saulėtekio ave. 7, LT-10257 Vilnius, Lithuania

[c]Institute of Data Science and Digital Technologies, Vilnius University,
Akademijos str. 4, LT-08663 Vilnius, Lithuania

[d]Vilnius University Hospital Santaros Klinikos,
Santariškių str. 2, LT-08661 Vilnius, Lithuania

Received: July 7, 2017 / **Revised:** February 12, 2018 / **Published online:** April 20, 2018

Abstract. Accurate methods of rapid medical diagnostics would obtain recognition among clinicians. In this paper, we present a ROC (receiver operating characteristic) analysis based approach to investigate intrinsic fluorescence spectra of medical samples. The approach provides researchers with capabilities for both spectroscopic feature selection and classification tasks. The method is illustrated using data obtained from samples of uterine cervix's tissues (normal, inflammatory, and precancerous conditions), but it is applicable for data from other types of medical specimens too. The results indicate the possibility to identify spectral ranges at which differences between various medical conditions are most evident as well as potential to apply the method for inflammation (cervicitis) vs. other medical conditions' diagnostics.

Keywords: receiver operating characteristic, classification, fluorescence spectroscopy, cervical smear, cervical cancer, medical data.

1 Introduction

In medicine, data come from various sources, including magnetic resonance imaging, thermovision, radiological, tomographic, ultrasound examination [2, 12]. Various data analysis methods and software tools find application in mining these data [14, 15]. Spectroscopic techniques may also be a convenient tool for medical diagnostics as they are non-invasive and can be used for many diagnostic purposes [19]. The measurements of different optical properties (mostly parameters of tissue fluorescence, Raman scattering,

[1] ORCID ID: 0000-0002-9500-5167

© Vilnius University, 2018

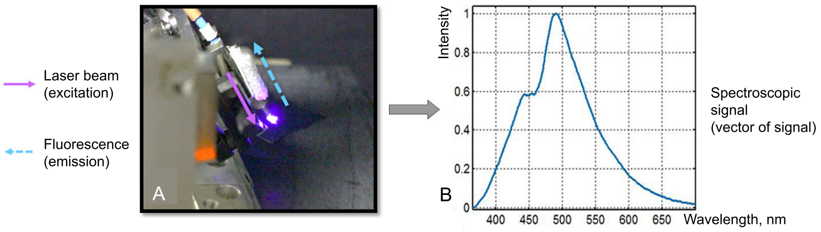

Figure 1. Spectroscopic measurements (A) and a fluorescence spectrum (B). (Online version in color.)

and light absorption) provide a possibility to analyze a contribution of key native tissue molecules to spectroscopic signals and to reveal differences between various pathologies.

Fluorescence is a type of physical phenomenon when light excites electrons in molecules or crystals, and they emit radiation that is different from the excitation due to intrinsic properties of the molecule or the crystal. This paper focuses on the type of fluorescence called autofluorescence that is native fluorescence of tissues (without any additional chemical substrates, medicines or fluorescent labels added to enhance the phenomenon) [4]. Fluorescence spectroscopy requires rather simple equipment: a light source for the excitation of molecules and rather low-resolution miniature spectrometers to register the fluorescence of the tissue directly or via the optical fibers.

Figure 1 demonstrates a measurement of fluorescence from one point of specimen and a fluorescence spectrum. Figure 1(A) shows a tip of a light guide of a fiber-based optical system, a light guide holder, and a sample holder. The light fiber is used to direct the laser beam to a sample. The sample is excited and begins to emit radiation called fluorescence, which is collected through a fiber and registered with a spectrometer. The solid violet arrow indicates the direction of a laser beam, while blue dashed line shows the direction of collected fluorescence. The bright spot near the center of Fig. 1(A) is a sample, which fluoresces. Figure 1(B) shows a fluorescence spectrum. The x-axis indicates wavelengths in nanometers (i.e., “colors” of radiation), and the y-axis indicates the intensity of each wavelength in arbitrary units.

In contrast to point monitoring (spectra registered from one point), there also exist other ways of fluorescence spectroscopy for the pathology diagnostics [4, 7] (e.g., hyperspectral imaging), but the implementation involves rather complicated and expensive equipment.

Fluorescence spectroscopy is thought to be a sensitive tool for tissue properties characterization [4, 16].

However, the fluorescent contribution of different native molecules and their complexes usually overlaps significantly. Therefore, difficulties arise either to analyze metabolism in tissues or to find diagnostically relevant information concerning certain medical conditions. The design of fluorescence-related diagnostic methods is based on clinical trials and statistical analyses that aim to reveal differences in properties related to pathological conditions. However, before it is possible to use these spectroscopic techniques in medical diagnostics, there is a challenge to develop an appropriate computational model that allows extracting relevant biomedical information about particular types of medical conditions from spectroscopic data.

Thus, the aim of this study was to implement an easy-to-interpret classification method, which allows extracting diagnostically relevant information from fluorescence spectra and revealing spectral ranges that are most useful for medical diagnostics. The scope of this paper was: (i) to develop procedures for spectroscopic data pre-processing and to search for classification rules that lead to the highest degree of agreement between spectroscopic classification (i.e., the results of spectroscopic data analysis) and the reference classification (e.g., the results of more conventional types of biomedical tests); as well as (ii) to disclose possibilities to apply the techniques to analyze spectroscopic data (autofluorescence) obtained from different types of medical specimens and illustrate our ideas with a case analysis of a particular type of specimens: liquid smears of uterine cervix tissues in various medical conditions [20]. The proposed method is based on techniques of signal pre-processing [11] and well-known receiver operating characteristic (ROC) analysis [6], but its novelty and strength lie in the multidisciplinarity of the research and the area of application.

2 Methods to process fluorescence spectra

2.1 The main steps of data analysis

The application of data analysis in biomedicine has two aims: (i) to reveal the parameters that allow recognizing pathological groups and (ii) to search for a decision rule that allows determining the correspondence of a spectrum to a pre-defined group of pathology/non-pathology.

Data analysis may be divided into two main phases: the research phase and the application phase (Fig. 2). The research phase includes steps in which appropriate methods to clean and pre-process data as well as to create and to select classification rules are developed. The result of this phase is a set of rules, which indicate how to arrive at a decision based on spectroscopic data. The performance of those rules is evaluated.

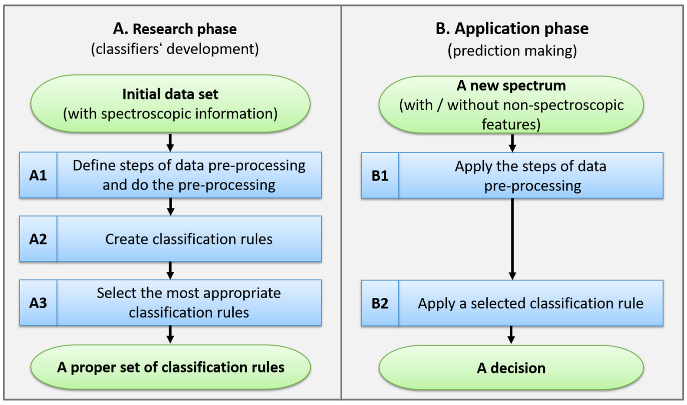
A. Research phase
(classifiers' development)
Initial data set
(with spectroscopic information)
A1
Define steps of data pre-processing and do the pre-processing
A2
Create classification rules
A3
Select the most appropriate classification rules
A proper set of classification rules
B. Application phase
(prediction making)
A new spectrum
(with / without non-spectroscopic features)
B1
Apply the steps of data pre-processing
B2
Apply a selected classification rule
A decision

Figure 2. The main steps of the analysis.

In the application phase, those rules are exploited to make a decision based on a new fluorescence spectrum. Smaller steps of these two phases are summarized in Fig. 2 and are described in the next subsections.

2.2 Classifiers' development

In this section, procedures, which are summarized in Fig. 2(A), are described in more detail.

2.2.1 Initial data set

An initial dataset (see Fig. 2(A) and Fig. 3) is a set of features related to patients' anamnesis (patient case history), their medical conditions (including results of medical and biochemical tests) as well as fluorescence spectra of medical specimens of those patients. The purpose of our analysis is to use this dataset to develop classification rules, which are based on spectroscopic and anamnestic information (such as patient's age), in order to predict results of medical, biochemical and another type of tests. In contrast to spectroscopy-based examination, these tests are usually more expensive and require more time (from several days to several weeks) before a patient hears a diagnosis.

2.2.2 Pre-processing of spectroscopic and non-spectroscopic data

Before classification rules can be created, the dataset must be prepared for the classificatory analysis. On the one hand, measurements of fluorescence spectra meet a difficulty in registering an absolute value of fluorescence intensity correctly. The main factors that lead to this difficulty are inhomogeneity in samples' composition, variability in the excited area of the specimen and in the angle at which light is collected. It is complicated (nearly impossible) to maintain all of these conditions constant. Also, effects related to

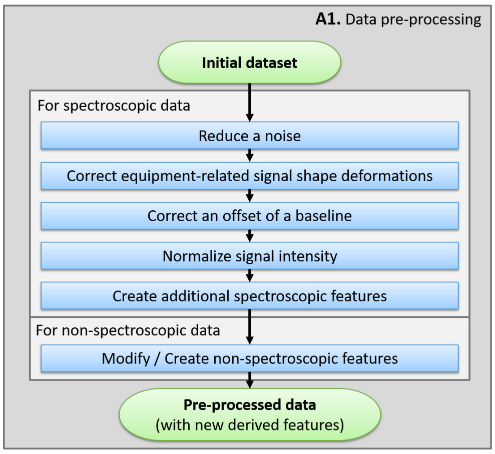
A1. Data pre-processing
Initial dataset
For spectroscopic data
Reduce a noise
Correct equipment-related signal shape deformations
Correct an offset of a baseline
Normalize signal intensity
Create additional spectroscopic features
For non-spectroscopic data
Modify / Create non-spectroscopic features
Pre-processed data
(with new derived features)

Figure 3. Scheme of data pre-processing.

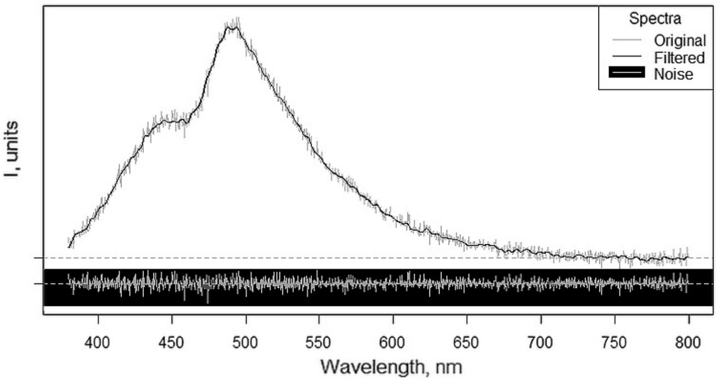

Figure 4. Fluorescence spectrum before (Original) and after (Filtered) noise reduction. Intensity is in relative units.

equipment, such as noise and detector sensitivity's dependence on wavelength, have to be accounted for. On the other hand, at least a few fluorophores' emission contribute to entire fluorescence spectrum of a biomedical tissue, and a relative change in neighboring components' contribution changes a shape of the spectrum, and these changes may be related to different medical conditions. Thus, several data pre-processing methods must be applied to eliminate or reduce shortcomings of fluorescence spectra measurement procedures and to highlight relevant differences between groups of spectra. Data pre-processing (Fig. 3) is one of the essential parts of the data analysis (e.g., Fig. 2(A)).

As there are three main sorts of distortions in signals of spectroscopic systems (noise, optical-system-related deformation of signal's shape and offset of signal's baseline), pre-processing begins with corrections of these distortions. Next follows procedures of signal intensity normalization as well as procedures that create new spectroscopic, non-spectroscopic and mixed features. The result of this stage is a pre-processed dataset augmented with new features.

○ *Noise reduction in spectroscopic signals.* An example of a spectrum before and after filter application is presented in Fig. 4. There are two dashed lines in the figure that indicate zero intensity. On the upper dashed line, there are two spectra: an original spectrum before pre-processing without noise reduction ("Original", light-grey line) and the same spectrum after the noise reduction ("Filtered", black line). The difference between these two spectra was called a noise spectrum. It is plotted on the lower dashed line as a dark grey line and in the legend is denoted as "Noise". Noise can be defined as a part of a measured signal that is not related to samples. Thermodynamic/thermoelectric effects are the most frequent cause of this noise, and the noise is equally common in all channels of the detector. Running medians filter can be applied to cut sudden relatively high-amplitude peaks that occasionally appear in signals of single channels, and Savitzky–Golay filter can reduce noise by minimally distorting the trend-line of the signal (the "real" signal). The parameters for signal filters were selected such that the noise was reduced significantly, but there were no visible trends in the noise spectrum for any of fluorescence spectra.

Next, the binning of several (e.g., ten) adjacent points on wavelength axis can smooth spectra even more as well as reduce the amount of data without losing too much information as adjacent points in the original signal are highly correlated.

○ *Correction of equipment-related signal shape deformations.* Equipment-related signal shape deformations arise because the sensitivity of different detector channels varies. It means that the same quantity of each different color (i.e., wavelength) is registered as a signal of different intensity, and the shape of the spectroscopic signal may be different if it is registered with another type of spectrometer. At this point, we have two options, which may be suitable for practical application.

Option I. As deformations for all signals produced by the same spectrometer are constant, it is possible to develop an analytic method, which is suitable to analyze signals of a concrete spectrometer and to avoid an additional step in which equipment-related signal shape deformations are corrected. This option allows researchers (or assistants that perform the diagnostic procedure) to prevent additional measurements and makes the procedure cheaper as an additional piece of equipment (calibrated lamp) is not necessary. It diminishes the risk to distort information in certain spectral regions in which the shape corrections may amplify the noise. On the other hand, in order to analyze the nature of spectra differences related to only biological phenomena in samples, adjustments of the shape of each spectrum by excluding the contribution of device function are required.

Option II. Usually, the correction of the signal shape is necessary to eliminate equipment-related distortions. This procedure results in a signal that allows doing not only classificatory analysis, but also delving into physical and biological nature of the signal as well as making the analysis more applicable for signals produced with devices that have various detector sensitivity functions. It also prevents from situations when the same set of equipment goes out of calibration.

○ *Correction of signal intensity baseline's offset.* An offset is a shift in signal intensity's baseline away from zero level. Usually, this distortion is corrected automatically before measurements of the fluorescence spectra by the software of a spectrometer (detector). The procedure is called nulling (reference signal is registered before the measurements and subtracted from the registered signal of an object), and the spectra in the initial dataset are usually offset-corrected. If this nulling was not carried out correctly, the correction of offset could be conducted after the noise, and equipment-related signal shape distortions are removed. In this case, it is needed to find a region in which level of the signal should be around zero and subtract this level from each channel of the measured spectrum.

○ *Normalization of spectra intensity.* Due to the nature of the investigated object (exact composition and concentration of fluorophores are unknown) and scheme of signal acquisition (the position of the probe is not fixed), it is unreasonable to compare the absolute intensities of spectra. Thus, the next step is to normalize each spectrum.

Traditionally, there are at least several ways to normalize spectroscopic signals such as to normalize to a maximum of spectra, to normalize to a particular wavelength, to normalize to mean intensity, to normalize to the area under the spectroscopic curve. This choice depends on peculiarities of data.

The way of normalization, which gives the best performance in the classificatory analysis, should be selected.

○ *Making additional spectroscopic features.* Spectroscopic features for classificatory analysis can be provided in several forms. The most straightforward way is to analyze spectra in the form of signal intensities at each wavelength. Still, spectra can be processed even further.

For instance, first-order gap-segment derivatives [10] ($\partial I/\partial\lambda$), where I is intensity and λ is wavelength) of spectroscopic signal intensities can be calculated for each spectrum. The contributions of different fluorophores highly overlap, thus, if relative intensity of adjacent fluorophore signals changes, this change may be more obviously reflected in the first-order derivative of a spectroscopic signal than in spectroscopic signal itself.

Moreover, various weighting functions may be applied to enhance the influence of some spectral ranges and to suppress the influence of others and in this way to reveal certain peculiarities of the spectra. Ratios between intensities at specific wavelengths can be calculated as an additional spectroscopic feature as well. Selection of these wavelengths can be based on prior knowledge about spectroscopic properties of a system of investigation or based on the exploratory analysis (e.g., various summary plots of spectra divided or not divided into classes according to other non-spectroscopic features).

○ *Create new non-spectroscopic and mixed features.* As the primary purpose of the analysis is to find subsets of features that lead to the best agreement between spectroscopic and medical information, relevant spectroscopic features, as well as relevant features of other kind, are needed to fulfilling this purpose. E.g., a new variable with derived class "pathology" that contains several certain pathologies may show a higher degree of association with spectroscopic data than a variable with only original medical classes. Thus, in parallel to pre-processing of spectroscopic data, the pre-processing of other non-spectroscopic features also takes place. The result of this pre-processing is either (i) modified versions of current features (*example 1A*: several pathologies of the same feature can be merged into one new class; *example 1B*: values of numeric variable such as "age" can be grouped to form a factor variable "age groups") or (ii) new mixed features that combine values of several spectroscopic and non-spectroscopic features (*example 2A*: a new feature that combines results of several medical tests to form classes "no pathology diagnosed" vs. "at least one pathology found"; *example 2B*: a new feature that combines values of age groups, results of medical tests and spectroscopic features such as ratio between particular intensities).

The investigator is the one who decides about the necessity of these features depending on the problem that should be solved and on the obtained performance of classification. If possible, less complicated and more easily interpreted features are preferred.

○ *Features in the pre-processed dataset.* To sum up, the pre-processed dataset will contain several types of features: medical, patient-related and specimen-related features and features derived from these non-spectroscopic data, spectroscopic features and features derived from spectroscopic data as well as mixed spectroscopic and non-spectroscopic features.

2.2.3 Classification rules

In this chapter, box A2 in Fig. 2 will be explained in more detail.

○ *ROC analysis and suitable classification performance measures.* In this study, we decided to carry out a classificatory analysis at each wavelength separately. To make a prediction, we need to select a threshold value. We can choose any value of fluorescence intensity as a threshold value, but we need an optimal one. Hence, to do the optimization, we chose a simple algorithm which results can be easily interpreted. The algorithm is called a *receiver operating characteristic* (ROC) [9].

In ROC analysis, a separate 2×2 classification table (confusion matrix) can be calculated for each threshold value. Based on this table, various performance measures may be calculated, including sensitivity (true positive rate, recall, Se), specificity (true negative rate, Sp), positive predictive value (precision, PPV), negative predictive value (NPV), Youden's index (J), F1 score, accuracy, balanced accuracy [1, 6], Cohen's kappa (κ) [13]. Thus, each threshold may be associated with several performance measures. A performance measure that is convenient for optimization should summarize all information that is in the classification table and take into account the issue of class size imbalances. Therefore, indexes κ, J and some other measures would fulfill these obligations. Nevertheless, we chose to use a balanced accuracy (BA), which can be defined as an average of diagnostic sensitivity and specificity [3, 6]:

$$\mathrm{BA} = \frac{\text{sensitivity} + \text{specificity}}{2}. \tag{1}$$

In classical version of ROC analysis, performance measure called area under the ROC curve (AUC) may also be calculated. One can claim, that AUC evaluates the overlap of numeric values of two classes. Classification tables at all possible threshold values are needed to calculate the AUC. Unfortunately, due to the nature of this calculation, AUC cannot be used to determine the optimal threshold. Thus, it was not included in our analysis as a suitable performance measure.

○ *The algorithm to create classification rules.* The algorithm that produces classification rules (box A2 in Fig. 2) is drawn in Fig. 5. Certain aspects of the algorithm will be described in this and several following sections.

The input for the algorithm of the classificatory analysis is the pre-processed dataset augmented with new features. Firstly, the algorithm takes a subset of data. It can involve only a particular age group, specific pathologies or certain values of derived features as well as the whole dataset with only missing values removed. The spectroscopic data in the subset can contain either spectra or derivatives of those spectra. Secondly, the algorithm selects one categorical feature and classes in this feature are treated as a reference for spectroscopic evaluation. This categorical variable can contain either (i) the results of the medical test, or (ii) derived non-spectroscopic classes or (iii) mixed classes based on both spectroscopic and non-spectroscopic features. The purpose of this procedure is to find non-spectroscopic and mixed features (and their values) that have the best agreement with spectroscopic information. Note: each pair of classes is investigated separately.

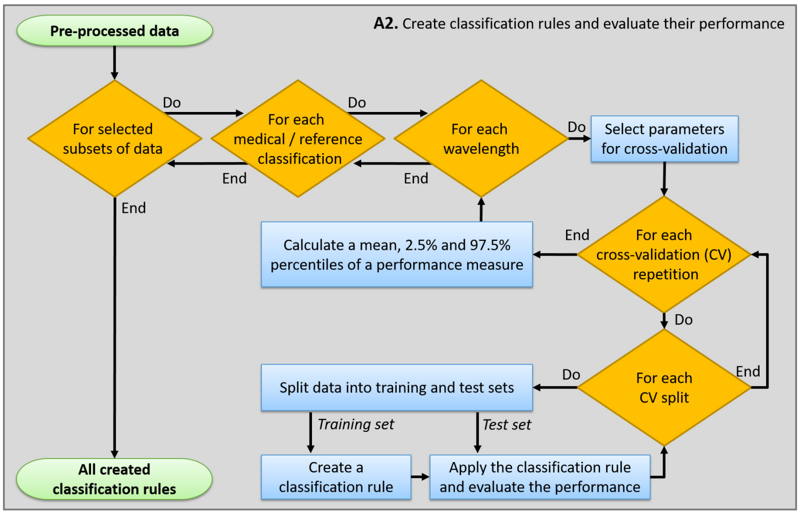
A2. Create classification rules and evaluate their performance
Pre-processed data
For selected subsets of data
Do
End
For each medical / reference classification
Do
End
For each wavelength
Do
Select parameters for cross-validation
End
For each cross-validation (CV) repetition
Calculate a mean, 2.5% and 97.5% percentiles of a performance measure
Do
For each CV split
End
Do
Split data into training and test sets
Training set
Test set
Create a classification rule
Apply the classification rule and evaluate the performance
All created classification rules

Figure 5. The algorithm to create classification rules.

Thirdly, the algorithm does the ROC analysis for intensities of spectroscopic signals at each wavelength separately. This procedure aims to aid finding wavelength ranges in spectroscopic data, which contain diagnostically most relevant information. Fourthly, the algorithm applies procedures of repeated k-fold cross-validation to split data into training and test sets and to create and evaluate classification rules based on these sets. As a visual aid, for each set of wavelengths, the result can be plotted as a performance spectrum, and the example is in Fig. 6(C).

○ *ROC analysis based procedures to create a classification rule and to assess its performance.* Let us describe the procedure how a classification rule is created and evaluated. The first step of the ROC analysis is applied on the training set. A threshold value (also known as a cut-off point), which maximizes BA criterion (mean of sensitivity and specificity), is calculated. This value separates values of spectra intensities into two classes: observations, which have values above the threshold, are classified as a class one and those, which have values below the threshold, are classified as the other class. These new classes are given the same labels as factor variable with reference classes has. A 10% trimmed means of each class are used for class name matching: if a reference class has label “A” and mean above the threshold, while the other reference class has label “B” and mean below the threshold, then all the cases with values above the threshold will be classified as “A” and all the remaining – as class “B”. Our classification rule is this determined threshold value and associated reference classes. It is passed to the next step.

The second step of the ROC analysis is to apply the rule on the test set with the purpose to evaluate the performance of classification rule out of sample. Thus, calculated threshold value is applied to classify the test data into classes, the classes created in the ROC analysis are compared to the reference classes, and degree of agreement is evaluated as a value of performance measure (BA). This BA value is stored for further calculations.

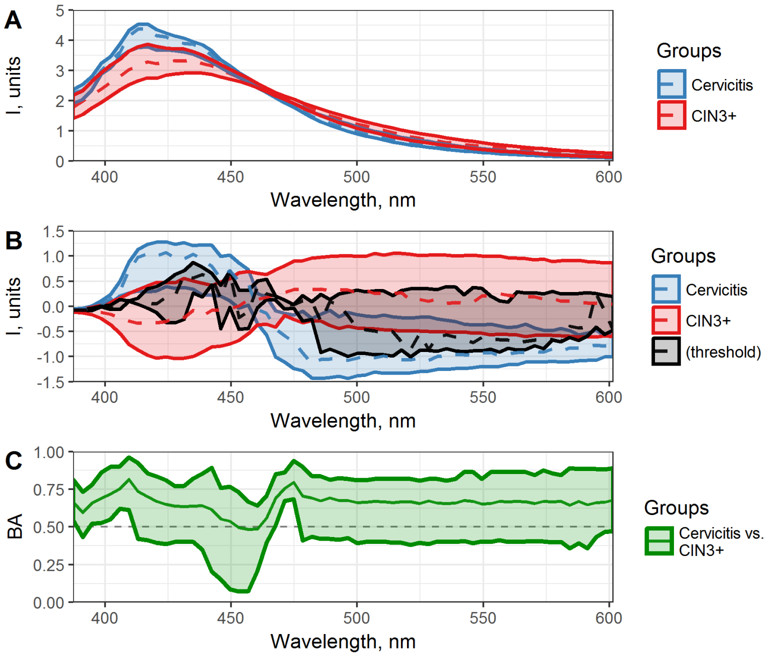

Figure 6. Statistically summarized spectra in Cervicitis and precancerous CIN3+ classes of feature *HistGr1* in the subset of whitish samples. "I, units" – intensity in relative units. (Online version in color.)

Additional performance measures can be calculated too and treated accordingly, but for the sake of simplicity, we will describe only BA in the following text. The procedures of ROC analysis are repeated in the subsequent repetitions of cross-validation, and a vector of BA values is collected.

○ *Cross-validation, stratification, blocking and confidence intervals.* To validate the results, in our calculations, we used m times repeated k-fold cross validation (CV) with stratification and blocking. In each repetition for each fold of test data, a separate estimate of BA value was calculated as described previously. The average of these $m \times k$ BA values was used as the main measure to evaluate the performance of the classification rule. The 2.5% and 97.5% percentiles of BA values, as well as difference of these two values (it was called a 95% width of BA values), were calculated to evaluate the spread of BA values.

To make us clear about terminology: in this paper, the term "stratification" means that proportions of some specimens in learning and test subsets were as similar as possible. The term "blocking" means that spectra of the same sample were treated as one block and all spectra of this block were assigned either only to training or only to test subset.

It should be mentioned, that if the dataset is very small, procedures of repeated k-fold cross-validation may be impossible. In these situations, leave-one-out cross-validation [18] may be a more appropriate method. On the one hand, if cross-validation is not applied at all, the results are more optimistic than they should be in reality and they have limited value in medical diagnostics due to lack of reliability. On the other hand, they may be used as a part of the exploratory (not predictive) analysis and help to define ideas and to allow raising questions that could be answered in future investigations with a larger amount of specimens.

2.3 Methodology to select the most appropriate classification rules

In this phase of analysis (box A3 in Fig. 2), there is a list of all created classification rules (Fig. 7). The number of rules is excessive. Furthermore, some of the rules are useful, but others are worthless. These rules are evaluated with BA values. Thus, the first step in selecting relevant rules is determining if 2.5% percentile of BA value is lower than or equal to 0.5 and if the mean of BA values is lower than 0.75. If at least one answer is "yes", the rule is eliminated. Next, the most appropriate rule per set of wavelengths (i.e., per subset of data, per selected reference classification, per pair of compared classes and per type of spectroscopic feature – spectrum, derivative, set principal components or ratios, etc.) is the rule associated with the highest mean BA value. As an additional part of the analysis, an expert investigates the plot of BA values (the performance spectrum, e.g., Fig. 6(C)) and determines the most appropriate wavelength ranges visually, i.e., finds the ranges around the best rules in which 2.5% percentile of BA values is above 0.5. The ranges with narrowest 95% widths should also be investigated more carefully as the results in these ranges are most reliable.

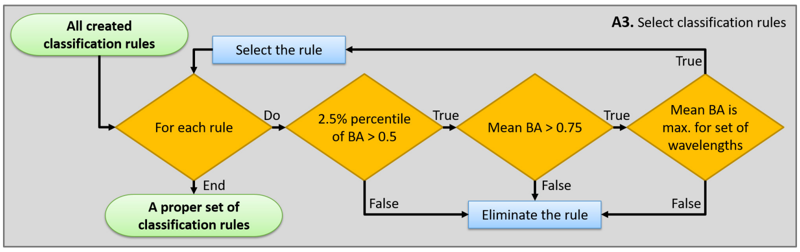
A3. Select classification rules
All created classification rules
Select the rule
True
For each rule
Do
2.5% percentile of BA > 0.5
True
Mean BA > 0.75
True
Mean BA is max. for set of wavelengths
End
A proper set of classification rules
False
False
Eliminate the rule
False

Figure 7. The algorithm to select the most appropriate classification rules. BA – balanced accuracy.

2.4 Prediction making

In the phase of prediction making (Fig. 2(B)), the rules of classification are already prepared and selected. When a new fluorescence spectrum is present and a particular rule is applicable to the situation, at first, the spectrum is pre-processed. Then an appropriate wavelength is selected and compared to the threshold value, and the spectrum (as well as the point of the specimen this spectrum is registered from) is assigned to a particular class in the same manner as described in the section about the ROC analysis and procedures how the rules are created.

It should be stressed that pieces of non-spectroscopic information, such as the age of a patient, may also be required in addition to fluorescence spectra to make a more precise prediction.

3 Case study: uterine cervix smears' autofluorescence spectra analysis

3.1 Software

From the analytical point of view, spectroscopic signals are multidimensional highly correlated data. Data analysis can be implemented in a freely available software called R [17], which is a programming language and an environment specialized for data analysis, visualization and statistical modeling. Functions from R packages ("hyperSpec" – for spectroscopic data management, "signal" – for signal-type data processing, "dplyr" – for data frame like data management, and other), as well as the code developed by our research group, took the key role in the analysis.

3.2 Samples and data acquisition

Lithuanian Bioethical Committee approved this investigation, and the investigation is in accordance with bioethical requirements.

The objects of this investigation are medical liquid-based smears of the uterine cervix (UC) in various medical conditions. The biological material of the research was collected in Vilnius University Hospital Santaros Klinikos (Vilnius, Lithuania). Further investigation of specimens, including spectroscopic measurements and data analysis, took place in Institute of Photonics and Nanotechnology (previously called Institute of Applied Research), Vilnius University (Vilnius, Lithuania).

The physical phenomenon we were measuring was autofluorescence: a type of fluorescence caused by only internal properties of materials with neither auxiliary substances nor labels added. The methods of spectroscopic measurements are similar to those described in [21] and [8]. Sediments of smears were deposited on a single-crystal silicon wafer (tray) and dried. This material of the tray was selected as it is not transparent in ultraviolet-visible range and does not fluoresce. Spectroscopic measurements were carried out through a fiber-based optical system. Ultraviolet 355 nm radiation (laser *STA-01-TH*, *Standa Ltd.*, Vilnius, Lithuania) was used to excite intrinsic fluorescence of samples, and spectrometer *OceanOptics USB2000* (*Ocean Optics, Inc.*, Dunedin, FL 34698, USA) was used as a detector to register spectroscopic signals. At least five spectra from different parts of the same specimen's surface were registered. To avoid additional procedures in signal pre-processing phase, spectra were nulled during the process of spectra registration. In addition to spectra of samples, a spectrum of calibrated *Bentham* light source with current-stabilized power supply *Bentham 607* (*Bentham Instruments, Ltd.*, Berkshire, United Kingdom) was registered to be used for spectra intensity corrections.

3.3 The dataset

The dataset, which we demonstrate the application of the method on, consisted of the following variables. *Spectra* – a numeric vector of spectroscopic signal intensities. *Derivatives* – first-order gap-segmented derivatives of spectroscopic signal intensities. *Color*

Table 1. Distribution of samples and spectra in the subset of whitish specimens.

Classification	Classes	Number		Percentage	
		Samples	Spectra	Samples	Spectra
HistGr1	Normal*	1	3	2.4%	1.5%
	Cervicitis	6	32	14.6%	15.7%
	CIN1*	2	12	4.9%	5.9%
	CIN2*	3	13	7.3%	6.4%
	CIN3+	23	115	56.1%	56.4%
	(Missing)	6	29	14.6%	14.2%
Hybrid-cervicitis	Cervicitis	6	32	14.6%	15.7%
	Other	35	172	85.4%	84.3%
Hybrid-cervicitis ($<$ 45 y.o.)	Cervicitis	4	22	14.3%	15.8%
	Other	24	117	85.7%	84.2%

*Sample sizes of these classes were too small and the classes were not included in the further analysis (see Table 2). $<$ 45 y.o. – patients under 45 years old only.

– a color of a specimen evaluated by an investigator as either whitish or non-whitish (e.g., yellowish, reddish, brownish). *Age* – the age of a patient on January 1st of the year when the cervical smear was taken. *HistGr1* – a medical feature defined by results of the histological examination. It is a factor variable with classes: Normal, Cervicitis (inflammation), pre-cancerous conditions CIN1, CIN2, CIN3+ (cervical intraepithelial neoplasia of grade 1, grade 2 as well as a mixture of CIN3 and carcinoma in-situ).

As we aim to search for a combination of features (medical, spectroscopic, mixed or other) that optimizes BA, additional features may lead to fulfilling this purpose. For example, if a result cytological test is negative, i.e., no pathology found, the histological test is not performed due to associated medical risks. Thus, some cases in *HistGr1* have missing values (see Table 1). *Hybrid-cervicitis* is a derived medical feature based on histological classification with the missing values filled according to the results of the cytological analysis. This feature is designed to compare spectra of cervicitis with all other conditions and consist of the following classes: Cervicitis and Other.

In this paper, we present only the results in the subset of samples that did not have expressed absorption of materials such as hemoglobin (i.e., their color was recognized as whitish) as this subset did not require any additional steps of pre-processing that have to take the influence of the absorption into account. For feature *Hybrid-cervicitis*, in addition to the set of all age groups, a subset of younger patients (under 45 years old) was investigated. There were insufficient patients to form a subset of older (above 45 years old) patients.

3.4 Pre-processing of the dataset

Running medians filter (width of the window is 11) and Savitsky–Golay filter (order of the polynomial is 9, the width of the window is 35) were applied, and ten adjacent points were binned (averaged) to reduce noise and dimensionality of data. We did not have the purpose of optimizing window width and other parameters of filters. However, values were chosen seeking to reduce noise without distorting the trends (see Fig. 4). Usually, five spectra per specimen were collected, but due to low intensity, some spectra were not suitable for further analysis, and they were excluded (see Table 1). Therefore, in some cases, there left less than five spectra per sample. The criterion for the exclusion: the maximum of the signal between 380–800 nm was below 2.5% of maximum in spectrometer's scale.

Next, the spectrum of calibrated light source "Bentham" was used as a benchmark to remove equipment-related signal shape deformations. Mean intensities between 410–700 nm of each spectrum were calculated and used to normalize spectra. As an additional spectroscopic feature, first-order gap segmented derivatives (gap size is 3 points, the segment size is 3 points) were calculated. Spectroscopic ranges above 380 nm, where the signal is uninformative due to equipment, and below 600 nm, where the signal is uninformative due to low intensity, were removed from further analysis. Both intensities and the derivatives of those intensities are numeric vectors of spectroscopic data. The pre-processing lowered the number of different wavelengths from 2048 per spectrum before pre-processing to 60–70 after pre-processing.

Finally, several new non-spectroscopic features (including *Hybrid-cervicitis*) were created. These features are described in the previous section.

3.5 Analysis of the dataset

In the phase of data analysis, a subset of only whitish specimens was selected. The number of samples $n = 41$. These specimens did not have visually clearly expressed absorption band in the visible part of the spectrum, and it was easier to relate spectroscopic features of this subset of samples to the fluorophores in the tissue. Agreement between classes in features *HistGr1* and *Hybrid-cervicitis* and spectroscopic information was investigated during ROC analysis at each wavelength. The distribution of samples and spectra in the examined subset of whitish specimens are presented in Table 1. The classes with only three or less unique medical specimens (Normal, CIN1, CIN2) as well as cases with no diagnosis (missing) were excluded from the analysis, therefore, Cervicitis vs. CIN3+ (in *HistGr1*) and Cervicitis vs. Other (in *Hybrid-cervicitis*) were used as the reference classifications. Both the mean-normalized spectra and their derivatives were analyzed. A 10-time repeated 3-fold cross-validation (with blocking and stratification) was applied. To ensure that the same spectra were included in the same folds throughout the whole analysis, seed for random number generator was reset to the same number before using each pair of compared classes. The criterion for the best agreement is maximum BA value.

4 Results

Classification rules performance measures are presented in Table 2. In the table, the values of wavelengths are rounded to whole numbers, the values of performance measures – to 2 decimal places, threshold – to 2 decimal places. The values with ± symbol indicate the mean value (in bold) and the standard deviation per cross-validation resamples. The threshold is optimized for the BA value, other performance measures (Cohen's κ, Youden's index J, sensitivity, specificity, positive and negative predictive values – PPV and NPV) are just calculated. Table shows at most one wavelength (per set of compared classes per type of spectrum) at which mean BA value is the highest. "BA$_{2.5\%}$"

Table 2. Examples of decision rule performance for mean-intensity-normalized spectra in the "whitish" specimens.

	Compared classes	Feature	Derivative	λ, nm
1	Cervicitis vs. **CIN3+**	HistGr1	No	410
2	Cervicitis vs. **CIN3+**	HistGr1	Yes	435
3	Cervicitis vs. **Other**	Hybrid-cervicitis	Yes	435
4	Cervicitis vs. **Other** ($<$ 45 y.o.)	Hybrid-cervicitis	Yes	435

	Cohen's κ	BA	$\text{BA}_{2.5\%}$	$\text{BA}_{97.5\%}$	$\text{BA}^{95\%}_{\text{width}}$	Youden's J
1	$\mathbf{0.46} \pm 0.23$	$\mathbf{0.82} \pm 0.11$	0.61	0.96	0.35	$\mathbf{0.63} \pm 0.21$
2	$\mathbf{0.54} \pm 0.18$	$\mathbf{0.85} \pm 0.06$	0.76	0.97	0.21	$\mathbf{0.71} \pm 0.13$
3	$\mathbf{0.34} \pm 0.13$	$\mathbf{0.80} \pm 0.08$	0.65	0.90	0.24	$\mathbf{0.59} \pm 0.16$
4	$\mathbf{0.38} \pm 0.16$	$\mathbf{0.82} \pm 0.07$	0.72	0.93	0.21	$\mathbf{0.64} \pm 0.13$

	Sensitivity	Specificity	PPV	NPV	Threshold
1	$\mathbf{0.66} \pm 0.15$	$\mathbf{0.97} \pm 0.18$	$\mathbf{0.99} \pm 0.05$	$\mathbf{0.47} \pm 0.18$	$\mathbf{3.49} \pm 0.11$
2	$\mathbf{0.71} \pm 0.13$	$\mathbf{1.00} \pm 0.00$	$\mathbf{1.00} \pm 0.00$	$\mathbf{0.52} \pm 0.15$	$(\mathbf{-8.36} \pm 0.08) \cdot 10^{-2}$
3	$\mathbf{0.62} \pm 0.11$	$\mathbf{0.97} \pm 0.15$	$\mathbf{0.99} \pm 0.03$	$\mathbf{0.33} \pm 0.08$	$(\mathbf{-8.45} \pm 0.49) \cdot 10^{-2}$
4	$\mathbf{0.64} \pm 0.13$	$\mathbf{1.00} \pm 0.00$	$\mathbf{1.00} \pm 0.00$	$\mathbf{0.36} \pm 0.14$	$(\mathbf{-8.36} \pm 0.09) \cdot 10^{-2}$

Legend: class name in bold is treated as *positive* $<$ 45 y.o. – a subset of patients under 45 years old; numeric values are written as mean ± standard deviation; BA – balanced accuracy; $\text{BA}_{2.5\%}$, $\text{BA}_{97.5\%}$ – a 2.5% and a 97.5% percentiles of BA; $\text{BA}^{95\%}_{\text{width}}$ – difference between 97.5% and 2.5% percentiles; PPV and NPV – positive and negative predictive values respectively; λ – wavelength.

and "$\text{BA}_{97.5\%}$" represent 2.5% and 97.5% percentiles of BA values obtained during the process of repeated cross-validation. Whereas column "$\text{BA}^{95\%}_{\text{width}}$" shows the difference between the percentiles, this difference indicates uncertainty of BA value. The interpretation of BA values and the difference should be as follows: the higher value in "BA" column is, the more accurate classification is. The lower value in "$\text{BA}^{95\%}_{\text{width}}$" column is, the more reliable BA value is. Class name, which is in bold, is treated as the positive case. This means that both specificity and NPV in this table is always related to the class "Cervicitis".

The results of row 1 in Table 2 are expanded and drawn in Fig. 6. Here the examples of wavelength-wise statistically summarized spectra of Cervicitis vs. precancerous CIN3+ classes (feature *HistGr1*) in the subset of whitish samples are presented. Figures 6(A) and (B) demonstrate a median (dashed line), first and third quartiles (solid lines) of preprocessed spectra. These sub-plots give the same information as a box in a box-and-whisker plot would provide for univariate data. The areas between the first and the third quartiles are shaded. The less overlapped ranges of classes Cervicitis and CIN3+ are, the better classification performance can be expected. Highly overlapping areas indicate that they are not suitable for discrimination of classes. The difference between Fig. 6(A) and (B) is that values in B are transformed by mean-centering and scaling intensities at each wavelength. Additionally, threshold values obtained in ROC analysis are added. Figure 6(C) shows mean (solid line in the middle of shaded area), 2.5% and 97.5% percentiles of balanced accuracy values obtained in ROC analysis. The area between the percentiles is shaded.

Table 2 indicates that the highest result was obtained comparing Cervicitis vs. CIN3+ (in *HistGr1*) using spectroscopic derivatives at 435 nm ($\text{BA} = 0.85 \pm 0.06$). The mean BA value obtained using non-derivatives is a bit lower but comparable (BA $= 0.82\pm0.11$), but the wavelength is different (410 nm). The further graphical analysis provides us with more insights. The spectroscopic features in spectral ranges, where the lower percentile is above the dashed line at BA $= 0.5$ (Fig. 6), are acceptable for classification. In the plot, there are two areas potentially suitable for Cervicitis vs. precancerous CIN3+ class discrimination. The range around 410 nm is the same as in Table 2, but there is another around 470 nm. Thus, besides the numerical analysis, plots are also useful. Moreover, spectral ranges around 410, 435 and 470 nm are noteworthy to finding out possible biological reasons for differences in spectra.

5 Discussion

In this research, the attention was focused on a fluorescence spectroscopy-based method due to the simplicity of devices for the signal registration and the potential to use these devices at a point of medical care. The data processing scheme was proposed to account for the peculiarities of fluorescence signal acquisition and creating a dataset using the pathology indications obtained by conventional medical tests. The possibility to apply the principles of data processing and diagnostic decision making for spectroscopic signals of medical objects (sometimes called “optical biopsy”) is pointed out.

In particular, we presented the approach to investigate intrinsic fluorescence spectra of medical samples, and the strength and novelty of the presented methods lie in the multidisciplinary area of application. The approach involves steps of data pre-processing and the ROC analysis based method to create classification rules as well as to evaluate their performance. The proposed method is flexible, and it is possible to extend it by using other performance measures, which were not calculated in this investigation. Employment of mixture of features, which vary in nature, is one of the novelty points in this research too.

The presented method creates classification rules at each available wavelength of the spectroscopic signal. The most appropriate rules are selected automatically. Moreover, the performance can be plotted as a function of wavelength. These two approaches (automatic and graphical selection) serve as a tool to reveal ranges of spectroscopic signal that are most relevant for medical diagnostics. The higher the performance (e.g., BA value) is, the more diagnostically useful the range is. Therefore, the method can be used either to create rules that serve directly for univariate binary classification tasks or as a way to highlight the most significant spectroscopic features and in this way facilitate biological and medical interpretation of signals’ origin and usefulness. Comparing to our previous pilot study [20], the presented method is more in alignment with the requirements of practical application due to more automatic calculations. Moreover, spectroscopic features chosen as the most relevant ones can be passed to more sophisticated classification algorithms that solve more complicated multivariate multi-class classification problems. The latter statement is out of the scope of this paper and gives the basis for future investigations.

In the case study, the analysis of cervical samples’ dataset is presented. The BA values above 0.75 for discrimination of both cervicitis vs. CIN3+ and cervicitis vs. other pathologies were achieved. Analysis of both spectra intensities and their derivatives showed relevance for medical diagnostics. In the present research, the object of the investigation was liquid cervical smears in medically approved conservation liquid and not tissues *in-vivo*, this is also one of the novelty points. As we used samples *in-vitro* and the results show high predictive value for non-cervicitis group (this result is in agreement with [20]), it suggests method's perspectives both in cytological laboratories when it is needed to confirm a non-cervicitis diagnosis and for cervical screening.

The spectral ranges, which are identified as relevant, are also biologically sensible as, e.g., at 410 nm biological materials such as elastin and collagen fluoresce. More information on possible fluorophores may be found in [5].

It should be noted that the research is an exploratory study, where data sample sizes were limited due to specifics of medical nature of the data, bioethical as well as patients' data protection requirements. However, positive results of this research allow possibilities for larger scale data collection. Furthermore, uterine cervix dataset was used as an illustration only. The method can be easily adapted for other types of specimens such as intervertebral disc degeneration related samples, endometrial washing samples from women's womb and other. More on the applications of the method in the field of medical diagnostics will be presented elsewhere shortly.

6 Conclusions

1. The procedures of ROC-based spectroscopic signals' analysis were implemented, and they are eligible to identify ranges of fluorescence spectra that are most relevant for diagnostic decision making.
2. The fluorescence spectra based method could be applied for cervicitis vs. other medical conditions' (e.g., CIN3+) diagnostics in whitish specimens, and the data analysis method has the capability to be applied to identify a broader range of medical conditions.

Acknowledgment. The authors would like to acknowledge the reviewers for valuable remarks and the proposed ideas to improve the quality of the paper.

References

1. A. K. Akobeng, Understanding diagnostic tests 3: Receiver operating characteristic curves, *Acta Paediatr.*, **96**(5):644–647, 2007.

2. M.J. Bilinskas, G. Dzemyda, M. Trakymas, Feature-based registration of thorax CT scan slices, *Informatica*, **28**(3):439–452, 2017.

3. K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The balanced accuracy and its posterior distribution, in *20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August, 2010*, IEEE, 2010, pp. 3121–3124.

4. A.C. Croce, G Bottiroli, Autofluorescence spectroscopy and imaging: A tool for biomedical research and diagnosis, *Eur. J. Histochem.*, **58**(4):2461, 2014.

5. T. Dramićanin, M. Dramićanin, Using fluorescence spectroscopy to diagnose breast cancer, in *Applications of Molecular Spectroscopy to Current Research in the Chemical and Biological Sciences*, InTech, London, 2016.
6. T. Fawcett, An introduction to ROC analysis, *Pattern Recogn. Lett.*, **27**(8):861–874, 2006.
7. L. Gao, R.T. Smith, Optical hyperspectral imaging in microscopy and spectroscopy – a review of data acquisition, *J. Biophotonics*, **8**(6):441–456, 2015.
8. V. Gėgžna, P. Sladkevičius, L. Valentin, A. Vaitkuvienė, Methods for autofluorescence analysis of uterine cavity washings, *Lith. J. Phys.*, **55**(1):63–70, 2015.
9. K. Hajian-Tilaki, Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation, *Casp. J. Intern. Med.*, **4**(2):627–35, 2013.
10. D. Hopkins, Shoot-out 2002: Transfer of calibration for content of active in a pharmaceutical tablet, *NIR News*, **14**(1):10, 2003.
11. Y. Jusman, S.C. Ng, N.A. Abu Osman, Intelligent screening systems for cervical cancer, *Sci. World J.*, **2014**:810368, 2014.
12. R. Kramme, K.-P. Hoffmann, R.S. Pozos (Eds.), *Springer Handbook of Medical Technology*, Springer, Berlin, Heidelberg, 2011.
13. R. Kwiecien, A. Kopp-Schneider, M Blettner, Concordance analysis: Part 16 of a series on evaluation of scientific publications, *Dtsch. Arztebl. Int.*, **108**(30):515–21, 2011.
14. V. Medvedev, O. Kurasova, Cloud technologies: A new level for big data mining, in F. Pop, J. Kołodziej, B. Di Martino (Eds.), *Resource Management for Big Data Platforms. Algorithms, Modelling, and High-Performance Computing Techniques*, Springer, Cham, 2016, pp. 55–67.
15. V. Medvedev, O. Kurasova, J. Bernatavičienė, P. Treigys, V. Marcinkevičius, G. Dzemyda, A new web-based solution for modelling data mining processes, *Simul. Model. Pract. Theory*, **76**:34–46, 2017.
16. M. Monici, Cell and tissue autofluorescence research and diagnostic applications, *Biotechnol. Annu. Rev.*, **11**:227–256, 2005.
17. R: A language and environment for statistical computing, 2017, `https://www.r-project.org/`.
18. C. Rushing, A. Bulusu, H.I. Hurwitz, A.B. Nixon, H. Pang, A leave-one-out cross-validation SAS macro for the identification of markers associated with survival, *Comput. Biol. Med.*, **57**:123–129, 2015.
19. R.K. Sahu, S. Mordechai, Spectroscopic techniques in medicine: The future of diagnostics, *Appl. Spectrosc. Rev.*, **51**(6):484–499, 2016.
20. A. Vaitkuviene, V. Gegzna, R. Kurtinaitiene, J.V. Vaitkus, Cervical smear photodiagnosis by fluorescence, *Photomed. Laser Surg.*, **30**(5):268–274, 2012.
21. D. Varanius, G. Terbetas, J.V. Vaitkus, A. Vaitkuviene, Spinal hernia tissue autofluorescence spectrum, *Lasers Med. Sci.*, **28**(2):423–30, 2013.