Neural Comput & Applic (2017) 28:753–763
DOI 10.1007/s00521-015-2103-9

ORIGINAL ARTICLE

Breast cancer diagnosis using GA feature selection and Rotation Forest

Emina Aličković[1] · Abdulhamit Subasi[2]

Received: 24 February 2015 / Accepted: 4 November 2015 / Published online: 18 November 2015

Abstract Breast cancer is one of the primary causes of death among the women worldwide, and the accurate diagnosis is one of the most significant steps in breast cancer treatment. Data mining techniques can support doctors in diagnosis decision-making process. In this paper, we present different data mining techniques for diagnosis of breast cancer. Two different Wisconsin Breast Cancer datasets have been used to evaluate the system proposed in this study. The proposed system has two stages. In the first stage, in order to eliminate insignificant features, genetic algorithms are used for extraction of informative and significant features. This process reduces the computational complexity and speed up the data mining process. In the second stage, several data mining techniques are employed to make a decision for two different categories of subjects with or without breast cancer. Different individual and multiple classifier systems were used in the second stage in order to construct accurate system for breast cancer classification. The performance of the methods is evaluated using classification accuracy, area under receiver operating characteristic curves and *F*-measure. Results obtained with the Rotation Forest model with GA-based 14 features show the highest classification accuracy (99.48 %), and when compared with the previous works, the proposed approach reveals the enhancement in performances. Results obtained in this study have potential to open new opportunities in diagnosis of breast cancer.

Keywords Logistic Regression · Decision Trees · Random Forest · Bayesian Network · Multilayer Perceptron (MLP) · Radial Basis Function Networks (RBFN) · Support Vector Machine (SVM) · Rotation Forest · Breast cancer diagnosis · Genetic algorithm (GA)

1 Introduction

Breast cancer is the most frequent cancer in females worldwide, encompassing 15 % of all female cancers. In 2012, 521,000 deaths were due to breast cancer. In spite of some risk, shortening might be accomplished with prevention; these approaches cannot lessen the most of breast cancers diagnosed in very late stages. As a result, early detection is the cornerstone of breast cancer control to improve breast cancer survival [57].

Mammography and fine-needle aspiration cytology (FNAC) are typically used diagnostic techniques, but these techniques have a lack of satisfying diagnostic performances. There is no hesitation that assessment of data obtained from patients' and doctors' decisions is the most valuable elements in diagnosis. Together with mammography and FNAC, different data mining techniques can be supportive tool in doctors' diagnosis and decision making; as a result, improved diagnosis system can be obtained. In regard to the above-mentioned requirements, data mining techniques can be utilized to facilitate improvement of the diagnostic systems. With using automatic diagnostic systems, the probable doctor mistakes during diagnosis can be eliminated, and the medical statistics can be analyzed in

✉ Abdulhamit Subasi
absubasi@effatuniversity.edu.sa

Emina Aličković
ealickovic@ibu.edu.ba

[1] Faculty of Engineering and Information Technologies, International Burch University, Francuske Revolucije bb. Ilidza, 71000 Sarajevo, Bosnia and Herzegovina

[2] Computer Science Department, College of Engineering, Effat University, Jeddah 21478, Saudi Arabia

more detail in a less amount of time. The purpose of this study is to establish an accurate automatic diagnostic system which can distinguish among benign breast tumors from malignant cancers. To solve this task, different data mining techniques were applied and their performances were evaluated and compared. These techniques include Logistic Regression, Decision Trees, Random Forest, Bayesian Network, Multilayer Perceptron (MLP), Radial Basis Function Networks (RBFN) and Support Vector Machine (SVM).

Selection of the most significant and informative features and removal of the remaining features (or in other words compression of original feature set to smaller set) are one of the most important tasks in design of the efficient classification model. Therefore, in order to construct an efficient breast cancer diagnosis dataset, there is a need for a method which will efficiently extract the most informative features given following constraint: Lack of any previous familiarity with information contained within the original data set and significance of the original features should be preserved. Genetic algorithms (GAs) may be employed as a tool to determine information dependencies and decrease the number of features in a dataset by simply structural techniques [37]. GA can be seen as data compression algorithm which eliminates unwanted features and chooses a feature subset having the equal discernibility as the initial set of features, resulting in better classification performances [9]. One of the main goals of this study is to use advantages of GA in feature reduction in the breast cancer data in constructing automatic diagnosis system. New dataset obtained after GA is fed as input to different classifiers. Our proposed approach has two stages. During the first stage, GA is employed as a feature reduction mechanism to determine the discriminative features. It serves to remove redundant data. In the second stage, the best feature subset is employed as the input to different data mining techniques. The accomplishment and efficiency of the methods are evaluated on breast cancer datasets. Experiments proved that data mining techniques have better predicative classification accuracy and performances with smaller number of attributes.

Numerous studies have proposed different systems for automatic diagnosis of breast cancer based on Wisconsin Breast Cancer datasets, and many of these studies reported high classification performances. Quinlan [41] used the C4.5 decision tree method and tenfold cross-validation. Hamilton et al. [17] used RIAC method, and Ster and Dobnikar [48] used linear discrete analysis method. Pena-Reyes and Sipper [38] used fuzzy-GA method, and Setiono [47] used a feed-forward neural network rule extraction algorithm. Albrecht et al. [3] obtained 98.80 % accuracy using learning algorithm that combined logarithmic simulated annealing with the perceptron [33]. Goodman et al. [15] used three distinct methods namely artificial immune recognition system (AIRS), big LVQ and optimized learning vector quantization, achieved 97.2, 96.8 and 96.7 % accuracies, respectively. Abonyi and Szeifert [2] used supervised fuzzy clustering method, and Hassanien [21] used rough set method. Sahan and Polat [44] used a novel hybrid technique based on fuzzy-artificial immune system and k-NN algorithm, and the accuracy was 99.14 % [9]. Maglogiannis and Zafiropoulos [32] used three different methods: SVM, Bayesian classifiers and artificial neural networks (ANNs). Peng et al. [39] used a hybrid method that joins filter and wrapper tools [9, 33]. In [49], support vector machine (SVM) and evolutionary algorithm were used, and obtained accuracy was around 97 %. Koloseni et al. [27] used differential evolution classifier with optimal distance measures applied on WDBC dataset, and obtained average classification accuracy was around 93.64 %. Astudillo et al. [4] applied tree-based topology-oriented SOM on WDBC dataset to discriminate between malign and benign cancer, and obtained classification accuracy was 93.32 %. Tabakhi et al. [51] proposed unsupervised feature selection algorithm based on ant colony optimization for feature selection and Naïve Bayes for classification, and obtained classification accuracy with this system for discriminating between benign and malign cancer was 92.42 % when applied on WDBC dataset. Saez et al. [43] proposed mutual information (MI) between features as a weighting factor for nearest neighbor (NN) classifier, and obtained classification accuracy for WDBC dataset was 96.14 %. Chen et al. [8] suggested system based on parallel time-variant particle swarm optimization (PTVPSO) for concurrent parameter optimization and feature selection for SVM, and obtained classification accuracy was 98.44 % when this system was applied on WDBC dataset. Zheng et al. [60] proposed breast cancer diagnosis system based on K-means and SVM (K-SVM), and in this study, proposed system was tested on WDBC dataset, and obtained classification results were 97.38 %. Lim et al. [30] extended Bandler–Kohout (BK) subproduct to interval-valued fuzzy sets (IVFS), and obtained classification accuracy with this approach was 95.26 % for WDBC dataset.

In this study, GA feature selection and different data mining techniques, namely Logistic Regression, Decision Trees, Random Forest, Bayesian Network, MLP, RBFN, SVM and Rotation Forest, have been investigated, in order to construct automated system which will distinguish between benign and malign tumor in breast cancer. Also the widely used datasets in the literature were used to evaluate performances of the proposed system. It is observed that the Rotation Forest which is a multiple classifier system (MCS) with GA feature selection achieved the highest classification accuracy (99.48 %) in breast cancer data classification.

This paper is organized as follows. In the next section, information is given about the Wisconsin Diagnostic Breast Cancer datasets, and the methods used in each step of the classification process are presented. Section 3 provides a complete experimental study of the different data mining techniques for diagnosis of breast cancer, in which the effect of feature set and algorithmic concerns are compared with respect to the classification performance. Finally, the conclusions are summarized in Sect. 4.

2 Materials and methods

2.1 Breast cancer database overview

Breast cancer is a malignant tumor arising from breast cells. Even though some of the risk factors (e.g., aging, genetic risk factors, family history, menstrual periods, not having children, obesity) that raise a woman's possibility of developing breast cancer are known, it is not known yet what causes most of the breast cancers and how various factors initiate cells to change into cancerous. Many studies are conducted to learn more, and scientists are having great improvement in understanding how certain alterations in DNA which can affect healthy breast cells to change into cancerous [25, 33].

In this study, two different Wisconsin Breast Cancer Datasets (obtained from UCI Machine Learning Repository) were studied. The first dataset is Wisconsin Breast Cancer (Diagnostic) (WBC (DIAGNOSTIC)) dataset. This dataset contains 569 different instances and 32 attributes. Three hundred and fifty-seven cases are benign, and 212 cases are malignant. All attributes are calculated from a digitized image of a fine-needle aspirate (FNA) of patients' breast tissues. All cell nuclei in breast tissues are described by ten real-valued features, and for all these features, the mean, the standard error and the "worst" (mean of the three largest values) are calculated. As a result, a total of 30 attributes for all images were obtained [52]:

- Radius (mean of distances from center to points on the perimeter)—$a_{1,1}$, $a_{1,2}$, $a_{1,3}$;
- Texture (standard deviation of grayscale values)—$a_{2,1}$, $a_{2,2}$, $a_{2,3}$;
- Perimeter—$a_{3,1}$, $a_{3,2}$, $a_{3,3}$;
- Area—$a_{4,1}$, $a_{4,2}$, $a_{4,3}$;
- Smoothness (local variation in radius lengths)—$a_{5,1}$, $a_{5,2}$, $a_{5,3}$;
- Compactness (perimeter^2/area − 1.0)—$a_{6,1}$, $a_{6,2}$, $a_{6,3}$;
- Concavity (severity of concave portions of the contour)—$a_{7,1}$, $a_{7,2}$, $a_{7,3}$;
- Concave points (number of concave portions of the contour)—$a_{8,1}$, $a_{8,2}$, $a_{8,3}$;
- Symmetry—$a_{9,1}$, $a_{9,2}$, $a_{9,3}$;
- Fractal dimension ("coastline approximation"-1)—$a_{10,1}$, $\mathrm{a}_{10,2}$, $\mathrm{a}_{10,3}$;

where $a_{i,1}$ refers to ith attribute mean, $a_{i,2}$ refers to ith attribute standard error, and $\mathrm{a}_{i,1}$ refers to ith attribute "worst" ($i = 1,\ldots30$).

The second dataset is Wisconsin Breast Cancer Original dataset and contains 699 samples obtained from a breast tissue. Subsequently, data with missing values are removed from dataset; as a result, 683 cases are used in our experiment. Every record in the database has nine attributes, with all values represented as integer numbers between 1 and 10, and was found to fluctuate notably among benign and malignant instances. The measured nine attributes are [53]:

- Clump thickness;
- Uniformity of cell size;
- Uniformity of cell shape;
- Marginal adhesion;
- Single epithelial cell size;
- Bare nuclei;
- Bland chromatin;
- Normal nuclei;
- Mitoses.

2.2 Genetic algorithm-based feature selection

Genetic algorithms (GA) have found broad range of applications. It is established on the resemblance to natural selection. GA operates with population, and the preeminent solution is received after a sequence of iterative steps. GA develops sequential populations of periodic solutions that are shown by a chromosome until satisfactory results are reached [55].

A fitness function estimates the importance of the answer in the evaluation step. Two major operators are crossover and mutation functions, and these have the key impact on the fitness value. Chromosomes for reproduction are selected by finding the fitness value, and the bigger fitness value is obtained, by selecting the chromosome with higher probability. The fitter chromosomes have higher likelihood to be selected into the recombination pool using either the roulette wheel or the tournament [55].

In mutation, the genes may be updated randomly. Crossover is genetic operator that joins distinct features from subsets pair into novel subset. Offspring substitutes the previous population using the elitism or variety replacement strategy to create a novel population in the upcoming generation [55]. To accomplish better performance, GA-based selected features are applied as an input to classifiers.

There are three criteria to model fitness function: model accuracy, number of selected features and cost. For any chromosome with acceptable classification accuracy rate, selection of only significant and informative features and reduced cost result in a satisfactory fitness value. The chromosome with higher fitness value has better chance to be used in the following generation, so these are properly expressed according to user's specifications. To get accurate feature selection based on GA, these steps are to be followed [55]:

1. Data preprocessing (scaling): Two advantages of scaling are evading of attributes in bigger numeric range to control attributes in lesser numeric range and avoiding of numerical difficulties in calculation [24, 55].
2. Conversion of genotype to phenotype: Here we convert each feature chromosome.
3. Feature subset
4. Fitness evaluation
5. Termination criteria (if it is met, process is stopped; otherwise, we continue with next generation.
6. Genetic operation: In this step, better solution is being searched by genetic operations.

The GA algorithm applied to feature selection is presented in Fig. 1.

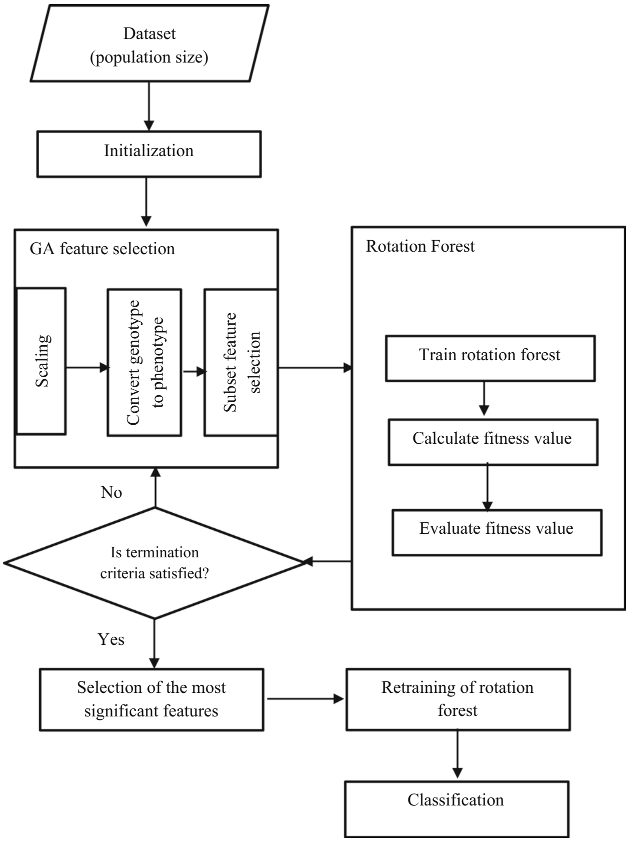

Fig. 1 Algorithm for proposed GA–Rotation Forest model

2.3 Logistic Regression

Logistic model originated as result of modeling the posterior probability of K classes via linear functions in x, while ensuring that they sum to one and remain in range [0, 1]. Model can be identified in terms of $K - 1$ *logit* transformations or log odds. Even though the model utilizes the last class as the denominator in the odds ratio, the selection of denominator is random in that the estimates are equally distributed under this choice. When $K = 2$, the model is straightforward because there is just a single linear function. In biostatical applications where binary response (only two classes) occurs repeatedly, this model is used extensively [16, 22, 56].

2.4 Bayesian Network

Bayesian Network illustrates the joint probability distribution for a set of variables by defining sets of local conditional probabilities together with a set of conditional independence assumptions. Every variable in the joint space is shown by a node in the Bayesian Network. For all variables, two types of information are specified. First, the variable is conditionally independent of its non-descendants in the network given its instant predecessors in the network. Second, a conditional probability table is given for every variable, telling probability distribution for that variable assumed the values of its instant antecedents. The joint probability for any desired assignment of values ($\boldsymbol{b_1}$, ..., $\boldsymbol{b_n}$) to the tuple of network variables ($\boldsymbol{B_1 \ldots B_n}$) can be computed by the formula:

$$P(b_1, \ldots, b_n) = \prod_{i=1}^{n} P(b_i|\text{Parents}(b_i)) \quad (1)$$

where Parents(B_i) denotes the set of immediate predecessors of B_i in the network. Values of $P(b_i|\text{Parents}(B_i))$ are the values stored in the conditional probability table associated with node B_i [46].

2.5 Multilayer Perceptron (MLP)

Multilayer Perceptrons (MLPs) are neural networks consisting of units that create the input layer, one or more hidden layers of computation nodes and output layer consisting of computation nodes. Input signal travels in forward direction on layer-by-layer basis. MLPs are successfully used to solve challenging and distinct problems by training them in supervised manner using well-known *back-propagation algorithm* [23].

Back-propagation learning constitutes of two passes through distinct layers: a forward pass and backward pass. In the forward pass, synaptic weights are all fixed, while, on the other hand, in the backward pass weights are adjusted. Error signal is created when the actual output of the network is subtracted from target data. This error signal propagates through the network in opposition to the direction of synaptic connections. Weights are tuned to build the real response more closely to the target. This learning process is named as *back-propagation learning* [23].

2.6 Radial Basis Function Networks (RBFN)

RBFN is popular substitute to Multilayer Perceptron (MLP) because it has more simple structure and more rapid training process. In RBFN, each neuron in the hidden layer uses RBF as its nonlinear activation function. A nonlinear transformation of the input is done in the hidden layer. Output layer of RBFN is a linear combiner and maps the nonlinearity into a new space. The output layer neurons' biases can be designed by adding extra neuron in the hidden layer, and constant activation function of the hidden layer is 1 [10].

2.7 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning method, and it chooses a modest amount of significant limit samples known as *support vectors* from all classes and constructs a linear discriminant function dividing them as broadly as it can be accomplished. These systems exceed the restrictions of linear limits by adjusting it to consist of an additional nonlinear function terms, preparing it to establish quadratic, cubic and higher-order decision limits [58].

Support Vector Machines (SVMs) are built on algorithm that develops a particular type of linear model called the *maximum margin hyper plane*. Hyper plane is different expression for a linear model. To illustrate a maximum margin hyper plane, it can be thought of a two-class dataset with linearly separable classes; in other words, there is a hyper plane in sample space classifying the entire training samples accurately. The greatest margin hyper plane is the one offering the most supreme division among the classes. It goes no nearer to any than it ought. To be precise, the *convex hull* of a group of points is the most stable enclosing convex polygon: It appears as soon as each point of the set is linked to each other point. Since it is assumed that two classes are linearly separable, their convex hulls never concatenate. Between every hyper plane dividing the classes, the maximum margin hyper plane is the one being the most furthest away from both convex hulls. It is the vertical bisector of the least distanced line linking the hulls [58]. With the selection of satisfactory mapping, the input examples become linearly or approximately linearly divisible in the high-dimensional plane. The SVM tries to find the optimal hyper plane that maximizes the distance between the instances of two different classes [54].

2.8 C4.5 Decision Tree

This algorithm is developed by J. Ross Quinlan, and it begins with big sets of samples being part of identified classes. The samples, defined by whichever combination of nominal and numeric characteristics, are considered for patterns that permit the classes to be accurately characterized. These patterns are then expressed as models, forming Decision Trees or sets of if–then rules that can be employed to classify novel samples, with special accent on making the models comprehensible and precise. C4.5 algorithm uses equations established on information theory to estimate the "goodness" of the test; particularly, they select the test that extracts the highest amount of data from a set of samples, given the restriction that just single attribute is to be tested [40].

In decision tree algorithms, problems are how to handle unknown values and overfitting. C4.5 is able to handle unknown values: Essentially, samples with unknown values are neglected, while calculating the data content and the data gain for an attribute $\boldsymbol{A}$ is subsequently multiplied by the fragment of samples where $\boldsymbol{A}$ value is already defined. Thus if A is unknown for a large fragments of samples, the data received by testing $\boldsymbol{A}$ at a node will be relatively petite. This matches the normal perception regarding how these attributes ought to be treated. A decision tree that accurately classifies all samples in a training set may not be as excellent classifier as a lesser tree not fitting all whole training data. In order to avoid this problem, pruning approach had been adopted for C4.5. This method is established on evaluation of error rate for all subtrees, and displacing the subtree with a leaf node in case when the evaluated error of the leaf is smaller. If the evaluation were ideal, this approach would the entire time guide to an improved decision tree. In reality, even though these are very crude, this approach frequently performs relatively fine [40].

2.9 Random Forests (RF)

Random Forests [5] is an important alteration of wrapping that constructs a large collection of de-correlated trees and then averages them. RF is very simple to train and adjust. As a consequence, it found wide range of applications. RF is used for both classification and regression, although there is difference when they are used for classification and when they are used for regression. When RF is employed to perform classification task, it receives a class vote from each tree and then using majority vote performs classification task. When RF is used for regression, predictions at a target point x from each tree are plainly averaged [22].

Usage of out-of-bag (OOB) samples is a significant characteristic of Random Forests. RFs utilize the OOB samples to construct a diverse variable rank measure and to compute the prediction strength of each variable. After the *bth* tree is developed, the OOB samples are sent to the tree and then prediction accuracy is recorded. After this, in the OOB samples, values for the *jth* variable are randomly selected, and the accuracy is calculated again and as a result of this random selection, accuracy is averaged over all trees and then used as a measure of the importance of variable j in the Random Forest [22].

2.10 Rotation Forest

Rotation Forest is a novel method for generation of group of classifiers. In the first step, the feature set is split into S subsets, and principal component analysis (PCA) runs independently on every subset, and after that a novel extracted feature set is reconstructed during which all the components are preserved. New features are obtained from linearly transformed data. A SVM with polynomial kernel is used in this study as base classifier for Rotation Forest. Distinct feature set splits direct to distinct rotations. As a consequence, distinct classifiers are acquired. But also, the evidence how the data are scattered is saved in the novel extracted feature space. Thus, individual classifiers with high performances are constructed. Therefore, achieving both diversity and accuracy together is the objective of Rotation Forest [42].

3 Results and discussion

In this study, we used two different WBC medical datasets to test the performances of models. These two datasets are WBC (Diagnostic) and WBC (Original) and are explained in Sect. 2.1. We used different data mining techniques namely Logistic Regression, Decision Trees, Random Forest (RF), Bayesian Network, Multilayer Perceptron (MLP), Radial Basis Function Networks (RBFN), Support Vector Machine (SVM) and Rotation Forest. Also we used genetic algorithm-based feature selection to find best attributes, and then, we applied data mining techniques for classification.

3.1 Experimental setup and dataset

Two different experiments were set up for the training data for two different WBC datasets. In the first case, the same training–testing dataset was applied as in [1, 13, 19]. In this work, publically available open-source machine learning software, called WEKA, was employed to implement algorithms and approach proposed in this study. In this training dataset, tenfold cross-validation was used. In the second case, we used GA feature selection, where the best attributes were selected, and then, tenfold cross-validation was used on these selected attributes. Numerous researches evaluating breast cancer classification using k-fold cross-validation can be found in the literature. In k-fold cross-validation technique, the dataset is separated into k subsets randomly. As a result, $k - 1$ subsets, in our case nine subsets, are used for training, and the rest is used for testing of the classifier efficiency [20]. We compared the efficiency of proposed techniques without GA feature selection and with GA feature selection.

Area under ROC [receiver operating characteristic (ROC)] curve (AUC) was also employed to assess the discrimination capability of the classifiers proposed in this study. ROC curves represent the performance of a classifier without taking into consideration class distribution or error overheads. A ROC curve is produced by plotting all sensitivity values (true-positive fraction), on the y-axis, adjacent to their equivalent (1-specificity) values (false-positive fraction) for all presented thresholds on the x-axis. The worth of the approximation to a curve is dependent on numerous thresholds tested. For all folds of a tenfold cross-validation, weight the samples for a selection of distinct overhead ratios, train the system on all weighted sets, calculate the true positives and false positives in the test set and plot the outcome point on the ROC axes [58]. The classification success is then calculated by AUC. The average AUC value provides a sign of a characteristic AUC values generated using the specified input data and displays how consistently result is predicted [18, 36, 50]. AUC is generally considered as the index of performance since it provides a single measure of total accuracy that does not depend on any specific threshold [34, 50]. Regardless of its positive sides, the ROC plot does not provide a rule for the case classification. However, there are approaches that can be employed to create decision rules from the ROC plot [12, 45]. As a guideline, Zweig and Campbell [45, 61] proposed that if the false-positive costs (FPCs) go beyond the false-negative costs (FNCs) the threshold should support specificity, while sensitivity should be supported if FPCs are bigger than the FNCs. Associating these overheads (costs) with the prevalence (p) of positive cases permits the computation of a slope [34, 45]:

$$m = (\mathrm{FPC}/\mathrm{FNC}) \times ((1 - P)/P) \tag{2}$$

where m refers to the slope of a tangent to the ROC plot. The sensitivity/specificity pair is positioned where the line and the curve first make contact [45]. An additional measure used to describe performance is F-measure, defined as:

$$F - \text{measure} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

3.2 Results without GA

The experimental results achieved for WBC (Diagnostic) dataset are given in Table 1. We get an average accuracy of 97.19 % for Logistic Regression, 93.32 % for Decision Tree (C 4.5), 96.13 % for Random Forest, 95.08 % for Bayes Net, 96.66 % for Multilayer Perceptron (MLP), 94.20 % for Radial Basis Function Network (RBFN), 96.89 % for SVM and 97.41 % for multiple classifier system (MCS) tool Rotation Forest.

The experimental results obtained for WBC (Original) dataset are given in Table 2. As shown in the Table 2, total accuracy achieved with the SVM classifier based on the polynomial kernel on the test set was equal to 96.78 %. The total accuracies are equal to 95.75 % for the RBFN classifier, 96.05 % for the Multilayer Perceptron (MLP), 96.05 % for the C4.5 Decision Tree, 96.34 % for the Random Forest, 97.22 % for the Bayes Net, 96.78 % for the Logistic Regression classifier and 96.78 % for Rotation Forest.

3.3 Results with GA

In the second test, we first used genetic algorithm (GA)-based feature selection to select the best attributes, and then, we used the same data mining techniques as in the previous section. Experimental results showed that highest classification performances are achieved when Rotation Forest is used as classifier. Therefore, the model proposed in this study is a model where GA is used for feature selection and Rotation Forest used for classification. GA–Rotation Forest structure is given in Fig. 1.

The experimental results obtained for WBC (Diagnostic) dataset are given in Tables 3, 4 and 5. To determine which of 30 attributes in WBC (Diagnostic) dataset is more

Table 1 Results for WBC diagnostic using different data mining techniques

	Logistic Regression	Decision Trees (C 4.5)	Random Forest	Bayes Net	ANN (MLP)	RBFN	SVM	Rotation Forest
MALIGN (%)	94.81	92.90	94.81	93.40	94.81	91.04	94.60	95.90
BENIGN (%)	98.59	93.56	96.92	96.08	97.76	96.08	98.30	98.30
AVERAGE (%)	97.19	93.32	96.13	95.08	96.66	94.20	96.89	97.41

Table 2 Results for WBC original using different data mining techniques

	Logistic Regression	Decision Trees (C 4.5)	Random Forest	Bayes Net	ANN (MLP)	RBFN	SVM	Rotation Forest
MALIGN (%)	95.00	95.40	95.00	97.90	96.70	95.80	96.20	96.20
BENIGN (%)	97.70	96.40	97.10	96.80	95.70	95.70	97.10	97.10
AVERAGE (%)	96.78	96.05	96.34	97.22	96.05	95.75	96.78	96.78

Table 3 Results for WBC diagnostic dataset with genetic algorithm feature selection

WBC (DIAGNOSTIC) data set genetic algorithm feature selection	Logistic Regression	Decision Trees (C 4.5)	Random Forest	Bayes Net	ANN (MLP)	RBFN	SVM	Rotation Forest
MALIGN (%)	98.65	93.90	94.80	94.59	97.30	91.00	97.3	98.65
BENIGN (%)	98.32	94.01	95.80	95.80	99.16	96.40	100	100.00
AVERAGE (%)	98.45	94.02	95.43	95.34	98.45	94.38	98.96	99.48

Table 4 AUC results for WBC diagnostic dataset with genetic algorithm feature selection

	Logistic Regression	Decision Trees (C 4.5)	Random Forest	Bayes Net	ANN (MLP)	RBFN	SVM	Rotation Forest
MALIGN	0.999	0.954	0.993	0.995	0.999	0.979	0.986	0.993
BENIGN	0.999	0.954	0.993	0.995	0.999	0.979	0.986	0.993
AVERAGE	0.999	0.954	0.993	0.995	0.999	0.979	0.986	0.993

Table 5 *F*-measure results for WBC diagnostic dataset with genetic algorithm feature selection

	Logistic Regression	Decision Trees (C 4.5)	Random Forest	Bayes Net	ANN (MLP)	RBFN	SVM	Rotation Forest
MALIGN	0.98	0.909	0.94	0.94	0.98	0.923	0.986	0.993
BENIGN	0.987	0.947	0.962	0.962	0.987	0.956	0.992	0.996
AVERAGE	0.984	0.932	0.953	0.953	0.984	0.944	0.990	0.995

important, GA is employed. Genetic algorithm-based feature selection gave us 14 attributes as important. These are $a_{1,2}$, $a_{2,1}$, $a_{3,1}$, $a_{3,2}$, $a_{4,2}$, $a_{5,2}$, $a_{6,3}$, $a_{7,1}$, $a_{7,3}$, $a_{8,2}$, $a_{8,3}$, $a_{9,1}$, $a_{9,3}$ and $a_{10,1}$. These 14 attributes noticeably differentiated between benign and malignant breast cells and tissues. As shown in the Tables 3, 4 and 5, total accuracy, AUC and *F*-measures achieved with the Rotation Forest classifier on WBC (Diagnostic) dataset were equal to 99.48 %, 0.993 and 0.995, respectively. These results were better than those achieved by the other classifiers. Indeed, the total accuracies are equal to 94.38 % for the RBFN classifier, 98.45 % for the Multilayer Perceptron (MLP), 94.02 % for the C4.5 Decision Tree, 95.34 % for the Random Forest, 95.34 % for the Bayes Net and 98.45 % for the Logistic Regression classifier. The AUCs were equal to 0.979 for the RBFN classifier, 0.999 for the Multilayer Perceptron (MLP), 0.954 for the C4.5 Decision Tree, 0.993 for the Random Forest, 0.995 for the Bayes Net and 0.999 for the Logistic Regression classifier. The *F*-measures were equal to 0.944 % for the RBFN classifier, 0.984 for the Multilayer Perceptron (MLP), 0.932 for the C4.5 Decision Tree, 0.953 for the Random Forest, 0.953 for the Bayes Net and 0.984 for the Logistic Regression classifier. Obtained results propose valuable information for the doctors and medical workers to pay much more attention to 14 attributes previously mentioned. This result confirms that Rotation Forest is superior as compared to other classifiers. Furthermore, it has reference classification accuracy in order to measure the ability of the suggested classification algorithm.

Applying genetic algorithm-based feature selection on WBC (Original) did not change results obtained without GA because WBC (Original) has very small number of attributes and GA-based feature selection gave as result of all these attributes.

4 Discussion

The performance demonstrated by the ensemble data mining techniques for breast cancer diagnosis lies in input variable choice and classification method selection. The parameters, which are most appropriate for breast cancer diagnosis, must be utilized as the inputs of the model. For this reason, GA is appropriate for classification of the WBC (Diagnostic) data in the breast cancer diagnosis. In the second test, where GA was applied, the highest obtained accuracy is 99.48 % with Rotation Forest classifier.

It can be observed from obtained performance results, two important observations can be obtained: (1) GA can correctly rank significant attributes since selected GA performs well in terms of classification performances, (2) Rotation Forest outperformed all other traditional linear and nonlinear classification methods by giving the highest accuracy. There are several results for superiority of Rotation Forest over other traditional methods employed in the literature for breast cancer classification. Rotation Forest is multiple classifier system, and because of this, it is more robust since it may all the time improve the performance results for individual classification methods and diversity in the groups. Every base classifier in Rotation Forest employs distinct subsets of WDB diagnostic and original datasets taking different features of these two datasets so that diversity can be achieved.

Accurate identification of breast cancer diagnosis is important for both diagnosis and treatment evaluation. The

Table 6 Comparison of accuracies with previous researches

References	Method	Classification accuracy (%)
Nauck and Kruse [35]	NEFCLASS	95.06
Goodman et al. [15]	Optimized-LVQ	96.70
Goodman et al. [15]	Big LVQ	96.80
Goodman et al. [15]	AIRS	97.20
Abonyi and Szeifert [2]	Supervised fuzzy clustering	95.57
Law et al. [28]	Mixture-based clustering	90.7
Gadaras and Mikhailov [14]	Fuzzy rule classification	96.08
Li and Liu [29]	SVM CPBK	93.26
Liu and Ren [31]	AFS	94.6
Cevikalp et al. [6]	SVM	97.6
Chang et al. [7]	CBFDT	98.4
Kim and Rattakorn [26]	Baseline	97.37
Fan et al. [11]	CBFDT	98.9
Zhao et al. [59]	GA with feature chromosome	99.0
Stoean and Stoean [49]	SVM and evolutionary algorithm	97.23
Koloseni et al. [27]	Differential evolution classifier	93.64
Astudillo and Oommenb [4]	Tree-based topology-oriented SOM	93.32
Tabakhi et al. [51]	Naïve Bayes	92.42
Saez et al. [43]	MI with k-NN	96.14
Chen et al. [8]	PTVPSO	98.44
Zheng et al. [60]	k-means and SVM	97.38
Lim and Chan [30]	BK with IVFS	95.26
Our method	GA with Rotation Forest	99.48

developed Rotation Forest model classifies WBC (Diagnostic) data using GA for feature selection with an accuracy of 99.48 %. This effect also resulted in an improvement of ROC area (AUC = 0.993), and *F*-measure (0.995) of Rotation Forest was higher than that of other classifiers. The Rotation Forest, as designated in this study, becomes as good as to other algorithms in breast cancer diagnosis. After applying different kinds of data mining techniques on our selected datasets, SVM with polynomial kernel also resulted in satisfactorily high accuracies of 98.96 %.

To summarize, the suggested expert system accomplished higher classification accuracy rate, decreased the number of attributes and obtained higher performance rate. Results obtained in this study prove that the suggested expert system is valuable in helping the doctors and other medical workers to make the correct breast cancer diagnosis and may demonstrate huge capacity in the area of medical decisions making.

To demonstrate the success of our approach, outcomes achieved in this research are compared with other results developed in the literature. To compare the breast cancer classification efficiency of the proposed model, numerous researches that employed the identical data but different classification techniques were used. For the sake of consistency with those researches, the same division of train–test dataset as explained previously was followed. To illustrate this, the classification performance with that of previous researches was compared. This is illustrated in Table 6. Most of these researches mentioned in Table 6 used the identical data division as our proposed model. For WBC (Original), both tests gave the same results because WBC (Original) with GA-based feature selection gave us all initial attributes (9 in totals) as important. It is worth of mentioning here that several systems evaluated on WBC (Original) dataset resulting in high classification performances are proposed in the literature. In [33], Multilayer Perceptron (AMMLP) algorithm was applied, and achieved classification accuracy was 99.26 %. In [9], rough set (RS)-based supporting vector machine classifier (RS_SVM) was proposed, and obtained classification accuracy was 96.87 %. In [3], LSA machine algorithm was applied, and obtained classification accuracy was near to 90 %. However, one of the main objectives of this study is to construct accurate classification system, but also to find the best-performing attribute selection algorithm. Therefore, WBC (Diagnostic) was employed to evaluate performances of system proposed in this study since it has more than threefold features when compared to WBC (Original).

5 Conclusion

A great number of researches have been conducted in the medical area to study medical disorders and find accurate diagnosis. Data mining techniques have been widely used for these purposes. In this study, we have proposed several different data mining methods with and without genetic algorithm-based feature selection to correctly classify medical data (data taken from Wisconsin Diagnostic Breast Cancer database). Random Forest and GA feature selection gave the highest accuracy of 99.48 %. In this research, one of the highest classification accuracies was obtained compared to all previous researches done in this field. We also achieved good classification accuracy by using SVM. Many powerful methods have been applied to WBC (Diagnostic) prediction problems. It is proved in this paper that instead of using complex methods based on strength classifiers to achieve good classification accuracies, an ensemble of more simple classifiers can be used as well, producing remarkable results. An ensemble of several methods offers us to use advantages of each method in order to achieve high classification accuracies for breast cancer diagnosis. We can use group of these rather simple methods to classify other medical diseases and to help doctors to make more precocious decisions in breast cancer diagnosis.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abbas HA (2001) An evolutionary artificial neural network approach for breast cancer diagnosis. Artif Intell Med 25: 265–281
2. Abonyi J, Szeifert F (2003) Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recogn Lett 24(14): 2195–2207
3. Albrecht AA, Lappas G, Vinterbo SA, Wong CK, Ohno-Machado L (2002) Two applications of the LSA machine. In: 9th international conference on neural information processing, pp 184–189
4. Astudillo CA, Oommenb BJ (2013) On achieving semi-supervised pattern recognition by utilizing tree-based SOMs. Pattern Recogn 46(1):293–304
5. Breiman L (2001) Random forests. Mach Learn 45:5–32
6. Cevikalp H, Triggs B, Yavuz HS, Kucuk Y, Kucuk M, Barkana A (2010) Large margin classifiers based on affine hulls. Neurocomputing 73:3160–3168
7. Chang PC, Fan CY, Dzan WY (2010) A CBR-based fuzzy decision tree approach for database classification. Expert Syst Appl 37:214–225
8. Chen HL, Yang B, Wang SJ, Liu DY, Li HZ, Wen BL (2014) Towards an optimal support vector machine classifier using a parallel particle swarm optimization strategy. Appl Math Comput 239:180–197
9. Chen H-L, Yang B, Liu J, Liu D-Y (2011) A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Syst Appl 38(7):9014–9022
10. Du K-L, Swamy M (2006) Neural networks in a softcomputing framework. Springer, New York
11. Fan CY, Chang PC, Lin JJ, Hsieh JC (2011) A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl Soft Comput 11:632–644
12. Fielding AH (2007) Cluster and classification techniques for the biosciences. Cambridge University Press, Cambridge
13. Fogel DB, Wasson EC, Boughton EM (1995) Evolving neural network for detecting breast cancer. Cancer Lett 96:49–53
14. Gadaras I, Mikhailov L (2009) An interpretable fuzzy rule-based classification methodology for medical diagnosis. Artif Intell Med 47(1):25–41
15. Goodman D, Boggess L, Watkins A (2002) Artificial immune system classification of multiple-class problems. In: Intelligent engineering systems through artificial neural networks: smart engineering system design: neural networks, fuzzy logic, evolutionary programming, complex systems and artificial life, vol 12, pp 179–184
16. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11(1):10–18
17. Hamilton HJ, Shan N, Cercone N (1996) RIAC: a rule induction algorithm based on approximate classification. In: International conference on engineering applications of neural networks
18. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36
19. Hassan MR, Begg R, Morsi Y, Lynch K (2006) HMM-fuzzy model for breast cancer diagnosis. In: 15th international conference on machines in medicine and biology
20. Hassan MR, Hossain MM, Begg RK, Ramamohanarao K, Morsi Y (2010) Breast-cancer identification using HMM-fuzzy approach. Comput Biol Med 40:240–251
21. Hassanien AE (2004) Rough set approach for attribute reduction and rule generation. J Am Soc Inf Sci Technol 55(11):954–962
22. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, California
23. Haykin S (2005) Neural networks: a comprehensive foundation. Pearson Education, New York
24. Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
25. Jerez-Aragones J, Gomez-Ruiz JA, Ramos-Jimenez G, Munoz-Perez J, Alba-Conejo E (2003) A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med 27(1):45–63
26. Kim SB, Rattakorn P (2011) Unsupervised feature selection using weighted principal components. Expert Syst Appl 38:5704–5710
27. Koloseni D, Lampinen J, Luukka P (2013) Differential evolution based nearest prototype classifier with optimized distance measures for the features in the data sets. Expert Syst Appl 40(10):4075–4082
28. Law M, Figueiredo M, Jain AK (2004) Simultaneous feature selection and clustering using mixture models. IEEE Trans Pattern Anal Mach Intell 26(9):1154–1166
29. Li DC, Liu CW (2010) A class possibility based kernel to increase classification accuracy for small data sets using support vector machines. Expert Syst Appl 37:3104–3110
30. Lim CK, Chan CS (2015) A weighted inference engine based on interval-valued fuzzy relational theory. Expert Syst Appl 42(7):3410–3419

31. Liu X, Ren Y (2010) Novel artificial intelligent techniques via AFS theory: feature selection, concept categorization and characteristic description. Appl Soft Comput 10:793–805
32. Maglogiannis I, Zafiropoulos E (2009) An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. Appl Intell 30(1):24–36
33. Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D (2011) WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst Appl 38(11):9573–9579
34. Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B (2001) An introduction to kernel-based learning algorithms. IEEE Trans Neural Netw 12(2):181–202
35. Nauck D, Kruse R (1999) Obtaining interpretable fuzzy classification rules from medical data. Artif Intell Med 16:149–169
36. Obuchowski NA (2003) Receiver operating characteristic curves and their use in radiology. Radiology 229:3–8
37. Pawlak Z (1982) Rough sets. Int J Parallel Prog 11(5):341–356
38. Pena-Reyes CA, Sipper M (1999) A fuzzy-genetic approach to breast cancer diagnosis. Artif Intell Med 17:131–155
39. Peng L, Yang B, Jiang J (2009) A novel feature selection approach for biomedical data classification. J Biomed Inform 179(1):809–819
40. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers, Los Altos
41. Quinlan JR (1996) Improved use of continuous attributes in C4.5. J Artif Intell Res 4:77–90
42. Rodriguez JJ, Kuncheva LI, Alonso CJ (2006) Rotation forest: a new classifier ensemble method. IEEE Trans Pattern Anal Mach Intell 28(10):1619–1630
43. Saez JA, Derrac J, Luengo J, Herrera F (2014) Statistical computation of feature weighting schemes through data estimation for nearest neighbor classifiers. Pattern Recogn 47(12):3941–3948
44. Sahan S, Polat K (2007) A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Comput Biol Med 3:415–423
45. Salzberg SL (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Disc 1:317–328
46. Sebe N, Cohen I, Garg A, Huang TS (2005) Machine learning in computer vision. Springer, New York
47. Setiono R (2000) Generating concise and accurate classification rules for breast cancer diagnosis. Artif Intell Med 18(3):205–217
48. 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
49. Stoean R, Stoean C (2013) Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. Expert Syst Appl 40:2677–2686
50. Swets JA (1979) ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 14:109–121
51. Tabakhi S, Moradi P, Akhlaghian F (2014) An unsupervised feature selection algorithm based on ant colony optimization. Eng Appl Artif Intell 32:112–123
52. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set (2012) Retrieved 15 Mar 2012, from UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data set: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
53. UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set (2012) Retrieved 16 Mar 2012, from UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data set: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)
54. Vapnik VN (2005) The nature of statistical learning theory. Spinger, New York
55. Wang CJ, Huang CL (2006) A GA-based feature selection and parameters optimization. Expert Syst Appl 31:231–240
56. Weka 3: Data Mining with Open Source Machine Learning Software in Java (2012) Retrieved 15 Mar 2012, from Weka 3—Data Mining with Open Source Machine Learning Software in Java. http://www.cs.waikato.ac.nz/~ml/weka/
57. WHO | Breast Cancer: Prevention and Control (2015) Retrieved 20 Jan 2015, from WHO | World Health Organization. http://www.who.int/cancer/detection/breastcancer/en/index1.html
58. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Elsevier, San Francisco
59. Zhao M, Fu C, Ji L, Tang K, Zhou M (2011) Feature selection and parameter optimization for support vector machines: a new approach based on genetic algorithm with feature chromosomes. Expert Syst Appl 38:5197–5204
60. Zheng B, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl 41(4): 1476–1482
61. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561–577