Neural Comput & Applic (2015) 26:437–446
DOI 10.1007/s00521-014-1731-9

ORIGINAL ARTICLE

# Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis

**A. H. El-Baz**

Received: 10 June 2014 / Accepted: 14 September 2014 / Published online: 2 October 2014


**Abstract** The effectiveness of classification and recognition systems has improved in a great deal to help medical experts in diagnosing diseases. Breast cancer is becoming a leading cause of death among women in the whole world; meanwhile, it is confirmed that the early detection and accurate diagnosis of this disease can ensure a long survival of the patients. This paper presents a hybrid intelligent system for recognition of breast cancer tumors. The proposed system includes two main modules: the feature extraction module and the predictor module. In the feature extraction module, rough set theory is used to preprocess the attributes on condition that the important information is not lost, deletes redundant attributes and conflicting objects from decision table. In the predictor module, a combined classifier is proposed based on K-nearest neighbor classifier. Experiments have been conducted on a widely used Wisconsin breast cancer dataset taken from University of California Irvine. Experimental results show that the proposed hybrid system can improve the rate of correct diagnosis of cases. The proposed combined classifier with rough set-based feature selection achieves 99.41 % classification accuracy and uses only 4 features which is the best shown to date. Different performance metrics are used to show the effectiveness of the proposed hybrid system. With these results, the proposed method is very promising compared to the previously reported results and can be used confidently for other breast cancer diagnosis problems.

A. H. El-Baz (✉)
Department of Mathematics, Faculty of Science,
Damietta University, New Damietta, Egypt
e-mail: ali_elbaz@yahoo.com

**Keywords** Breast cancer diagnosis · Hybrid intelligence system · Ensemble classifier · Rough set · K-Nearest neighbor · Feature selection

## 1 Introduction

Cancer is a group of diseases that causes cells in the body to change and grow out of control. Most types of cancer cells eventually form a lump or mass called a tumor and are named after the part of the body where the tumor originates.

Breast cancer begins in the breast tissue that is made up of glands for milk production, called lobules, and the ducts that connect the lobules to the nipple. The remainder of the breast is made up of fatty, connective and lymphatic tissues [1].

Breast cancer typically is detected either during a screening examination, before symptoms have developed, or after symptoms have developed, when a woman feels a lump. Most masses seen on a mammogram and most breast lumps turn out to be benign; that is, they are not cancerous, do not grow uncontrollably or spread, and are not life-threatening. When cancer is suspected based on clinical breast exam or breast imaging, microscopic analysis of breast tissue is necessary for a definitive diagnosis and to determine the extent of spread (in situ or invasive) and characterize the pattern of the disease. The tissue for microscopic analysis can be obtained via a needle or surgical biopsy. Selection of the type of biopsy is based on individual patient clinical factors, availability of particular biopsy devices, and resources (American Cancer Society, 2014) [1].

Most breast cancers are invasive or infiltrating. These cancers have broken through the ductal or glandular walls

where they originated and grow into surrounding breast tissue. The prognosis (forecast or outcome) of invasive breast cancer is strongly influenced by the stage of the disease—that is, the extent or spread of the cancer when it is first diagnosed.

Medical diagnostic decision support systems have become an established component of medical technology. The main concept of the medical technology is an inductive engine that learns the decision characteristics of the diseases and can then be used to diagnose future patients with uncertain disease states.

In general, given a pattern recognition problem, the traditional approach is to evaluate a set of different learning algorithms against a representative validation set and select the best one. In order to achieve the best possible classification performance, we need to design many algorithms. It is now recognized that the key to recognition problems does not lie wholly in any particular solution. No single model exists for all pattern recognition problems and no single technique is applicable to all problems. Furthermore, the sets of patterns misclassified by the different algorithms would not necessarily overlap, which suggested that different algorithms potentially offered complementary information [2]. This led to the emergence of ensemble learning. Ensemble learning is a learning method where a collection of a finite number of classifiers is trained for the same classification task, and thus, it can gain better performance at the cost of computation. In recent years, ensemble learning has been employed to increase the accuracy in classification beyond the level achieved by individual classifiers [3, 4]. Typically, ensemble learning involves either statistical parametric classifiers or neural networks trained on the same data, and a method that combines their outputs into a single one. If one could pick the best classifier to use for every sample, the misclassified samples in the output would be the ones that were wrongly classified by all methods [5].

The main contribution of the paper is to build a hybrid intelligent system which combine two methodologies: rough set theory as a preprocessing step for selecting the most discriminatory features and a combined classifier using K-nearest neighbor (KNN) as base classifier so as to automatically produce a diagnostic system. We find that the proposed hybrid approach produces a system exhibiting two prime characteristics: first, it attains high classification performance which is the best shown to date; second, the resulting systems involve a few set of discriminatory features, only four features, and are therefore (human-) interpretable.

The rest of the paper is organized as follows. Section 2 gives the background information including breast cancer classification problem and previous research in corresponding area. The proposed hybrid intelligent system is explained in Sect. 3. In Sect. 4, different performance measurements are introduced which are commonly used for testing the effectiveness of automatic diagnosis system. The results obtained are given in Sect. 5. This section also includes the discussion of these results. Consequently in Sect. 6, the conclusion is given with summarization of results by emphasizing the importance of this study and suggesting some future work.

## 2 Background

### 2.1 Breast cancer dataset

Cancer begins with uncontrolled division of one cell and results in a visible mass named tumor, see Fig. 1. Tumor can be benign or malignant. Malignant tumor 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 color 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. As for the other cancer types, early diagnosis in breast cancer can be lifesaving. A more recent study gives the estimated new female breast cancer cases and deaths by age, US, 2013 as shown in Table 1.

In this study, the Wisconsin breast cancer database [6] taken from fine needle aspirates from human breast tissue was analyzed. They have been collected by Dr. William H. Wolberg at the University of Wisconsin-Madison Hospitals, USA. The data consist of 683 records of

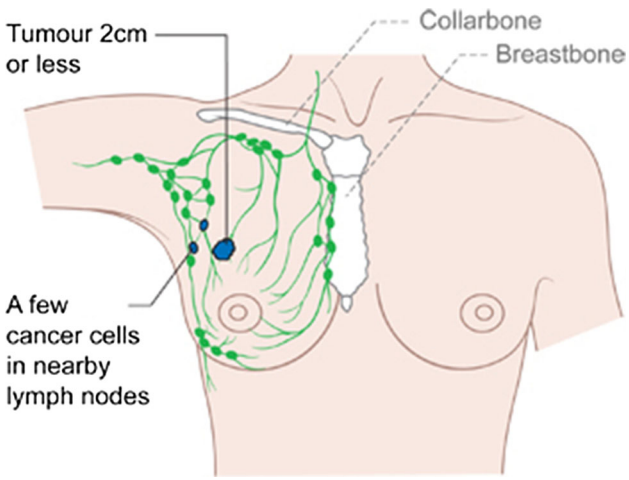


**Fig. 1** Breast cancer copyright © Cancer Research UK. http://www.cancerresearchuk.org

**Table 1** Estimated new female breast cancer cases and deaths by age, US, 2013[a]

| Age (years) | In situ cases | Invasive cases | Deaths |
|---|---|---|---|
| <40 | 1,900 | 10,980 | 1,020 |
| <50 | 15,650 | 48,910 | 4,780 |
| 50–64 | 26,770 | 84,210 | 11,970 |
| 65+ | 22,220 | 99,220 | 22,870 |
| All ages | 64,640 | 232,340 | 39,620 |

*Source*: Total estimated cases are based on 1995–2009 incidence rates from 49 states as reported by the North American Association for Central Cancer Registries. Total estimated deaths are based on data from US Mortality Data, 1995–2009, National Center for Health Statistics, Centers for Disease Control and Prevention

American Cancer Society, Surveillance and Health Services Research, 2013. http://www.cancer.org/research/cancerfactsstatistics/index

[a] Rounded to the nearest 10

**Table 2** Description of the breast cancer datasets

| Databases | No. of attributes | No. of instances | No. of malignant | No. of benign | No. of classes |
|---|---|---|---|---|---|
| Wisconsin breast cancer (WBC) | 9 | 683 | 239 | 444 | 2 |

**Table 3** Wisconsin breast cancer dataset attributes

| Attribute number | Attribute | Domain | Mean | Standard deviation |
|---|---|---|---|---|
| 1 | Clump thickness | 1–10 | 4.44 | 2.82 |
| 2 | Uniformity of cell size | 1–10 | 3.15 | 3.07 |
| 3 | Uniformity of cell shape | 1–10 | 3.22 | 2.99 |
| 4 | Marginal adhesion | 1–10 | 2.83 | 2.86 |
| 5 | Single epithelial cell size | 1–10 | 3.23 | 2.22 |
| 6 | Bare nuclei | 1–10 | 3.54 | 3.64 |
| 7 | Bland chromatin | 1–10 | 3.45 | 2.45 |
| 8 | Normal nucleoli | 1–10 | 2.87 | 3.05 |
| 9 | Mitoses | 1–10 | 1.60 | 1.73 |

virtually assessed nuclear features of fine needle aspirates taken from patients' breasts. Each record in the database has nine attributes. A brief description of this dataset is presented in Table 2. The nine attributes detailed in Table 3 are graded on an interval scale from a normal state of 1–10, with 10 being the most abnormal state. There are 239 malignant cases and 444 benign cases. A malignant label is confirmed by performing a biopsy on the breast tissue. Either a biopsy or a periodic examination is used to confirm a benign label.

In the Clump thickness, benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayer [6]. While in the Uniformity of cell size/shape, the cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not. In the case of Marginal adhesion, the normal cells tend to stick together, where cancer cells tend to lose this ability. So loss of adhesion is a sign of malignancy. In the Single epithelial cell size, the size is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell. The Bare nuclei is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors. The Bland chromatin describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells, the chromatin tends to be coarser. The Normal nucleoli are small structures seen in the nucleus. In normal cells, the nucleolus is usually very small if visible. In cancer cells, the nucleoli become more prominent, and sometimes, there are more of them. Finally, Mitoses are nuclear division plus cytokines and produce two identical daughter cells during prophase. It is the process in which the cell divides and replicates. Pathologists can determine the grade of cancer by counting the number of mitoses.

### 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 WBCD taken from UCI machine learning repository. Among these, Quinlan [7] reached 94.74 % classification accuracy using 10-fold cross-validation with C4.5 decision tree method. Hamilton et al. [8] obtained 96 % accuracy with RIAC method while Ster and Dobnikar [9] obtained 96.8 % with linear discreet analysis (LDA) method. The accuracy obtained by Bennett and Blue [10] who used support vector machine (SVM) (5× CV) method was 97.2 % while by Nauck and Kruse [11] was 95.06 % with neuro-fuzzy techniques and by Pena-Rayes and Sipper [12] was 97.36 % using fuzzy-GA method. Moreover, Setiono [13] reached 98.1 % using neuro-rule method. Goodman et al. [14] 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. Nevertheless, Abonyi and Szeifert [15] applied supervised fuzzy clustering (SFC) technique and obtained 95.57 % accuracy. In

[16], Least Square SVM (LS-SVM) was used and 98.53 % accuracy was obtained. A learning algorithm applying linear least squares reaching a classification accuracy of 96.0 % over the entire WBCD is presented in [17]. In [18], the combination of further division of partition space (FDPS) and flexible neural tree (FNT) is proposed to improve the neural network classification performance and the obtained result is 98.25 %. A method based on gravitational potential energy between particles [19] is applied for the whole WBCD dataset, and the best result obtained is 98.81 %. More recently in [20], a new resampling method called SUNDO is proposed combining an oversampling and an undersampling technique. Four classifiers based, respectively, on Support Vector Machine, Decision Tree, labeled Self-Organizing Map and Bayesian Classifiers have been developed and applied for WBCD. The best results obtained for these four classifiers are 97.2, 94.3, 96.7 and 96.4 %, respectively.

## 3 Method

### 3.1 Proposed hybrid system

K-Nearest neighbor (KNN) 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 datasets, 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 rough set theory. The most discriminatory set of features are obtained using rough set technique. Then, these features are used in classification phase as input to a combined classifiers using KNN as base classifier. The final results are combined using a majority voting (MV) technique. The block diagram of the whole classification system can be seen in Fig. 2.

#### 3.1.1 Rough set-based feature selection

Rough set theory (RST) is a new intelligent mathematical tool proposed by Pawlak in 1982 to deal with uncertainty and incompleteness [21]. Over the past few years, RST has become a topic of great interest to researchers and has been applied to many domains. It is based on the concept of an upper and a lower approximation of a set, the approximation space and models of sets. The main advantage of RST is that it does not need any preliminary or additional information about data: like probability in statistics or basic probability assignment in Dempster–Shafer theory and membership grade in fuzzy set theory [22]. One of the major applications of RST is the attribute reduction that is possible to find a minimal subset. The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Using the dependency degree as a measure, attributes are removed and reduced set provides the same dependency degree as the original. This section recalls some essential definitions from RST that are used for feature selection. Detailed description and formal definitions of the theory can be found in [23, 24].

*3.1.1.1 Information system* Knowledge representation in rough sets is done via information system [23], which is denoted as 4-tuple $S = \langle U, A, V, f \rangle$, where $U$ is the closed universe, a finite set of $N$ objects $\{x_1, x_2, \ldots, x_n\}$, $A$ is a finite set of attributes $\{a_1, a_2, \ldots, a_n\}$, which can be further divided into two disjoint subsets of $C$ and $D$, $A = \{C \cup D\}$ where $C$ is condition attributes and $D$ is a set of decision attributes. $V = U_{a \in A} V_a$ and $V_a$ is a domain of the attribute a, and $f : U \times A \rightarrow V$ is the total decision function called the information function such that $f(x, a) \in V_a$ for every $a \in A, x \in U$.

*3.1.1.2 Indiscernibility relation* One of the most significant aspects of RS theory is its indiscernibility relation. The $R$-indiscernibility relation is denoted by IND($R$), is defined as [23]:

$$\mathrm{IND}(R) = \{(x, y) \in U \times U | \forall a \in R, a(x) = a(y)\}$$

where $a(x)$ denotes the value of attribute a of object $x$. If $(x, y) \in \mathrm{IND}(R)$, $x$ and $y$ are said to be indiscernible with respect to $R$. The equivalence classes of the $R$-

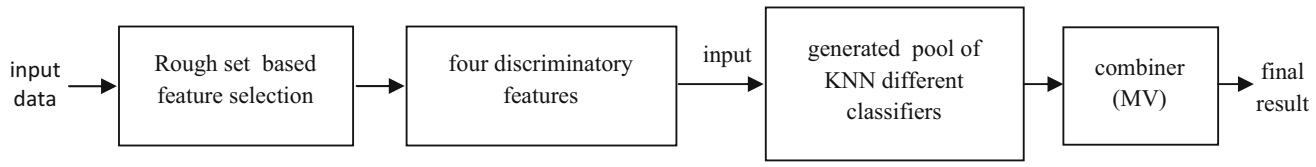


**Fig. 2** Block diagram of the proposed hybrid intelligent system

indiscernibility relation are denoted by $[x]_R$. The indiscernibility relation is the mathematical basis of RS theory.

*3.1.1.3 Lower and upper approximation* In RS theory, the lower and upper approximations are two basic operations, for any concept $X \subseteq U$ and attribute set $R \subseteq A$, $X$ could be approximated by the lower approximation and upper approximation. The lower approximation of $X$ is the set of objects of $U$ that are surely in $X$, defined as [23]:

$$\underline{R}(X) = \{x \in U | [x]_R \subseteq X\}$$

The upper approximation of $X$ is the set of objects of $U$ that are possibly in $X$, defined as [23]:

$$\bar{R}(X) = \{x \in U | [x]_R \cap X \neq \varnothing\}$$

and the R-boundary region of X is defined as [23]:

$$\mathrm{BND}(X) = \bar{R}(x) - \underline{R}(x)$$

A set is said to be rough if its boundary region is non-empty, otherwise the set is crisp.

*3.1.1.4 Attribute reduction and core* There often exist some condition attributes that do not provide any additional information about the objects in U in the information system [25]. So, these redundant attributes can be eliminated without losing essential classificatory information. Reduct and core attribute sets are two fundamental concepts of rough set theory. A reduct attribute set is a minimal set of attributes from A (the whole attributes set) that provided that the object classification is the same as with the full set of attributes. Given $C$ and $D \subseteq A$, a reduct is a minimal set of attributes such that $\mathrm{IND}(C) = \mathrm{IND}(D)$. Let $\mathrm{RED}(A)$ denote all reducts of $A$. The intersection of all reducts of $A$ is referred to as a core of A, i.e., $\mathrm{CORE}(A) = \cap\, \mathrm{RED}(A)$, the core is common to all reducts.

*3.1.1.5 Dependency degree* Various measures can be defined to represent how much $C$, a set of decision attributes, depends on $D$, a set of condition attributes. One of the most common measure is the dependency [25] denoted as $\gamma_c(D)$, is defined as: $\gamma_c(D) = |\mathrm{POS}_C(D)|/|U|$ where $|U|$ is the cardinality of set $U$, $\mathrm{POS}_C(D)$ called positive region, is defined by $\mathrm{POS}_C(D) = \cup_{x \in U/D}(X)$. Note that $0 \leq \gamma_C(D) \leq 1$, If $\gamma_c(D) = 1$ we say that $D$ depends totally on $C$, if $0 < \gamma_c(D) < 1$, we say that $D$ depends partially on $C$, and if $\gamma_c(D) = 0$ means that $C$ and $D$ are totally independent of each other.

*3.1.1.6 Reduction process for feature selection* We adapted Johnson's heuristic [26] to compute reducts as follows. It sequentially selects features by finding those that are most discernible for a given decision feature (see Fig. 3) [27]. It computes a discernibility matrix $M$, where

$m_{ij} = \{\{f \in F_p : f(c_i) \neq f(c_j)\}$ for $f_d(c_i) \neq f_d(c_j)$, and $\varnothing$ otherwise$\}$

JOHNSONS_REDUCT($F_p$, $f_d$, C)

Input $F_p$: conditional features, $f_d$: decision feature, C: cases

Output R: Reduct R $\subseteq$ $F_p$

1. $R \leftarrow \varnothing$, $F' \leftarrow F_p$
2. M $\leftarrow$ computeDiscernibilityMatrix(C, $F'$, $f_d$)
3. do
4. $f_h \leftarrow$ selectHighestScoringFeature(M)
5. $R \leftarrow R \cup \{f_h\}$
6. for ( i=0 to |C|, j=i to |C|)
7. $m_{i,j} \leftarrow \varnothing$ if $f_h \in m_{ij}$
8. $F' \leftarrow F' - \{f_h\}$
9. until $m_{ij} = \varnothing\ \forall$ i,j
10. Return R

**Fig. 3** Pseudocode for Johnson's heuristic algorithm

each cell $m_{i,j}$ of the matrix corresponding to cases $c_i$ and $c_j$ includes the conditional features in which the two cases' values differ. Formally, strict discernibility is defined as:

$$m_{ij} = \{\{f \in F_p : f(c_i) \neq f(c_j)\} \text{ for } f_d(c_i) \neq f_d(c_j), \text{and } \varnothing \text{ otherwise}\}$$

This is a greedy heuristic algorithm that is often applied to discernibility functions to find a single reduct [27]. The algorithm begins by setting the current reduct candidate, $R$, to the empty set. Then, each conditional attribute appearing in the discernibility function is evaluated according to the heuristic measure. For the standard Johnson algorithm, this is typically a count of the number of appearances an attribute makes within clauses; attributes that appear more frequently are considered to be more significant. The attribute with the highest heuristic value is added to the reduct candidate, and all clauses in the discernibility function containing this attribute are removed. As soon as all clauses have been removed, the algorithm terminates and returns the reduct $R$. $R$ is assured to be a reduct as all clauses contained within the discernibility function have been addressed. Variations of the algorithm involve alternative heuristic functions in an attempt to guide the search down better path. For breast cancer data (WBCD), the discriminatory set of features obtained by Johnson reduct is {Clump thickness, Uniformity of cell size, Bare nuclei, Bland chromatin}.

### *3.1.2 K-Nearest neighbor (KNN) algorithm*

KNN 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 $R^N$ [28]. Nearest neighbors 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 $d(x_i, x_j)$ is the Euclidean distance between $x_i$ and $x_j$ samples is defined by

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

KNN 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 KNN algorithm can be summarized as follows [28]: KNN 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 $d(x_i, x_j)$. 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:

1. Training phase: For each training example $\langle x, f(x)\rangle$, add the example to list training_examples.
2. Classification phase: Given a query instance $x_q$ to be classified,

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

   2.2 return $\hat{f}(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i))$ where $\delta(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$

### 3.1.3 Combining rules

Fixed combiners are heavily studied in the literature on combining classifiers, e.g., see [29–31]. The new confidence $q_j(x)$ for class $j$ is computed by:

$$h_j(x) = \text{rule}_i(p_{ij}(x)) \quad (1)$$

$$q_j(x) = \frac{h_j(x)}{\sum_j h_j(x)} \quad (2)$$

The following combiners are used for rule in (1): Maximum, Median, mEan, Minimum, Product. Note that the final classification is made by

$$w(x) = \arg\max_j (q_j(x)) \quad (3)$$

The Maximum rule selects the classifier producing the highest estimated confidence, which seems to be noise sensitive. In contrast, the Minimum rule selects by (3) the classifier having the least objection. Median and Mean average the posterior probability estimates thereby reducing estimation errors. This is good, of course, if the individual classifiers are estimating the same quantity. This probably will not hold for some of the classifiers.

In this study, a majority voting, which is a popular way of combining classifiers, is used. Majority counts the votes for each class over the input classifiers and select the majority class. This fits in the above framework if this rule is substituted in (1):

$$h_j(x) = \sum_i I\left(\arg\max_i (p_{ij}(x)) = i\right) \quad (4)$$

in which $I(\cdot)$ is the indicator function defined as follow:

$$I(y) = \begin{cases} 1, & \text{if } y \text{ is true} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

## 4 System performance measurements

### 4.1 Sensitivity, precision, $F$-measure, accuracy and specificity

There are different metrics to measure the performance of the classification methods which are commonly and widely used in automatic medical diagnosis systems. These metrics are True positive (TP), true negative (TN), false positive (FP) and false negative (FN). TP is the number of correct predictions in which an instance is positive; FN is the number of incorrect predictions in which an instance is negative; FP is the number of incorrect predictions in which an instance is positive; and TN is the number of correct predictions in which an instance is negative.

The dataset WBCD has two classes (positive and negative). Recall (sensitivity) is the percentage of real positive cases that are correctly predicted positive [Eq. (6)]. In a Medical context Recall is moreover regarded as primary, as the aim is to identify all real positive cases. Conversely, precision (confidence) indicates the percentage of predicted positive cases that are correctly real positives [Eq. (7)]. A measure ($F$-measure) that combines precision and recall is the harmonic mean of precision and recall [Eq. (8)]. Accuracy is the proximity of measurement results to the true value [Eq. (9)]. Specificity indicates the percentage of samples that were classified as normal and which were labeled as normal (Eq. (10). A measurement system can be accurate but not precise, precise but not accurate, neither,

or both. A measurement system is considered valid if it is both accurate and precise.

$$\text{Recall (sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$F\text{-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{8}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{10}$$

### 4.2 Youden's index, positive and negative likehood and discriminant power

Sokolova et al. [32] have shown that the accuracy measure does not distinguish between the number of correct labels of different classes. Sensitivity and specificity separately estimate a classifier's performance on different classes. It has been shown [32] that higher accuracy does not guarantee overall better performance of an algorithm and that a combination of measures gives a balanced evaluation of the algorithm's performance. In this paper, the Youden's index, likelihoods and discriminant power (DP) given in [32] are used to evaluate the performance of our system:

$$\text{Youden's index}: \ \gamma = \text{sensitivity} - (1 - \text{specificity}),$$

$$\text{Positive likehood}: \ \rho_+ = \text{sensitivity}/(1 - \text{specificity}),$$

$$\text{Negative likehood}: \ \rho_- = (1 - \text{sensitivity})/(1 - \text{specificity}),$$

$$\text{Discriminant power}: \ \text{DP} = \sqrt{3}/\pi \left[ \log\left( \frac{\text{sensitivity}}{1 - \text{sensitivity}} \right) + \log\left( \frac{\text{specificity}}{1 - \text{specificity}} \right) \right]$$

Youden's index evaluates the classifiers performance to a finer degree with respect to both class. A higher positive value of $\rho_+$ means a better performance on the positive (abnormal) class. A higher negative value of $\rho$-means a better performance on the negative (normal) class. The DP evaluates how well a classifier discriminates between normal and abnormal cases. The classifier performance is poor if DP $< 1$, limited if DP $< 2$, fair if DP $< 3$, good in other cases [32].

### 4.3 Confusion matrix

A confusion matrix [33] 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 4 shows the confusion matrix for a two class classifier.

**Table 4** Representation of confusion matrix

| Actual | Predicted | |
|---|---|---|
| | Negative | Positive |
| Negative | a | b |
| Positive | c | d |

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

### 4.4 Receiver operating characteristics (ROC)

Curves are used for analyzing the prediction performance of predictor [34]. The information of ROC curves is helpful in selection of appropriate classifier under certain decision criteria. The improvement in ROC curves represents low values of false-positive rate and high values of true-positive rate. These values help the points shifting toward upper left corner of ROC and thus providing better decision. This kind of behavior is desirable in those applications where the cost of false-positive rate (FPR) is too important. For example, a weak patient cannot afford high FPR. Minor damage of healthy tissues may be a matter of life and death. On the other hand, when attempts are made to reduce FPR by simply adjusting decision threshold, the risk of false negative cases might rise in a poor prediction model. This kind of prediction model, specifically in medical applications, might cause high misclassification cost in various fatal diseases such as lungs, liver and breast/colon cancer. The values of area under curve (AUC) and area under convex hull (AUCH) are calculated for assessing performance of the classifier on imbalanced/balanced dataset and examination of classifier consistency.

## 5 Results and discussion

To evaluate the effectiveness of the proposed hybrid intelligent system for diagnosis of breast cancer, experiments are conducted on the WBCD database mentioned above. First, rough set technique was used as a preprocessing step for feature selection. The Johnson algorithm is used to obtain the minimum reduct which it contains only four features. The reduced set of features is {Clump thickness, Uniformity of cell size, Bare nuclei and Bland chromatin}. In the prediction phase, these four features are used as input to a pool of KNN classifier. For the sake of checking the effectiveness of the proposed hybrid intelligent method, the whole dataset is divided into two disjoint subsets, namely 50 % for training and 50 % for testing for all the conducted experiments.

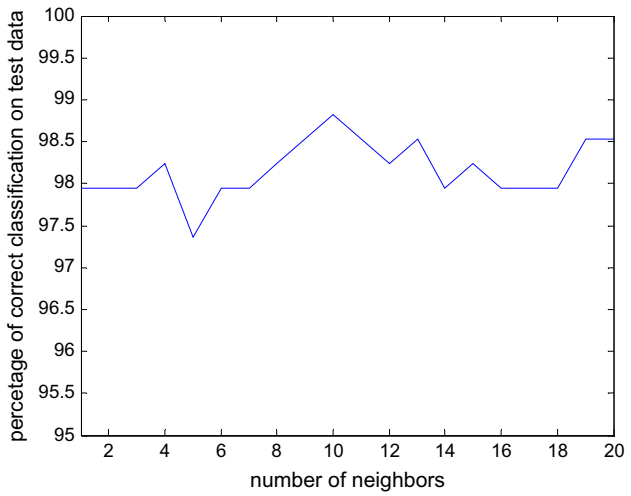


**Fig. 4** Accuracies of the base KNN classifiers versus the number of neighbors

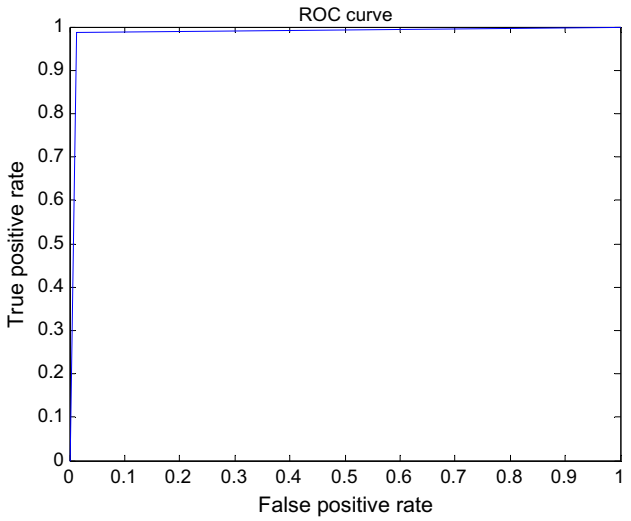


**Fig. 5** The ROC curve for the best individual KNN classifier

Figure 4 gives the accuracies of the pool of KNN classifiers versus the number of neighbors. It is observed that, in the case of using individual classifiers, there is an oscillation in the obtained results and the best individual classifier is obtained at $k = 10$ and its accuracy is 98.83 %.

A combined classifier is built using the majority voting technique to combine the results obtained by the pool of the individual classifiers. The accuracy of the ensemble classifier is 99.41 % which is superior over all the individual classifiers.

Also, comparison of our results with the previous results reported by earlier and state of the art methods indicates that the proposed hybrid system obtains the highest classification test accuracy reported so far.

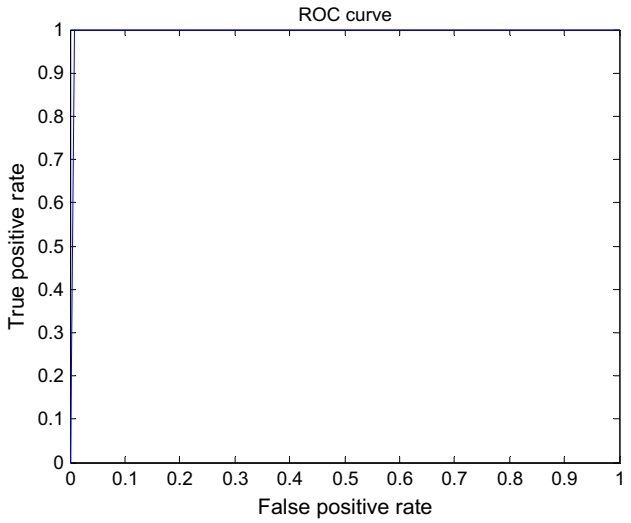


**Fig. 6** The ROC curve for the hybrid classifier

Moreover, different performance measures are used to test the effectiveness of the proposed system. Figures 5 and 6 give the ROC curve for the best individual KNN classifier and the proposed hybrid classifier. The area under the ROC curve is obtained which gives the accuracy of the classifier.

The obtained classification accuracy and the values of different performance measures such as sensitivity, precision, *F*-measure and specificity are given in Table 5 and Fig. 7 for both the best individual KNN classifier and the proposed hybrid system.

Also, classification results of the proposed hybrid system are displayed by using a confusion matrix, see Table 6. In a confusion matrix, each cell contains the raw number of exemplars classified for the corresponding combination of desired and actual outputs, as described above. It is clear that, for the proposed system, all the malignant cases are classified correctly and only two benign cases out of 300 cases are misclassified may be this is due to the overlap between the two regions.

Beside, another well-known performance metrics namely, Youden's index, positive and negative likehood and discriminant power are obtained see Table 7. The Discriminant Power for both the best individual classifier and the proposed hybrid intelligence system is greater than 3 which indicates that good performance for both of them. The Discriminant Power of the proposed system is Infinity which shows how well the proposed system discriminates between malignant and benign cases.

From the obtained results, we conclude that the proposed hybrid intelligence system 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

**Table 5** System performance measurements of the best individual KNN classifier and proposed hybrid system

| Proposed method | Accuracy | Precision | Recall or sensitivity | Specificity | *F*-measure |
|---|---|---|---|---|---|
| Best individual classifier (%) | 98.83 | 96.39 | 98.77 | 98.85 | 97.57 |
| Proposed hybrid method (%) | 99.41 | 97.59 | 100 | 99.23 | 98.78 |

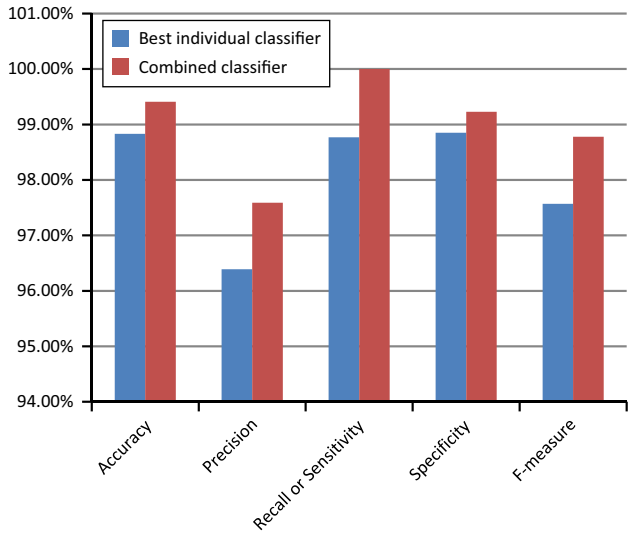


**Fig. 7** System performance measurements of the best individual KNN classifier and proposed hybrid intelligence method on the test data

**Table 6** Confusion matrix for the best individual KNN classifier and proposed hybrid system

| Classifier | Desired result | Output result | |
|---|---|---|---|
| | | Benign | malignant |
| Best individual classifier | Benign | 257 | 3 |
| | malignant | 1 | 80 |
| Proposed hybrid method | Benign | 258 | 2 |
| | malignant | 0 | 81 |

using such an efficient tool, they can make accurate decisions with only the four discrimination features obtained.

## 6 Conclusion

This study aims at diagnosing breast cancer with a hybrid intelligent system. By hybridizing a rough set theory with a combined classifier based on *k*-nearest neighbor algorithm as base classifier, a method was obtained to solve this diagnosis problem via classifying Wisconsin breast cancer dataset (WBCD). This dataset is a very commonly used dataset 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 hybrid intelligent system with regard to other studies. Using only four discriminatory features obtained by rough set, a classification accuracy of 99.41 % is obtained, which is the highest one reached so far. The effectiveness of the proposed hybrid system is shown using different performance measurements which are commonly used in testing the performance of automatic medical systems. These results are for WBCD, but it states that this proposed hybrid intelligent system can be used confidently for other breast cancer diagnosis problems, too. Also, besides of breast cancer problem, other medical diagnosis applications can also be conducted by this system.

**Table 7** Another system performance measurements of the best individual KNN classifier and proposed hybrid method

| Proposed method | Youden's index | Positive likehood ratio | Negative likehood ratio | Discriminant power |
|---|---|---|---|---|
| Best individual classifier | 0.9762 | 85.8870 | 1.0696 | 4.8735 |
| Proposed hybrid method | 0.9923 | 129.87 | 0 | Infinity |

## References

1. American Cancer Society Homepage (2014) Citing internet sources available from: http://www.cancer.org. Accessed 10 May 2014
2. Ghosh J (2002) Multiclassifier systems: back to the future. In: Roli F, Kittler J (eds) Multiple classifier systems. Lect Notes Comput Sci 2364:1–15
3. Zhang C, Ma Y (2012) Ensemble machine learning: methods and applications. Springer, Berlin
4. Etemad SA, Arya A (2014) Classification and translation of style and affect in human motion using RBF neural networks. Neurocomputing 129:585–595
5. Meynet J, Thiran JP (2010) Information theoretic combination of pattern classifiers. Pattern Recogn 43(10):3412–3421
6. Wolberg WH, Mangasarian OL (1990) Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci USA 87(23):9193–9196
7. Quinlan JR (1996) Improved use of continuous attributes in C4.5. J Artif Intell Res 4:77–90
8. Hamilton HJ, Shan N, Cercone N (1996) RIAC: a rule induction algorithm based on approximate classification. Technical Report CS 96-06, University of Regina
9. Ster B, Dobnikar A (1996) Neural networks in medical diagnosis: comparison with other methods. In: Proceedings of the

international conference on engineering applications of neural networks, pp 427–430
10. Bennet KP, Blue JA (1997) A support vector machine approach to decision trees, Math Report, vols. 97–100, Rensselaer Polytechnic Institute
11. Nauck D, Kruse R (1999) Obtaining interpretable fuzzy classification rules from medical data. Artif Intell Med 16:149–169
12. Pena-Reyes CA, Sipper M (1999) A fuzzy-genetic approach to breast cancer diagnosis. Artif Intell Med 17:131–155
13. Setiono R (2000) Generating concise and accurate classification rules for breast cancer diagnosis. Artif Intell Med 18:205–219
14. Goodman DE, Boggess L, Watkins A (2002) Artificial immune system classification of multiple-class problems. In: Proceedings of the artificial neural networks in engineering ANNIE, pp 179–183
15. Abonyi J, Szeifert F (2003) Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recogn Lett 24:2195–2207
16. Polat K, Günes S (2007) Breast cancer diagnosis using least square support vector machine. Digit Signal Proc 17(4):694–701
17. Guijarro-Berdias B, Fontenla-Romero O, Perez-Sanchez B, Fraguela P (2007) A linear learning method for multilayer perceptrons using leastsquares. Lect Notes Comput Sci 365–374
18. Yang B, Wang L, Chen Z, Chen Y, Sun R (2010) A novel classification method using the combination of FDPS and flexible neural tree. Neurocomputing 73:690–699
19. Shafigh P, Yazdi Hadi S, Sohrab E (2013) Gravitation based classification. Inf Sci 220:319–330
20. Cateni S, Colla V, Vannucc M (2014) A method for resampling imbalanced data sets in binary classification tasks for real-world problems. Neurocomputing 135:32–41
21. Pawlak Z (1982) Rough sets. Int J Parallel Prog 11(5):341–356
22. Chen HL, Yang B, Liu J, Liu DY (2011) A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Syst Appl 38:9014–9022
23. Pawlak Z (1996) Why rough sets. In: Proceedings of the fifth IEEE international conference on fuzzy systems, vol 2, 8–11 September 1996, New Orleans, LA, USA, pp 738–743
24. Rami N, Khushaba N, Al-Ani A, Al-Jumaily A (2011) Feature subset selection using differential evolution and a statistical repair mechanism. In: Expert systems with applications. Elsevier, pp 11515–11526
25. Pawlak Z (1997) Rough set approach to knowledge-based decision support. Eur J Oper Res 99(1):48–57
26. Johnson DS (1974) Approximation algorithms for combinatorial problems. J Comput Syst Sci 9:256–278
27. Jensen R, Shen Q (2008) Computational intelligence and feature selection: rough and fuzzy approaches. Wiley
28. Mitchell TM (1997) Machine learning. The McGraw-Hill
29. Kittler J, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20(3):226–239
30. Xu L, Krzyzak A, Suen CY (1992) Methods of combining multiple classifiers and their application to handwriting recognition. IEEE Trans SMC 22:418–435
31. Schapire RE (1990) The strenght of weak learnability. Mach Learn 5:197–227
32. Sokolova M, Japkowicz N, Szpakowicz S (2006) Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. Adv Artif Intell 1015–1021
33. Kohavi R, Provost F (1998) Glossary of terms. Editorial for the Special Issue on Appl Mach Learn the Knowl Discov Process 30(2–3)
34. Tom F (2004) ROC graphs: notes and practical considerations for researchers. Mach Learn 31:1–38