Available online at www.sciencedirect.com
ScienceDirect
Digital Signal Processing 17 (2007) 694–701
Digital Signal Processing
www.elsevier.com/locate/dsp

Breast cancer diagnosis using least square support vector machine

Kemal Polat *, Salih Güneş

Electrical and Electronics Engineering Department, Selcuk University, 42075 Konya, Turkey

Available online 27 November 2006

Abstract

The use of machine learning tools in medical diagnosis is increasing gradually. This is mainly because the effectiveness of classification and recognition systems has improved in a great deal to help medical experts in diagnosing diseases. Such a disease is breast cancer, which is a very common type of cancer among woman. In this paper, breast cancer diagnosis was conducted using least square support vector machine (LS-SVM) classifier algorithm. The robustness of the LS-SVM is examined using classification accuracy, analysis of sensitivity and specificity, k-fold cross-validation method and confusion matrix. The obtained classification accuracy is 98.53% and it is very promising compared to the previously reported classification techniques. Consequently, by LS-SVM, the obtained results show that the used method can make an effective interpretation and point out the ability of design of a new intelligent assistance diagnosis system.

Keywords: Breast cancer diagnosis; Wisconsin breast cancer diagnosis data; Least square support vector machine; Confusion matrix; k-Fold cross validation; Medical diagnosis

1. Introduction

There is a considerable increase in the number of breast cancer cases in recent years. It is reported in [1] that breast cancer was the second one among the most diagnosed cancers. It is also stated that breast cancer was the most prevalent cancer in the world by the year 2002. Breast cancer outcomes have improved during the last decade with development of more effective diagnostic techniques and improvements in treatment methodologies. A key factor in this trend is the early detection and accurate diagnosis of this disease. The long-term survival rate for women in whom breast cancer has not metastasized has increased, with the majority of women surviving many years after diagnosis and treatment [2].

The use of classifier systems in medical diagnosis is increasing gradually. There is no doubt that evaluation of data taken from patient and decisions of experts are the most important factors in diagnosis. But, expert systems and different artificial intelligence techniques for classification also help experts in a great deal. Classification systems, helping possible errors that can be done because of fatigued or inexperienced expert to be minimized, provide medical data to be examined in shorter time and more detailed.

In this study, LSSVM was used to diagnose the breast cancer. 100, 96.30, and 94.44% classification accuracies were obtained by using LSSVM for 50–50% of training-test partition, 70–30% of training-test partition, and 80–20%

* Corresponding author. Fax: +90 332 241 0635.
E-mail address: kpolat@selcuk.edu.tr (K. Polat).

1051-2004/$ – see front matter
doi:10.1016/j.dsp.2006.10.008

of training-test partition, respectively. Also, k-fold cross validation, confusion matrix, and sensitivity and specificity analysis was used to show the diagnostic performance of LS-SVM.

The used data source is Wisconsin breast cancer dataset (WBCD) taken from the University of California at Irvine (UCI) machine learning repository [3]. This dataset is commonly used among researchers who use machine learning (ML) methods for breast cancer classification and so it provides us to compare the performance of our system with other conducted studies related with this problem.

The rest of the paper is organized as follows. Section 2 gives the background information breast cancer classification problem and previous research in literature. We present the proposed method in Section 3. In Section 4, we give the experimental data to show the effectiveness of our method. Finally, we conclude this paper in Section 5 with future directions.

2. Background

2.1. Breast cancer dataset

In a study conducted by Parkin et al., the cancer statistic was obtained relating 20 large 'areas' of the world [4]. According to this research, finished by the year 2002, the most prevalent cancer type was found to be breast cancer. As can be seen from Fig. 1 [1], the incidence of new cases for breast cancer is the most encountered cancer type for women in both developed and developing countries. The mortality of breast cancer is also very high with regard to the other cancer types.

Cancer begins with uncontrolled division of one cell and results in a visible mass named tumour. Tumour can be benign or malignant. Malignant tumour grows rapidly and invades its surrounding tissues through causing their damage. Breast cancer is a malignant tissue beginning to grow in the breast. The abnormalities like existence of a breast mass, change in shape and dimension of breast, differences in the colour of breast skin, breast aches, etc. are the symptoms of breast cancer. Cancer diagnosis is performed based on the nonmolecular criterions like tissue type, pathological properties and clinical location [5]. As for the other cancer types, early diagnosis in breast cancer can be life saving. The used data source in this study was taken from UCI machine learning repository [3].

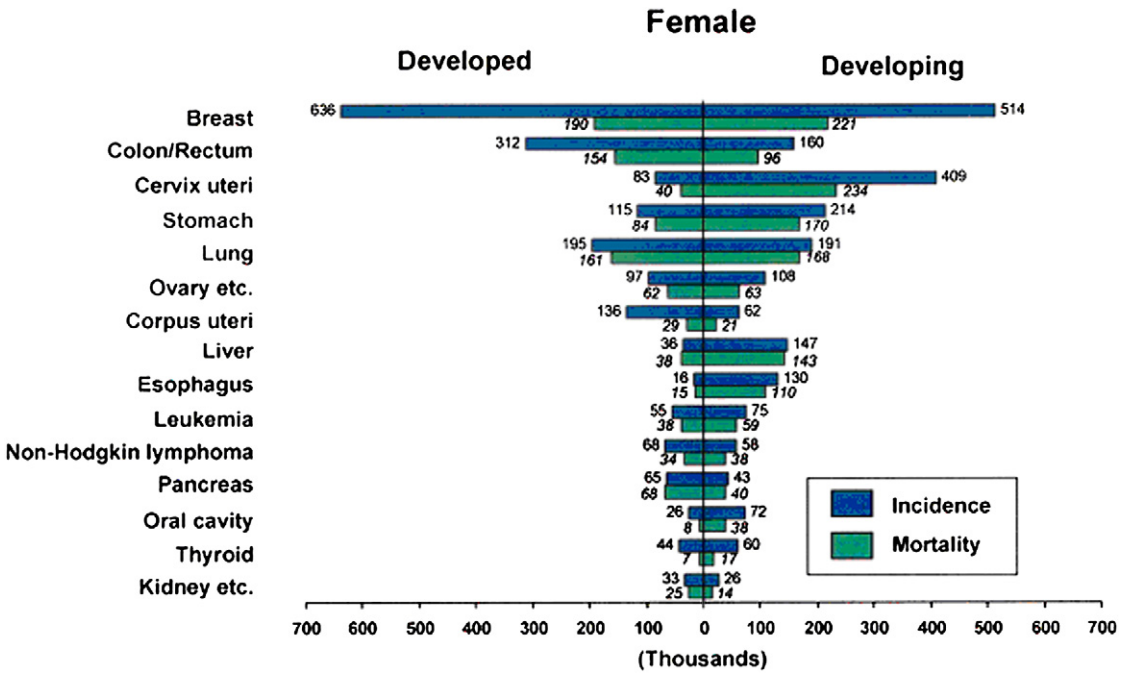

Fig. 1. Estimated numbers of new cancer cases (incidence) and deaths (mortality) in 2002 [1].

The name of the dataset for breast cancer problem is WBCD (Wisconsin breast cancer dataset). The dataset consists of 683 samples that were collected by Dr. W.H. Wolberg at the University of Wisconsin–Madison Hospitals taken from needle aspirates from human breast cancer tissue [6]. The WBCD database consists of nine features obtained from fine needle aspirates, each of which is ultimately represented as an integer value between 1 and 10. The measured variables are as follows:

(1) clump thickness (x_1);
(2) uniformity of cell size (x_2);
(3) uniformity of cell shape (x_3);
(4) marginal adhesion (x_4);
(5) single epithelial cell size (x_5);
(6) bare nucleoi (x_6);
(7) bland chromatin (x_7);
(8) normal nuclei (x_8);
(9) mitoses (x_9).

444 samples of the dataset belong to benign, and remaining 239 data is of malignant.

2.2. Previous research in diagnosis of breast cancer

As for other clinical diagnosis problems, classification systems have been used for breast cancer diagnosis problem, too. When the studies in the literature related with this classification application are examined, it can be seen that a great variety of methods were used which reached high classification accuracies using the dataset taken from UCI machine learning repository. Among these, Quinlan reached 94.74% classification accuracy using 10-fold cross validation with C4.5 decision tree method [7]. Hamilton et al. obtained 96% accuracy with RIAC method [8] while Ster and Dobnikar obtained 96.8% with linear discreet analysis (LDA) method [9]. The accuracy obtained by Bennett and Blue who used support vector machine (SVM) ($5 \times$ CV) method was 97.2% [10] while by Nauck and Kruse was 95.06% with neuro-fuzzy techniques [11] and by Pena-Rayes and Sipper was 97.36% using fuzzy-GA method [12]. Moreover, Setiono was reached 98.1% using neuro-rule method [13]. Goodman et al. applied three different methods to the problem which were resulted with the following accuracies: optimized-LVQ method's performance was 96.7%, big-LVQ method reached 96.8% and the last method, AIRS, which he proposed depending on the artificial immune system, obtained 97.2% classification accuracy [14]. Nevertheless, Abonyi and Szeifert applied supervised fuzzy clustering (SFC) technique and obtained 95.57% accuracy [15].

3. Least square support vector machine (LSSVM)

In this section we firstly mention about SVM classifier after that LSSVM related to SVM.

3.1. Support vector machines (SVMs)

SVM is a reliable classification technique, which is based on the statistical learning theory. This technique was firstly proposed for classification and regression tasks by [16].

As shown in Fig. 2, a linear SVM was developed to classify the data set which contains two separable classes such as $\{1, -1\}$. Let the training data consist of n datum $(x_1, y_1), \ldots, (x_n, y_n)$, $x \in R^n$ and $y \in \{1, -1\}$. To separate these classes, SVMs have to find the optimal (with maximum margin) separating hyperplane so that SVM has good generalization ability. All of the separating hyperplanes are formed with

$$D(x) = (w * x) + w_0 \tag{1}$$

and provide the following inequality for both $y = 1$ and -1.

$$y_i\big[(w * x_i) + w_0\big] \geqslant 1, \quad i = 1, \ldots, n. \tag{2}$$

The data points which provide above formula in case of equality are called the support vectors. The classification task in SVMs is implemented by using these support vectors.

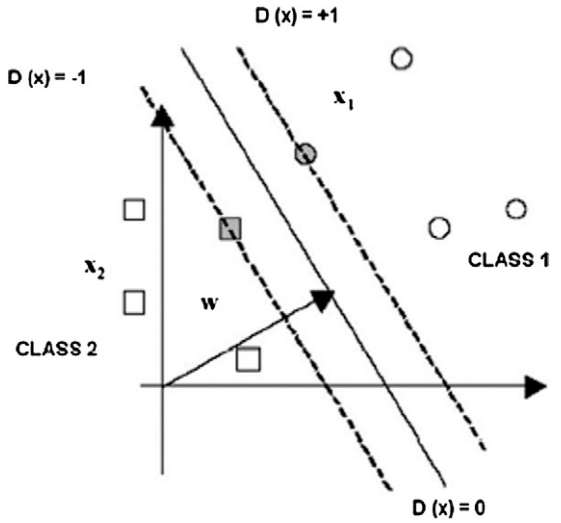

Fig. 2. The structure of a simple SVM.

Margins of hyperplanes obey the following inequality.

$$\frac{yk \times D(xk)}{\|w\|} \geqslant \Gamma, \quad k = 1, \ldots, n. \tag{3}$$

To maximize this margin (Γ), the norm of w is minimized. To reduce the number of solutions for norm of w, the following equation is determined.

$$\Gamma \times \|w\| = 1. \tag{4}$$

Then formula (5) is minimized subject to constraint (2).

$$1/2\|w\|^2. \tag{5}$$

When we study the nonseparable data, slack variables ξ_i are added into formulas (2) and (5). Instead of formulas (2) and (5), new formulas (6) and (7) are used.

$$y_i\big[(wx_i) + w_0\big] \geqslant 1 - \xi_i, \tag{6}$$

$$C\sum_{i=1}^{n}\xi_i + 1/2\|w\|^2. \tag{7}$$

Since originally SVMs classify the data in linear case, in the nonlinear case SVMs do not achieve the classification tasks. To overcome this limitation on SVMs, kernel approaches are developed. Nonlinear input data set is converted into high dimensional linear feature space via kernels. The SVM is used the RBF kernels. RBF kernels are as follows:

$$\text{RBF kernels:} \quad K(x, x') = \exp\big(-\|x - x'\|^2/\sigma^2\big),$$

where σ is a positive real number.

In our experiments σ is selected to be 1000.

3.2. Least squares support vector machines

LSSVMs are proposed by [17]. The most important difference between SVMs and LSSVMs is that LSSVMs use a set of linear equations for training while SVMs use a quadratic optimization problem [4]. While formula (7) is minimized subject to formula (6) in Vapnik's standard SVMs, in LSSVMs formula (9) is minimized subject to formula (8).

$$y_i\big[(wx_i) + w_0\big] = 1 - \xi_i, \quad i = 1, \ldots, n, \tag{8}$$

$$1/2\|w\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2. \tag{9}$$

According to these formulas, their dual problems are built as follows.

$$(w, b, \alpha, \xi) = 1/2\|w\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\big\{y_i\big[(wx_i) + w_0\big] - 1 + \xi_i\big\}. \tag{10}$$

Another difference between SVMs and LSSVMs is that α_i (Lagrange multipliers) are positive or negative in LSSVMs but they must be positive in SVMs. Information in detail is found in [17,18].

4. Performance evaluation

In this section, we present the performance evaluation methods used to evaluate the proposed method. Finally, we give the experimental results and discuss our observations from the obtained results.

4.1. Performance evaluation methods

We have used four methods for performance evaluation of breast cancer diagnosis. These methods are classification accuracy, analysis of sensitivity and specificity, confusion matrix, and k-fold cross validation. We explain these methods in the following sections.

4.1.1. Classification accuracy

In this study, the classification accuracies for the datasets are measured using the equation:

$$\text{accuracy}(T) = \frac{\sum_{i=1}^{|T} \text{assess}(t_i)}{|T|}, \quad t_i \in T, \tag{11}$$

$$\text{assess}(t) = \begin{cases} 1, & \text{if classify}(t) \equiv \text{correctclassification}, \\ 0, & \text{otherwise}, \end{cases}$$

where T is the set of data items to be classified (the test set), $t \in T$, $t.c$ is the class of item t, and classify(t) returns the classification of t by LSSVM classifier.

4.1.2. Sensitivity and specificity

For sensitivity and specificity analysis, we use the following expressions.

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}\ (\%), \tag{12}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}\ (\%), \tag{13}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

True positive (TP): An input is detected as a patient with atherosclerosis diagnosed by the expert clinicians.

True negative (TN): An input is detected as normal that is labeled as a healthy person by the expert clinicians.

False positive (FP): An input is detected as a patient that is labeled as a healthy by the expert clinicians.

False negative (FN): An input is detected as normal with atherosclerosis diagnosed by the expert clinicians.

4.1.3. Confusion matrix

A confusion matrix [19] contains information about actual and predicted classifications done by a classification system. Performance of such a system is commonly evaluated using the data in the matrix. Table 1 shows the confusion matrix for a two class classifier.

We can explain the entries of our confusion matrix:

Table 1
Representation of confusion matrix

Actual	Predicted	
	Negative	Positive
Negative	*a*	*b*
Positive	*c*	*d*

Table 2
Classification accuracies obtained with our proposed system and other classifiers from literature

Author (year)	Method	Classification accuracy (%)
Quinlan (1996)	C4.5 (10 × CV)	94.74
Hamilton et al. (1996)	RIAC (10 × CV)	94.99
Ster and Dobnikar (1996)	LDA (10 × CV)	96.80
Bennett and Blue (1997)	SVM (5 × CV)	97.20
Nauck and Kruse (1999)	NEFCLASS (10 × CV)	95.06
Pena-Reyes and Sipper (1999)	Fuzzy-GA1 (train: 75%–test: 25%)	97.36
Setiono (2000)	Neuro-rule 2a (train: 50%–test: 50%)	98.10
Goodman et al. (2002)	Optimized-LVQ (10 × CV)	96.70
Goodman et al. (2002)	Big-LVQ (10 × CV)	96.80
Goodman et al. (2002)	AIRS (10 × CV)	97.20
Abonyi and Szeifert (2003)	Supervised fuzzy clustering (10 × CV)	95.57
Our study (2006)	LS-SVM (10 × CV)	98.53

- a is the number of *correct* predictions that an instance is *negative*,
- b is the number of *incorrect* predictions that an instance is *positive*,
- c is the number of *incorrect* of predictions that an instance is *negative*,
- d is the number of *correct* predictions that an instance is *positive*.

4.1.4. k-Fold cross validation

k-Fold cross validation is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other $k - 1$ subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set $k - 1$ times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over [20].

4.2. Results and discussion

To evaluate the effectiveness of LS-SVM, we made experiments on the WBCD database mentioned above. We compare our results with the previous results reported by earlier methods. Table 2 gives the classification accuracies of our method and previous methods. As we can see from these results, our method using 10-fold cross validation obtains the highest classification accuracy, 98.53%, reported so far.

The LSSVM classification of breast cancer was classified as 10-fold cross validation, 50–50%, 70–30%, and 80–20%, respectively, due to training and test of all the WBCD dataset. The obtained test classification accuracies were 98.53, 95.89, 96.59, and 97.08%, respectively.

The obtained classification accuracy and values of sensitivity and specificity by LSSVM classifier for diagnosis of breast cancer with 50–50% training-test partition, 70–30% training-test partition, and 80–20% training-test partition were shown in Table 3.

Table 3
The obtained classification accuracy and values of sensitivity and specificity by LSSVM classifier for diagnosis of breast cancer with 50–50% training-test partition, 70–30% training-test partition, and 80–20% training-test partition

Measures	50–50% training-test partition	70–30% training-test partition	80–20% training-test partition
Sensitivity (%)	94.87	94.52	97.87
Specificity (%)	96.42	97.72	97.77
Accuracy (%)	95.89	96.59	97.08

Table 4
Confusion matrixes using LSSVM on the with 50–50% training-test partition, 70–30% training-test partition, and 80–20% training-test partition for detection of breast cancer

Output/desired	Result (benign)	Result (malignant)	Partitions
Result (benign)	216	6	50–50% training-test partition
Result (malignant)	8	111	
Result (benign)	129	4	70–30% training-test partition
Result (malignant)	3	69	
Result (benign)	88	1	80–20% training-test partition
Result (malignant)	2	46	

In this study, there were two classes as benign and malignant. Classification results of the network were displayed by using a confusion matrix. In a confusion matrix, each cell contains the raw number of exemplars classified for the corresponding combination of desired and actual network outputs. The confusion matrix showing the classification results of this network is given in Table 4.

From the above results, we conclude that the LS-SVM obtains very promising results in classifying the possible breast cancer patients. We believe that the proposed system can be very helpful to the physicians for their final decision on their patients. By using such an efficient tool, they can make very accurate decisions.

5. Conclusions

With the improvements in expert systems and machine learning tools, the effects of these innovations are entering to more application domains day-by-day and medical field is one of them. Decision making in medical field can be a trouble sometimes. Classification systems that are used in medical decision making provide medical data to be examined in shorter time and more detailed.

According to the statistical data for breast cancer in the world, this disease is among the most prevalent cancer types. In the same time, this cancer type is also among the most curable ones if it can be diagnosed early.

The LSSVM structure that we have built had given very promising results in classifying the breast cancer. Classification systems that are used in medical decision making provide medical data to be examined in shorter time and more detailed. In this study, for the diagnosis of breast cancer, a medical decision making system based on LSSVM is proposed.

In the research reported in this paper, a medical decision making system based on LSSVM was applied on the task of diagnosing breast cancer and the most accurate learning methods was evaluated. Experiments were conducted on the WBCD dataset to diagnose breast cancer in a fully automatic manner using LSSVM. The results strongly suggest that LSSVM can aid in the diagnosis of breast cancer. It is hoped that more interesting results will follow on further exploration of data. Although developed method is built as an offline diagnosing system, it can be rebuilt as an online diagnosing system in the future.

Acknowledgment

This study is supported by the Scientific Research Projects of our University (project no. 05401069).

References

[1] http://caonline.amcancersoc.org/cgi/content/full/55/2/74, last accessed August 2006.
[2] D. West, P. Mangiameli, R. Rampal, V. West, Ensemble strategies for a medical diagnosis decision support system: A breast cancer diagnosis application, Eur. J. Oper. Res. 162 (2005) 532–551.
[3] ftp://ftp.ics.uci.edu/pub/machine-learning-databases, last accessed August 2006.
[4] D.M. Parkin, F. Bray, J. Ferlay, P. Pisani, Global cancer statistics, Can. J. Clin. 55 (2005) 74–108.
[5] T. Kıyan, T. Yıldırım, Breast cancer diagnosis using statistical neural Networks, in: XII TAINN Symposium Proceedings, vol. E8, Çanakkale, Turkey, 2003, p. 754.
[6] R. Setiono, Generating concise and accurate classification rules for breast cancer diagnosis, Artific. Intell. Med. 18 (2000) 205–219.
[7] J.R. Quinlan, Improved use of continuous attributes in C4.5, J. Artific. Intell. Res. 4 (1996) 77–90.
[8] H.J. Hamilton, N. Shan, N. Cercone, RIAC: A rule induction algorithm based on approximate classification, Technical Report CS 96-06, University of Regina, 1996.
[9] B. Ster, A. Dobnikar, Neural networks in medical diagnosis: Comparison with other methods, in: Proceedings of the International Conference on Engineering Applications of Neural Networks (EANN '96), 1996, pp. 427–430.
[10] K.P. Bennet, J.A. Blue, A support vector machine approach to decision trees, Math. Report, vols. 97–100, Rensselaer Polytechnic Institute, 1997.
[11] D. Nauck, R. Kruse, Obtaining interpretable fuzzy classification rules from medical data, Artific. Intell. Med. 16 (1999) 149–169.
[12] C.A. Pena-Reyes, M. Sipper, A fuzzy-genetic approach to breast cancer diagnosis, Artific. Intell. Med. 17 (1999) 131–155.
[13] R. Setiono, Generating concise and accurate classification rules for breast cancer diagnosis, Artific. Intell. Med. 18 (2000) 205–219.
[14] D.E. Goodman, L. Boggess, A. Watkins, Artificial immune system classification of multiple-class problems, in: Proceedings of the Artificial Neural Networks in Engineering ANNIE 02, 2002, pp. 179–183.
[15] J. Abonyi, F. Szeifert, Supervised fuzzy clustering for the identification of fuzzy classifiers, Pattern Recogn. Lett. 24 (2003) 2195–2207.
[16] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[17] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (3) (1999) 293–300.
[18] D. Tsujinishi, S. Abe, Fuzzy least squares support vector machines for multi-class problems, Neural Netw. Field 16 (2003) 785–792.
[19] R. Kohavi, F. Provost, Glossary of terms. Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, vol. 30, nos. 2–3, 1998.
[20] Jeff Schneider's home page, http://www.cs.cmu.edu/~schneide/tut5/node42.html, last accessed August 2006.

Kemal Polat graduated from Electrical–Electronics Engineering Department of Selcuk University with B.Sc. degree in 2000 and from Electrical–Electronics Engineering Department of Selcuk University with M.Sc. degree in 2004. Subsequently, he is pursuing his Ph.D. degree at the Electrical–Electronics Engineering Department of Selcuk University. His current research interests are artificial immune systems, biomedical signals and digital signal processing, pattern recognition and classification.

Salih Günes graduated from Erciyes University in 1989. He took his M.S. degree in 1993 all in electrical and electronic engineering at Erciyes University. He took her Ph.D. degree in Electrical and Electronic Engineering at Selcuk University in 2000. He is an Assistant Professor at the Department of Electrical and Electronics Engineering of Selcuk University. His interest areas are biomedical signal processing, artificial immune system, logic circuits, and artificial intelligence.