Turkish Journal of Electrical Engineering and Computer Sciences

Volume 27 | Number 3 Article 15

1-1-2019

Classification of the likelihood of colon cancer with machine learning techniques using FTIR signals obtained from plasma

SUAT TORAMAN

MUSTAFA GİRGİN

BİLAL ÜSTÜNDAĞ

İBRAHİM TÜRKOĞLU

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 1765 – 1779

doi:10.3906/elk-1801-259

Classification of the likelihood of colon cancer with machine learning techniques using FTIR signals obtained from plasma

Suat TORAMAN[1,*], **Mustafa GİRGİN**[2], **Bilal ÜSTÜNDAĞ**[3], **İbrahim TÜRKOĞLU**[4]

[1]Department of Informatics, Faculty of Science, Fırat University, Elazığ, Turkey
[2]Department of General Surgery, Fırat University Hospital, Elazığ, Turkey
[3]Department of Medical Biochemistry, Faculty of Medicine, Fırat University, Elazığ, Turkey
[4]Department of Software Engineering, Faculty of Technology, Fırat University, Elazığ, Turkey

Received: 29.01.2018 • **Accepted/Published Online:** 29.11.2018 • **Final Version:** 15.05.2019

Abstract: Colon cancer is one of the major causes of human mortality worldwide and the same can be said for Turkey. Various methods are used for the determination of cancer. One of these methods is Fourier transform infrared (FTIR) spectroscopy, which has the ability to reveal biochemical changes. The most common features used to distinguish patients with cancer and healthy subjects are peak densities, peak height ratios, and peak area ratios. The greatest challenge of studies conducted to distinguish cancer patients from healthy subjects using FTIR signals is that the signals of cancer patients and healthy subjects are similar. In the current study, a method in which the area and height ratios of the FTIR signal, as well as various statistical features, are proposed in order to overcome this difficulty. Blood samples (plasma) were collected from 30 colon cancer patients and 40 healthy subjects, and FTIR measurements were performed. A total of 16 features were obtained, including five height ratios, five area ratios, and six statistical features, from each FTIR signal. The 16 features were classified with a multilayer perceptron neural network and support vector machines using cross-validation and their performances were then compared. The current study demonstrated that different features obtained from plasma FTIR spectra can be used together in order to distinguish colon cancer patients from healthy individuals.

Key words: Colon cancer, plasma, FTIR signal, feature extraction, pattern recognition, artificial neural network, support vector machines, classification

1. Introduction

According to data from the Department of Cancer of the Turkish Ministry of Health, colorectal cancer is ranked fourth among the types of cancer most presented in adult males and females. While some markers in the blood are examined in diagnoses, treatment, and follow-up of colon cancers, patient-disturbing methods such as colonoscopy and computed tomography are also used [1, 2]. For example, colonoscopy is an invasive method that is frequently used to diagnose cancer. Biopsy is both invasive and carries a risk of infection [3, 4]. Also, the sensitivity and specificity of some tumor markers such as carcinoembryonic antigen (CEA) and cancer antigen (CA) 19-9 are low [4]. Therefore, there is a need for simpler and faster noninvasive methods instead of the existing conventional methods. One of the methods employed to distinguish between the existence of cancer and healthy subjects is FTIR spectroscopy. Many studies have been conducted on the distinction of cancerous and normal tissue by FTIR in prostate cancer [5], cervical cancer [6–8], pancreatic cancer [9], gastric tissues

*Correspondence: storaman@firat.edu.tr

[10–12], esophagus cancer [13], breast cancer [14, 15], and colon cancer [16–19]. In addition, various studies have recently been carried out for the separation of patients and healthy subjects using blood samples and FTIR signals in leukemia [20, 21], Parkinson disease [22], gastric cancer [23], and other types of cancer such as breast, gastrointestinal, lung [3, 24], and colon cancer [4, 25]. Studies carried out on blood samples are presented in Table 1.

In previous studies, the plasma or serum samples obtained from blood were dried for FTIR measurement. Then the spectral measurement of the samples was performed. As can be seen in Figure 1, peaks in the samples appear after performing the drying process for a certain period of time. In previous studies, a distinction was then made using these peak values [7, 9, 10, 12, 13, 15, 23, 24, 26]. In the current study, FTIR measurements of plasma samples were performed in liquid form. Thus, the waiting period required for the time-consuming drying process (30–60 min) was not required and the analysis process could be performed more quickly. In the literature, only certain peaks in the signal have been examined. In the current study, all changes of the FTIR signal within a certain range (1800–1000 cm^{-1}) were investigated using various statistical features. These features were then classified by artificial neural network (ANN) and support vector machine (SVM). Data from the colon cancer patients and healthy subjects were also evaluated by independent t-test. The aim of the current study is to investigate the possibility of distinguishing colon cancer patients from healthy subjects using the features obtained from liquid plasma FTIR spectra. A flowchart of the current study is presented as Figure 2.

Table 1. Studies on cancer detection and classification from blood samples.

Authors	Year/Type	Algorithms/features used	Results
Erukhimovitch et al. [20]	2006, leukemia	Cluster analysis	Significant reduction observed at spectral peaks of 1056, 1270, and 1592 cm^{-1}
Ostrovsky et al. [3]	2013, breast, lung, gastro, other	Principal component analysis, Fisher's linear discrimination analysis	Sensitivity: 93.33% Specificity: 87.80% Accuracy: 90.70%
Sheng et al. [21]	2013, leukemia	Curve fitting	H2959/H2931 ratio and RNA/DNA (A1115/A1028) defined as a distinctive feature.
Sheng et al. [23]	2013, gastric cancer	Curve fitting	H2959/H2931 ratio defined as a distinctive feature.
Wang et al. [24]	2014, lung cancer	Curve fitting	A1080/A1170 ratio defined as a distinctive feature

The remainder of the paper is organized as follows: in Section 2 sample preparation and FTIR measurement, feature extraction, classifier models, and parameters for performance are explained. Sections 3 and 4 provide the results of the study and the discussion. Finally, Section 5 concludes the paper and briefly outlines our future work plans.

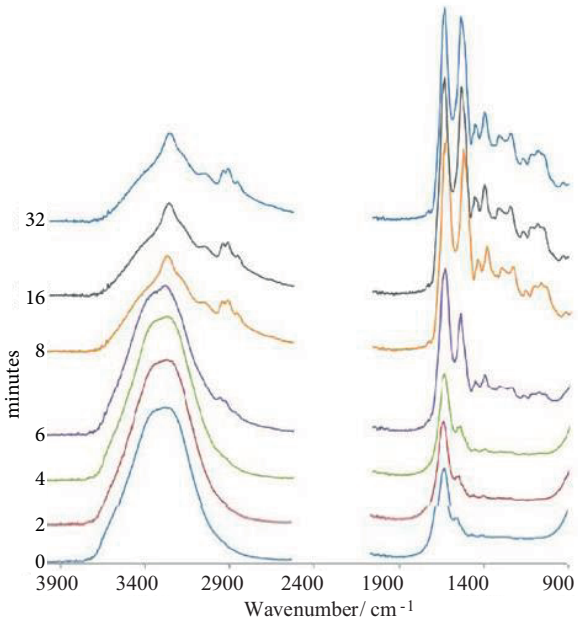

Figure 1. ATR-FTIR spectra showing the drying of human serum [27].

2. Materials and methods

2.1. Sample preparation and ATR-FTIR measurement

Blood samples of 30 colon cancer patients and 40 healthy subjects were obtained from the Department of General Surgery of Fırat University, Turkey (see Table 2). The samples were centrifuged at 4000 rpm for a period of 4 min in order to separate the plasma from other cellular material. The plasma samples were transferred to clean tubes and the samples were then stored at -20 °C. FTIR measurements of the samples were performed using a PerkinElmer Spectrum 100 FTIR spectrometer with a zinc selenide attenuated total reflection (ATR) accessory. The samples were measured in the region of 4000–450 cm^{-1} with 32 scans (for a high signal-to-noise ratio) and 4 cm^{-1} resolution.

Table 2. Statistics of cancer patients and healthy subjects.

	Mean age ± SD	Sex		Stage of cancer			
		Male	Female	I	II	III	IV
Colon cancer	58.37 ± 14.35	23	7	4	11	10	5
Healthy	52.18 ± 18.15	17	23	-	-	-	-

For all spectra, baseline correction was performed and max–min normalized to the amide II band [3]. In addition, the 1800–1000 cm^{-1} region was investigated in the study as the important spectra in cancer detection are within this range [3, 4]. The study was reviewed and approved by the Noninvasive Research Ethics Committee of Fırat University [28].

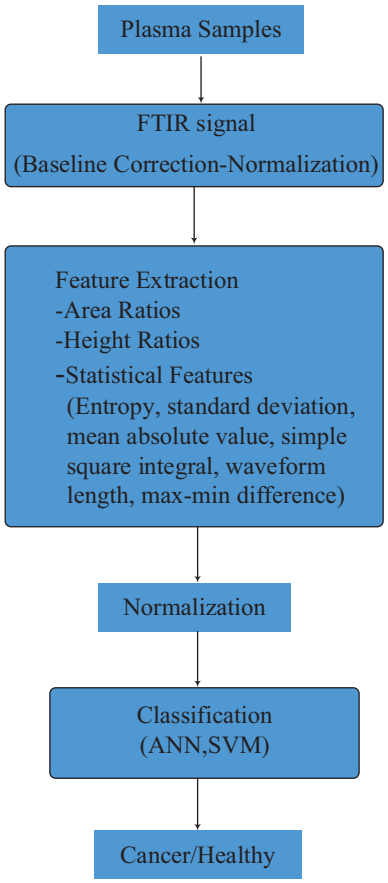

Figure 2. Flowchart of the study.

2.2. Feature extraction

Feature extraction is significant for pattern recognition. The effect of features extracted from the pattern plays an important role in the performance of the pattern recognition system. Feature extraction provides a representation of the data within a small dimension. This process, which is called dimensionality reduction, is performed by removing redundant information [29]. In the current study, feature extraction was performed from ATR-FTIR signals. Thus, each FTIR signal is represented with fewer features. A total of 16 features extractions, encompassing five height ratios, five area ratios, and six statistical features, the mathematical expressions of which are presented in Table 3, were performed from each FTIR signal [30–35] (see Figure 3).

Here, N represents the length of the signal, x_i represents a part of the FTIR signal, and p_i represents the probability value of the ith wavenumber. Entropy can be used to measure irregularity of a nonstationary signal [36]. Standard deviation is a measure used to quantify the distribution of the data in statistics. Moreover, it is also the square root of the variance [37, 38]. Mean absolute value is used to determine the mean value of the examined range of the signal. Simple square integral is a feature used to express the energy of the signal. Waveform length is the cumulative length of the signal in a particular range [31–33]. Max–min difference is the difference between the maximum and minimum value of the signal in the specified range [35].

In previous studies, various spectral area ratios and height ratios were used in order to distinguish cancer patients from healthy subjects [10, 11, 18, 20, 21, 23, 24]. The height ratios in the current study were examined

Table 3. Features extracted from the FTIR signal.

Features	Mathematical definition	
Entropy	$H(p) = \sum_{i=1}^{N} p_i log p_i$	(1)
Standard deviation	$STD = \sqrt{\frac{1}{N+1}} \sum_{i=1}^{N} (x - x_i)^2$	(2)
Mean absolute value	$MAV = \frac{1}{N} \sum_{i=1}^{N} \lvert x_i \rvert$	(3)
Simple square integral	$SSI = \sum_{i=1}^{N} \lvert x_i \rvert^2$	(4)
Waveform length	$WL = \sum_{i=1}^{N-1} \lvert x_{i+1} - x_i \rvert$	(5)
Max–min difference	$MMD = max(x_i) - min(x_i)$	(6)

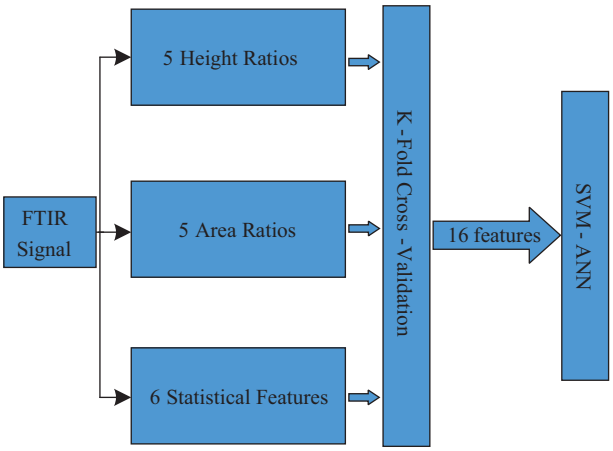

Figure 3. Feature extraction from FTIR signal and classification.

according to [23] as follows: H1453/H1400, H1080/H1550, H1314/H1243, H1646/H1550, and H2959/H2931 [21, 23, 39]. Bands that are of interest include the following: 1080 cm^{-1} is the (PO_2^-) symmetric stretching of nucleic acids, 1314 cm^{-1} indicates amide III, 1550 cm^{-1} is amide II (N-H bending and C-N stretching), 1646 cm^{-1} is amide I (C=O stretching), 1400 cm^{-1} is due to symmetric CH3 bending, 1453 cm^{-1} is due to CH2 bending, 1170 cm^{-1} is C-O bands from glycomaterials and proteins, 1243 cm^{-1} is asymmetric stretching vibrations of phosphodiesters, 2931 cm^{-1} is due to asymmetric CH2 stretching vibrations, and 2959 cm^{-1} is due to asymmetric CH3 stretching vibrations [23, 24, 39, 40]. The height ratios were obtained by proportioning the intensity values of the signal.

The area ratios were calculated as the area between the signal and the wavenumber axis according to the given range (see Figure 4). In the current study, the area ratios were examined according to [24] as follows: A2959/A1545, A1650/A1545, A1080/A1170, A1080/A1545, and A1080/A1243. The baselines of the selected areas are A2959 (2997–2887 cm^{-1}), A1650 (1725–1593 cm^{-1}), A1545 (1595–1480 cm^{-1}), A1170 (1184–1140 cm^{-1}), A1243 (1258–1203 cm^{-1}), and A1080 (1140–1000 cm^{-1}) [24]. The ratio of A2959/A1545 lipids/proteins and the ratio of A1650/A1545 can indicate changes in protein components and structures. The A1080/A1170

ratio can indicate some amino acids in the protein structure (serine residues, tyrosine, threonine) and nucleic acids. The A1080/A1545 ratio can provide information about DNA levels. The A1080/A1243 ratio can identify structural changes in nucleic acids [7, 24].

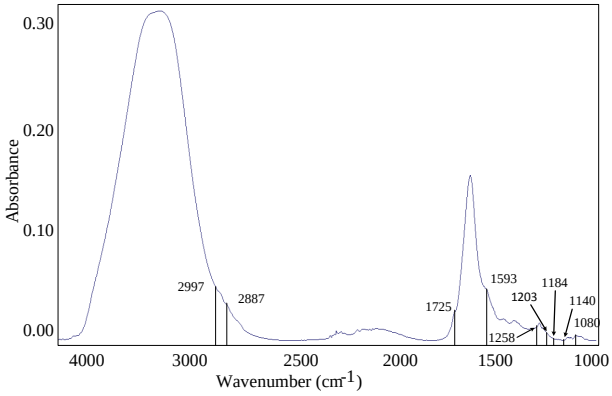

Figure 4. Areas of IR spectrum of colon cancer patients and healthy subjects.

2.3. Support vector machines

SVM is an instructive machine-learning algorithm that uses structural risk minimization. The purpose of SVM is to find the optimal hyperplane that would separate the data of different classes from each other. For a two-class classification problem, presume the training data were defined as $(x_j, y_j), j = 1, 2, 3, ...n, x_j \in R^d$ and class labels as $y_j \in \{-1, +1\}$. The SVM will try to find the hyperplane where the distance between the two classes is the greatest. In this case, w is weight vector and b is bias, and the optimal hyperplane is expressed as follows [41–43]:

$$w.x_j + b \geq +1 \quad \forall x_j \in ClassA, \tag{7}$$

$$w.x_j + b \leq -1 \quad \forall x_j \in ClassB. \tag{8}$$

The maximization of the optimal hyperplane boundaries is performed by solving the following optimization problem:

$$argmin\frac{1}{2}||w||^2. \tag{9}$$

Accordingly,

$$y_j(w.x_j + b) \geq 1 \quad j = 1....n. \tag{10}$$

Lagrange multipliers are used for the solution to this problem:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{j=1}^{n} \alpha_j[y_j(w.x_j + b) - 1]. \tag{11}$$

Here, α is the Lagrange multiplier and x_j is the support vector. For a two-class problem that can be separated linearly, the decision function is defined as:

$$f(x) = sign(\sum_{j=1}^{n} y_j \alpha_j (x.x_j) + b). \tag{12}$$

The SVM's kernel functions (linear, radial basis function (RBF), polynomial) are as follows:

$$Linear : K(x_j, y_j) = (x_j.y_j), \tag{13}$$

$$RBF : K(x_j, y_j) = exp(\frac{-||x_j - y_j||^2}{2\sigma^2}), \tag{14}$$

$$Polynomial : K(x_j, y_j) = (x_j y_j + 1)^d. \tag{15}$$

σ, d are the parameters of the kernel functions. The SVM parameter C was searched in the range of $[10^{-2}, ..., 10^3]$.

2.4. Multilayer perceptron neural network

The ANN has been developed based on the structure and learning characteristics of biological neuron cells. ANNs are used for various applications such as pattern recognition and classification. In the current study, a feedforward multilayer perceptron neural network (MLPNN) model was used. In general, the performance of the feedforward neural network depends on the number of neurons and hidden layers, the activation function, and the learning algorithm [44–47]. The number of neurons, the learning algorithm, and the activation function in the MLP model are shown in Table 4. The parameters of the MLP network, such as number of neurons in the hidden layer and the number of layers, are selected empirically during the experimental works.

Table 4. Structure of the MLPNN model.

Model	Feedforward multilayer perceptron
Number of layers	3
Number of neurons in layer	Hidden: 20, 10; Output: 1
Initial weights	Random
Activation functions	Logarithmic sigmoid
Learning rule	Backpropagation algorithm

The backpropagation method was used for the learning procedure. The weights in the network were updated by calculating the root mean square error. The mean square error was repeated until reaching the desired error value [44, 48]. All data were normalized using $x_i = (x_i - x_{min})/(x_{max} - x_{min})$ prior to classification. x_i represents the ith element of the data, and x_{max} and x_{min} are the maximum and minimum elements of the data.

2.5. Parameters for classification and performance

K-fold cross-validation was used to estimate model performances objectively. In the current study, 4, 5, and 10-fold cross-validation was applied to the dataset. For example, for k = 5, the dataset is divided into 5 parts. Four parts are used for training and the remaining part for testing. This process is performed for all parts and the mean error is then calculated. Classifier performances were evaluated by sensitivity, specificity, and accuracy indicators [3, 49]. Three indicators were calculated with a confusion matrix. Calculations of the indicators are as follows:

- True positive (TP): Number of colon cancer patients correctly defined.
- False negative (FN): Number of colon cancer patients incorrectly defined.
- True negative (TN): Number of healthy subjects correctly defined.
- False positive (FP): Number of healthy subjects incorrectly defined.

Sensitivity, specificity, and accuracy are defined as:

$$Sensitivity = \frac{TP}{TP + FN} \times 100, \tag{16}$$

$$Specificity = \frac{TN}{TN + FP} \times 100, \tag{17}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100. \tag{18}$$

Differences between the healthy subjects and the colon cancer patients were also tested using the independent t-test. Values of P found to be lower than 0.05 were considered significantly different. MATLAB was used to process all data.

3. Results

The area ratio and height ratio values of each FTIR signal were calculated separately. Feature extraction was performed for the 1800–1300 cm^{-1} and 1300–1000 cm^{-1} ranges of each signal using the features specified in Table 3. All features obtained were first compared by independent t-test.

The average of H1314/H1243, which indicates the change of nucleic acids and proteins [23], was 0.528 in healthy subjects and 0.708 in patients with colon cancer (for H1314/H1243, P = 0.014). If 0.71 is used as the standard, 73.33% of the patients with colon cancer ($n = 22$) are above this value, while 65% of healthy subjects ($n = 26$) are below this value (see Table 5).

The average of H1645/H1550 was 0.597 in 40 healthy subjects and 0.760 in 30 patients with colon cancer. The P-value of the H1646/H1550 ratio is 0.003. However, the H1645/H1550 ratios of 19 of the 30 colon cancer patients and the H1645/H1550 ratios of 22 of the 40 healthy individuals were in the range of 0.77–0.58. Hence, the H1646/H1550 ratio, although statistically significant, could not exactly distinguish healthy individuals from patients with colon cancer. In addition, other spectral ratios and statistical features exhibit similar features in terms of the number of patients and control groups. For this reason, properties other than the above spectral ratio could not be used to evaluate colon cancer and healthy subjects (see Tables 6 and 7).

The H1314/H1243 ratio is low in terms of sensitivity and specificity, even though it is significant compared to the t-test results. As a result, none of the features alone attained sufficient sensitivity and specificity for

effectively distinguishing colon cancer patients from healthy individuals. Therefore, a machine learning approach using all of the features was applied. As a result of this approach, 95.71% accuracy, 93.33% sensitivity, and 97.50% specificity were obtained with MLP and 94.29% accuracy, 93.33% sensitivity, and 97.50% specificity with SVM (linear kernel function).

The six features from Table 3 were calculated separately for the 1800–1300 cm^{-1} and 1300–1000 cm^{-1} ranges. Classification was performed in two ways: five height ratios, five area ratios, and six features were calculated for 1800–1300 cm^{-1}, and five height ratios, five area ratios, and six features were calculated for 1300–1000 cm^{-1} were each classified separately. The results obtained as a result of both classification processes performed using a total of 16 features are presented in Tables 8, 9, and 10.

The dataset was evaluated for different folds (k = 4, 5, 10). The best results and average accuracy of both classifiers (10 times) are shown in Tables 8, 9, and 10. The best result in the classification with MLP was obtained in the 1300–1000 cm^{-1} range (see Table 8). Tables 9 and 10 show the SVM classification results in the ranges of 1300–1000 cm^{-1} and 1800–1300 cm^{-1}, respectively. The best achieved accuracy in the classification with SVM was obtained with the linear kernel function in the range of 1300–1000 cm^{-1}. Both the MLP and SVM classifiers returned the best accuracy in the range of 1300–1000 cm^{-1}. According to these results, it can be said that the discrimination of this region is better.

The best accuracy achieved in the current study was 95.71%. This value was reached with the MLP model where 75% of the dataset was used for training. When 80% and 90% of the dataset was used for training, the classification accuracy decreased. When the test data were reduced, the accuracy of the MLP was also reduced. The best accuracy in SVM was achieved when 75% and 90% of the dataset was used for training. SVM remained largely unaffected by changes in the test data.

In addition, when the models were compared for training and testing times, SVM showed much faster training and testing times. MLP completed training and testing in 11.25 s, 21 s, and 45.5 s, respectively, for 4, 5, and 10 folds. The average duration of SVM for training and testing was approximately 1 s for 4, 5, and 10 folds). These results show that SVM requires less time and less computational power for training and testing. Models were trained using the system with Intel Core i3-2120 CPU 3.3 GHz and 8 GB RAM.

Table 5. Statistical data of height ratios.

Ratios	Colon cancer patients	Healthy subjects	t-test (P-value)
H1080/H1550	0.0592 ± 0.1218	0.2486 ± 0.3833	0.011
H1646/H1550	0.7603 ± 0.0987	0.5972 ± 0.2760	0.003
H1314/H1243	0.7083 ± 0.2307	0.5288 ± 0.3338	0.014
H1453/H1400	0.4973 ± 0.2181	0.4018 ± 0.2494	0.990
H2959/H2931	0.8427 ± 0.0807	0.7170 ± 0.3011	0.029

4. Discussions

Although there have been many studies on colon cancer and other types of cancer tissues, few have been conducted on FTIR signals obtained from blood samples. These studies can be divided into two: those obtained from cancerous tissues and those obtained from blood samples (plasma/serum). The results of the studies carried out on the separation of colon cancer patients and healthy subjects using tissue samples are as follows: Cheng et

Table 6. Statistical data of area ratios.

Ratios	Colon cancer patients	Healthy subjects	t-test (P-value)
A2959/A1545	0.6322 ± 0.1590	0.5378 ± 0.2718	0.095
A1650/A1545	0.8130 ± 0.1028	0.6459 ± 0.3081	0.006
A1080/A1545	0.0516 ± 0.1310	0.2449 ± 0.3935	0.012
A1080/A1243	0.1534 ± 0.2204	0.3253 ± 0.4108	0.042
A1080/A1170	0.1552 ± 0.2393	0.1555 ± 0.1282	0.995

Table 7. Statistical data of six features.

	1800–1300 cm^{-1}			1300–1000 cm^{-1}		
Features	Colon cancer	Healthy	P	Colon cancer	Healthy	P
ENT	0.712 ± 0.163	0.739 ± 0.134	0.440	0.488 ± 0.317	0.547 ± 0.238	0.374
WL	0.775 ± 0.085	0.679 ± 0.238	0.030	0.192 ± 0.229	0.240 ± 0.179	0.329
STD	0.818 ± 0.070	0.698 ± 0.256	0.160	0.613 ± 0.170	0.614 ± 0.141	0.991
MAV	0.057 ± 0.053	0.219 ± 0.323	0.008	0.024 ± 0.068	0.197 ± 0.347	0.009
SSI	0.052 ± 0.036	0.203 ± 0.302	0.009	0.063 ± 0.027	0.159 ± 0.311	0.009
MMD	0.771 ± 0.083	0.678 ± 0.225	0.030	0.430 ± 0.159	0.407 ± 0.090	0.452

Table 8. Best accuracy and average results obtained with MLP.

Number of folds	Range (cm^{-1})	TP	FN	FP	TN	Sen (%)	Spe (%)	Acc (%)	Mean acc (%) ± Std
4	1300–1000	28	2	1	39	93.33	97.50	**95.71**	89.75 ± 3.56
	1800–1300	25	5	0	40	83.33	100.00	92.85	87.27 ± 3.34
5	1300–1000	25	5	0	40	83.33	100.00	92.85	89.00 ± 2.43
	1800–1300	26	4	1	39	86.66	97.50	92.85	87.49 ± 3.80
10	1300–1000	23	7	0	40	76.66	100.00	90.00	85.00 ± 3.38
	1800–1300	25	5	1	39	83.33	97.50	**91.42**	88.14 ± 3.01

al. [50] obtained 100%, 97.5%, 95%, and 100% accuracy, respectively, in the classification of normal, dysplasia, early carcinoma, and advanced cancer tissues using SVM. Then Cheng et al. [17] classified the normal, dysplasic, early carcinoma, and advanced cancer tissues and obtained 100%, 94%, 97.5%, and 100% accuracy, respectively. Xie et al. [19] obtained more than 90% classification sensitivity and specificity using colon spectral data. Dong et al. [18] were able to distinguish colorectal malignant tissues from normal tissues with a sensitivity of 96%.

The studies that were carried out using blood samples are presented in Table 11. The studies in the literature conducted on colon cancer using blood samples were examined. Barlev et al. [4] proposed a method for colorectal cancer screenings, with each subject represented by two feature vectors (11 principal components) from plasma and 13 principal components from peripheral blood mononuclear cells. Second derivatives of certain bands was taken so as to distinguish patients from healthy subjects, and the discriminant analysis method was applied. In conclusion, 81.5% sensitivity and 71.4% specificity levels were obtained (see Table 11).

In the current study, each sample is represented by a total of 16 features including five height ratios, five

Table 9. Best accuracy and average results obtained with SVM for 1300–1000 cm^{-1} range.

Number of folds	Kernel function	TP	FN	FP	TN	Sen (%)	Spe (%)	Acc (%)	Mean acc (%) ± Std
4	Polynomial	24	6	7	33	80.00	82.50	81.43	74.71 ± 3.30
	Linear	28	2	2	38	93.33	95.00	**94.29**	90.00 ± 2.33
	RBF	20	10	7	33	66.67	82.50	75.71	72.85 ± 2.51
5	Polynomial	25	5	9	31	83.33	77.50	80.00	75.71 ± 1.90
	Linear	27	3	2	38	90.00	95.00	**92.85**	89.28 ± 1.81
	RBF	19	11	8	32	63.33	80.00	72.86	71.43 ± 1.16
10	Polynomial	25	5	11	29	83.33	72.50	77.14	75.14 ± 2.04
	Linear	28	2	2	38	93.33	95.00	**94.29**	90.57 ± 2.45
	RBF	20	10	9	31	66.67	77.50	72.86	71.28 ± 1.25

Table 10. Best accuracy and average results obtained with SVM for 1800–1300 cm^{-1} range.

Number of folds	Kernel function	TP	FN	FP	TN	Sen (%)	Spe (%)	Acc (%)	Mean acc (%) ± Std
4	Polynomial	26	4	8	32	86.66	80.00	82.86	80.00 ± 2.13
	Linear	25	5	4	36	83.33	90.00	**87.14**	85.14 ± 1.37
	RBF	20	10	7	33	66.67	82.50	75.71	73.71 ± 2.15
5	Polynomial	27	3	7	33	90.00	82.50	85.71	83,42 ± 2.25
	Linear	24	6	3	37	80.00	92.50	**87.14**	85.71 ± 2.12
	RBF	19	11	8	32	63.33	80.00	72.86	71.00 ± 1.91
10	Polynomial	24	6	4	36	80.00	90.00	85.71	82.28 ± 2.53
	Linear	25	5	3	37	83.33	92.50	**88.57**	85.00 ± 1.54
	RBF	19	11	8	32	63.33	80.00	72.86	70.00 ± 2.13

area ratios, and six statistical features. Furthermore, the statistical values in certain ranges (1800–1300 cm^{-1} and 1300–1000 cm^{-1}) were calculated separately in order to distinguish cancer patients from healthy subjects, and the data were classified with higher accuracy using MLP and SVM (see Tables 8, 9, and 10).

In the studies on cancer detection using FTIR, analyses were performed especially within the range of 1800–750 cm^{-1}. The reason is that many important molecules used in cancer detection studies are to be found in this range [3]. In addition, the FTIR signals of healthy subjects and patients with colon cancer are very similar. For this reason, in addition to the height and area ratios examined in previous studies, the current study also examined entropy, standard deviation, absolute mean value, simple square integral, waveform length, and max–min difference features of the FTIR signal.

5. Conclusions

The data obtained in the current study showed that although the t-test was meaningful, none of the features were sufficient to make a distinction between patients and healthy subjects alone. For this reason, the current study classified all the features obtained from the FTIR signal together. The results obtained with the performed classification process are promising in the distinguishing of patients with colon cancer from healthy subjects.

Table 11. Comparison of studies on colon cancer using blood samples

Authors	Type (number of patients, healthy subjects)	Algorithms/features used	Results	
Barlev et al. [4]	Colorectal cancer (34+10, 18)	Principal component analysis, quadratic discrimination analysis, Fisher linear discrimination analysis	AUC: 0.77 Sensitivity: 81.50% Specificity: 71.40%	
			MLP	SVM
Proposed method	Colon cancer (30, 40)	Statistical feature extraction, SVM, MLP	Sen: 93.33% Spe: 97.50% Acc: 95.71%	93.33% 95.00% 94.29%

A large number of features of the FTIR signal were revealed by means of the feature extraction performed. Effective accuracy values were achieved using all of these features together. Results showed that these 16 features could be useful in the classifying of patients with colon cancer and healthy subjects.

Also, time for drying was not needed as the samples were measured in liquid state, and therefore a faster measurement was achieved. In the future, a subsequent study is planned to better understand the effect of the water in the plasma on the signal and to compare it with the results of the current study.

Acknowledgments

The authors would like to thank the Fırat University Hospital Central Laboratory and Bingöl University Central Laboratory Application and Research Center for their assistance and support in conducting this study.

References

[1] Bruening W, Sullivan N, Carter Paulson E, Zafar H, Mitchell M et al. Imaging Tests for the Staging of Colorectal Cancer. Rockville, MD, USA: Agency for Healthcare Research and Quality, 2014.

[2] Lech G, Słotwiński R, Słodkowski M, Krasnodebski IW. Colorectal cancer tumour markers and biomarkers: recent therapeutic advances. World Journal of Gastroenterology 2016; 22: 1745-1755.

[3] Ostrovsky E, Zelig U, Gusakova I, Ariad S, Mordechai S et al. Detection of cancer using advanced computerized analysis of infrared spectra of peripheral blood. IEEE Transactions on Biomedical Engineering 2012; 60: 343-353.

[4] Barlev E, Zelig U, Bar O, Segev C, Mordechai S et al. A novel method for screening colorectal cancer by infrared spectroscopy of peripheral blood mononuclear cells and plasma. Journal of Gastroenterology 2015; 51: 56-78.

[5] Baker MJ, Gazi E, Brown MD, Shanks JH, Gardner P et al. FTIR-based spectroscopic analysis in the identification of clinically aggressive prostate cancer. British Journal of Cancer 2008; 99: 1859-1866.

[6] Sindhuphak R, Issaravanich S, Udomprasertgul V, Srisookho P, Warakamin S et al. A new approach for the detection of cervical cancer in Thai women. Gynecologic Oncology 2003; 90: 10-14.

[7] Jusman Y, Sulaiman SN, Isa NAM, Yusoff IA, Adnan R et al. Capability of new features from FTIR spectral of cervical cells for cervical precancerous diagnostic system using MLP networks. In: Tencon 2009 IEEE Region 10 Annual International Conference; 23–26 January 2009; Singapore. New York, NY, USA: IEEE. pp. 1-6.

[8] Njoroge E, Alty SR, Gani MR, Alkatib M. Classification of cervical cancer cells using FTIR data. In: EMBS 2006 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 30 August–3 September 2006; New York, NY, USA: IEEE. pp. 5338-5341.

[9] Wan C, Cao W, Cheng C. Research of recognition method of discrete wavelet feature extraction and PNN classification of rats FT-IR pancreatic cancer data. Journal of Analytical Methods in Chemistry 2014; 2014: 564801.

[10] Li QB, Sun XJ, Xu YZ, Yang LM, Zhang YF et al. Diagnosis of gastric inflammation and malignancy in endoscopic biopsies based on Fourier transform infrared spectroscopy. Clinical Chemistry 2005; 51: 346-350.

[11] Li QB, Sun XJ, Xu YZ, Yang LM, Zhang YF et al. Use of Fourier-transform infrared spectroscopy to rapidly diagnose gastric endoscopic biopsies. World Journal of Gastroenterology 2005; 11: 3842-3845.

[12] Fujioka N, Morimoto Y, Arai T, Kikuchi M. Discrimination between normal and malignant human gastric tissues by Fourier transform infrared spectroscopy. Cancer Detection and Prevention 2004; 28: 32-36.

[13] Maziak DE, Do MT, Shamji FM, Sundaresan SR, Perkins DG et al. Fourier-transform infrared spectroscopic study of characteristic molecular structure in cancer cells of esophagus: an exploratory study. Cancer Detection and Prevention 2007; 31: 244-253.

[14] Eckel R, Huo H, Guan HW, Hu X, Che X et al. Characteristic infrared spectroscopic patterns in the protein bands of human breast cancer tissue. Vibrational Spectroscopy 2001; 27: 165-173.

[15] Fabian H, Thi NAN, Eiden M, Lasch P, Schmitt J et al. Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy. Biochimica et Biophysica Acta 2006; 1758: 874-882.

[16] Andronie L, Pânzaru SC, Cozar O, Domşa I. FT-IR spectroscopy for human colon tissue diagnostic. Romanian Journal of Biophysics 2011; 21: 85-91.

[17] Cheng C, Xiong W, Tian Y. Classification of rat FTIR colon cancer data using wavelets and BPNN. Chinese Journal of Chemistry 2009; 27: 911-914.

[18] Dong L, Sun X, Chao Z, Zhang S, Zheng J et al. Evaluation of FTIR spectroscopy as diagnostic tool for colorectal cancer using spectral analysis. Spectrochimica Acta A 2014; 122: 288-294.

[19] Xie YB, Liu Q, He F, Guo CG, Wang CF et al. Diagnosis of colon cancer with Fourier transform infrared spectroscopy on the malignant colon tissue samples. Chinese Medical Journal 2011; 124: 2517-2521.

[20] Erukhimovitch V, Talyshinsky M, Souprun Y, Huleihel M. FTIR spectroscopy examination of leukemia patients plasma. Vibrational Spectroscopy 2006; 40: 40-46.

[21] Sheng D, Liu X, Li W, Wang Y, Chen X et al. Distinction of leukemia patients' and healthy persons' serum using FTIR spectroscopy. Spectrochimica Acta A 2013; 101: 228-232.

[22] Ahmed SSSJ, Santosh W, Kumar S, Thanka Christlet TH. Neural network algorithm for the early detection of Parkinson's disease from blood plasma by FTIR micro-spectroscopy. Vibrational Spectroscopy 2010; 53: 181-188.

[23] Sheng D, Wu Y, Wang X, Huang D, Chen X et al. Comparison of serum from gastric cancer patients and from healthy persons using FTIR spectroscopy. Spectrochimica Acta A 2013; 116: 365-369.

[24] Wang X, Shen X, Sheng D, Chen X, Liu X. FTIR spectroscopic comparison of serum from lung cancer patients and healthy persons. Spectrochimica Acta A 2014; 122: 193-197.

[25] Toraman S, Türkoğlu I. A new automatic recognition method for determine colon cancer from FTIR sign patterns. In: ICNASE 2016 International Conference on Natural Science and Engineering; 19–20 March 2016; Kilis, Turkey. pp. 2354-2361 (in Turkish with an abstract in English).

[26] Sun X, Xu Y, Wu J, Zhang Y, Sun K. Detection of lung cancer tissue by attenuated total reflection-Fourier transform infrared spectroscopy - a pilot study of 60 samples. Journal of Surgical Research 2013; 179: 33-38.

[27] Hands JR, Dorling KM, Abel P, Ashton KM, Brodbelt A et al. Attenuated total reflection Fourier transform infrared (ATR-FTIR) spectral discrimination of brain tumour severity from serum samples. Journal of Biophotonics 2014; 7: 189-199.

[28] Toraman S. Feature extraction using infrared spectroscopy from blood samples related to colon cancer. PhD, Fırat University, Elazığ, Turkey, 2016 (in Turkish with an abstract in English).

[29] Poston WL, Marchette DJ. Recursive dimensionality reduction using Fisher's linear discriminant. Pattern Recognition 1998; 31: 881-888.

[30] Toraman S, Türkoğlu I. Automatic determination of desired cell fields on histopathologic images. In: IATS 2009 5th International Advanced Technologies Symposium; 13–15 May 2009; Karabük, Turkey. pp. 83-85 (in Turkish with an abstract in English).

[31] Phinyomark A, Limsakul C, Phukpattaranont P. A novel feature extraction for robust EMG pattern recognition. Journal of Computing 2009; 1: 71-80.

[32] Phinyomark A, Hirunviriya S, Limsakul C, Phukpattaranont P. Evaluation of EMG feature extraction for hand movement recognition based on Euclidean distance and standard deviation. In: ECTI-CON 2010 Electrical Engineering/Electronics Computer Telecommunications and Information Technology; 19–21 May 2010; Chiang Mai, Thailand. pp. 856-860.

[33] Phinyomark A, Nuidod A, Phukpattaranont P, Limsakul C. Feature extraction and reduction of wavelet transform coefficients for EMG pattern classification. Elektronika ir Elektrotechnika 2012; 122: 27-32.

[34] Vigneshwari C, Vimala V, Vignesh SV, Sumithra G. Analysis of finger movements using EEG signal. International Journal of Emerging Technology and Advanced Engineering 2013; 3: 583-588.

[35] Kharat PA, Dudul SV. Daubechies wavelet neural network classifier for the diagnosis of epilepsy. WSEAS Transactions on Biology and Biomedicine 2012; 9: 103-113.

[36] Hu X, Miller C, Vespa P, Bergsneider M. Adaptive computation of approximate entropy and its application in integrative analysis of irregularity of heart rate variability and intracranial pressure signals. Medical Engineering and Physics 2008; 30: 631-639.

[37] Martinez WL, Martinez AR. Computational Statistics Handbook with MATLAB. Boca Raton, FL, USA: CRC Press, 2002.

[38] Toraman S, Türkoğlu I. Efficient feature selection using sequential backward selection algorithm in the classification of histopathological images. In: ISERD 2015 16th International Society for Engineering Research and Development International Conference; 10 November 2015; Prague, Czech Republic. pp. 10-13.

[39] Yano K, Ohoshima S, Gotou Y, Kumaido K, Moriguchi T et al. Direct measurement of human lung cancerous and noncancerous tissues by Fourier transform infrared microscopy: can an infrared microscope be used as a clinical tool? Analytical Biochemistry 2000; 287: 218-225.

[40] Movasaghi Z, Rehman S, Rehman IU. Fourier transform infrared (FTIR) spectroscopy of biological tissues. Applied Spectroscopy Reviews 2008; 43: 134-179.

[41] Khazaee A, Ebrahimzadeh A. Classification of electrocardiogram signals with support vector machines and genetic algorithms using power spectral features. Biomedical Signal Processing and Control 2010; 5: 252-263.

[42] Arjmandi MK, Pooyan M. An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. Biomedical Signal Processing and Control 2012; 7: 3-19.

[43] Kavzoğlu T, |l|ölkesen İ. Investigation of the effects of kernel functions in satellite image classification using support vector machines. Harita Dergisi 2010; 144: 73-82 (in Turkish with an abstract in English).

[44] Elhaj FA, Salim N, Harris AR, Swee TT, Ahmed T. Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals. Computer Methods and Programs in Biomedicine 2016; 127: 52-63.

[45] Colak C, Karaman E, Turtay MG. Application of knowledge discovery process on the prediction of stroke. Computer Methods and Programs in Biomedicine 2015; 119: 181-185.

[46] Acır N, Öztura I, Kuntalp M, Baklan B, Güzelis C. Automatic detection of epileptiform events in EEG by a three-stage procedure based on artificial neural networks. IEEE Transactions on Biomedical Engineering 2005; 52: 30-40.

[47] Selver MA, Taygur MM, Seçmen M, Zoral EY. Hierarchical reconstruction and structural waveform analysis for target classification. IEEE Transactions on Antennas and Propagation 2016; 67: 3120-3129.

[48] Duda RO, Hart PE, Stork DG. Pattern Classification. New York, NY, USA: Wiley-Interscience Publication, 2001.

[49] Visa S, Ramsay B, Ralescu A, Knaap EVD. Confusion matrix-based feature selection. CEUR Workshop Proceedings 2011; 710: 120-127.

[50] Cheng CG, Tian YM, Jin WY. A study on the early detection of colon cancer using the methods of wavelet feature extraction and SVM classifications of FTIR. Spectroscopy 2008; 22: 397-404.