PCA and DWT with Resilient ANN based Organic Compounds Charts Recognition

Rabah N. Farhan
College of Computer,
University of Anbar
Ramadi, Iraq

Salah A. Aliesawi
College of Computer,
University of Anbar
Ramadi, Iraq

Zahraa Z. Abdulkareem
College of Computer,
University of Anbar
Ramadi, Iraq

ABSTRACT

A supervised learning depending on the resilient propagation neural network (RPROP) procedure has been used to solve the problem of FTIR charts recognition of the organic materials by training features extracted from two methods; principal component analysis (PCA) and discrete wavelet transform (DWT). During the testing process, it was found that; the best results are obtained from features that obtained from the principal component analysis, which in turn achieve a higher accuracy rate as well as the lowest false positive rate (where it gets accuracy rate about 97.22%, where the false positive rate about 2.7 %), where DWT get an accuracy rate about 91.6%, where the false positive rate about 8.3 %.

Keywords

FTIR spectrum; Discrete Wavelet Transform; Principal Component Analysis; Resilient Propagation Neural Network

1. INTRODUCTION

IR is a spectroscopy of the vibrations of the molecules that form the material and records absorptions of IR light that caused because of all the chemical bonds that found among all the molecules. Infrared (IR) spectral data plays an important role in many areas of applications such as applications depending on computer-based approaches for the interpretation of IR spectrum and can be classified into three fields [1]:

1) Knowledge-based systems in which depend on chemical expertise is encoded to assist in spectra interpretation;
2) Pattern recognition methods that used different data analysis tools such as statistics, and neural networks;
3) Search in spectral libraries by taking unknown spectra and comparing it with known spectra in library to know it.

Fourier Transform Infrared (FTIR) Spectroscopy is a non-destructive analytical technique that used for identifying and analyzing organic materials. It can also be used in the analysis of solids, liquids and gasses. The name of FTIR comes from the way in which the data is measured and converted into a spectrum in the frequency domain by using a mathematical technique that known as Fast Fourier transform(FFT). FTIR spectrum is equivalent to the "fingerprint" of the material. Therefore, it can be utilized in either quantitative or qualitative analysis. There are many researchers used the FTIR spectrum in their research to explore information from it in many areas where **Plamen N. Penchev et al. (1999)** [2] had classified the IR spectrum by determining the presence or absence of 20 chemical substructures in the spectrum. By deriving two types of features, the first one measures the intensity of the spectral band between two intervals, and the second one is measured from the logarithmic absorbance ratio. **R. Linker et al. (2007)** [3] presented a method for identifying of five Mediterranean soil types by PCA scores as features, linear discriminant analysis and probabilistic neural networks as classification methods. The results show that very high percentages of correct classification are achieved the use of the two methods. **Congo Cheng et al. (2007)** [4] classified FTIR cancer data of lung cancer as normal, early and advanced cancer, by using continuous wavelet analysis as feature extracted method and BPNN model for identification process. The identification accurate rate is 100% 100% 90% for normal, advanced cancer and early cancer respectively. **Marjana Noviˇc. (2008)**[5] they used hadamard transformation and Kohonen and counter propagation neural networks for prediction of structural fragments of an unknown compound from its infrared spectrum. They found that more than 90% of structural fragments were predicted correctly. **Yessi Jusman et al. (2009)** [6] diagnose normal and abnormal Cervical Precancerous cells by extract eight new features of frequency ranges and used some kind of ANN for classification. The best result was obtained from Levenberg-Marquardt Back-propagation (trainlm) algorithm because it gives 96.7 %. **XIE Yi-bin et al. (2011)** [7] diagnosed normal and malignant colon tissue of colon cancer by use principal component analysis (PCA) and support vector machine for classification. The classification result was 92.9% and 92.3% for sensitivity and specificity respectively.

2. METHODOLOGY

The block diagram that explains the whole methodology of this research is shown in Fig 1 for training and Fig 2 for testing. It can be seen that the proposed system is divided into three main parts: Preprocessing, feature extraction and classification.

The preprocessing is the first stage where original FTIR signals prepared for the next step by removing the possible noise that may occur. It's divided into three substages: smoothing, normalization and choice of region of interest, where the fingerprints of all the samples are in it. The next stage is the feature extraction stage, where (DWT) and (PCA) are used separately for the purpose of feature extraction. The last step is using (RPROP) for the purpose of classification. In this research, nine different materials spectrums of simple organic materials are used. Each spectrum of each material presented by 2 axes (X and Y) and each spectrum is represented by 934 points, where the X axis represents wavenumbers measured by (cm^{-1}) and it is the same for all the different materials and Y axis represents the transmittance of the material and it's different from one material to another.

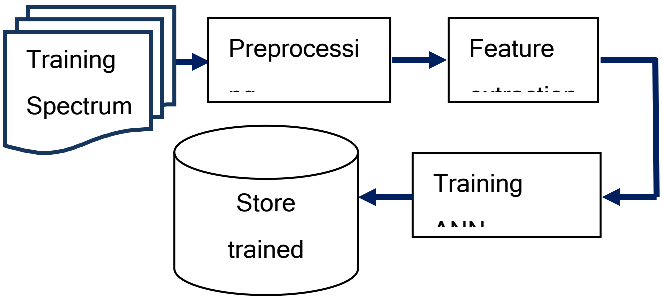


Fig 1:Block diagram of proposed training method of FTIR signal recognition.

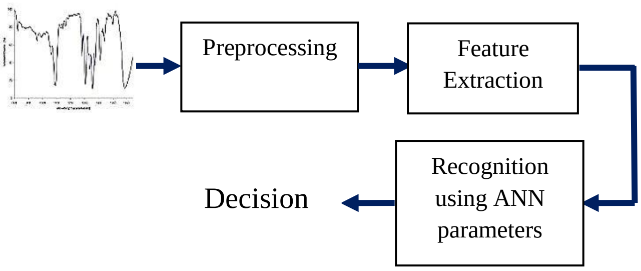


Fig 2: block diagram of testing step.

2.1 Database collection

The data were collected in college of education for woman in university of Anbar by using the Thermo scientific device of serial number AFM0900533 IR 100. A few milligrams of samples were mixed with a few milligrams of KBr powder .By putting the mixture in a die and pressure it, a thin disc (Pellet) is made for using it for examination. After that put this pellet in the FTIR spectroscopy that measures the spectrum.

2.2 Preprocessing

The digital filter technique was used for noise removal (such as small unwanted peaks that caused because of environmental effects). It is also used for smoothing signals. We used Savitzky–Golay filter that is used for smooth the chemistry signals by using a third degree polynomial and a window size of 25 points. The next step is to normalize the data because some samples are needed for normalization to reduce the testing error and to organize data efficiently. The third step is to choose the region of interest and because most the functional groups that give the graph it's fingerprint by their vibration manner are not exceeded these two periods (550-3500) so we can only choose this region as a region of interest.

2.3 Feature extraction

It is important after the preprocessing stage to extract the features from the FTIR waveform for signal analysis by using discrete wavelet transform or PCA for this purpose.

2.3.1 Wavelet transforms

Wavelet is time frequency idea that comes in the eighties and becomes widely used in many fileds such as signal processing, image processing, engineering applications, fluid dynamics, and other fields because it provides a reconstruction of the signal without redundancy [8]. There are two types of wavelet methods: discrete wavelet transform (DWT) and continuous Wavelet Transform (CWT). The ability of WT to decompose any given signal into a multiscale presentation makes it suitable to analyze a given signal on different frequency bands, and helps in defining the most essential scales of that signal [9]. In this work discrete wavelet transform was used.

Discrete wavelet transforms is a linear transformation that works on a data, these data have a length equal to an integer number of power two [10]. The discrete wavelet transform (DWT) is the standard wavelet transform algorithm, which uses the set of dyadic scales and translates of the mother wavelet to form an orthogonal basis for signal analysis. The equation of discrete wavelet transform and it's inverse can be expressed as [11].

$$\boldsymbol{a_{j,k} = \int_{-\infty}^{\infty} f(t)\psi_{j,k}(t)dt} \qquad (1)$$

$$f(t) = K_{\psi}\ \sum_j \sum_k a_{j,k}\ \psi_{j,k}(t) \qquad (2)$$

where function ψ(t) called the generating or mother wavelet.
In the proposed method, the signals were passed into two filters (low pass filter and high pass filter) as shown in Fig 3. The result of this process is two subsets of coefficients these two subsets are detailed coefficients that result from high pass (HPF) filter and the approximation coefficients that result from low pass filter (LPF).

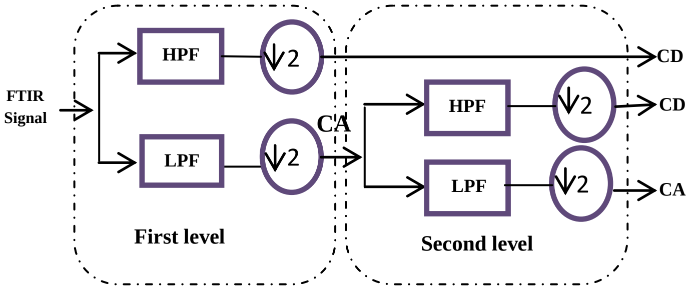


Fig 3: Diagram of wavelet implementation

So by applying multi_decomposition through passing the approximation subset to low pass filter and high pass filter, the filter that used is simple haar transform filter. Fig.4 show results of this step.

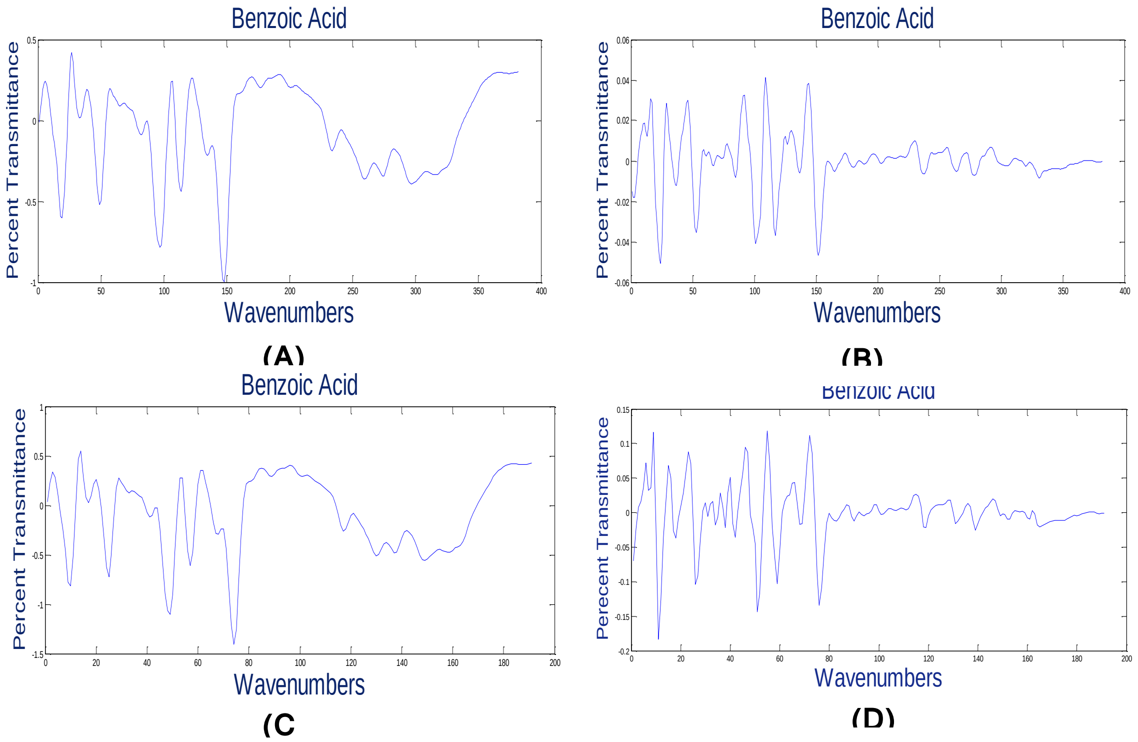


Fig 4: Applying wavelet transform(a) approximation coefficients(L) (b) detailed coefficients(H) (c) approximation coefficients(LL) (d) detailed coefficients(HL).

2.3.2 Principal component analysis (PCA)

The feature extraction algorithm can be achieved based on the ability of the PCA, where it is one of the oldest techniques, and has been rediscovered many times for finding a map from the original feature space to a lower dimensional feature space [12]. In the next way X (Wavenumbers) and Y (Transmittance) axis signals were used to compute PCA for each spectrum as shown in Fig.5.

step1 Data Adjusting: Is done by subtracting the mean of each axis from the original axis data for centering data by making it's mean equal to zero .

$$x_{new(i)} = x_{(i)} - \bar{x} \qquad (3)$$

$$y_{new(i)} = y_{(i)} - \bar{y} \quad , \qquad (4)$$

where $\bar{x}, \bar{y}$ are mean of x and y axis respectively.

step2 Find the Covariance Matrix: The covariance can be calculated from mean centered data using the equation

$$cov = \frac{1}{m-1} AA^T, \qquad (5)$$

where A is the array that needs to compute covariance for it, A^T is the transpose of array A. The basic formula for Covariance is expressed as

$$cov(x, y) = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{m-1} \qquad (6)$$

$$cov(x, y) = \frac{\sum_{i=1}^{m} x_{new(i)} y_{new(i)}}{m-1} \qquad (7)$$

Where x and y represent two separate dimensions of data and $\bar{x}, \bar{y}$ represent the mean of x and y axes respectively.

step3 Compute Eigen Values & Eigen Vectors (Feature vector): For the purpose of computing eigenvalues and eigenvectors we must apply this equation

$$Av = \lambda v \qquad (8)$$

where A, is $(M \times M)$ matrix, V is $(M \times 1)$ non-zero vector, λ is scalar.

where the scalar **λ** is an eigenvalue of A if there exists a non-zero vector v. So v can be denoted as eigenvector and λ as their corresponding eigenvalues. All the vectors v satisfying Av = **λ**v is called Eigenspace of A corresponding to eigenvalue **λ** .We can rewrite the condition Av = **λ**v as

$$A.v = \boldsymbol{\lambda}.v \qquad (9)$$

$$A.v - \boldsymbol{\lambda}.v.\ I=0 \qquad (10)$$

$$(A- \boldsymbol{\lambda} I)\ v =0 \qquad (11)$$

where I, is n x n identity matrix.

By finding the roots of |**A-λ.I**| that will give the eigenvalues and for each of these eigenvalues there will be an eigenvector, this eigenvector can be considered as the weights in a linear transformation when computing principal component scores.

step4 Compute Row Feature Vector: it can be found by transpose of Eigenvectors matrix.

step5 Compute New Data Set

New (Final) Data =Row Feature Vector × RowData Adjust.

After finding the PCA, the results show that the first 70 features of the second principal component analysis are enough for recognition, where it will be fed into the neural as inputs.

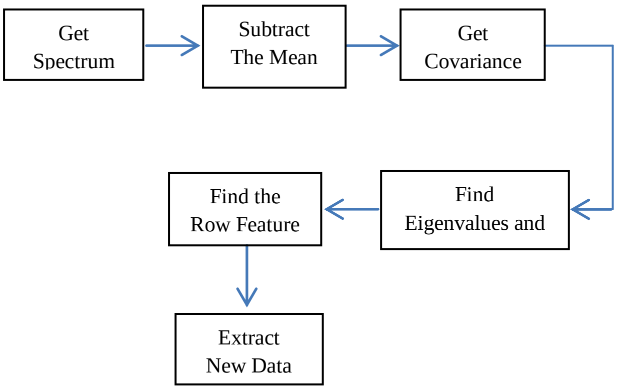


Fig 5: PCA Algorithm Steps.

3. CLASSIFICATION

For the purpose of classifying the algorithm of resilient propagation neural network (**RPROP**) was used, where it was produced by Riedmiller and Braun in 1993 as an adaptive learning process [13-15] and it's one of the best training methods for neural networks, where it's used for feed-forward neural network learning method of supervised learning. In this algorithm the weights adjusted using the following equation

$$w_i(t+1) = w_i(t) + \Delta w_i(t) \qquad (12)$$

The weight is updated $\Delta w_i(t)$ depending on the sign of the derivative. When the derivative sign is positive that the mean error increased, the weight is decreased by its update-value, but if the derivative is negative, and the update-value is added [14, 15]:

$$\Delta w_i(t) = \begin{cases} -\Delta(t), \text{ if } \frac{\partial E \partial(t)}{\partial w_i} > 0 \\ -\Delta(t), \text{ if } \frac{\partial E \partial(t)}{\partial w_i} < 0 \\ 0, \quad \text{else} \end{cases}$$

So we can write the previous weights adjusted equation as

$$w_i(t+1) = w_i(t) - sign\left(\frac{\partial E(t)}{\partial w_i}\right).\Delta_i(t) \quad , \qquad (14)$$

Where the t, is the iteration number, $\frac{\partial E(t)}{\partial w_i}$ partial derivative and $\Delta_i(t)$ is updated-value. Each weight has its individual update-value.

Where update-value follows the next learning rule [14, 15]

$$\Delta_i(t) = \begin{cases} \eta^+ * \Delta(t-1), \text{ if } \frac{\partial E \partial(t-1)}{\partial w_i} * \frac{\partial E \partial(t)}{\partial w_i} > 0 \\ \eta^- * \Delta(t-1), \text{ if } \frac{\partial E \partial(t-1)}{\partial w_i} * \frac{\partial E \partial(t)}{\partial w_i} < 0 \\ \Delta(t-1), \text{else} \end{cases} \qquad (15)$$

Where $0 < \eta^- < 1 < \eta^+$

There is one exception, if the partial derivative changes sign, i.e. the previous step was too large and the minimum was missed, the previous weight-update is reverted [15]

$$\Delta w_i(t+1) = \Delta w_i(t-1), \text{ if } \frac{\partial E \partial(t-1)}{\partial w_i} * \frac{\partial E \partial(t)}{\partial w_i} < 0$$

(16)

4. RESULTS

The topology of (RPROP) that used for this research is: input layer with 191 neuron for Wavelet features and 70 neuron features for PCA, one hidden layer with 50 neurons. The output layer has 9 neurons, which produces nine classes of the output represented as (000000001, 000000010, ……., 1000000000).

This section will be compared between the experimental result obtained from (DWT) and experimental result obtained from (PCA). Results are based mainly on calculating the accuracy rate and False Positive Rate (FPR) for implementing spectrum recognition system. They are given by

$$\textit{Accuracy Rate (AR)} = \frac{N_R}{N_T} * 100 \qquad (17)$$

$$\textit{False Positive Rate (FR)} = \frac{N_F}{N_T} * 100 \ , \qquad (18)$$

Where N_T is the total number of all samples of the material, N_R is the number of correctly recognized as the correct material, N_F is the total number of spectrums that are not recognized well as the correct material. Tables (1) and (2) show the results of a test of each of the nine materials for (DWT) and (PCA) feature extraction methods, where (PCA) has an accuracy rate equal to 97.22% compared with the accuracy rate obtained from (DWT), where it was equal to 91.6%, while the false positive rate(FPR) for the (PCA) and (DWT) were 2.7%, 8.3% respectively.

Table 1. Results with (DWT) features.

Metrial Name	*Training set*	*Testing Set*	N_R	N_F	*AR%*	*FR%*
Aminobenzoic Acid	5	4	3	1	75	25
Benzoic Acid	5	4	3	1	75	25
Caffeine	5	4	4	0	100	0
Cholesterol	5	4	4	0	100	0
Glucose	5	4	4	0	100	0
Glycine	5	4	4	0	100	0
Indol	5	4	3	1	75	25
Phthalic Acid	5	4	4	0	100	0
Picolinic Acid	5	4	4	0	100	0
Total	45	36	33	3	91.6	8.3

Table 2. ***Results with PCA features***

Metrial Name	*Training set*	*Testing Set*	*N_R*	*N_F*	*AR %*	*FR %*
Aminobenzoic Acid	5	4	3	1	75	25
Benzoic Acid	5	4	4	0	100	0
Caffeine	5	4	4	0	100	0
Cholesterol	5	4	4	0	100	0
Glucose	5	4	4	0	100	0
Glycine	5	4	4	0	100	0
Indol	5	4	4	0	100	0
Phthalic Acid	5	4	4	0	100	0
Picolinic Acid	5	4	4	0	100	0
Total	45	36	35	1	97.22	2.7

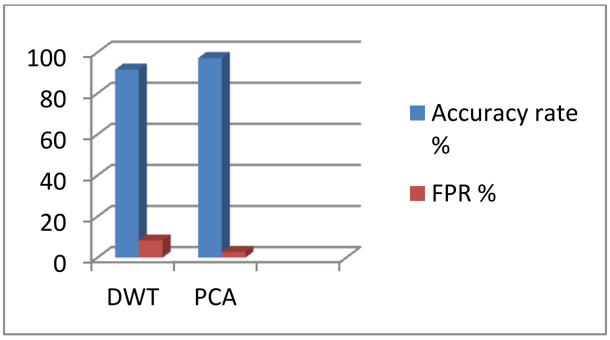
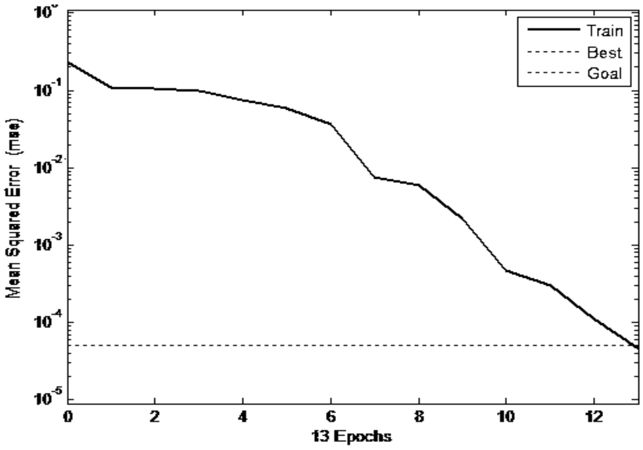
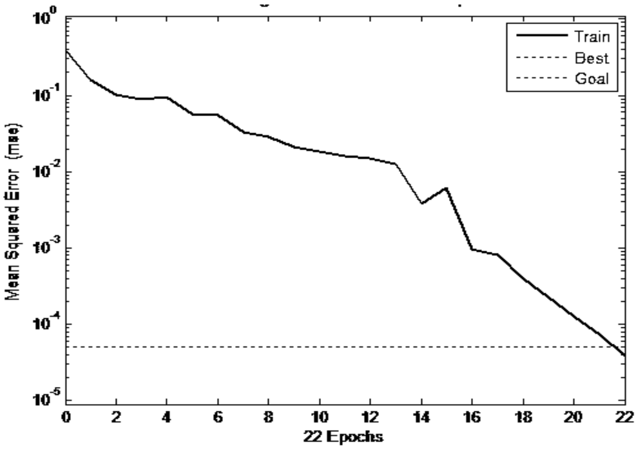


Fig 6: Experimental results of total accuracy and (FPR) for both approaches



(a



(b)

Fig 7: Mean squared error with the epoch for features from (a) DWT approach (b) PCA approach

So the number of epochs of (DWT) training are smaller than for PCA but the gradient for PCA is smaller than for DWT, where the training time is equal for the two approaches.

Table 3. Comparison results for two feature extraction approaches.

Approach	Training Time(S)	Epoch	Gradient	Accuracy	FPR
DWT	0.00.02	13	0.00016302	91.6%	8.3%
PCA	0.00.02	22	0.05357304	97.22%	2.7%

5. CONCUSIONS

In this paper, some samples of simple organic compounds are collected, preprocessed and their features are extracted by using DWT and PCA approaches where the second principal component analysis was used rather than the first principal component analysis because the first principal component analysis will be having the same results because it occurres by applying this law

New (Final) Data =Row Feature Vector × RowData Adjust.

Where (RowData Adjust) represent x and y axes values. So the first principal component will obtaine by multiplying (Row Feature Vector) by x axis data, so that will not provide variation because x axis is equal to all the nine materials where it represents wavenumber. So the features extracted by using one of these two previous DWT and PCA approaches are trained using resilient propagation neural network (RPROP). Table (3) shows the difference between the two feature extracted methods (DWT and PCA), where it provides that, where this approach will extract better features than (DWT), where the accuracy rate of PCA was 97.22%, with a false positive rate equal to 2.7%, where the accuracy of DWT was 91.6%, with a false positive rate equal to 8.3% even if the epoch of DWT is smaller than that of PCA.

6. REFERENCES

[1] K. Varmuza, P. N. Penchev, and H. Scsibrany, 1998. "Maximum Common Substructures of Organic

Compounds Exhibiting Similar Infrared Spectra", American Chemical Society, 19 February.

[2] Plamen N. Penchev, George N. Andreev, Kurt Varmuza, 1999. "Automatic classification of infrared spectra using a set of improved expert-based features", Elsevier Science B.V, January.

[3] C. Du, R. Linker, A. Shaviv, 2007. "Identification of agricultural Mediterranean soils using mid-infrared photoacousticspectroscopy", Elsevier.B.V, December.

[4] Cungui Cheng, YumeiTian, Changjiang Zhang, 2007. "Classification of FTIR Cancer Data Using Wavelets and BPNN", Proc. of SPIE Vol. 6826 682633-1.

[5] Marjana Noviˇc, 2008. chapter 4,"Kohonen and Counter propagation Neural Networks Applied for Mapping and Interpretation of IR Spectra", from the book (Artificial Neural Networks: Methods and Applications (Methods in Molecular Biology, by David J. Livingstone)), Springer.

[6] Yessin Jusman, Siti Noraini Sulaiman, Nor Ashidi,Mat Isa and Intan Aidha Yusoff, Rohana Adnan, Ahmad Zaki, Nor Hayati Othman, 2009. "Capability of New Features from FTIR Spectral of Cervical Cells for Cervical Precancerous Diagnostic System Using MLP Networks", IEEE.

[7] XIE Yi-bin, LIU Qian, HE Fei, GUO Chun-guang, WANG Cheng-feng and ZHAO Ping, 2011"Diagnosis of colon cancer with Fourier transform infrared spectroscopy on the malignant colon tissue samples", Chinese Medical Journal.

[8] Hannu Olkkonen, 2011. "Discrete Wavelet Transforms – Biomedical Applications", InTech, Croatia, ISBN 978-953-307-654-6, available free at www.intechopen.com.

[9] M. D. Pawar, S. M. Badave, 2011. "Speaker Identification System Using Wavelet Transformation and Neural Network", International Journal of Computer Applications in Engineering Sciences, ISSN: 2231-4946, VOL I, SPECIAL ISSUE ON CNS.

[10] Marcin Kociołek, Andrzej Materka, Michał Strzelecki, Piotr Szczypiński, 2001. "Discrete Wavelet Transform – Derived Features For Digital Image Texture Analysis", International Conference on Signals and Electronic Systems, Lodz, Poland, pp. 163-168.

[11] Daniel T.L.Lee, Akio Yamamoto, 1994." wavelet analysis: Theory and applications", Hewlett –Packard Journal.

[12] Yijuan Lu**,** Ira Cohen, Xiang Sean Zhou, Qi Tian, 2007**.** "Feature Selection Using Principal Feature Analysis", ACM Multimedia, Augsburg, Bavaria, Germany.

[13] Martin Riedmiller, Heinrich Braun, 1993. "A Direct Adaptive Method for Faster Back propagation Learning: The RPROP Algorithm", IEEE.

[14] Fawzi M. Al_Naima, Ali H. Al_Timemy, "Resilint Back Propagation Algorithem For Breast Biopsy Classification Based On Artificial Neural Networks", www.intechopen.com.

[15] Frauke Günther, Stefan Fritsch, 2010. "neuralnet: Training of Neural Networks", The R Journal, Vol. 2/1, ISSN 2073-4859.