ELSEVIER

Computers in Biology and Medicine 37 (2007) 415–423

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

# A new hybrid method based on fuzzy-artificial immune system and $k$-nn algorithm for breast cancer diagnosis

Seral Şahan[a,*], Kemal Polat[a], Halife Kodaz[b], Salih Güneş[a]

[a]Department of Electrical and Electronics Engineering, Engineering Faculty, Selcuk University, 42075 Konya, Turkey
[b]Computer Engineering, Selcuk University, Konya, Turkey

Received 14 July 2005; accepted 8 May 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. As the incidence of this disease has increased significantly in the recent years, machine learning applications to this problem have also took a great attention as well as medical consideration. This study aims at diagnosing breast cancer with a new hybrid machine learning method. By hybridizing a fuzzy-artificial immune system with $k$-nearest neighbour algorithm, a method was obtained to solve this diagnosis problem via classifying Wisconsin Breast Cancer Dataset (WBCD). This data set is a very commonly used data set in the literature relating the use of classification systems for breast cancer diagnosis and it was used in this study to compare the classification performance of our proposed method with regard to other studies. We obtained a classification accuracy of 99.14%, which is the highest one reached so far. The classification accuracy was obtained via 10-fold cross validation. This result is for WBCD but it states that this method can be used confidently for other breast cancer diagnosis problems, too.

*Keywords:* $k$-Nearest neighbour classification system; Artificial immune systems; Fuzzy weighting; Breast cancer diagnosis; Wisconsin Breast Cancer Diagnosis Data; $k$-Fold cross validation

## 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 in more detail.

In this study a new hybrid method was proposed to be used in breast cancer diagnosis problem as a classifier. This method involves a two-stage system in which a fuzzy-artificial immune system and $k$-nearest neighbour ($k$-nn) classification system are hybridized. The first stage of the whole system conducts a data reduction process for $k$-nn algorithm of the second stage. This

[*] Corresponding author. Tel.: +90 332 2232113; fax: +90 332 2410635.
*E-mail addresses:* ssahan09@yahoo.com (S. Şahan), kpolat@selcuk.edu.tr (K. Polat), hkodaz@selcuk.edu.tr (H. Kodaz), sgunes@selcuk.edu.tr (S. Güneş).

0010-4825/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2006.05.003

provides less training data for *k*-nn and so classification time of the algorithm can be reduced in a great deal.

The used data source is Wisconsin Breast Cancer Dataset (WBCD) taken from the University of California at Irvine (UCI) Machine Learning Repository [3]. This data set 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 performance of the system was analysed with regard to the classification accuracy, sensitivity and specificity. The values of these performance criterions were obtained via 10-fold cross validation, which is a very common performance evaluation method in ML literature. Our proposed system reached 99.14% classification accuracy in test phase and this result is the highest one among the studies applied for WBCD classification problem so far. It also indicates that the proposed method can be applied confidently to other breast cancer problems with different data sets especially with ones that have higher number of training data.

The rest of the paper is organized as follows. Section 2 gives the background information including breast cancer classification problem, previous research in corresponding area and brief introduction to natural and artificial immune systems. We explained the method in Section 3 with subtitles of proposed hybrid system and measures for performance evaluation. In each subsection of that section, the detailed information is given. The results obtained in applications are given in Section 4. This section also includes the discussion of these results in specific and general manner. Consequently in Section 5, we conclude the paper with summarization of results by emphasizing the importance of this study and mentioning about some future work.

## 2. Background

### 2.1. Breast cancer classification problem

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, completed 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 the 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 non-molecular 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 ML repository [3].

The name of the data set for breast cancer problem is WBCD. The data set consist of 683 samples that were collected by Dr. W.H. Wolberg at the University of Wisconsin-Madison

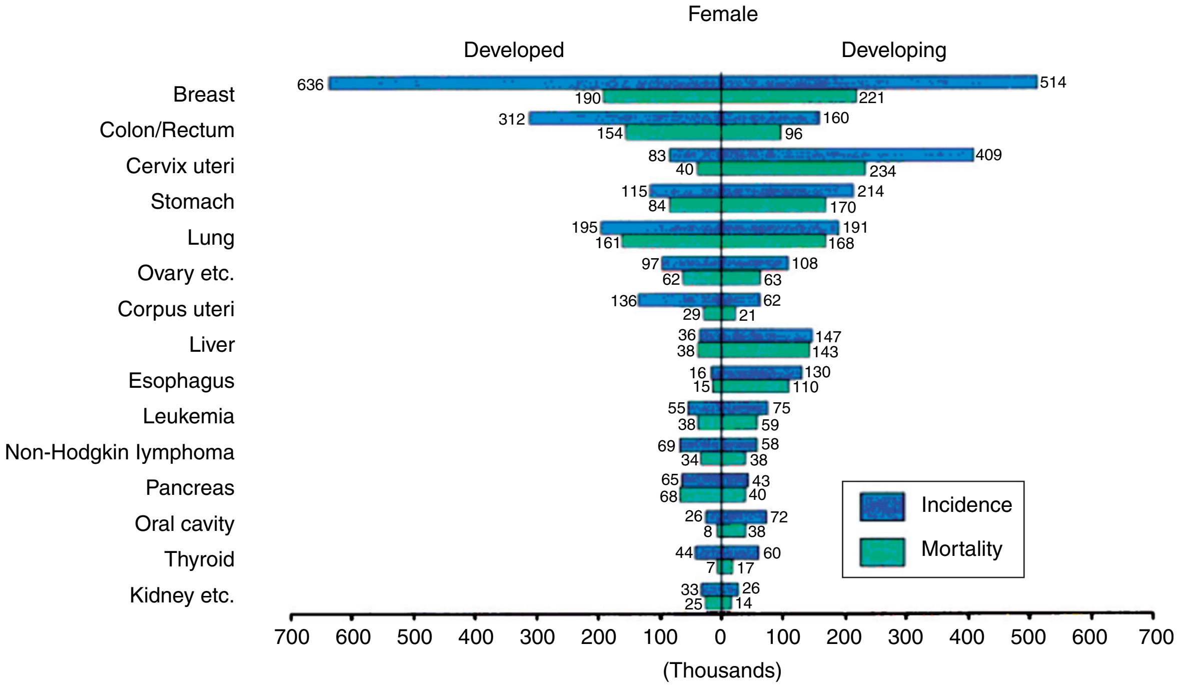

Fig. 1. Estimated numbers of new cancer cases (incidence) and deaths (mortality) in 2002 [1].

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 nucleoli ($x_6$);
(7) bland chromatin ($x_7$);
(8) normal nucleoli ($x_8$); and
(9) mitoses ($x_9$).

A total of 444 samples of the data set belong to benign, and remaining 239 data are malignant.

### 2.2. Previous research

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 data set 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 × 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 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].

### 2.3. Natural and artificial immune systems

The natural immune system is a distributed novel-pattern detection system with several functional components positioned in strategic locations throughout the body. Immune system regulates defence mechanism of the body by means of innate and adaptive immune responses. Between these, adaptive immune response is more important for us because it contains metaphors like recognition, memory acquisition, diversity, self-regulation, etc. The main architects of adaptive immune response are lymphocytes, which divide into two classes as T and B lymphocytes (cells), each having its own function. Especially, B cells have a great importance because of their secreted antibodies (*Abs*) that take very critical roles in adaptive immune response.

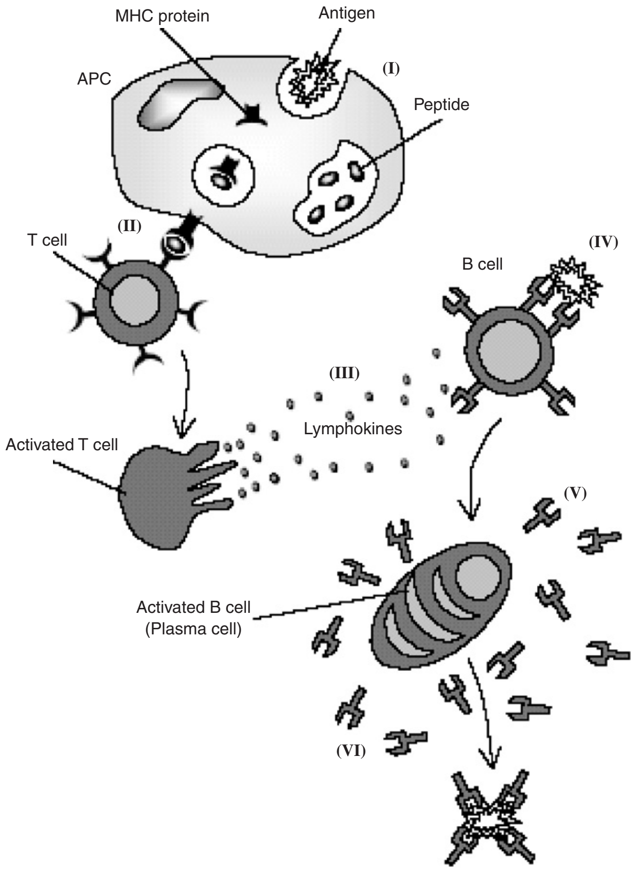

Fig. 2. Immune response [16].

The simplified working procedure of our immune system is illustrated in Fig. 2. Specialized antigen presenting cells (APCs) called macrophages circulates through the body and if they encounter an antigen, they ingest and fragment them into antigenic peptides (I). The pieces of these peptides are displayed on the cell surface by major histocompatibility complex (MHC) molecules existing in the digesting APC. The presented MHC–peptide combination on the cell surface is recognized by the T-cells causing them to be activated (II). Activated T cells secrete some chemicals as alert signals to other units in response to this recognition. B cells, one of the units that take these signals from the T cells, become activated with the recognition of antigen by their antibodies occurring at the same time (IV). When activated, B cells turn into plasma cells that secrete bound antibodies on their surfaces (V). Secreted *Abs* bind the existing antigens and neutralize them signalling other components of immune system to destruct the antigen–antibody complex (VI) [16]. For detailed information about immune system refer to [17].

Artificial immune systems (AISs) emerged in the 1990s as a new computational research area. AISs link several emerging computational fields inspired by biological behaviour such as artificial neural networks and artificial life.

In the studies conducted in the field of AIS, B cell modelling is the most encountered representation type. Different representation methods have been proposed in that modelling. Among these, shape-space representation is the most commonly used one [18].

The shape-space model ($S$) aims at quantitatively describing the interactions among antigens (*Ags*), the foreign elements that enter the body like microbe, etc., and antibodies (*Ag–Ab*). The set of features that characterize a molecule is called its *generalized shape*. The *Ag–Ab* representation (binary or real-valued) determines a distance measure to be used to calculate the degree of interaction between these molecules. Mathematically, the generalized shape of a molecule ($m$), either an antibody or an antigen, can be represented by a set of coordinates $m = \langle m_1, m_2, \ldots, m_L \rangle$, which can be regarded as a point in an $L$-dimensional real-valued shape-space ($m \in S^L$). In this work, we used real strings to represent the molecules. *Ags* and *Abs* were considered of same length $L$. The length and cell representation depend upon the problem [16].

## 3. Method

### 3.1. Proposed hybrid system

$K$-nn algorithms are known especially with their simplicity in machine learning literature. They are also advantageous in that the information in training data is never lost. But, there are some problems with them. First of all, for large data sets, these algorithms are very time consuming because each sample in training set is processed while classifying a new data and this requires longer classification times. This cannot be problem for some application areas but when it comes to a field like medical diagnosis, time is very important as well as classification accuracy. So, an attempt has been made in this study to reduce the size of training data. This data reducing stage was realized by using an artificial intelligence algorithm, an AIS algorithm. Immune cells have the capability of grouping foreign invaders with regard to their origin and they produce immune responses according to this grouping. This clustering capability of immune system has been used as inspiration source to design the data reduction algorithm in our purpose. One more stage has been added to our system as a weighting stage. This was done to remove the negative effects of using a distance-based algorithm as will be explained in the following sections. The block diagram of the whole classification system can be seen in Fig. 3.

#### 3.1.1. K-nn algorithm

$K$-nn algorithm is among the instance-based classifiers. In instance-based methods, system parameters or classifying system units simply consist of the samples that are presented to the system. This algorithm assumes that all instances correspond to points in the $n$-dimensional space $R^N$ [19]. Nearest neighbours of a sample in this space are determined by standard Euclidean distance.

Let $x$ be a sample and it is defined by a feature vector of:

$$\langle a_1(x), a_2(x), \ldots, a_n(x) \rangle$$

here $a_r(x)$ is the $r$th feature of $x$ sample. The $d(x_i, x_j)$ Euclidean distance between $x_i$ and $x_j$ samples is defined by

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}. \tag{1}$$

$k$-nn algorithm uses an $f(\cdot)$ function. For classification applications, this function is the class of presented sample. If we denote this by $f(x_i)$, the procedure of $k$-nn algorithm can be summarized as follows: $k$-nn algorithm stores all training data and corresponding classes of this training data as system units. Let $\langle x_i, f(x_i) \rangle$ be a vector indicating individual training sample and the corresponding class of this sample. During the classification in the system, $k$-nearest system units to presented $x_i$ sample are determined via Eq. (1). The class of presented sample is approximated according to the number of these $k$-nearest units. The class of nearest samples that have the highest percentage in $k$-nearest units be this class estimation; $\hat{f}(x_i)$. If we state this procedure in terms of algorithmic base: *Training phase*:

- For each training example $\langle x, f(x) \rangle$, add the example to list *training_examples*.

*Classification phase*:

- Given a query instance $x_q$ to be classified,

  *Let $x_1, x_2, \ldots, x_k$ denote $k$ instances from *training_samples* that are nearest to $x_q$.

  *return

$$\hat{f}(x_q) \leftarrow \underset{v \in V}{\text{arg max}} \sum_{i=1}^{k} \delta(v, f(x_i)), \tag{2}$$

where $\delta(a, b) = 1$ if $a = b$ and where $\delta(a, b) = 0$ otherwise.

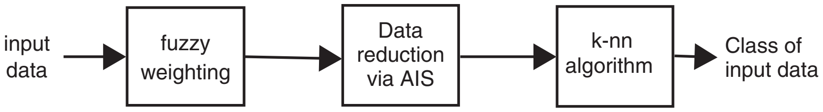

Fig. 3. Block diagram of the proposed system.

As stated previously, this method is quite simple but suffers from the disadvantages caused by using presented training samples simply as classification units. Our proposed method was adapted to the training phase of this system. It will be no longer valid for training samples to be used as system units, instead with a training procedure in the data reduction phase memory units will be formed and these memory units will be used as system units in the classification phase of the algorithm. The formation of memory units by using an AIS algorithm in the data reduction stage is explained in the following paragraphs.

### 3.1.2. *Used AIS data reduction algorithm*

Immune system cells have large clustering capability and this property has been modelled in AIS literature in a great deal. The proposed algorithm, inspired from this property, used the learning strategy in immune system as the following way. The presented training samples are named as *Ags* while formed memory units which will be then used as classifying units for $k$-nn algorithm are called as *Abs* in the algorithm.

The training procedure of the algorithm conducts the following steps:

(1) For each $Ag_i$ do: $(i : 1, \ldots, N)$

  (1.1) Determine the class of $Ag_i$. Call memory *Abs* of that class and calculate the distances between $Ag_i$ and these memory *Abs* with Eq. (3):

$$D = \sqrt{\sum_{k=1}^{L} w_{j,k}(Ab_{j,k} - Ag_{i,k})^2}. \quad (3)$$

  Here $Ab_{j,k}$ and $Ag_{i,k}$ are the $k$th attribute of $Ab_j$ and $Ag_i$, respectively, $j = 1, \ldots, Mc$, where $Mc$ is the number of memory *Abs* for related class. $w_{j,k}$ is the weight of $k$th attribute that belongs to the class of $Ab_j$.

  (1.2) If the minimum distance among the calculated distances above is less than a threshold value named as suppression value (*supp*) then return to step 1.

  (1.3) Form a memory *Ab* for $Ag_i$:

  At each iteration do:

    (1.3.1) Make a random *Ab* population with *Ab* = [*Ab_mem*; *Ab_rand*] and calculate the distances of these *Abs* to $Ag_i$.

    (1.3.2) Select $m$ nearest *Abs* to $Ag_i$; clon and mutate these *Abs* (*Ab_mutate*).

    (1.3.3) Keep the $m$ nearest *Abs* in the *Ab_mutate* population to $Ag_i$ as *Ab_mem* temporary memory population.

    (1.3.4) Define the nearest *Ab* to $Ag_i$ as *Ab_cand*, candidate memory *Ab* for $Ag_i$ and stop iterative process if the distance of *Ab_cand* to $Ag_i$ is less than a threshold value named as stopping criterion (*sc*).

    (1.3.5) Concatenate *Ab_cand* as a new memory *Ab* to memory matrix of the class of $Ag_i$.

  (1.4) Stop training.

The mutation mechanism in the algorithm which is used in many AIS algorithms and named as *hypermutation* is performed proportional to distance between two cells:

$$Ab'_{j,k} = Ab_{j.k} \pm D^*_{j,I}(Ab_{j.k}). \quad (4)$$

Here, $Ab'_{j.k}$ is the new value and $Ab_{j.k}$ *is the old value of* $k$th attribute of $j$th *Ab*. $D_{j,i}$ stands for the distance between $Ag_i$ and $Ab_j$.

The resulted memory *Abs* form *training_samples* in the algorithm of $k$-nn. These memory *Abs* carry the same class information in training samples if their number and location in sample space is well-adjusted in the training algorithm above. This adjustment is realized through supp parameter and mutation mechanism, respectively. Especially, the number of memory *Abs* affects the classification performance in a great deal since with a few number of memory units it is hard to represent the class information hidden in training data while a high number of these memory units prevent us to reduce the data.

With or without a data-reduction stage, this classification system uses a distance criteria that takes into account all of the features in same degree. However, we know that such distance-based approaches have a disadvantage of this property since distance can be dominated by irrelevant features. Sometimes this bottleneck is referred as ‘curse of dimensionality’ and it arises in situations when one attribute value in shape space can cause two data in the same class to be distant from each other. In such situations, the presented samples can be recognized and classified by different system units by distance-based approaches. To prevent this possible situation from causing any classification error, the following feature weighting procedure was conducted prior to the data reduction stage.

### 3.1.3. *Fuzzy weighting procedure*

In the fuzzy weighting procedure, each feature takes new feature value according to its old value. Two types of membership functions are defined in this procedure known as input and output membership functions. These are selected as triangular membership functions as shown in Figs. 4 and 5, respectively.

The formation of these membership functions is realized as follows: as a first step, the mean values of each feature are calculated through using all of the samples’ corresponding feature values in

$$m_i = \frac{1}{N}\sum_{k=1}^{N} x_{k,i}, \quad (5)$$

here $x_{k,i}$ represents the $i$th feature value of sample $x_k$, $k = 1, 2, \ldots, N$. After calculation of these sample means for each feature, the input membership function is formed by triangles as in Fig. 4. The supports of these triangles are determined by $m_i/8$, $m_i/4$, $m_i/2$, $m_i$, $2 \times m_i$, $4 \times m_i$, $8 \times m_i$ as shown in the figure. The lines of input membership functions are named as $mf1, mf2, \ldots, mf8$. For the output membership function formation, again eight parts formed membership functions (Fig. 5). The interval [0,1] is divided into eight equal part and the corresponding lines are again named but in this case

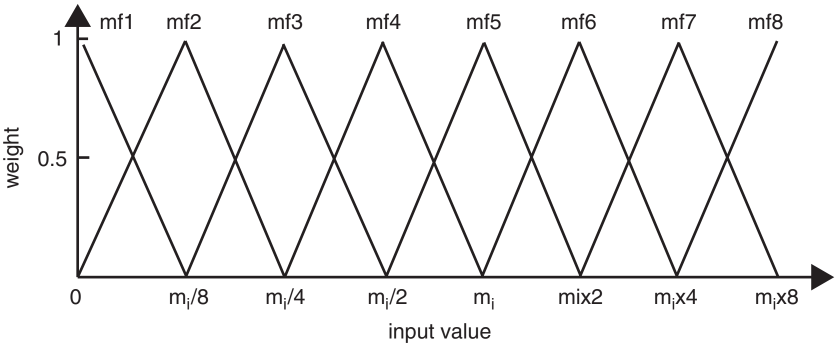

Fig. 4. Input membership functions.

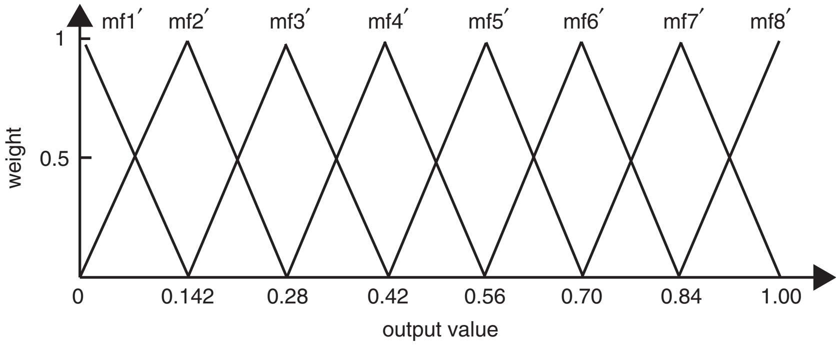

Fig. 5. Output membership functions.

Table 1
Fuzzy rule base for our system

1. if Input_value cuts $mf1$ and $mf2$ then Output_value $= (mf1'(y) + mf2'(y))/2$
2. if Input_value cuts $mf2$ and $mf3$ then Output_value $= (mf2'(y) + mf3'(y))/2$
3. if Input_value cuts $mf3$ and $mf4$ then Output_value $= (mf3'(y) + mf4'(y))/2$
4. if Input_value cuts $mf4$ and $mf5$ then Output_value $= (mf4'(y) + mf5'(y))/2$
5. if Input_value cuts $mf5$ and $mf6$ then Output_value $= (mf5'(y) + mf6'(y))/2$
6. if Input_value cuts $mf6$ and $mf7$ then Output_value $= (mf6'(y) + mf7'(y))/2$
7. if Input_value cuts $mf7$ and $mf8$ then Output_value $= (mf7'(y) + mf8'(y))/2$

as $mf1', mf2', \ldots, mf8'$. Before continuing it is worth mentioning that these input and output membership functions are formed for each feature so there will exist different input–output membership function configuration for each feature since the sample means of each feature differ.

After determination of input and output membership functions for each feature, the weighting procedure comes into scene. For a feature value, say $x_{k,i}$, that is for $i$th feature value of $x_k$ sample, this value is taken as in the $x$-axis of input membership function and $y$ values of the points at which $x_{k,i}$ cuts the input membership functions are determined. For example if this feature value is between 0 and $m_i/8$, then this point will cut both line $mf1$ and $mf2$. The $y$ values at these intersection points, say $y_1$ and $y_2$, are known as membership values ($\mu$) and they will then be used in a fuzzy rule base in the following manner: firstly, the input membership value, $\mu(i)$, is determined by using the above intersection points:

$$\mu(i) = \mu_{A\cap B}(x_{k,i}) = \min(\mu_A(x_{k,i}), \mu_B(x_{k,i})), \quad x \in X. \tag{6}$$

Here, $\mu_A(x_{k,i})$ and $\mu_B(x_{k,i})$ membership values correspond to the intersection points as mentioned above. The rule base for our system is used as presented in Table 1. After this $\mu(i)$ value is determined through using Eq. (6) for our $x_{k,i}$ feature value, the output weight value is then determined by using output membership functions and the rules in Table 1. Here in determining weight as a last step, firstly the input membership value, $\mu(i)$, is presented to output membership function to determine the corresponding weighted value of our original feature value. This membership value is now taken as a point in $y$-axis of the output membership functions and again as for the case in input membership functions, the intersection points are determined which are cut by this membership value. It is apparent from output membership functions that there will be more than one intersection points. That which of them will be used is decided through the rules in Table 1. For example, if input feature value cuts $mf1$ and $mf2$ lines in input membership functions then the output value for this feature will be the mean of two points that $\mu(i)$ cuts $mf1'$ and $mf2'$ at the output membership functions.

### 3.2. *Measures for performance evaluation*

#### 3.2.1. *Classification accuracy, specificity and sensitivity*

In this study, the classification accuracy for the data set was measured according to [20]:

$$\text{accuracy}(T) = \frac{\sum_{i=1}^{|T} \text{assess}(t_i)}{|T|}, \quad t_i \in T, \tag{7}$$

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

where $T$ is the set of data items to be classified (the test set), $t_0T$, $t.c$ is the class of the item $t$, and classify$(t)$ returns the classification of $t$ by AIRS.

Besides classification accuracy, sensitivity and specificity measures are also given in two-class problems as

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \tag{9}$$

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

#### 3.2.2. *k-Fold cross validation*

For test results to be more valuable, $k$-fold cross validation is used among the researchers. It minimizes the bias associated with the random sampling of the training [21]. In this method, whole data are randomly divided to $k$ mutually exclusive and approximately equal size subsets. The classification algorithm is trained and tested $k$ times. In each case, one of the folds is taken as test data and the remaining folds are added to form training data. Thus $k$ different test results exist for each training-test configuration. The average of these results gives the test accuracy of the algorithm [21]. We used this method as 10-fold cross validation in our applications.

## 4. Results and discussion

As stated in Section 3.1.2, the key parameter to determine in proposed system is supp parameter since it determines the number of memory *Abs* so the classification accuracy. The value of this parameter is selected in the [0,1] range. If this value is selected too high, the number of *Abs* will be too low and in contrary if this value is too low, there will be more memory *Abs*. The number of memory *Abs* highly affects the classification performance. In our system this value was selected in the range of 0–1 and the supp value that gave the highest performance was tried to be found. The variation of classification accuracy with regard to the supp parameter is shown in Fig. 6.

As can be seen from the figure, the appropriate supp value to obtain highest classification accuracy lies in the range 0.10–0.18. In this range the number of memory *Abs* varies between 250 and 20. The maximum classification accuracy was reached for the 0.15 value of supp parameter which was obtained as 99.14%. $k$ value was taken as 5 for this accuracy and the number of memory *Abs* was recorded as approximately 100 at this point. The performance parameters for this value are shown in Table 2.

The classification accuracy obtained by Fuzzy-AIS-kNN for WBCD is the highest one among the classifiers reported in the literature. The comparison of our method with these classifiers with respect to the classification accuracy is shown in Table 3

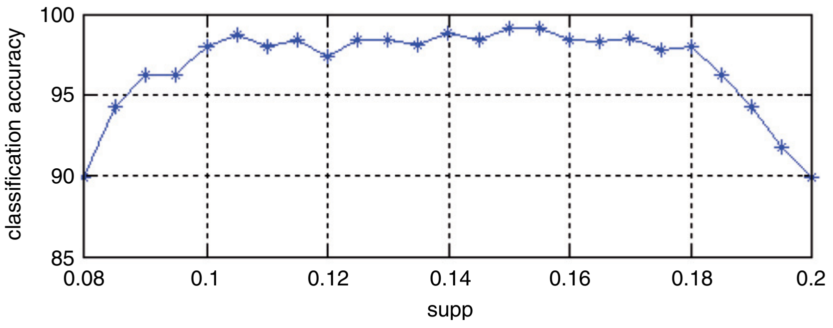

Fig. 6. Classification accuracy of the proposed system versus supp parameter.

Table 2
Obtained performance parameters for highest classification accuracy

| Supp | $k$ | Number of memory antibodies | Accuracy (%) | Specificity (%) | Sensitivity (%) |
|---|---|---|---|---|---|
| 0.15 | 5 | ~ 100 | 99.14 | 100 | 99.56 |

Table 3
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 (2005) | Fuzzy-AIS-knn (10 × CV) | 99.14 |

Our proposed method has reached the highest classification accuracy among the classifiers in the table. If the methods that used cross validation are taken for comparison, because cross validation method gives more reliable results, the second highest classification accuracy is 97.20% and an increase of 1.94% has been reached by our system which is not negligible for such medical problems.

Certainly, if more data are used in training phase the result may be more promising, even it may be 100%. Then the system can be used confidently to help experts for decision making in their diagnosis problems.

## 5. Conclusion

With the improvements in expert systems and ML 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.

In this study, for the diagnosis of breast cancer, a new ML method was proposed. The method is a hybrid of $k$-nn algorithm and a data reduction stage which was developed as an AIS. Also, a fuzzy-weighting procedure was used prior to this data reduction algorithm. We used WBCD in our study because it is a commonly used data set among researchers who applied ML methods to breast cancer problem. By selecting this data set, we could compare our classification accuracy with other methods. We have reached 99.14% classification accuracy via 10-fold cross validation. This accuracy is the highest one reached so far for WBCD. It also states that our system can be used for any breast cancer diagnosis problem and it would give high classification accuracies especially for large data sets. Also, besides of breast cancer problem, other medical diagnosis applications can also be conducted by this system.

## Acknowledgement

This study has been supported by Scientific Research Project of Selcuk University (Project No: 05401069).

## References

[1] ⟨http://caonline.amcancersoc.org/cgi/content/full/55/2/74⟩ (last accessed: 25 April 2005).

[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: 7 April 2005).

[4] D.M. Parkin, F. Bray, J. Ferlay, P. Pisani, Global cancer statistics, 2002, a Cancer, J. Clinicians 55 (2005) 74–108.

[5] T. Kıryan, T. Yıldırım, Breast cancer diagnosis using statistical neural networks, XII. TAINN Symposium Proceedings, E(8): 754, Çanakkale, Turkey, 2003.

[6] R. Setiono, Generating concise and accurate classification rules for breast cancer diagnosis, Artif. Intell. Med. 18 (2000) 205–219.

[7] J.R. Quinlan, Improved use of continuous attributes in C4.5, J. Artif. 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, Artif. Intell. Med. 16 (1999) 149–169.

[12] C.A. Pena-Reyes, M. Sipper, A fuzzy-genetic approach to breast cancer diagnosis, Artif. Intell. Med. 17 (1999) 131–155.

[13] R. Setiono, Generating concise and accurate classification rules for breast cancer diagnosis, Artif. 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 (2002), 2002, pp. 179–183.

[15] J. Abonyi, F. Szeifert, Supervised fuzzy clustering for the identification of fuzzy classifiers, Pattern Recognition Lett. 24 (2003) 2195–2207.

[16] L.N. De Castro, J. Timmis, Artificial Immune Systems: A New Computational Intelligence Approach, Springer, UK, 2002.

[17] A.K. Abbas, A.H. Lichtman, Cellular and Molecular Immunology, fifth ed., W.B. Saunders Company, 2003.

[18] A.S. Perelson, G.F. Oster, Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-nonself discrimination, J. Theoret. Biol. 81 (1979) 645–670.

[19] T.M. Mitchell, Machine Learning, The McGraw-Hill Companies Press, 1997.

[20] A. Watkins, AIRS: A resource limited artificial immune classifier, Master Thesis, Mississippi State University, 2001.

[21] D. Delen , G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2) (2005) 113–127.

**Seral Şahan** graduated from Electrical-Electronics Engineering Department of Ege University with B.Sc. degree in 2002 and from Electrical-Electronics Engineering Department of Selcuk University with M.Sc. degree in 2004. Now, she is pursuing her Ph.D. degree at the same department of Selcuk University. Her research interests are artificial immune systems, classifier systems and pattern recognition.

**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.

**Halife Kodaz** graduated from Computer Engineering Department of Selcuk University with B.Sc. degree in 1999 and from Computer Engineering Department of Selcuk University with M.Sc. degree in 2002. Now, he is pursuing his Ph.D. degree at the Electrical-Electronics Engineering Department of Selcuk University. His research interests are artificial immune systems and machine learning.

**Salih Günes** graduated from the Erciyes University in 1989. He took his M.S. degree in 1983 in Electrical and Electronic Engineering at the Erciyes University. He took his Ph.D. degree in Electrical and Electronic Engineering at the Selcuk University in 2000. He is an Assistant Prof. Dr. at the Department of Electrical and Electronics Engineering at the Selcuk University. His interest areas are biomedical signal processing, artificial immune system, logic circuits and artificial intelligence.