Artificial Neural Network in Predicting Cancer Based on Infrared Spectroscopy

Yaniv Cohen, Arkadi Zilberman, Ben Zion Dekel, and Evgenii Krouk

Abstract In this work, we present a Real-Time (RT), on-site, machine-learning-based methodology for identifying human cancers. The presented approach is reliable, effective, cost-effective, and non-invasive method, which is based on Fourier Transform Infrared (FTIR) spectroscopy—a vibrational method with the ability to detect changes as a result of molecular vibration bonds using Infrared (IR) radiation in human tissues and cells. Medical IR Optical System (IROS) is a tabletop device for real-time tissue diagnosis that utilizes FTIR spectroscopy and the Attenuated Total Reflectance (ATR) principle to accurately diagnose the tissue. The combined device and method were used for RT diagnosis and characterization of normal and pathological tissues ex vivo/in vitro. The solution methodology is to apply Machine Learning (ML) classifier that can be used to differentiate between cancer, normal, and other pathologies. Excellent results were achieved by applying feedforward backpropagation Artificial Neural Network (ANN) with supervised learning classification on 76 wet samples. ANN method shows a high performance to classify; overall, 98.7% (75/76 biopsies) of the predictions are correctly classified and 1.3% (1/76 biopsies) is wrong classification.

Y. Cohen (✉) · E. Krouk
National Research University Higher School of Economics, 20 Myasnitskaya ul, Moscow 101000, Russian Federation
e-mail: yanivcohen79@gmail.com

E. Krouk
e-mail: ekrouk@hse.ru

A. Zilberman
Ben Gurion University of the Negev, 8410501 Beer-Sheva, Israel
e-mail: arkadiz@gmail.com

B. Z. Dekel
Ruppin Academic Center, 4025000 Emek Hefer, Israel
e-mail: benziond@ruppin.ac.il

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 193, https://doi.org/10.1007/978-981-15-5925-9_12

1 Introduction

Tumor detection at initial stages is a major concern in cancer diagnosis [1–9]. Cancer screening involves costly and lengthy procedures for evaluating and validating cancer biomarkers. Rapid or one-step method preferentially non-invasive, sensitive, specific, and affordable is required to reduce the long diagnostic processes. IR spectroscopy is a technique routinely used by biochemists, material scientists, etc., as a standard analysis method. The observed spectroscopic signals are caused by the absorption of IR radiation that is specific to functional groups of the molecule. These absorption frequencies are associated with the vibrational motions of the nuclei of a functional group and show distinct changes when the chemical environment of the functional group is modified [2, 3]. IR spectroscopy essentially provides a molecular fingerprint and IR spectra contain a wealth of information on the molecule. In particular, they are used for the identification and quantification of molecular species, the interactions between neighboring molecules, their overall shape, etc. IR spectra can be used as a sensitive marker of structural changes of cells and of reorganization occurring in cells [2–9] and most biomolecules give rise to IR absorption bands between 1800 and 700 cm^{-1}, which are known as the "fingerprint region" or primary absorption region. The medical IROS device [2] relates to methods employing Evanescent Wave FTIR (EW-FTIR) spectroscopy using optical elements and sensors operated in the ATR regime in the MIR region of the spectrum.

Therefore, as recently shown, Fourier transform IR (FTIR) spectroscopy coupled with computational methods can provide fingerprint spectra of benign tissues and their counterpart malignant tumors with a high rate of accuracy [3].

Our aim was to use FTIR spectroscopy combined with machine learning methods for the primary evaluation of the characteristic spectra of colon and gastric tissue from patients with healthy and cancer tissue, thus creating a novel platform for the application of FTIR spectroscopy for real-time, on-site early diagnosis of colon cancer.

In Fig. 1, the circle of data transfer of patients' data to the medical IROS [2] for machine learning is presented, whereas Fig. 2 presents full coupled system of data transfer from each patient of different hospitals into the center of collection information and its decision made by medical personnel after analyzing results of machine learning occurring there.

The remainder of this paper is organized as follows. Section 2 presents a short summary of medical IROS. Basic definitions of ANN are discussed in Sect. 3. Section 4 provides a description of network training algorithms. Section 5 includes preliminary practical results. Section 6 concludes the paper.

Fig. 1 Circle of data transfer

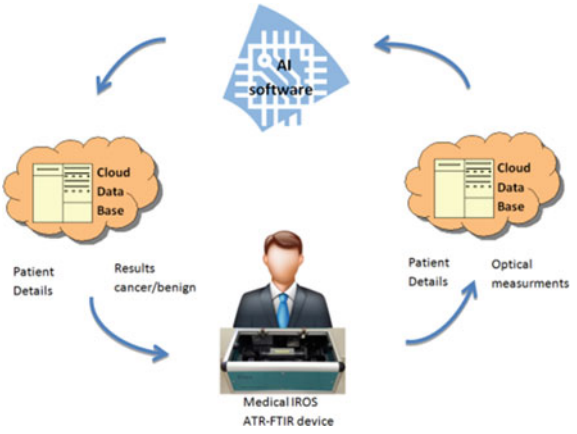
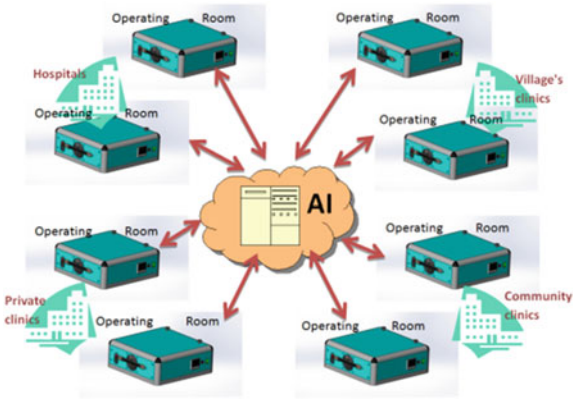


Fig. 2 Transfer of data from each patient and each hospital to the center of final decision



2 Short Summary of Medical IROS

The aim is to develop a dedicated combined apparatus suitable for biological tissue characterization via FTIR spectroscopic measurement during clinical practice [2]. According to the teachings of the device, it relates to combined device and method for the in vitro analysis of tissue and biological cells which may be carried out in a simple and, preferably, automated manner. The device and method produce result rapidly (up to minutes) and permit the determination/detecting of structural changes between a biological specimen and a reference sample.

In accordance with the teachings of the medical IROS, the human's tissue is applied to unclad optical element (crystal, etc.) working in ATR regime. A beam of mid-IR (infrared) radiation is passed through a low loss optical element and interacts with the tissue via ATR effect. In this process, the absorbing tissue is placed in direct

contact with the optical element. The novel combined apparatus (FTIR spectrometer with opto-mechanical elements and software) adopts an integrative design in appearance, and it is a bench top device.

3 ANN Concept—Basic Definitions

ANN [10] is the mathematical structure, which consists of interconnected artificial neurons that mimic the way a biological neural network (or brain) works. ANN has the ability to "learn" from data, either in a supervised or an unsupervised mode and can be used in classification tasks [11, 12]. In Multi-layer Feedforward (MLFF) networks, the neurons (nodes) are arranged in layers with connectivity between the neurons of different layers. Figure 3 is the schematic representation of a simple artificial neural network model. The artificial neurons have input values, which are the output product of other neurons or, at the initial level, the input variables (input $p = 1, 2, \ldots, n$). These values are then multiplied by a weight W and the sum of all these products (Σ) is fed to an activation function F. The activation function alters the signal accordingly and passes the signal to the next neuron (s) until the output of the model is reached. Each node is connected by a link with numerical weights and these weights are stored in the neural network and updated through the learning process.

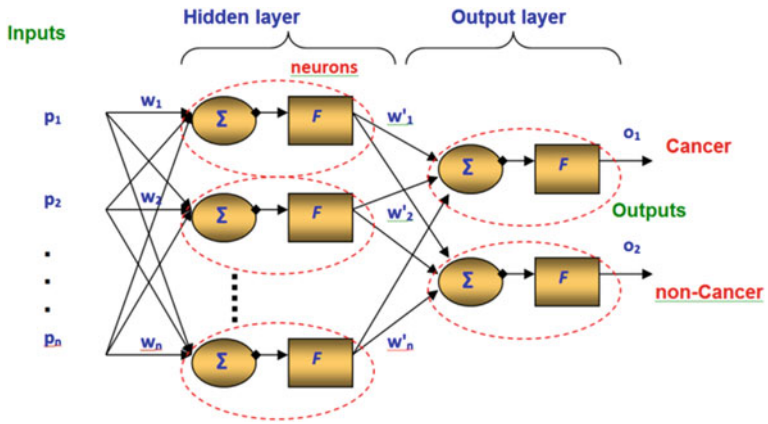


Fig. 3 Multi-layer feedforward network, $p_1, \ldots, p_n$ measured spectral signatures

4 Network Training Algorithms

Levenberg–Marquardt (LM) backpropagation method is a network training function that updates weight and bias values according to LM optimization. It is often the fastest backpropagation algorithm, and is highly recommended as a first-choice supervised algorithm, although it does require more memory than other algorithms.

LM algorithm is an iterative technique that locates a local minimum of a multivariate function that is expressed as the sum of squares of several non-linear, real-valued functions. The algorithm changes current weights of the network iteratively such that objective function, $F(w)$, is minimized as shown in Eqs. 1 or 2:

$$F(w) = \sum_{i=1}^{P} \sum_{j=1}^{M} \left(d_{ij} - o_{ij}\right)^2 \tag{1}$$

$$F(w) = EE^T \tag{2}$$

where $w = [w_1, w_2, \ldots w_N]^T$ is a vector of all weights, N is the number of weights, P is the number of observations or inputs (signatures), M is the number of output neurons, and d_{ij} and o_{ij} are the desired value ("target value") and the actual value ("predicted value") of the ith output neuron and the jth observation.

LM method is very sensitive to the initial network weights. Also, it does not consider outliers in the data, what may lead to overfitting noise. To avoid those situations, *Bayesian regularization* technique can be used.

5 Preliminary Practical Results

Acknowledgment: Data base presented in this paper with a special permission from P.I.M.S (PIMS LTD, Beer Sheva, Israel). The goal is to analyze the influence of ANN structure on the results of classification. After choosing the better structure, the performance of different ANN training methods was compared. Spectral data used for analysis are presented in Table 1.

Thereinafter, pre-processing, number of inputs selection, ANN design, training, testing, validation, and selection of training algorithms and optimal amount of neurons in hidden layer are discussed in Sects. 5.1–5.5, respectively.

Table 1 Spectral data used for analysis (~5.7–11 um waveband)

Spectral interval, cm^{-1}	Resolution, cm^{-1}	Number of spectral signatures
950–1750	4	200

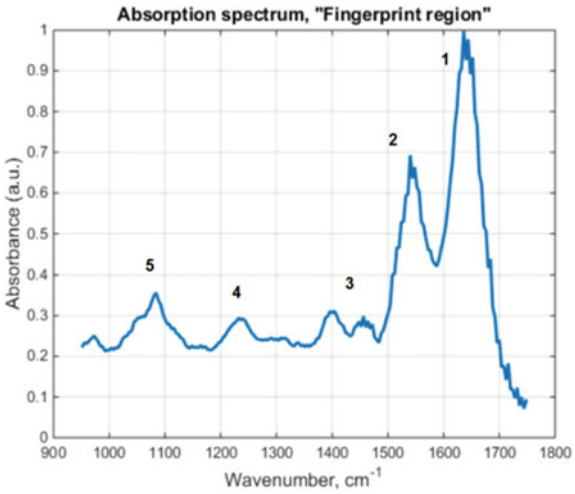


Fig. 4 Typical molecular absorption positions (molecular bonds and spectral signatures), where 1—protein Amide I, 2—protein Amide II, 3—lipids and protein (CH3), 4—phospholipids and Amide III, 5—PO2 phospholipids and nucleic acids. The strength of spectral signatures is changed depending on the tissue features/pathologies [2–4]

5.1 Pre-processing

The data is extracted and formatted in accordance with ANN demands:

(1) The measured FTIR-ATR signal is converted to a spectral absorbance $A(\lambda)$ defined by Eq. 3:

$$A_\lambda = -\log_{10}\left[\frac{I_\lambda - I_{dark,\lambda}}{I_{ref,\lambda} - I_{dark,\lambda}}\right] \tag{3}$$

where I_λ is the spectral intensity measured with the sample, $I_{ref,\lambda}$ is the reference signal (without sample) for source correction, $I_{dark,\lambda}$ is the dark counts, and λ is the wavenumber, cm^{-1}.

(2) Peak normalization (Fig. 4). The absorption spectrum $A(\lambda)$ is normalized by a maximal value at 1640 cm^{-1} (Amide I absorption) provided by Eq. 4:

$$Y(\lambda) = A_\lambda / A\left(1640\,\mathrm{cm}^{-1}\right). \tag{4}$$

(3) First derivative of the spectral absorbance is depicted in Fig. 5.

5.2 Number of Inputs Selection

Measured spectral signatures at given wavelengths, $p_n(\lambda)$, $n = 200$, are used as the inputs (input layer) to ANN. To reduce amount of inputs, the criterion of "min

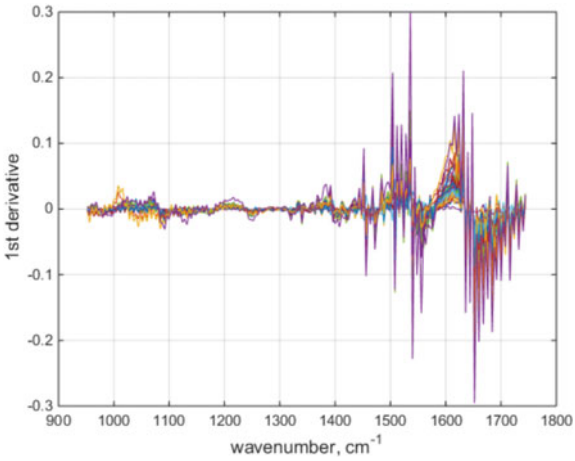


Fig. 5 The graph of first derivative of the spectral absorbance

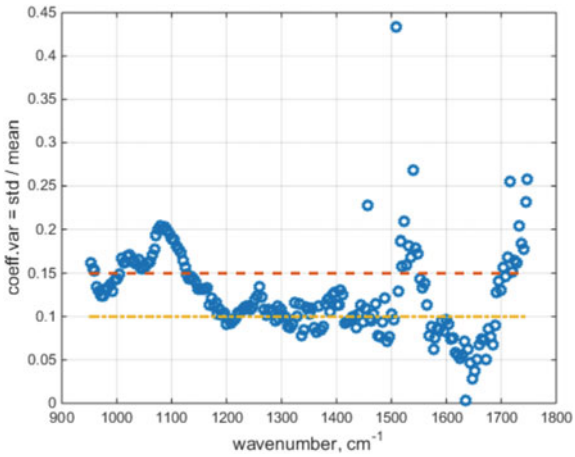


Fig. 6 The graph of CV

variance" was used. The variance and Coefficient of Variations (CV) were calculated at each wavelength for the data matrix [76 × 200]. Then the threshold was applied to the CV vector (Fig. 6):

$$CV = \frac{\sqrt{\text{var}(p)}}{\bar{p}} \geq \text{threshold}. \tag{5}$$

The spectral signatures at appropriate CV values (CV ≥ threshold) were used as the inputs to ANN.

Table 2 Dataset for ANN training and validation, dataset = 76 samples

Class labels (biopsy)	Count	Percent
Norm	72	94.74
Polyp	2	2.63
Cancer	2	2.63

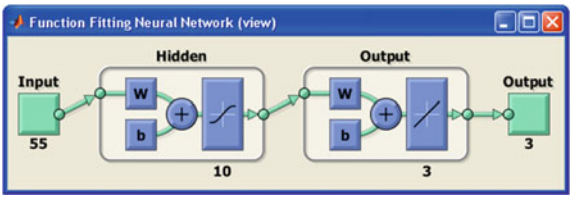


Fig. 7 Example of ANN structure for classification of the dataset with 10 hidden layer neurons and 3 output neurons: 55 spectral signatures (Input); 3 outputs: Cancer, Normal, Polyp

5.3 ANN Design

The data partitioning is the following: training set 60%, testing set 20%, and validation set 20%. The experimental data used for ANN models development are given in Table 2. 45 samples were used for training set and the rest are used for testing and validation (31 samples). The selection of data for training and testing was made in such a way that at least one sample of polyp and one sample of cancer will be in the training and testing sets.

The selected ANN structure is a three-layer feedforward, fully connected hierarchical network consisting of one input layer, one hidden layer, and one output layer. Different iterative backpropagation algorithms have been implemented to determine errors for the hidden layer neurons and subsequent weight modification. To define the number of neurons in the hidden layer of the network, Mean Square Error (MSE) and R^2 were analyzed. In order to avoid undesirably long training time, a termination criterion has been adopted. This criterion may be either completion of a maximum number of epochs (training cycles) or achievement of the error goal (Fig. 7).

5.4 Training, Testing, and Validation

The training stopped when the validation error is starting to increase (occurred at 15 training cycle (epoch) which is presented in Fig. 8.

To evaluate the performance of the network and indicate the error rate of presented model, statistical error estimation methods are used. The basic error estimation method is MSE provided by Eq. 6:

$$MSE = \sum_{t=1}^{n} \left(X_t - X_t^o\right)^2 / n. \tag{6}$$

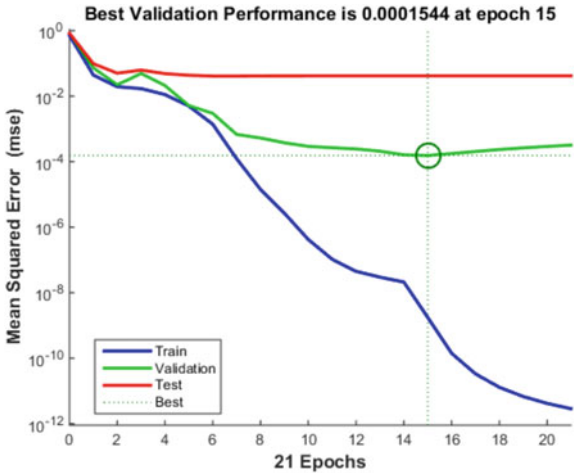


Fig. 8 Training, testing, and validation

5.5 *Selection of Training Algorithms and Optimal Amount of Neurons in Hidden Layer*

The best results obtained with LM training algorithm are presented in Table 3.

Performance evaluation examines the confusion matrix between target classes (True) and output classes (predicted). The confusion matrix shows the percentages of correct and incorrect classifications. Classification accuracy is the percentage of the number of the correctly classified samples over the total number of samples in each group or class (Table 4). Figure 9 shows an example of using LM algorithm for ANN training and validation.

Overall, 98.7% (75/76 biopsies) of the predictions are correctly classified, while 1.3% (1/76 biopsies) is wrong classification.

Table 3 Network training backpropagation algorithms

Algorithm	Number of inputs	Transfer functions	Number of hidden neurons	Network performance MSE	Best validation performance	R^2
LM	**138**	Tansig-pureline	2	0.01	3.3×10^{-3}	0.95
			5	**0.0088**	$\mathbf{1.5\times10^{-4}}$	**0.96**
LM	**55**	Tansig-pureline	2	0.014	3.7×10^{-3}	0.92
			5	0.009	4.87×10^{-4}	0.96
			8	**0.0088**	$\mathbf{2.4\times10^{-4}}$	**0.96**
			11	0.009	8×10^{-4}	0.95

Table 4 Classification accuracy for different numbers of hidden neurons

Figure	Number of hidden neurons	Normal classification %	Polyp classification, %	Cancer classification, %
9a	2	72 biopsies are correctly classified as "Normal" 0% misclassified	2 cases (Polyp) are incorrectly classified as "Normal" 100% misclassified	2 cases (Cancer) are incorrectly classified as "Normal" 100% misclassified
9b	5	72 biopsies are correctly classified as "Normal" 0% misclassified	2 cases (Polyp) are incorrectly classified as "Normal" 100% misclassified	2 cases (Cancer) are correctly classified as "Cancer" 0% misclassified
9c	8	72 biopsies are correctly classified as "Normal" 0% misclassified	1 case is correctly classified as "Polyp" 50% misclassified	2 cases are correctly classified as "Cancer" 0% misclassified

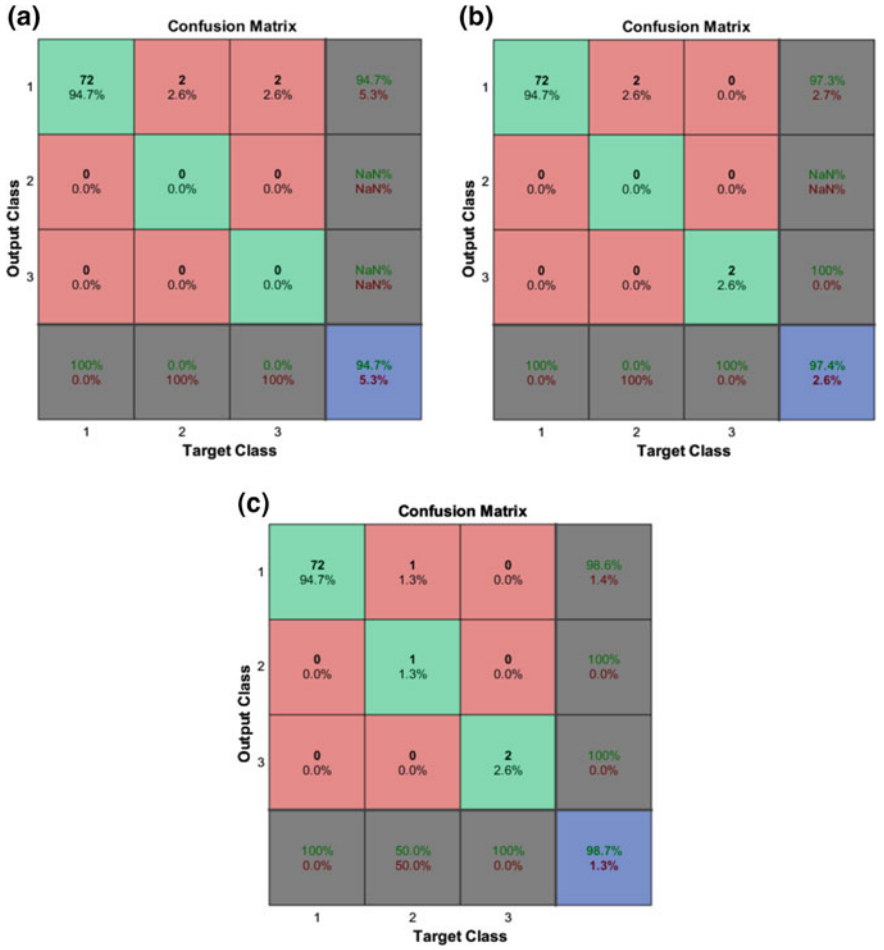


Fig. 9 Output class (predicted) and Target class (desired): 1—Normal; 2—Polyp; 3—Cancer **Green**—correctly classified; **Red**—misclassified; **Blue**—total percent of correctly and misclassified

6 Conclusions

This report aims to evaluate ANN in predicting cancer and other pathologies based on measurements by FTIR-ATR device. The feedforward backpropagation neural network with supervised learning is proposed to classify the disease: cancer/non-cancer or cancer-polyp-normal. The reliability of the proposed neural network method is examined on the data collected through Medical IROS (FTIR-ATR) device and obtained by a biopsy.

Choosing the optimal ANN architecture is followed by selection of training algorithm and related parameters. The selected ANN structure is a three-layer feed-forward, fully connected hierarchical network consisting of one input layer, one hidden layer, and one output layer. Six iterative backpropagation algorithms have been implemented to determine errors for the hidden layer neurons and subsequent weight modification. The determination of the number of layers and neurons in the hidden layers is done by the trial-and-error method. In order to determine optimal ANN model, a number of hidden neurons (2–11) in single hidden layer were considered and varied. The transfer functions tansig in hidden and linear in output layer were found to be optimal. After training, each ANN model is tested with the testing data, and optimal ANN architecture was found by minimizing test error with testing data and Mean Square Error (MSE) for training data. The final network structure in the first strategy has 55 inputs, 8 neurons in the hidden layer, and 3 neurons in the output layer. The best performance was obtained with LM training algorithm. Overall, 98.7% (75/76 biopsies) of the predictions are correctly classified and 1.3% (1/76 biopsies) is wrong classification. Using ATR-FTIR with ANN software with large database may have an important role for the development of next-generation real-time techniques for ex vivo identification tests of tumors.

References

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clinic. **68**, 394–424 (2018)
2. Dekel, B., Zilberman, A., Blaunstein, N., Cohen, Y., Sergeev, M.B., Varlamova, L.L., Polishchuk, G.S.: Method of infrared thermography for earlier diagnostics of gastric colorectal and cervical cancer. In: Chen, Y.W., Tanaka, S., Howlett, R., Jain, L. (eds.) Innovation in Medicine and Healthcare—InMed 2016, SIST, vol. 60, pp. 83–92. Springer, Cham (2016)
3. Zlotogorski-Hurvitz, A., Dekel, B.Z., Malonek, D., Yahalom, R., Vered, M.z: FTIR-based spectrum of salivary exosomes coupled with computational-aided discriminating analysis in the diagnosis of oral cancer. J. Cancer Res. Clin Oncol. **145**, 685–694 (2019)
4. Simonova, D., Karamancheva, I.: Application of Fourier transform infrared spectroscopy for tumor diagnosis. Biotechnol. Biotechnol. Equip. **27**(6), 4200–4207 (2013)
5. Theophilou, G., Lima, K.M., Martin-Hirsch, P.L., Stringfellow, H.F., Martin, F.L.: ATR-FTIR spectroscopy coupled with chemometric analysis discriminates normal and malignant ovarian tissue of human cancer. R. Soc. Chem. **141**, 585–594 (2016)
6. Paraskevaidi M., Martin-Hirsch P.L., Martin F.L.: ATR-FTIR spectroscopy tools for medical diagnosis and disease investigation. In: Kumar, C.S.S.R. (ed.) Nanotechnology Characterization Tools for Biosensing and Medical Diagnosis, pp. 163–211. Springer, Cham (2019)
7. Lei, L., Bi, X., Sun, H., Liu, S., Yu, M., Zhang, Y., Weng, S., Yang, L., Bao, Y., Wu L., Xu, Y., Shen K.: Characterization of ovarian cancer cells and tissues by Fourier transform infrared spectroscopy. J. Ovarian Res. **11**, 64.1–64.10 (2018)
8. Dong, L., Sun, X., Chao, Z., Zhang, S., Zheng, J., Gurung, R., Du, J., Shi, J., Xu, Y., Zhang, Y., Wu, J.: Evaluation of FTIR spectroscopy as diagnostic tool for colorectal cancer using spectral analysis. Spectrochim Acta Part A Mol. Biomol. Spectrosc. **122**, 288–294 (2014)

9. Rehman, S., Movasaghi, Z., Darr, J.A., Rehman, I.U.: Fourier transform infrared spectroscopic analysis of breast cancer tissues; identifying differences between normal breast, invasive ductal carcinoma, and ductal carcinoma in situ of the breast. Appl. Spectrosc. Rev. **45**(5), 355–368 (2010)
10. Yang, H., Griffiths, P.R., Tate, J.D.: Comparison of partial least squares regression and multi-layer neural networks for quantification of non-linear systems and application to gas phase Fourier transform infrared spectra. Anal. Chim. Acta **489**, 125–136 (2003)
11. Lasch, P., Stämmler, M., Zhang, M., Baranska, M., Bosch, A., Majzner, K.: FT-IR hyperspectral imaging and artificial neural network analysis for identification of pathogenic bacteria. Anal. Chem. **90**(15), 8896–8904 (2018)
12. Lasch, P., Diem, M., Hänsch, W., Naumann, D.: Artificial neural networks as supervised techniques for FT-IR microspectroscopic imaging. J. Chemom. **20**(5), 209–220 (2006)