Spectroscopy 26 (2011) 155–165
DOI 10.3233/SPE-2011-0535
IOS Press

Study on the early detection of gastric cancer based on discrete wavelet transformation feature extraction of FT-IR spectra combined with probability neural network

Tao Hu [a,b,*], Yu-Hui Lu [c], Cun-Gui Cheng [d] and Xiao-Chen Sun [e]

[a] *Faculty of Life Science and Chemical Engineering, Huaiyin Institute of Technology, Huaian, China*
[b] *National Special Superfine Powder Engineering Center, Nanjing University of Science and Technology, Nanjing, China*
[c] *The First College of Clinic Medicine, Zhejiang Chinese Medical University, Hangzhou, China*
[d] *Department of Chemistry, Zhejiang Normal University, Jinhua, China*
[e] *Department of Physics, Zhejiang Normal University, Jinhua, China*

Abstract. This paper introduces a new method for the early detection of gastric cancer using a combination of feature extraction based on discrete wavelet transformation (DWT) for horizontal attenuated total reflectance–Fourier transform infrared spectroscopy (HATR–FT-IR) and classification using probability neural network (PNN). 344 FT-IR spectra were collected from 172 pairs of fresh normal and abnormal stomach tissue's samples. After preprocessing, 5 features were extracted with DWT analysis. Based on the PNN classification, all FT-IR spectra were classified into three categories. The accuracy of identifying normal gastric tissue, early gastric cancer tissue and gastric cancer tissue samples were 100.00, 97.56 and 100.00%, respectively. This result indicated that FT-IR with DWT and PNN could effectively and easily diagnose gastric cancer in its early stages.

Keywords: HATR–FT-IR, discrete wavelet transformation, probability neural network, earlier stage of gastric cancer, diagnose

1. Introduction

Cancer (medical term: *malignant neoplasm*) is a large, heterogeneous class of *diseases* in which a group of *cells* display uncontrolled growth, invasion that intrudes upon and destroys adjacent tissues, and often *metastasizes*, wherein the tumor cells spread to other locations in the body via the *lymphatic system* or through the *bloodstream*. Gastric cancer, commonly referred to as stomach cancer, is one of the deadliest diseases, and ranks second among cancers with at least 800,000 deaths worldwide per year. Gastric cancer is often *asymptomatic* or causes only *nonspecific symptoms* in its early stages. By the time symptoms occur, the cancer has often reached an advanced stage [8].

To find the cause of symptoms, the patient would be asked about his medical history, and may be ordered a physical exam and laboratory studies. The patient may also have one or all of the following

[*] Corresponding author: Tao Hu, Faculty of Life Science and Chemical Engineering, Huaiyin Institute of Technology, Huaian 223003, China. Tel.: +86 517 8359 1044; E-mail: hutao126@126.com.

exams: (1) gastroscopic exam is a frequently-used diagnostic method of choice; (2) upper GI series (called barium roentgenogram); (3) computed tomography or CT scanning of the abdomen may reveal gastric cancer, but is more useful to determine invasion into adjacent tissues, or the presence of spread to local lymph nodes. All of the techniques in the early detection of gastric cancer, gastroscopic exam is the uppermost method. In the exam, abnormal tissue seen in a gastroscope examination will be biopsied by the surgeon or gastroenterologist. This tissue is then sent to a pathologist for histological examination under a microscope to check for the presence of cancerous cells. A biopsy, with subsequent histological analysis, is the only sure way to confirm the presence of cancer cells. Various gastroscopic modalities have been developed to increased yield of detect mucosa with a dye that accentuates the cell structure and can identify areas of dysplasia. Endocytoscopy involves ultra-high magnification to visualize cellular structure to better determine areas of dysplasia. Other gastroscopic modalities such as optical coherence tomography are also being tested investigationally for similar applications [9].

A number of cutaneous conditions are associated with gastric cancer. A condition of darkened hyperplasia of the skin, frequently of the axilla and groin, known as acanthosis nigricans, is associated with intra-abdominal cancers such as gastric cancer. Other cutaneous manifestations of gastric cancer include *tripe palms* (a similar darkening hyperplasia of the skin of the palms) and the sign of Leser–Trelat, which is the rapid development of skin lesions known as seborrheic keratoses. Various blood tests may be done; including: complete blood count (CBC) to check for anemia. Also, a stool test may be performed to check for blood [11].

Fourier transform infrared spectroscopy (FT-IR) is a technique which is used to obtain an infrared spectrum of *absorption*, emission, photoconductivity or Raman scattering of a solid, liquid or gas. For an FT-IR spectrometer simultaneously collects spectral data in a wide spectral range, this confers a significant advantage over a dispersive spectrometer which measures intensity over a narrow range of wavelengths at a time. The term Fourier transform infrared spectroscopy originates from the fact that a Fourier transform (a mathematical algorithm) is required to convert the raw data into the actual spectrum. In another word, computer processing which is a common algorithm called the Fourier transform, is required to turn the raw data (light absorption for each mirror position) into the desired result (light absorption for each wavelength). FT-IR has the advantages of high distinguishing ability, rapid scanning time, large radiation fluxes, wide spectral range and trace research. Because of the above advantages of FT-IR, there are so many applications in analytical chemistry using FT-IR spectroscopy. The presence of grape seed oil in *Nigella sativa* L. seed oil had been analyzed Fourier transform infrared spectroscopy. The results had been disposed with some methods of chemometrics [10]. The oligomeric content of Aβ samples in the presence of apolipoprotein E isoforms had also been analyzed by the method of FT-IR. And form the research it was found that FT-IR spectroscopy could discriminate between Aβ42 oligomers and fibrils [1]. Some researchers used FT-IR as a sensitive and effective assay for the detection and discrimination between different fungal genera [13]. Our team had did some work in the applications of FT-IR spectroscopy and all the results showed the technique of Fourier transform make the applications of infrared spectroscopy much more extensive [3–5].

The wavelet transform can provide the information in local time and frequency scales together. It decomposes a signal into localized contributions labeled by a scale and a position parameter; each of the contributions represents the information of different frequency. The mainly effect of the applications of wavelet transform of FT-IR analysis is denoising, data compression, model transfer and background deduction. Because of its properties including orthogonality, orientation and flexible time–frequency windows, et al., wavelet transform could be widely used in the identifying traditional Chinese medicines, plant taxonomy and could also been used in early detection of gastric cancer [6].

Discrete wavelet transform (DWT) is one kind of wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information (location in time) [7].

This article focuses on how to efficiently identify normal gastric tissues, early gastric cancer tissues and gastric cancer tissues. The identification between normal tissue and early cancer tissue is difficult because their FT-IR spectra are very similar. In order to improve the classification accuracy rate, some important features were extracted in the DWT domain. These features were inputted to the probability neural network (PNN) to identify the normal gastric tissues, early gastric cancer tissues and gastric cancer tissues. Experimental results show that the new method is more efficient and much better in identification than traditional method solely based on HATR–FT-IR.

2. Theoretical section

2.1. DWT

In numerical analysis and functional analysis, DWT is a wavelet transform that the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information. Based on this advantage, DWT has a huge number of applications in science, engineering, mathematics and computer science. Most notably, it is used for signal coding to represent a discrete signal in a more redundant form, often as a preconditioning for data compression. DWT is originated from the discretization of continuous wavelet transformation (CWT) and the common discretization is dyadic. The CWT of a function or signal, for example, can be defined as

$$W_f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \Psi^* \left(\frac{t-b}{a} \right) \mathrm{d}t, \tag{1}$$

where $\Psi^*(t)$ denotes the mother wavelet function. The parameters a named as scale parameter and b named as translation parameter are respectively used to control the dilation and position of the mother function.

After the dyadic discretization, the function of DWT is accordingly expressed as

$$W_{\mathrm{dwt}}(j,k) = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} f(t) \Psi^* \left(\frac{t-2^j k}{2^j} \right) \mathrm{d}t, \tag{2}$$

where a and b are replaced by 2^j and $2^j k$. An efficient way to implement this scheme using filters was developed in 1989 by Mallat. The original signal $f(t)$ passes through two complementary filters and emerges as low frequency and high frequency signals. The decomposition process can be iterated, with successive approximations being decomposed in turn, so that a signal can be broken down into many lower-resolution components [7,12].

2.2. PNN

PNN is the feed-forward network model of artificial neural network based on the theory of statistics with the Parzen window function as the activate function. PNN absorbs the advantages of RBF

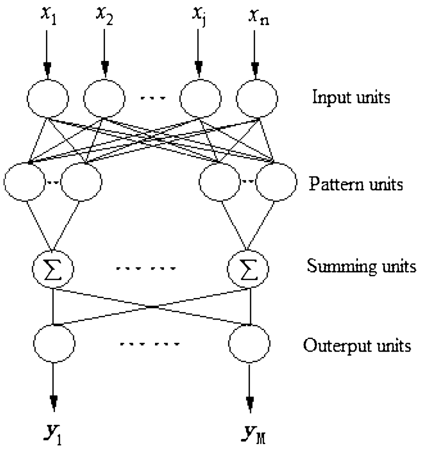


Fig. 1. The topological structure of PNN.

neural network and classical theory of probability density estimation. Compared with the traditional feed-forward neural network, it especially has the remarkable advantage in the pattern classification.

In the topological structure of PNN (Fig. 1), the first layout and last layout denote input neuro-units and output neuro-units respectively; the two middle layouts are hidden neuro-units. The first hidden layout is the pattern unit layout with the Parzen window function as the activate function, the second layout is summing unit layout which alternatively sums the output of the first layout. The training methods of PNN mostly have Recursion Orthogonal Least Square algorithm (ROLS) and Recursion Least Square (RLS). The two methods have quick convergence rate. In contrast, the training rate of the second one is faster and has higher training precision.

To the input vector x, output value Y_j of the jth nerve cell in the output layout of PNN can be written as:

$$Y_j = \sum_{k=1}^{M} w_{jk} H_k(x), \quad j = 1, 2, \ldots, M, \tag{3}$$

$$H_{k(x)} = \sum_{i=1}^{n_k} P_i(\|x - c_{kj}\|), \tag{4}$$

where x – input vector with dimension; $H_k(x)$ – output of the kth unit in the second hidden layout; w_{jk} – connected weight between the jth nerve cell in the second hidden layout and the kth nerve cell of output layout; $P(*)$ – Parzen window function; c_{kj} – the kth center vector of the jth class in the first hidden layout; n_k – the number of the center vectors of the kth class in the first hidden layout; $\|\cdot\|$ – Euclidean norm; M – the number of nerve cells in the output layout [2].

3. Experimental section

3.1. Tissue preparation

172 pairs of fresh normal and cancerous stomach tissues were collected from 172 patients. The patient group included 101 male and 71 female, aged between 42 and 70 years old. The fresh samples were

obtained from the Anatomical Pathology Laboratory at Jinhua Municipal Central Hospital (Zhejiang, China). The normal part of the gastric tissue was taken from the same stomach 5 to 10 cm away of the cancerous gastric tissues. Each tissue was divided into two equal parts, one was washed with aqueous NaCl on room temperature between 18 and 22°C, and the other was fixed with 10% formalin, embedded in paraffin and stained with hematoxylin and eosin for pathological examination. The collected sample tissues were put in the HATR (Horizontal Attenuated Total Reflection) for analysis.

3.2. Spectral measurements

FT-IR spectra have been collected in Thermo NEXUS 670 FT-IR spectrophotometer (Thermo, Madison, WI) and a single-bounce HATR accessory equipped with DTGS detector in the region 4000–650 cm^{-1} with fully computerized data storage. All data were processed with OMNIC 5.2 software. All spectra were recorded as 64-scans with 2 cm^{-1} resolution. The FT-IR spectrum background was recorded before collecting sample's FT-IR spectrum. Reference spectra were recorded using a blank HATR germanium wafer. Single beam spectra were obtained for all the samples and ratioed against the background spectra of air to present the spectra in absorbance units. The tissue sample was put on germanium wafer, and then impacted using pressure tower. FT-IR spectra were collected according to the instrument test requirement. After each experiment the HATR germanium wafer was thoroughly washed with distilled water and dried with nitrogen, and its spectra were examined to ensure that no residue from the previous experiment was retained on the Ge crystal surface. All of the tissue samples dried with absorbent cotton and further dried with nitrogen. The tissue samples cover the whole area of the HATR element that contributes to the spectral measurement. To ensure good contact with Ge crystal surface, all tissue samples were pressed against it using a pressure tower to give the same mechanical pressure on all the samples. All spectra were automatic baseline corrected. All experiments were repeated three times and the averaged spectra used for further analysis.

3.3. Data analysis

The FT-IR spectra of samples were obtained by measurement. The absorption values from different wave bands were based on the characters of the absorption value. The absorption values from different wave bands based on the characters of the absorption value were obtained by copying data method. Matlab 7.0 software was used for wavelet transformation. Daubechies wavelet, which possesses better exploration ability for signal singularity, is acted as analysis wavelet. One-dimension discrete wavelet transform were taken to different samples. The FT-IR spectra of the samples were decomposed into five layers in the DWT domain. Through a comparative analysis, two layers (3 and 4) were selected to extract eigenvector. The selected character variables were used for PNN training and identification.

4. Results and discussion

4.1. FT-IR analysis

The normal tissue, early cancer and malignant gastric tissue samples from 172 patients had been studied. The result of pathology reports of the typical normal gastric tissue, early gastric cancer tissue and gastric cancer tissue samples are shown as Fig. 2.

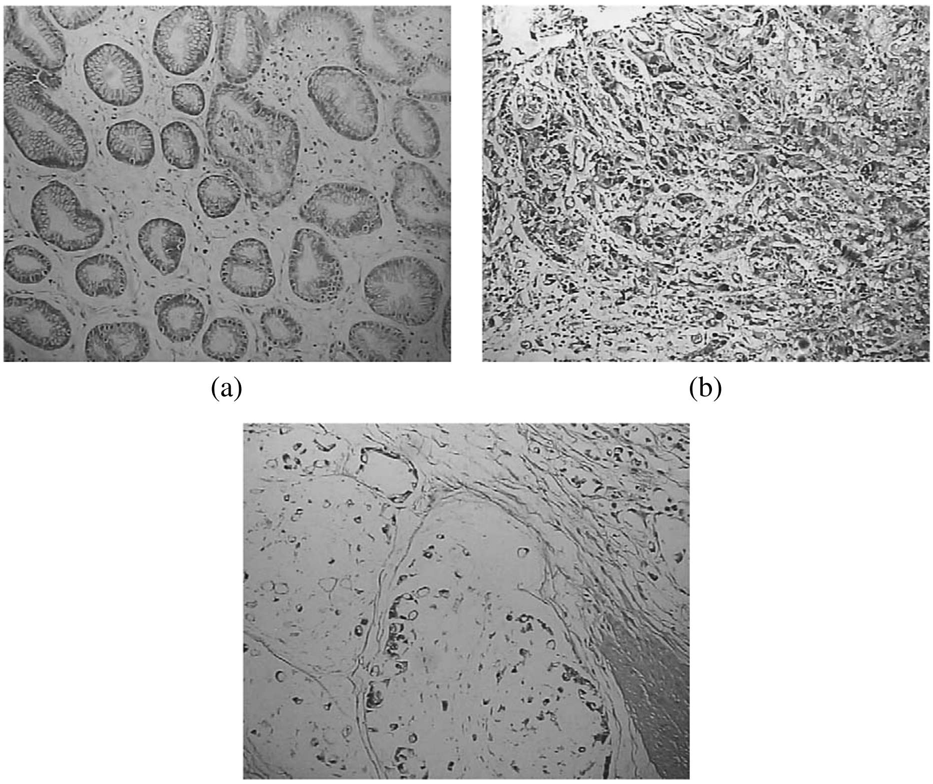
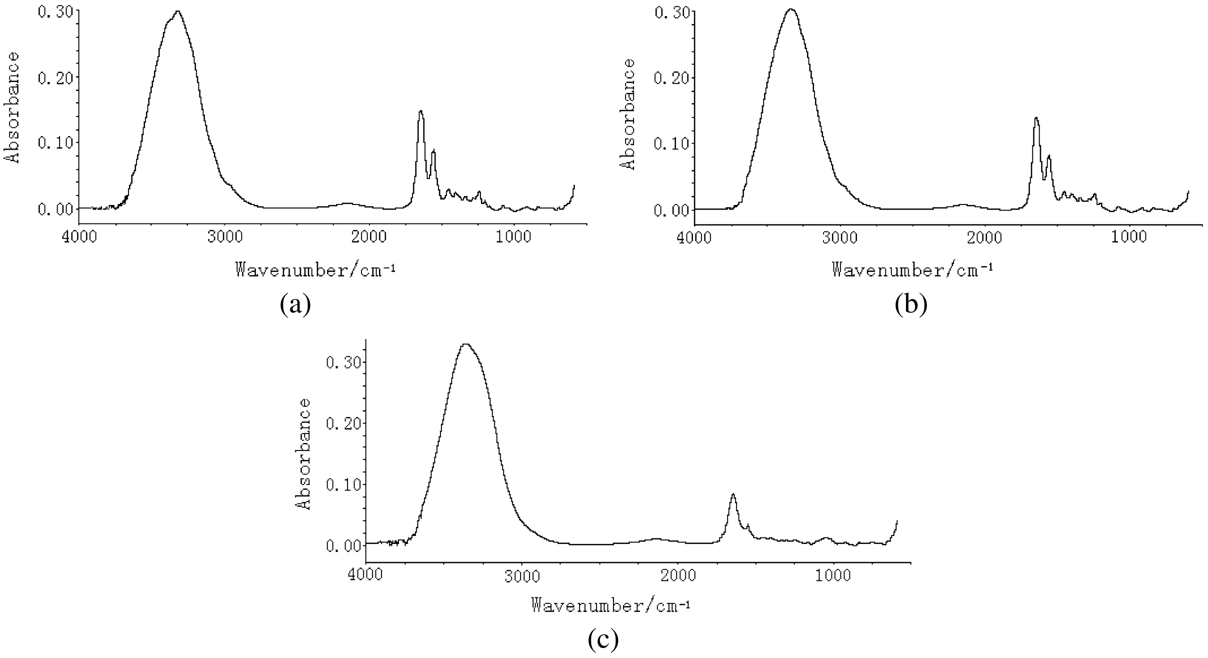


Fig. 2. The report of pathological section of (a) normal gastric tissue, (b) early gastric cancer tissue and (c) gastric cancer tissue samples.

The Fourier transform infrared spectra of normal gastric tissue, early gastric cancer tissue and gastric cancer tissues were collected by FT-IR spectrophotometer with HATR accessory directly. The FT-IR spectra of them are shown as Fig. 3.

From Fig. 3, we can see that the absorption band around 3500 cm^{-1} was attributed to –OH on stretching. In the region 1645–836 cm^{-1}, the absorption bands at 1645 and 1556 cm^{-1}, which are the amide I and amide II vibrations of the protein, and include water bending mode at 1640 cm^{-1} existing in tissues. The protein absorption peak of the cancerous gastric tissue has lower value than the normal tissue. The band at 1457 cm^{-1} is symmetric and asymmetric CH_3 bending of the methyl groups of lipid, protein, nucleic acid and glucide matter etc., respectively, and absorption band at 1339 cm^{-1} is the connective tissue of collagen. FT-IR spectra of normal gastric tissue had higher intensity of the absorption band in the range of 1457–1283 cm^{-1}, where the cancerous gastric tissue has no absorption. The absorption bands in the range of 1246–836 cm^{-1} is the absorption peak of stretching vibration of phosphodiester. The bands at 1206 and 1036 cm^{-1} are the connective tissue of collagen. The band at 1246 cm^{-1} is the asymmetric phosphate [PO^{2-} (asym)] stretching. The band at 1084 cm^{-1} is the symmetry phosphate [PO^{2-} (sym)] stretching. Glycogen is also a main contribution to the intensity of this band. The vibrations of the PO^{2-} groups are mainly in the phosphodiester groups on nucleic acids. The band at 1036 cm^{-1} which frequently found in the glycogen-rich tissues, is assigned as –CH_2OH vibrations and



(a)

(b)

(c)

Fig. 3. The typical FT-IR spectra of (a) normal gastric tissue, (b) early gastric cancer tissue and (c) gastric cancer tissue samples in the region 650–4000 cm^{-1}.

the C–O stretching coupled with C–O bending of the C–OH carbohydrates, respectively. The absorption peak of the cancerous tissue had been reduced or disappeared [5].

The difference of FT-IR spectrum between gastric cancer and normal gastric tissue is obvious, so they can be easily identified. The FT-IR spectrum between early gastric cancer, normal gastric tissue is similar (Fig. 3(a) and (b)), the accurate identification is difficult. In this paper, DWT is introduced to identify them.

The higher information quantities exist in 3600–2800 and 1800–650 cm^{-1}. Because the former includes the absorbability of OH and N–H stretching bands, and the character is not obvious. The later includes fingerprint region which contains more molecule structure information. Thus we select the region (1800–650 cm^{-1}) to analyze their spectra.

4.2. The result of DWT

The proper wavelet base and detail were decided by analyzing the signal spectra property and the comparison of decomposition results with different wavelet bases and details. We chose two feature peaks in the DWT domain to extract the features of FT-IR. Daubechies wavelet was selected as the analysis wavelet. The wavelet decomposition level was set as 5.

Figure 4 shown pre-processed Fourier transform infrared spectra of cancerous gastric tissue and its DWT coefficients, where, d1–d5 indicates detail information after decomposition.

From Fig. 4, we can see that the detailed signal d1 and d2 are a high frequency signal while d3, d4 and d5 are more sensitive to the changes of the spectra. Even the d5 approximation looks very similar with the original FT-IR spectral data, but it is smoother than the original FT-IR spectra after noise is removed. The d3 and d4 had been used as feature vector spaces. Figure 5 shows a diagram of a divided

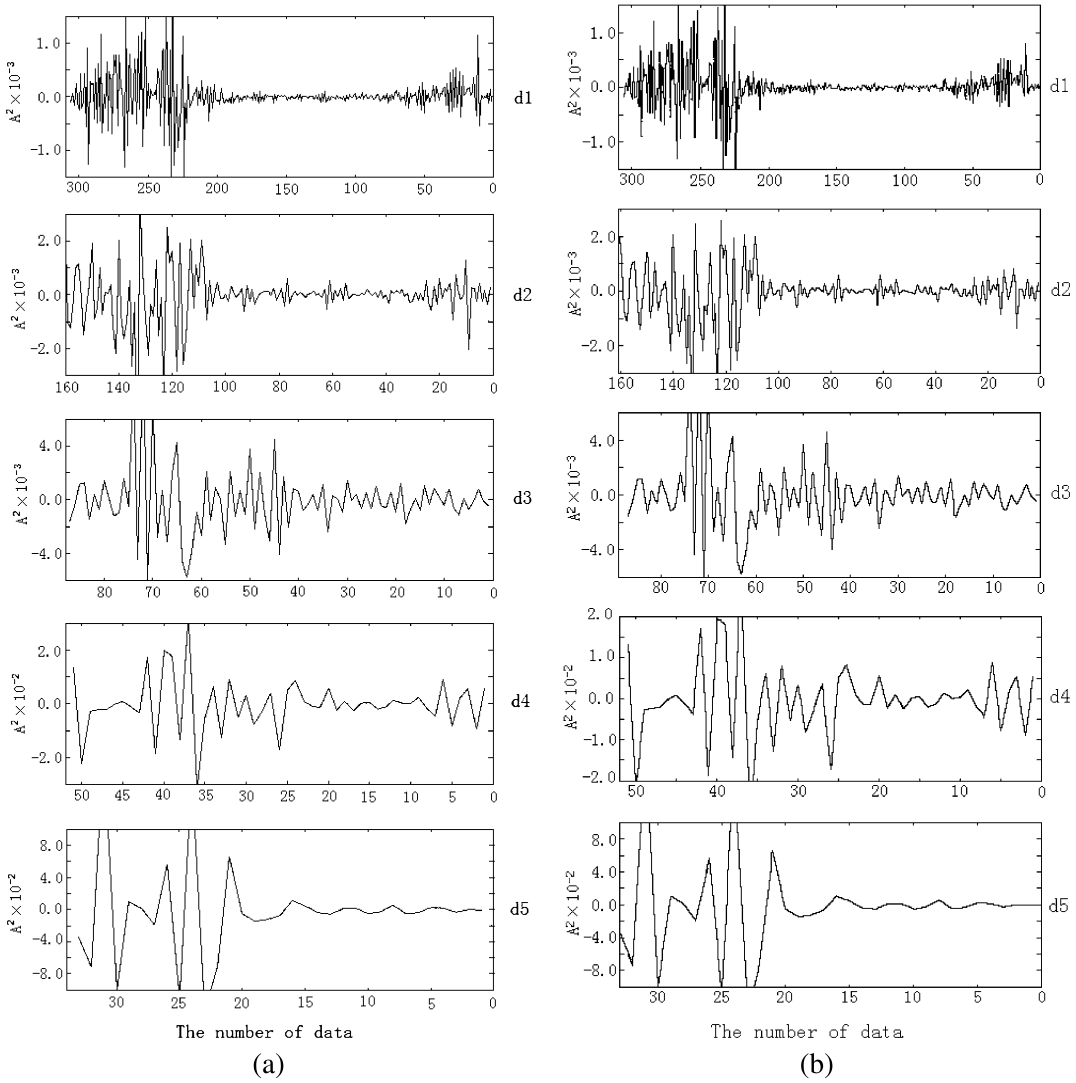


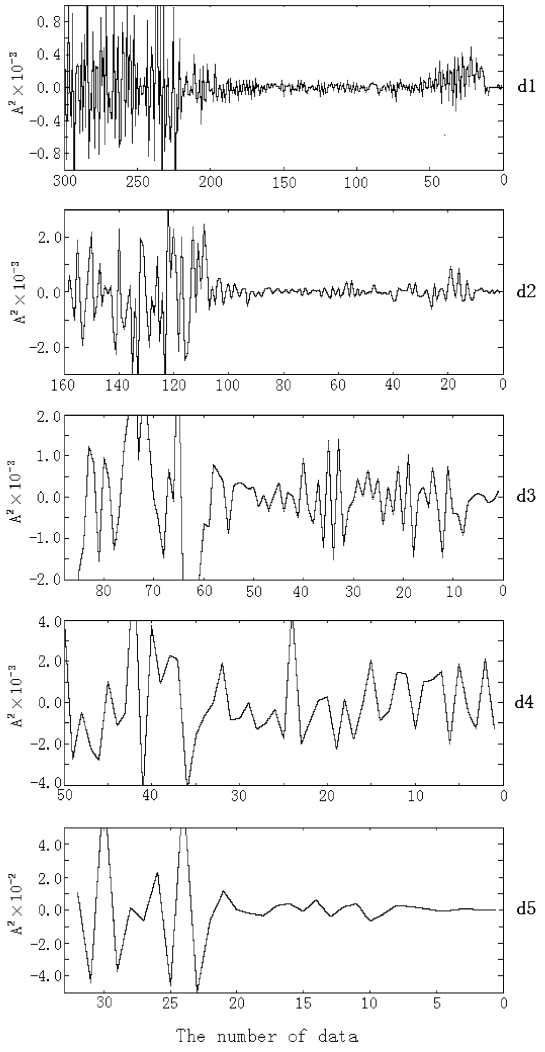
Fig. 4. The result of pre-processed spectra with DWT.

feature space. Three feature bands were selected from d3 and two feature bands were selected from d4. The feature vectors were defined as the energy (the sum of wavelet coefficient square) at each feature band. Thus, five feature variances were generated from two detailed signals.

4.3. The recognition results with PNN

After testing, we determine the structure of PNN as five nodes in the input layer, four nodes in the hidden layer and three nodes in the output layer, the error recognition rate is set as 1%.

PNN is used to identify the three kinds of tissue samples. For the training process, we use five input layer nodes of PNN structure, followed by normalized five feature vectors. The output layer nodes are divided into category 1 – normal tissues, category 2 – early cancer tissues and category 3 – cancer tissues.



(c)

Fig. 4. (Continued.)

The trained network is used to verify the 172 different sample data. The input data is the eigenvector extracted from the wavelet transformation of the original FT-IR spectra. The results are shown in Table 1.

As shown in Table 1, PNN was able to correctly identify all but two cases. The accuracy of identifying early gastric cancer tissues, and gastric cancer tissues from normal gastric tissues were 97.56, 100.00 and 100.00%, respectively.

5. Conclusions

The pathogenesis and development of gastric carcinoma are a multi-step process that is controlled by many genes and affected by many factors. FT-IR could provide information about molecular structure, which makes it possible to reflect the changes of protein, nuclear acid, sugar and fat in cells and the

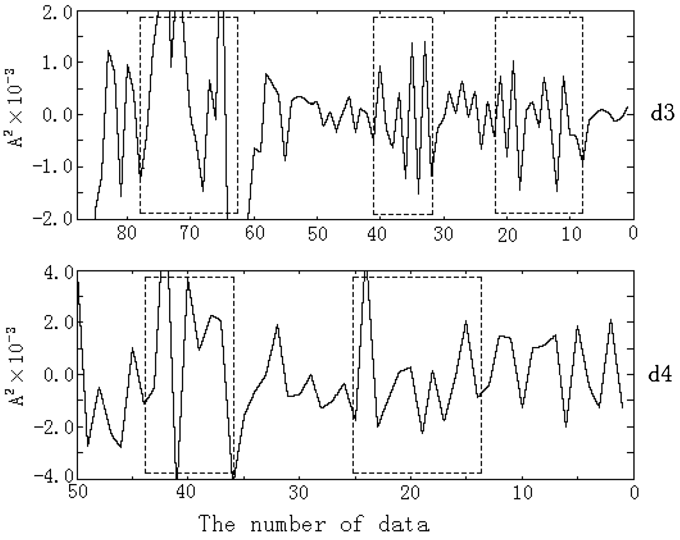


Fig. 5. Division of three feature regions of detail signal in the DWT domain.

Table 1

Indicates the recognition accuracy of normal gastric tissue, early gastric tissue and gastric cancer tissue samples (%)

	Training samples	Testing samples
Normal	100.00	100.00
Early carcinoma	97.56	97.56
Advanced carcinoma	100.00	100.00

structure changes of space array of the molecule, thus the diagnosis of cancer cells could be made at the molecular level. As cancerous degree is very small, normal tissues and early cancerous contain similar chemical composition. The FT-IR spectra between early gastric cancer tissue, normal gastric tissue are similar. It is difficult to identify them. By applying the wavelet feature extraction of the FT-IR data and PNN classification, better results can be achieved to distinguish the normal, early staged cancerous tissue and make the early detection of cancer become possible.

Acknowledgement

Thank the Anatomical Pathology Laboratory of Jinhua Municipal Central Hospital, Jinhua, Zhejiang province, China, for providing the samples.

References

[1] E. Cerf, J.-M. Ruysschaert, E. Goormaghtigh and V. Raussens, *Spectroscopy – Biomed. Appl.* **24** (2010), 245–249.
[2] C.G. Cheng, J. Liu, W.Q. Cao, R.W. Zheng, H. Wang and C.J. Zhang, *Vib. Spectrosc.* **54** (2010), 50–55.
[3] C.G. Cheng, J. Liu, H. Wang and W. Xiong, *Appl. Spectrosc. Rev.* **45** (2010), 165–178.
[4] C.G. Cheng, J. Liu, C.J. Zhang, M.Z. Cai, H. Wang and W. Xiong, *Appl. Spectrosc. Rev.* **45** (2010), 148–164.
[5] C.G. Cheng, Y.M. Tian and W.Y. Jin, *Spectroscopy – Biomed. Appl.* **22** (2008), 397–404.
[6] Q.H. Hong, R.P. Yu, R.W. Zheng, H. Wang and C.G. Cheng, *Chinese J. Chem.* **29** (2011), 1024–1030.

[7] T. Hu, W.Y. Jin and C.G. Cheng, *Spectroscopy – Biomed. Appl.* **25** (2011), 271–285.

[8] M. Huleihel, A. Salma, V. Erukhimovitch, J. Ramesh, Z. Hammody and S. Mordechai, *Biochem. Biophys. Methods* **50** (2002), 111–121.

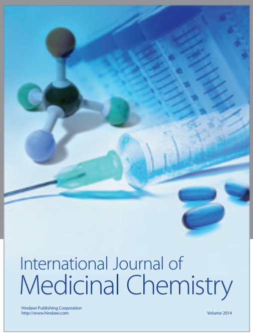
[9] H. Inoue, S.-ei Kudo and A. Shiokawa, *Nat. Clin. Pract. Gastr. Hepatol.* **2**(1) (2005), 31–37.

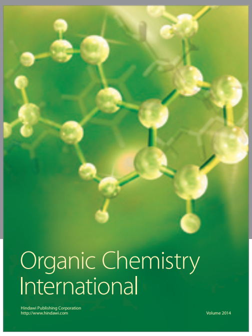
[10] A.F. Nurrulhidayah, Y.B. Che Man, H.A. Al-Kahtani and A. Rohman, *Spectroscopy – Biomed. Appl.* **25** (2011), 243–250.

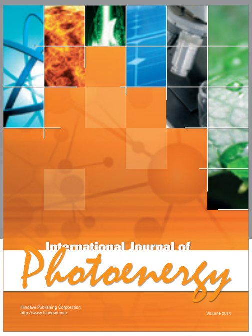
[11] M. Pentenero, M. Carrozzo, M. Pagano and S. Gandolfo, *Int. J. Dermatology* **43** (2004), 530–532.

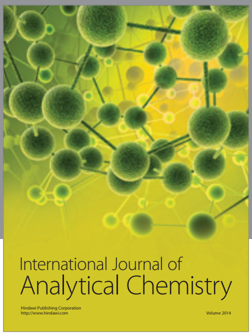
[12] S. Prabhakar, A.R. Mohanty and A.S. Sekhar, *Tribology Int.* **35** (2002), 793–800.

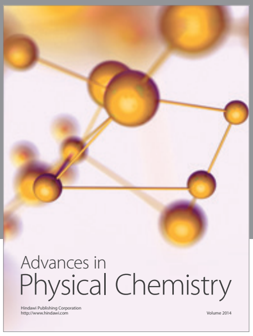
[13] A. Salman, L. Tsror, A. Pomerantz, R. Moreh, S. Mordechai and M. Huleihel, *Spectroscopy – Biomed. Appl.* **24** (2010), 261–267.

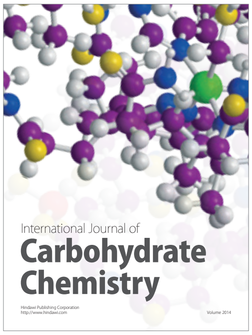

International Journal of
Medicinal Chemistry

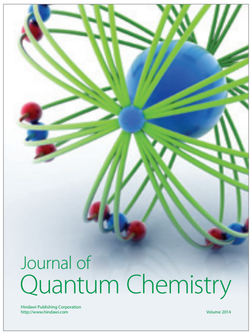

Organic Chemistry
International


International Journal of
Photoenergy

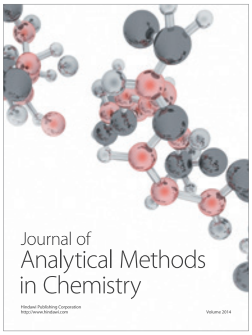

International Journal of
Analytical Chemistry


Advances in
Physical Chemistry

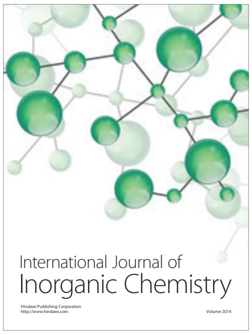

International Journal of
Carbohydrate
Chemistry


Journal of
Quantum Chemistry

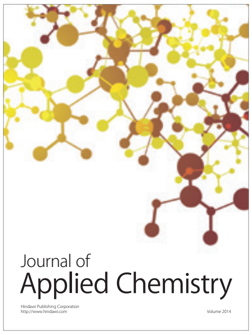
The Scientific
World Journal

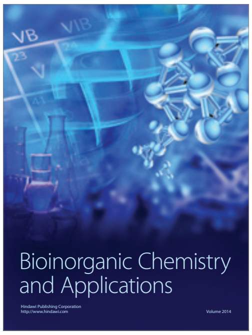

Journal of
Analytical Methods
in Chemistry

Journal of
Spectroscopy

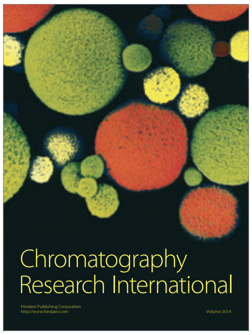

International Journal of
Inorganic Chemistry

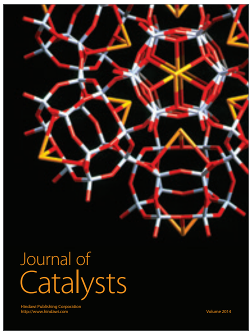
International Journal of
Electrochemistry

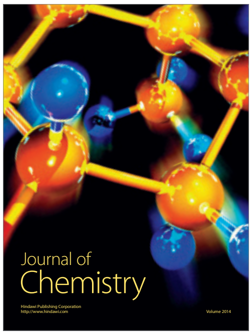

Journal of
Applied Chemistry


Bioinorganic Chemistry
and Applications

Journal of
Theoretical Chemistry


Chromatography
Research International


Journal of
Catalysts


Journal of
Chemistry

International Journal of
Spectroscopy