REAL-TIME, ON-SITE, MACHINE LEARNING IDENTIFICATION METHODOLOGY OF INTRINSIC HUMAN CANCERS BASED ON INFRA-RED SPECTRAL ANALYSIS – CLINICAL RESULTS

Yaniv COHEN
School of Electronic Engineering Institute of Electronics and Mathematics (MIEM HSE), National Research University Higher School of Economics, Moscow 101000 Russia

Arkadi ZILBERMAN
School of Electrical & Computer Engineering, Ben Gurion University of the Negev, Beer-Sheva 8410501 Israel.

Ben Zion DEKEL
Dept. of Electrical & Computer Engineering, Ruppin Academic Center, Emek Hefer 4025000 Israel.

Evgenii KROUK
School of Electronic Engineering Institute of Electronics and Mathematics (MIEM HSE), National Research University Higher School of Economics, Moscow 101000 Russia

ABSTRACT

In this work we present a real-time (RT), on-site, machine-learning based methodology for identifying intrinsic human cancers. The presented approach is reliable, effective, cost-effective and non-invasive and based on the 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 table-top device for real-time tissue diagnosis that utilizes FTIR spectroscopy and the attenuated total reflectance (ATR) principle to accurately diagnose the tissue. The ATR measurement principle is performed utilizing a radiation source and a Fourier transform (FT) spectrometer. Information acquired and analyzed in accordance with this method provides accurate details of biochemical composition and pathologic condition of the tissue.
The combined device and method were used for RT diagnosis and characterization of normal and pathological tissues ex-vivo/in-vitro. Therefore, the presented device can be used in close conjunction with a surgical procedure
The solution methodology is to select a set of "features" that can be used to differentiate between cancer, normal and other pathologies using an appropriate classifier. These features serve as spectral signatures (intensity levels) at specific values of measured FTIR-ATR spectral responses.
Excellent results were achieved by applying the following three machine learning (ML) based classification methods to 76 wet samples: Partial least square regression (PLSR) and Principal component regression (PCR)
Both of the methods (PCR & PLSR) show a high performance to classify "Cancer" or "non-Cancer"; Correct Classification: 100 %; Incorrect Classification: 0.0 %.
Naive Bayesian classifier (NBC); Shows a high performance to classify "Cancer" or "non-Cancer" (benign); Correct Classification: 100 %; Incorrect Classification: 0.0 %.

Keywords: Machine learning, FTIR, ATR, Stomach cancer and Colorectal cancer.

1. INTRODUCTION

Tumor detection at initial stages is a major concern in cancer diagnosis [1-10]. Cancer screening involves costly and lengthy procedures for evaluating and validating cancer biomarkers. Rapid or one step method preferentially noninvasive, 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 [4]. 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 [5, 10]. Organic applications of IR spectroscopy are almost entirely concerned with spatial frequencies in the range of 4000 cm -1 to 400 cm-1 (2.5 um to 25um), which is known as mid-infrared (MIR) region of the spectrum. The range of spatial frequencies lower than 400 cm-1 is called far-infrared (FIR) and those greater than 4000 cm-1 are called near-infrared (NIR) [11-15]. Most of the fundamental molecular vibrations and many of the first overtones and combinations occur in the MIR range. The bands in the MIR tend to be sharp and have very high absorption, with both characteristics being desirable. Because the bands are sharp, most small molecules have distinctive spectral “fingerprints” that can be readily identified in mixtures. Also, because individual peaks can often be associated with individual functional groups, it is possible to see changes in the spectrum of an individual "objects" due to a specific reaction. Most biomolecules give rise to IR absorption bands between 1800 cm^{-1} and 700 cm^{-1}, which is known as the "fingerprint region" or primary absorption region. The medical IROS device [2] relates to methods employing Evanescent Wave Fourier Transform Infrared (EW-FTIR) spectroscopy using optical elements and sensors operated in the ATR regime in the MIR region of the spectrum. FTIR can be used to detect vibration in chemical bonds [8] and, as such, it is used to sense the biochemical composition of tissues [3, 4]. Although not capable of detecting specific molecules because many bond vibrations are shared among biomolecules, FTIR can be used to quantify classes of molecules (i.e. glycogen, protein, fat or nucleic acid etc.). FTIR has largely been performed on excised tissues and used to demonstrate that the overall biomolecular composition of diseased tissues is altered in a predictable manner relative to that of adjacent normal tissue [16]. Unlike conventional

methods, FTIR-ATR spectroscopy in the MIR region of the spectrum probes tissue biochemistry at a molecular level and the observed MIR spectra exhibit superimposed or composite vibrational bands. Large biomolecules are represented in FTIR-spectra by groups of characteristic IR-bands from which valuable information can be gained regarding the structure of the molecule and its interactions depending on position, form (shape), and intensity. The medical IROS device provides a method to detect functional molecular groups to elucidate complex structure within tissue, to characterize, distinguish and diagnose healthy, tumorous, precancerous, and cancerous tissue at an early stage of development. Typically, cancer occurs when a normal cell undergoes a change which causes the cell to multiply at a metabolic rate for exceeding that of its neighboring cells. Continued multiplication of the cancerous cell frequently results in the creation of a mass of cells called a tumor. Cancerous tumors are harmful [1] because they grow at the expense of normal neighboring cells, ultimately destroying them. In addition, cancerous cells are often capable of traveling throughout the body via the lymphatic and circulatory systems and of creating new tumors where they arrive. It should be noted that in addition to tumors which are cancerous (also referred to as malignant tumors) there are tumors which are non-cancerous. Non-cancerous tumors are commonly referred to as benign tumors. It is useful to be able to determine whether a tumor is cancerous or benign. [16]. The device uses IR spectroscopic method for determining if a tissue is a malignant tumor tissue, a benign tumor tissue, or a normal or benign tissue. Includes also nontoxic, ex-vivo, and fast (real-time) characterization of normal and abnormal tissue from breast, stomach, lung, prostate, kidney and other body parts during surgery, allowing an alternative first step of spectral histopathological examination and disease state characterization.

2. BASIC PRINCIPLES OF FTIR-ATR DETECTION

FTIR is a method of obtaining infrared spectra by first collecting an interferogram of a sample signal using an interferometer, and then performing Fourier Transform (FT) on the interferogram to obtain the spectrum. The detection scheme is based on Michelson interferometer, where a moving mirror varies the length of one optical path relative to the other, and creates an interferogram that is mathematically converted to an absorbance spectrum by a Fourier transform. As the optical path difference (OPD) in the interferometer grows, different wavelengths produce peak readings at different positions. FTIR spectroscopy is based on the interaction between the radiation and the sample, which absorbs the IR wavelengths causing transitions between vibrational energetic levels; therefore, vibrational modes of different chemical bonds can be detected and allow to identify different molecules. EW-FTIR spectroscopy is based on the phenomenon of attenuated total reflection (ATR) [15]. ATR spectroscopy utilizes total internal reflection phenomenon. In ATR spectroscopy a crystal with a high refractive index and IR transmitting properties is used as internal reflection element (ATR crystal). The ATR element is placed in contact with the sample. The beam of radiation propagating in ATR undergoes total internal reflection at the interface ATR-sample, provided the angle of incidence at the interface exceeds the critical angle θc. Total internal reflection of the light at the interface between two media of different refractive index creates an "evanescent wave" that penetrates into the medium of lower refractive index. "Evanescent" means "tending to vanish", which is appropriate because the intensity of evanescent waves decays exponentially with distance from the interface at which they are formed. This distance is typically in the 1-10 um range. Based on ATR spectrum, typical IR absorbance positions can be mentioned:

1) The bands around ~1640 cm^{-1} and ~1550 cm^{-1} - protein absorption region (Amide I and Amide II);
2) The bands around ~1480 cm^{-1} and ~1400 cm^{-1} - lipids and protein absorption region (CH3);
3) The bands between 1000–1300 cm^{-1} , PO2 symmetrical and asymmetrical stretching vibrations, indicate changes for phospholipids and nucleic acids:
4) Phospholipids and Amide III at ~1240 cm^{-1};
5) CO stretching at ~ 1160 cm^{-1} .
6) The bands between 2800-3100 cm^{-1} (the stretching vibrations of lipid hydrocarbons): the peaks around ~2850 cm^{-1} and ~2923 cm^{-1} indicate enhancement in lipid contents;
7) The peak around 2350 cm^{-1} is the carbon dioxide absorption (CO2).
8) The peak around ~3150-3600 cm^{-1} - strong water absorption;

3. 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. 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 produces result rapidly (up to minutes) and permits 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 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 the 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 (Figure 1).

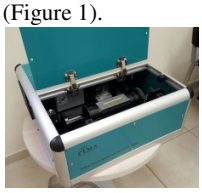


Fig. 1. The benchtop device for tissue characterization ex-vivo.

3.1. Methods for tissue diagnosis

Since the peak positions, peak widths, band shapes and relative intensities of spectra for tumor tissue may be different from healthy tissue, from a large number of spectrum data, the regularity of variations and the judgment criteria for diagnosis can be obtained. Statistical analysis can be used to assist in the identification of cellular types. For instance, several multivariate classification methods partial component regression (PCR) have been shown to provide satisfactory results [11-14]. In one embodiment, providing a diagnosis of the tissue includes forming an intensity spectrum. A diagnosis probability is computed based on intensities at particular wavelengths in the intensity spectrum. The diagnosis probability is compared to a threshold probability to characterize the tissue. a) Main molecular bonds: amide-I (~1650 cm^{-1}), amide-II (~1550 cm^{-1}), amide-III (~1240 cm^{-1}), symmetric phosphate (~1080 cm^{-1}) , glycogen (~1030 cm^{-1}), CH2 & CH3 of lipids (~2852 cm^{-1}, ~2923 cm^{-1}, ~2960 cm^{-1}).

b) Main absorbance ratios as malignancy indicators: glucose/phosphate (1030/1080), glycogen/amide II (1045/1545), Amide I / Amide II (1650/1550), CH2/CH3 (2922/2960).

Fig. 2. Circle of data transfer.

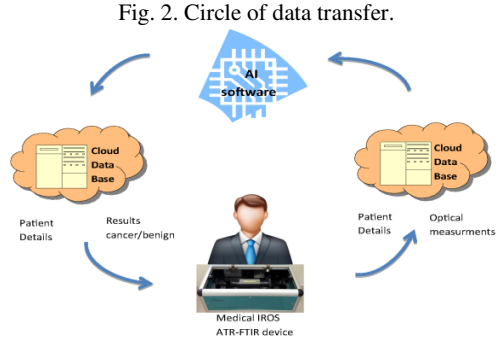
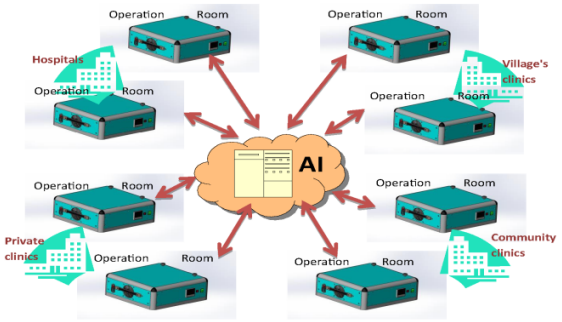


Fig. 3. Transfer of data from each patient and each hospital to the Center of final decision.



3.2. Data Base and Cloud

In Fig. 2, the circle of data transfer of patients' data to the medical IROS for machinery learning is presented, whereas Fig. 3 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 machinery learning occurring there.

4. FTIR-ATR DATA CLASSIFICATION

4.1. Problem description: Cancer Detection

The aim of the presented analysis is to choose and build a classifier that can distinguish between cancer, normal and other tissue pathologies from the measured FTIR spectroscopy data. The solution methodology is to select a set of "features" that can be used to distinguish between cancer and other control patients using an appropriate classifier. These features are the spectral signatures (intensity levels) at specific values of measured FTIR-ATR spectral response. The suggested approach - Machine Learning takes a known set of input data (spectral signatures) and known responses or class labels (e.g. "polyp", "cancer", "colitis", etc.) to the data, and seeks to build a Predictive Model that generates reasonable predictions / classifications for the response to new data. The classification methods and results presented based on: 1) Partial least square regression (PLSR); 2) Principal component regression (PCR); 3) Naive Bayesian classifier (NBC);

4.2. Data preparation and pre-processing

Acknowledgment: Data base presented in this article with a special permission from PIMS LTD, Data from IRB-approved human research. All participants in this research were checked by a MD or Gastroenterologist and referred by them to colonoscopy or gastroscopy. During the colonoscopy, a tissue biopsies taken from colon/stomach lesion/polyp and from adjacent normal appearing tissue. A minute part of the sample being used for conventional diagnosis (Histopathological evaluation), where the other wet part of the specimen (not in "formalin") is diverted to the Medial I.R.O.S system. The technician, operating the system, scanned the designated lesions without knowing the clinical diagnosis.(Blind). The data set was separated into two groups: training (calibration) set and validation (test) set. The training and validation sets include the observations (samples) with a known class labels (see Table 1 and 2).

Table 1. Calibration (training): 46 samples (spectral signals).

Class labels	Count	Percent
Norm	40	86.96%
Polyp	1	2.17%
Colitis	2	4.35%
LGD	1	2.17%
Cancer	2	4.35%

Table 2. Validation (test): 30 samples (spectral signals).

Class labels	Count	Percent
Norm	25	83.33%
Pathology*	2	6.67%
Crohn*	1	3.3%
Inflammation*	1	3.3%
Polyp	1	3.3%

* The calibration (training) does not include such class labels as "Inflammation", "Crohn", "Pathology" and they are classified as "Normal" in the training.

Two types of classification have been performed:

1.Binary: 0/1 or "Cancer"/"non-Cancer". Here all other types of pathologies (Polyp, LGD, Colitis, etc.) were represented as "non-Cancer";

2.Full classification following to the class labels in the training (Table 1).

4.2.1 Pre-Processing

1) The measured FTIR-ATR signal is converted to a spectral absorbance, A(λ), defined as: $A(\lambda) = -\log_{10}[R(\lambda)]$ (1)

Where: $$R(\lambda) = \frac{I(\lambda) - Dark(\lambda)}{REF(\lambda) - Dark(\lambda)} \quad (2)$$

I(λ) is the spectral intensity measured with the sample placed on HATR; REF(λ) is the reference signal (without sample) for source spectrum correction; Dark(λ) is the dark counts;
λ is the wavenumber, cm^{-1};

2) Peak normalization.

The absorption spectrum A(λ) is normalized by a maximal value at 1640 cm^{-1} (Amide I absorption):

$$Y(\lambda) = A(\lambda) / A(1640cm^{-1}), \quad (3)$$

Measured spectral absorbance according to Eq. (1) The normalized absorption spectrum A(λ) is shown in Fig.4 computed according to spectral data presented in Table 3.

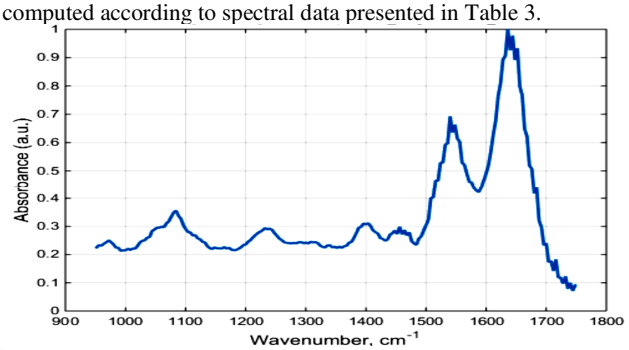


Fig. 4. Normalized spectrum,Eq.(3); peak normalization.

Table 3. Spectral data used for analysis.

Spectral interval, cm^{-1}	Resolution, cm^{-1}	Number of spectral signatures ("features")
950 - 1750	4	200

Selected feature parameters are organized in an n×p data matrix Y[n,p], where the n objects (samples) constitute the rows; and the p variables (feature parameters or spectral signatures at a specific wavenumber) the columns. In training set, the data matrix Y has n = 46 rows representing 46 patients and p = 200 columns representing 200 spectral signatures (see Table 1 and Table 3).

4.3. Machine Learning approach for classification

The data set contains samples with measurements of different variables (predictors or spectral signatures), i.e. signal responses at a specific wavenumber Y[n, pλ], and their known class labels, e.g. "polyp", "cancer", "colitis", etc.

The problem of classification can be formulated as following: If we obtain data for new samples K[n, pλ], could we determine to which classes those samples probably belong?

The steps in Machine Learning and classification / prediction are presented in a flowchart below (see Fig. 5).

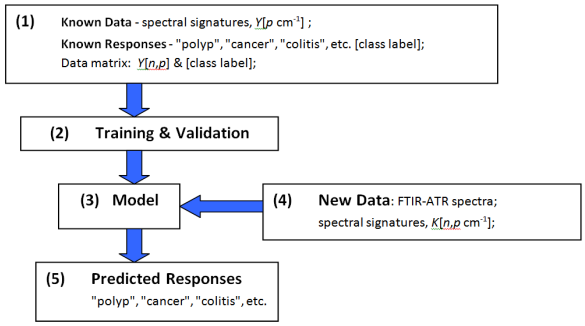


Fig. 5. Flowchart of machine learning approach.

5. PARTIAL LEAST SQUARE REGRESSION (PLSR) AND PRINCIPAL COMPONENT REGRESSION (PCR)

The measurement variables (responses) are the dependent (Y) variables. The rest of the variables are the independent (X) variables (spectral signatures). The purpose of a multiple regression is to find an equation that best predicts the Y variable as a linear function of the X variables.

The general equation for the regression is:

$$Y = b_0 + b_1 x_1 + .. b_k x_k + f = b_0 + \sum_{i=1}^{k} b_i x_i + f = b_0 + f + BX \quad (4)$$

where Y is the response (sample), xk are the predictors (spectral signatures or "features"), bk are the regression coefficient to be determined, b0 is the offset and a constant factor, and f is the residual. If X and Y are mean-centered, then b0 =0.

Equation (4) can be written in matrix form: **Y=Xb+f.**

The parameters **b** can be estimated by a least squares (LS) fit minimizing the sum of squared residuals. Multiple linear regression (MLR) is used for estimating the regression vector **b**.

The solution for regression coefficient for the LS is

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \quad (5)$$

where T means transpose of the matrix. The LS may not work since the inverse of $\mathbf{X}^T\mathbf{X}$ might not exist or may be unstable. It is also sensitive to noise.

Principal component regression (PCR) is a type of regression analysis, which considers principle components (PC) as independent variables, instead of adopting original variables.

The basic idea behind PCR is to calculate the principal components and then use some of these components as predictors in a linear regression model fitted using the typical least squares procedure. The PCs are the linear combination of the original variables which can be obtained by PCA.

The equation for regression can be formulated as

$$\mathbf{Y=Vd+f,\ d} = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\mathbf{Y}, \quad (6)$$

where **V** is the principal components and **d** contains coefficients. The number of components needs to be determined by testing and checking. The principal components are latent variables. PLS is an alternative for PCR. The latent variables in PLS are also linear combinations of the descriptive variables in the data set, but instead of maximizing the variance in the matrix with descriptive variables like in PCA, the covariance with the response variable is maximized. The scores on the PLS factors are used as input for multiple linear regression after selection of the optimal number of PLS-factors to be considered. In PLSR method, data is compressed into orthogonal factors, which have similar properties to PCs in PCA. The prediction performance can be evaluated using a) the coefficient of determination (R^2) of the linear regression of predicted against measured values; b) the root mean square errors of calibration (RMSEC); c) the root mean square errors of prediction (RMSEP).

In the present analysis the class labels were defined as:

1) Ordinal variables (Table 4), i.e. each sample was assigned a dummy variable for calibration modeling;

2) Binary variables (Table), which have values +1 for samples, belonging to class "Cancer", and –1 for samples, which are not from the class, i.e., "non-Cancer".

Table 4. Ordinal variables as a response in calibration (training) set.

Class labels	**ordinal variables**	**Count**
Norm	10	40
Polyp	30	1
Colitis	50	2
LGD	70	1
Cancer	90	2

Table 5. Binary variables as a response in calibration set.

Class labels	**binary variables**	**Count**
Norm	-1	40
Polyp	-1	1
Colitis	-1	2
LGD	-1	1
Cancer	+1	2

5.1. Training and Calibration

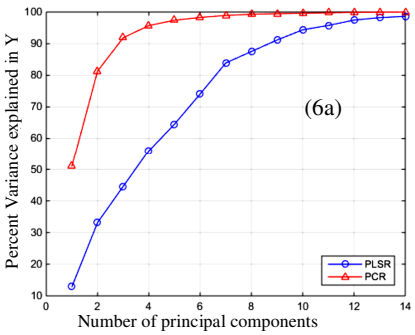
Figures 6a and 6b presents numerical analysis of error as a function of the number of PCs and PLS components. Figure from 7 present the 14-components PLSR model, of PCR model, and of PCR/PLSR model, based on data presented in Table 3 to Table 5, respectively.

Fig. 6. (a) Percent of Variance explained in the data vs principal components; (6) Prediction error performance with the number of PLS components (ordinal variables).

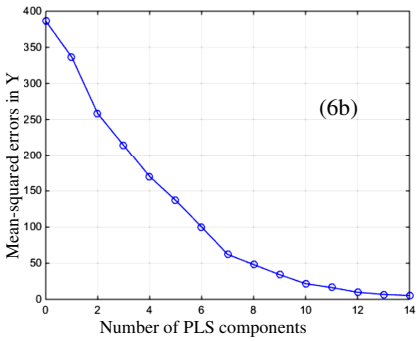
Fig. 7. The 14 components PCR/PLSR model trained with 46 samples and binary variables (Table 5): +1 - "Cancer"; -1 - "non-Cancer". R^2(PLSR) =0.9918 ; R^2 (PCR) =0.9085.

5.2. Validation

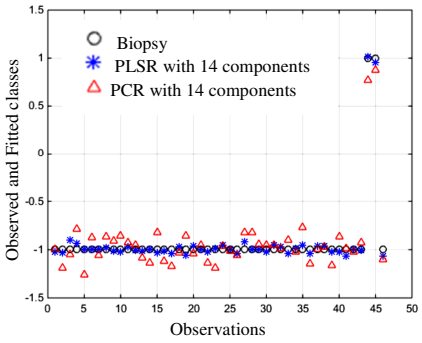
After the learning phase, 30 test values have been analyzed when the class labels are represented as ordinal and binary variables (according to Table 4 and Table 5), and presented in



(6a)

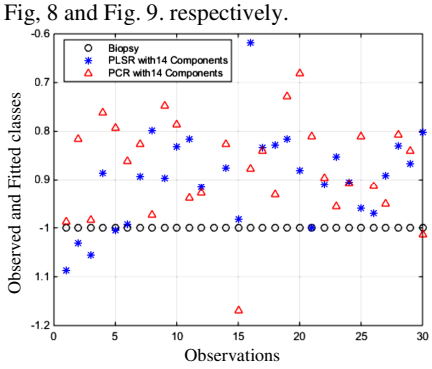


(6b)



(7)

Fig, 8 and Fig. 9. respectively.



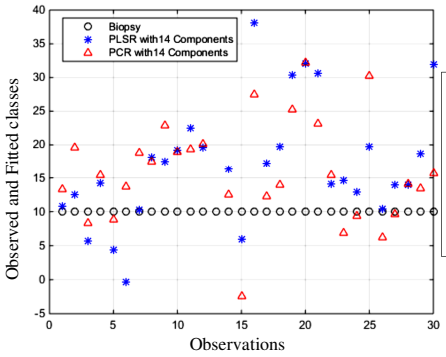
(8)

Biopsy

PLSR with 14 components

PCR with 14 components

Fig. 8. Observed and Predicted classes: ordinal variables in Table 4. Fig. 9. Observed and Predicted classes: binary variables in Table 5.



(9)

Biopsy

PLSR with 14 components

PCR with 14 components

5.3. PCR/PLSR Summary

a) In the case of binary variables (Table 5), if a predicted response value is above or equal to 0, the corresponding object or sample is considered to be a member of the class "Cancer". If not, the object is rejected as a non-member.

Both of the methods (PCR & PLSR) show a high performance to classify "Cancer" or "non-Cancer" (Fig.7):

Correct Classification: 100 %; Incorrect Classification : 0.0 %.

b) Classification with Ordinal classifiers (Fig.6): following to the class labels (Table 4), the values "10" and "30" correspond to "normal" and "polyp" respectively. There are values between [10...30] that should be defined and classified as belong to an appropriate class. The threshold should be determined, e.g. "25", where below this value all data are classified as "normal". To prevent misclassification, the additional analysis should be performed (e.g. PLS-DA method). Partial least squares Discriminant Analysis (PLS-DA) is a variant used when the class label is categorical (nominal) as in our case (Table 1).

c) The PCR method shows reduced performance in training (R^2 = 0.85) for the ordinal responses . Increasing database will improve the performance of PCR method.

5.4. Naive Bayes classifier (NBC)

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors.

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

5.4.1 Training

The details of the misclassifications are shown in Table 9 - the confusion matrix between target classes (True) and output classes (predicted).

Table 9. Confusion matrix between True and predicted classes (training)

Predicted class True class (biopsy)	Normal	Polyp	Colitis	LGD	Cancer
Normal	**40**	0	0	0	0
Polyp	0	**1**	0	0	0
Colitis	0	0	**2**	0	0
LGD	0	0	0	**1**	0
cancer	0	0	0	0	**2**

Percentage Correct Classification: 100 %

Percentage Incorrect Classification: 0.0 %

The confusion matrix in Fig.10 shows the percentages of correct and incorrect classifications in the case of "Cancer" or "non-Cancer" derivation (binary). Correct classifications are the green squares on the matrix diagonal. Incorrect classifications form the red squares.

If the method has learned to classify properly, the percentages in the red squares should be very small, indicating few or zero misclassifications. If this is not the case then further training would be advisable.

In the Fig.10, the first two diagonal cells show the number and percentage of correct classifications by the training: 44 biopsies are correctly classified as "non-Cancer", benign (or normal). This corresponds to 95.7% of all 46 biopsies. Similarly, 2 cases are correctly classified as "Cancer" or malignant. This corresponds to 4.3% of all biopsies. Overall, 100% of the predictions are correct and 0% are wrong classifications.

5.4.2 Validation

The confusion matrix in Fig.11 shows the percentages of correct and incorrect classifications in validation.

Percentage Correct Classification : 100 %

Percentage Incorrect Classification : 0.0 %

Table 10. Confusion matrix between True and predicted classes (validation).

Predicted class True class (biopsy)	Normal	Polyp
Normal	29	0
Polyp	1	0

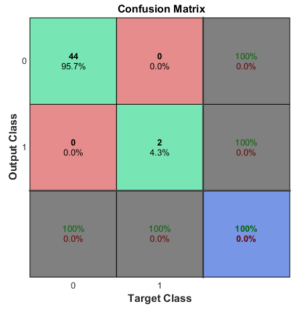


Fig. 10. Training: "Cancer"/ "non-Cancer"

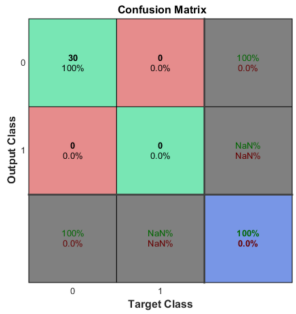


Fig.11. Validation: "Cancer"/ "non-Cancer"

Here: [29 0] means Naive Bayes classifier (NBC) classified 29 "Normal" correctly;
[1 0] means Naive Bayes classifier (NBC) classified 0 "Polyp" correctly, and misclassified one "Polyp" as "Normal"

Assess Classifier Performance:

The out-of-sample misclassification rate is	**3.3%.**
Percentage Correct Classification:	96.67 %
Percentage Incorrect Classification:	3.33 %

6. SUMMARY

In this work, we presented and built an ATR-FTIR system for a fast colon cancer/non-cancer detection. The decision algorithm is based on Machine Learning classifier that can distinguish between cancer and other tissue pathologies from the measured FTIR spectroscopy data taken by a top table device Medical IROS (by PIMS LTD).
The solution methodology was to select a set of "features" that used to distinguish between cancer and other control patients using an appropriate classifier. These features are the spectral signatures (intensity levels) at specific values of measured FTIR-ATR spectral response.
The classification methods and results, based on 76 wet samples, presented:
Partial least square regression (PLSR);
Principal component regression (PCR);
1) Both of the methods (PCR & PLSR) show a high performance to classify "Cancer" or "non-Cancer"
2) Correct Classification: 100%; Incorrect Classification: 0.0%.
Naive Bayesian classifier (NBC);
1) Shows a high performance to classify "Cancer" or "non-Cancer" (benign)
2) Correct Classification: 100%; Incorrect Classification: 0.0%.
During the next steps, more machine learning classifiers, should be investigates, for example, LDA and ANN. to choose the best classifier for real time, on site, cloud based diagnosis using Artificial Intelligence (AI) together with the medical device IROS. We plan to increase database to improve the performance of the machine learning classifier.

7. REFERENCES

[1] Bray, F. , Ferlay, J. , Soerjomataram, I. , Siegel, R. L., Torre, L. A. and Jemal, A., "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", CA: **Cancer Journal for Clinicians**, vol. 68, pp. 394-424, 2018, doi:10.3322/caac.21492 (2018).

[2] B. Dekel, A. Zilberman, N. Blaunstein, Y. Cohen, M.B. Sergeev, L.L. Varlamova and G.S. Polishchuk, "Method of Infrared Thermography for Earlier Diagnostics of Gastric Colorectal and Cervical Cancer", in: Chen YW., Tanaka S., Howlett R., Jain L. (Editors) Innovation in Medicine and Healthcare 2016**. InMed 2016. Smart Innovation, Systems and Technologies**, vol 60. Springer, Cham, pp. 83-92, 2016, DOI: https://doi.org/10.1007/978-3-319-39687-3_8 (2016).

[3] S. Christodolos and C. Christina, "Thermal Heterogeneity Constitutes A Marker for the Detection of Malignant Gastric Lesions In Vivo ", **Journal of Clinical Gastroenterology** Vol. 36 pp 215-218, 2003.

[4] S. Christodolos and C. Christina, "Thermal Heterogeneity Constitutes A Marker for the Detection of Malignant Gastric Lesions In Vivo ", **Journal of Clinical Gastroenterology**, Vol. 36 pp 215-218, 2003.

[5] B. Z. Dekel, A. Zlotogorski-Hurvitz, D. Malonek, R. Yahalom, M. Vered, "Diagnosis of oral cancer based on FTIR-ATR spectra of salivary exosomes – Preliminary study", **Proc. of NBC Conference, 2017**, pp. 3.

[6] N. Shussman and Y. Mintz, "Laparoscopic Infrared Imaging — The Future Vascular Map", 2010,**Laparoendoscopic & Advanced Surgical Techniques,** doi:10.1089/lap.2010.0474 (2010).

[7] A. Rogalski, "Next decade in infrared detectors," Proc. SPIE 10433, **Electro- Optical and Infrared Systems: Technology and Applications XIV**, 104330L (9 October 2017); doi: 10.1117/12.2300779

[8] P. Robert1 et al , "Low power consumption infrared thermal sensor array for smart detection and thermal imaging applications", **AMA Conferences 2013 - SENSOR 2013, OPTO 2013, IRS 2 2013**, DOI 10.5162/irs2013/i2.1

[9] Xie, W. , McCahon, P. , Jakobsen, K. and Parish, C. (2004), Evaluation of the ability of digital infrared imaging to detect vascular changes in experimental animal tumours. **Int. J. Cancer**, 108: 790-794. doi:10.1002/ijc.11618

[10] Song, C. , Appleyard, V. , Murray, K. , Frank, T. , Sibbett, W. , Cuschieri, A. and Thompson, A. (2007), Thermographic assessment of tumor growth in mouse xenografts**. Int. J. Cancer**, 121: 1055-1058. doi:10.1002/ijc.22808

[11] Berz, Reinhold and Claus E.E. Schulte-Uebbing. "**MammoVision** (Active Functional Infrared Breast Thermography) Compared to X-Ray Mammography-114 Cases Evaluated." (2010).

[12] B.B. Lahiri, S. Bagavathiappan, T. Jayakumar, John Philip, **Infrared Physics & Technology**, Volume 55, Issue 4, July 2012, Pages 221-235, https://doi.org/10.1016/j.infrared.2012.03.007

[13] Ćurković, S. et al. Medical thermography (digital infrared thermal imaging – DITI) in paediatric forearm fractures – **A pilot study Injury**, Volume 46, S36 - S39 DOI: https://doi.org/10.1016/j.injury.2015.10.044

[14] Sivanandam, S., Anburajan, M., Venkatraman, B. et al. Medical thermography: a diagnostic approach for type 2 diabetes based on non-contact infrared thermal imaging **Endocrine** (2012) 42: 343. https://doi.org/10.1007/s12020-012-9645-8

[15] Magalhaes C, Vardasca R, Mendes J. Recent use of medical infrared thermography in skin neoplasms. **Skin Res Technol**. 2018;00:1–5. https://doi.org/10.1111/srt.12469

[16] Saira Chaudhry et al. Art Tucker & Dylan Morrissey (2016): The use of medical infrared thermography in the detection of tendinopathy: a systematic review**, Physical Therapy Reviews**, DOI:10.1080/10833196.2016.1223575