Neural Comput & Applic (2013) 23:2387–2403
DOI 10.1007/s00521-012-1196-7

ORIGINAL ARTICLE

Decision tree classifiers for automated medical diagnosis

Ahmad Taher Azar · Shereen M. El-Metwally

Received: 14 July 2012 / Accepted: 22 September 2012 / Published online: 7 October 2012

Abstract Decision support systems help physicians and also play an important role in medical decision-making. They are based on different models, and the best of them are providing an explanation together with an accurate, reliable and quick response. This paper presents a decision support tool for the detection of breast cancer based on three types of decision tree classifiers. They are single decision tree (SDT), boosted decision tree (BDT) and decision tree forest (DTF). Decision tree classification provides a rapid and effective method of categorizing data sets. Decision-making is performed in two stages: training the classifiers with features from Wisconsin breast cancer data set, and then testing. The performance of the proposed structure is evaluated in terms of accuracy, sensitivity, specificity, confusion matrix and receiver operating characteristic (ROC) curves. The results showed that the overall accuracies of SDT and BDT in the training phase achieved 97.07 % with 429 correct classifications and 98.83 % with 437 correct classifications, respectively. BDT performed better than SDT for all performance indices than SDT. Value of ROC and Matthews correlation coefficient (MCC) for BDT in the training phase achieved 0.99971 and 0.9746, respectively, which was superior to SDT classifier. During validation phase, DTF achieved 97.51 %, which was superior to SDT (95.75 %) and BDT (97.07 %) classifiers. Value of ROC and MCC for DTF achieved 0.99382 and 0.9462, respectively. BDT showed the best performance in terms of sensitivity, and SDT was the best only considering speed.

Keywords Computer-aided diagnosis (CAD) · Decision support systems (DSS) · Decision tree classification · Single decision tree · Boosted decision tree · Decision tree forest · *k*-fold cross-validation

1 Introduction

Breast cancer is the most commonly diagnosed cancer among women, and early diagnosis of breast cancer plays a leading role in reducing the mortality and improving the prognosis of this disease [47]. Worldwide, breast cancer comprises 22.9 % of all cancers in women [7]. In 2008, breast cancer caused 458,503 deaths worldwide (13.7 % of cancer deaths in women) [7]. In the west, earlier research has demonstrated that 1 in 9 women will develop breast cancer in their life, and this risk has been further stratified according to age, with patient up to 25 years, 1 in 15,000; up to age 30, 1 in 1900; and up to 40, 1 in 200 [61, 64]. In Egypt, breast cancer is the most common cancer among women, representing 18.9 % of total cancer cases (35.1 % in women and 2.2 % in men) among the Egypt National Cancer Institute (NCI) series of 10,556 patients during the year 2001 [27, 66], with an age-adjusted rate of 49.6 per 100,000 population. Breast cancer is a malignant tumor that develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division [56]. The cells can invade nearby tissue and can spread through the bloodstream and lymphatic system (lymph nodes) to other parts of the body. Early breast cancer usually does not cause pain and may exhibit no noticeable symptoms. As the cancer progresses, signs and

A. T. Azar (✉)
Faculty of Engineering, Misr University for Science
and Technology (MUST), 6th of October City, Egypt
e-mail: ahmad_t_azar@ieee.org

S. M. El-Metwally
Systems and Biomedical Engineering Department,
Cairo University, Giza, Egypt
e-mail: shereen.elmetwally@k-space.org

symptoms can include a lump or thickening in or near the breast; a change in the size or shape of the breast; nipple discharge, tenderness, or retraction (turning inward); and skin irritation, dimpling, or scaliness. These changes can occur as part of many different conditions, however. Currently, there are no methods to prevent breast cancer, which is why early detection represents a very important factor in cancer treatment and allows reaching a high survival rate. There is no denying the fact that after more than 40 years of experience [75], systematic screening with mammograms reduces breast cancer mortality for women over 40 years of age. Mammography is the most sensitive technique currently available for the detection of non-palpable lesions and is therefore the method of choice [43, 79]. Mammography has been shown to reduce breast cancer mortality by 18–29 % [22, 68, 88]. Two types of mammography are known: screen-film mammography (SFM) (also known as conventional mammography) and digital mammography (DM). Screening mammograms are used to look for breast disease in women who are asymptomatic, that is, those who appear to have no breast problems. Screening mammograms usually take 2 views (X-ray pictures taken from different angles) of each breast. SFM has its known limitations [38]. Approximately 10–20 % of palpable breast cancers are not visible on mammograms mainly as a result of insufficient contrast between normal and abnormal breast tissues. Only 5–40 % of lesions recommended for biopsy proves to be malignant [13, 14, 50]. In addition, technical factors such as limited exposure range, film processing and management, and film artifacts further limit conventional mammography. Furthermore, film serves as the medium for image acquisition, display and storage with conventional mammography systems, which means there is no opportunity to intervene in each of these processes to improve image quality. Hence, DM was introduced as an alternative diagnostic technique in order to overcome the problems of SFM [6, 21, 80, 86, 87]. DM incorporates a new technique called computer-aided diagnosis (CAD) which employs the tools of image processing for image enhancement and diagnosis [5, 16, 25, 26, 41, 53]. Digital mammograms are recorded and stored on a computer [44, 63]. It reduces the number of patients recalled for additional mammograms, reduces the number of false-positive breast biopsy results, and can potentially enable detection of breast cancer at an earlier stage [52, 65, 74, 78]. This new technology may also be more effective in detecting cancer in women with radiodense fibroglandular tissue, which is less effectively imaged by conventional screen-film mammograph.

1.1 Problem statement

Subjectivity among screening radiologists in the interpretation of mammograms results in a high percentage of misdiagnosed cancer cases and a high percentage of missed cancer cases. This subjectivity is the end product of several factors including radiologists' fatigue, incompetence and lack of training to name a few. Digital mammograms are among the most difficult medical images to be read due to their low contrast and differences in the types of tissues. Important visual clues of breast cancer include preliminary signs of masses and calcification clusters. Also, tumors are of different shapes and some of them have the characteristics of the normal tissue. All these reasons make the decisions that are made on such images more difficult. Unfortunately, in the early stages of breast cancer, these signs are very subtle and varied in appearance, making diagnosis difficult, challenging even for specialists. A false-positive detection may cause an unnecessary biopsy. Statistics show that only 20–30 % of breast biopsy cases are proved cancerous. In a false-negative detection, an actual tumor remains undetected that could lead to higher costs or even to the cost of a human life. Here is the trade-off that appears in developing a classification system that could directly affect human life. In addition, the tumors existing are of different types. Nowadays, medical decision support systems have become the cornerstones of medical technology. Decision support systems help physicians to diagnose the type of diseases in case of uncertain illnesses, by learning the basic characteristics which are used in the decision-making processes of diseases. For this purpose, many studies have been conducted by constructing various and numerous models.

The use of data mining techniques like decision tree (DT) has shown great potential in this field. With the involvement of soft computing, the pattern matching, classification and detection of algorithms which have direct applications in many medical problems have become much easier to be implemented and diagnosed. Hence, this paper tries to find out the efficiency of the single decision tree (SDT), boosted decision tree (BDT) and decision tree forest (DTF) techniques in the breast cancer classification and detection purposes. These systems classify the digital mammograms in two categories: normal and abnormal. The normal ones are those characterizing a healthy patient. The abnormal ones include both benign cases, representing mammograms showing a tumor that is not formed by cancerous cells, and malignant cases, those mammograms taken from patients with cancerous tumors. Benign tumors are not considered cancerous: their cells are close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous. Left unchecked, malignant cells eventually can spread beyond the original tumor to other parts of the body.

The rest of this paper is organized as follows. Section 2 depicts the related work. Subjects and methods are presented in Sect. 3. The description of BDT classifier is presented in Sect. 4. The performances analysis of the proposed classifier systems is introduced in Sect. 5. Finally, the conclusion is presented in Sect. 6.

2 Related work

A Decision Tree is one of the most popular classification algorithms in current use in Data Mining and Machine Learning [9]. Data mining technique has become a popular research tool for medical researchers who seek to identify and exploit patterns and relationships among large number of variables and be able to predict the outcome of a disease using the historical cases stored within data sets [12, 49, 71]. For many problems of classification where large data sets are used and the information contained is complex and may contain errors, decision trees provide a useful solution. A definition of a decision tree was given in Russell and Norvig [72], as a construct which takes as input an object or situation described by a set of properties and outputs a yes/no decision. In terms of ability, decision trees are a rapid and effective method of classifying data set entries and can provide good decision support capabilities. Mehta et al. [60] emphasized the importance of classification in mining of large data sets and also discussed the wide range of uses that classification can be put to in economic, medical and scientific situations. Decision trees classifiers can perform automatic feature selection and complexity reduction, and their tree structure provides easily understandable and interpretable information regarding the predictive or generation ability of the classification. The decision tree is then constructed by recursively partitioning a data set into purer, more homogenous subsets on the basis of a set of tests applied to one or more attribute values at each branch or node in the tree.

Dietterich [24] discussed improvement to decision tree design methods up to the end of the 1980s and provided a good background to these and more classical decision tree development methods. Lim et al. [54] compared several decision trees, statistical and neural network methods on a variety of data sets. Both of these works showed that a wide range of speed and accuracies can be obtained from the different decision tree algorithms commonly used and that the effectiveness of different algorithms varies greatly with the data set. One of the most common methods of inducting decision tree structure is C4.5, designed by Quinlan [69], which also deals with data sets in which variables are continuous or integer, or where there is missing data. This suite of algorithms provides a large amount of information concerning the data which has been manipulated and the decision trees developed to handle that data. However, the results provided by Llora and Garrell [55] of an application of C4.5 to training data sets available on the internet allow comparison of novel methods with this benchmark system. Two main processes are common to all decision tree development methods. The first is the growth of the tree to enable it to accurately categorize the data set being used, and the second is the pruning stage, whereby superfluous nodes and branches are removed. Bradford et al. [8] used a method of pruning classifier trees that minimized misclassification errors, while Ferri et al. [30] analyzed several pruning algorithms for estimation trees, leading to the determination of the best algorithm to be applied in a specific situation. Friedman et al. [34] discussed the problems of constructing decision trees, the main one of which is what question to ask at each node in order to divide and conquer the data set optimally. They showed that this problem becomes harder as one deals with larger and larger data sets, and with more and more variables. Fulton et al. [36], in a related analysis of the problems of generating decision trees capable of dealing with large, complex data sets, showed that it is simpler to construct decision trees that can deal with a small subset of the original data set. Alsabti et al. [1] discussed the problems of scaling decision trees up to large data sets, with the loss of accuracy that often occurs as a result. Garofalakis et al. [37] discussed methods for constructing decision trees with user-defined constraints such as size limits or accuracy. These limits are often important for users to be able to understand or use the data sets adequately or to avoid overfitting the decision tree to the data that is available. Ankerst et al. [3] used an interactive approach, with the user updating the decision tree through the use of a visualization of the training data. This method resulted in a more intuitive decision tree and one that the user was capable of implementing according to their existing knowledge about the system in question. Kuo et al. [46] described a novel computer-aided diagnosis (CADx) system using data mining with decision tree for classification of breast tumor to increase the levels of diagnostic confidence and to provide the immediate second opinion for physicians. In this study, the accuracy of data mining with decision tree for classifying malignancies was 96 %, the sensitivity was 93.33 %, the specificity was 96.67 %, the positive predictive value (PPV) was 93.33 %, and negative predictive value (NPV) was 96.67 % for the proposed CADx system. Jerez-Aragonés et al. [48] presented a decision support tool for the prognosis of breast cancer relapse based on specific topologies of neural networks for different time intervals during the follow-up time of the patients, considering the events occurring in different intervals as different problems, and decision trees, in understanding the underlying relationships in breast cancer data, for selecting the most important prognostic factors corresponding to every time interval. Delen et al. [20] used two popular data mining algorithms (artificial neural networks and decision trees) along with a most commonly used statistical method (logistic regression) to develop prediction models for breast cancer survivability using a large data set (more than 200,000 cases). The results indicated that the decision tree (C5) is the best predictor with 93.6 % accuracy on the holdout sample (this prediction accuracy is better than any reported in the literature), artificial neural networks came out to be the second with 91.2 % accuracy, and the logistic regression models came out to be the worst of the three with 89.2 % accuracy. Ibrahim et al. [42] developed the decision tree method for competing risks survival time breast cancer data based on proportional hazards for subdistribution of competing risks. The results show that the proposed method performs quite well in the identification of correct data structure. However, the performance decreases as the censoring percentage increases. Endo et al. [28] presented optimal models to evaluate the predictions of 5-year survival rate for breast cancer patients from the perspectives of accuracy, sensitivity and specificity. The accuracy was 85.8 ± 0.2, 84.5 ± 1.4, 83.9 ± 0.2, 83.9 ± 0.2, 84.2 ± 0.2, 82.3 ± 0.2, 85.6 ± 0.2 % for the logistic regression model, ANN, naive Bayes, Bayes net, decision trees with naive Bayes, decision trees (ID3) and decision trees (J48), respectively. Logistic regression model showed the highest accuracy. Decision Trees (J48) had the highest sensitivity, and artificial neural network showed the highest specificity. Ture et al. [83] investigated different decision tree decision tree methods (CART, CHAID, QUEST, C4.5 and ID3) used additionally to the well-known Kaplan–Meier estimates to investigate their predictive power in determining recurrence-free survival of breast cancer patients. In this study, C4.5 showed a better degree of separation with calculated sensitivity, specificity and predictive rates of 43.8, 91 and 77.4 %, respectively. Lee and Yang [51] proposed a case-based computer-assisted entropy-based feature extraction and decision tree induction protocol for breast cancer diagnosis using thermograph images. Both image processing and data mining techniques were combined to find the diagnosis rules. This algorithm was proven to successfully condense positive rules with very few tree paths in a relatively high percentage of cancer patients. Shanthi and Bhaskaran [76] presented a novel technique to detect and classify the breast cancers based on Intuitionistic fuzzy C-means and decision tree approach. The classification performance of each model is evaluated using three statistical measures: classification accuracy, precision and recall. The verification result showed that the proposed algorithm gave better results. In another work [29], a hybrid model integrating a case-based data clustering method and a fuzzy decision tree (CBFDT) was developed for medical data classification. Initially, a case-based clustering method was applied to preprocess the data set to obtain a more homogeneous data within each cluster; then, a fuzzy decision tree was applied to the data in each cluster together with genetic algorithms (GAs) in order to construct a decision-making system based on the selected features and diseases identified. The average forecasting accuracy obtained for breast cancer using CBFDT model was 98.4 %. Štajduhar and Dalbelo-Bašic [81] proposed a pre-processing method for uncensoring censored survival data to be used with various machine learning algorithms. This is done by pre-assigning censored instances a positive or negative outcome according to their features and observation time. The proposed procedure calculates the goodness of fit of each censored instance to both the distribution of positives and the spoiled distribution of negatives in the entire data set and relabels that instance accordingly. A thorough empirical testing using the naïve Bayes classifier and decision trees was performed in a simulation study and on two real-world medical data sets: the Wisconsin prognostic breast cancer data set and the malignant skin melanoma data set. This method provided good results especially when applied to heavily censored data.

3 Classifier model description

This section provides detailed description for three common DT classifiers; SDT, BDT and DTF.

3.1 Single decision tree (SDT)

A decision tree is a classification tool that uses a tree-like graph structure. The feature vector is split into unique regions, corresponding to the classes, in a sequential manner [9, 15]. Presenting a feature vector, the region to which the feature vector will be assigned, is searched via a sequence of decisions along a path of nodes of an appropriately constructed tree. Given an input feature vector X, $X \in R^n$, a binary decision tree is built with the following steps.

3.1.1 Binary questions

A set of binary (true/false) questions are asked, of the form: $X \subset A$, $A \subseteq X$ for categorical queries, or $X > C_j$ where C_j is a proper threshold value. For each feature, every possible value of the threshold C_j defines a specific split of the subset X.

3.1.2 Splitting criterion

Every binary split of a node generates two descendant nodes. A criterion for tree splitting t is based on a node impurity function $I(t)$. A variety of node impurity measures is defined, as shown in Eq. (1).

$$I(t) = \varphi(P(\omega_1|t), P(\omega_2|t), \ldots, P(\omega_M|t)) \quad (1)$$

where φ is an arbitrary function and $P(\omega_i|t)$ denotes the probability that a vector X_t belongs to the class ω_i: $i = 1, 2, \ldots, M$. A usual choice for φ is the entropy function from Shannon's Information Theory, as shown in Eq. (2).

$$I(t) = -\sum_{i=1}^{M} P(\omega_i|t)\log_2 P(\omega_i|t) \quad (2)$$

where $\log_2$ is the logarithm with base 2 and M is the total number of classes. The decrease in node impurity is defined as shown in Eq. (3).

$$\Delta I(t) = I(t) - a_R I(t) - a_L I(t) \quad (3)$$

with a_R, a_L the proportions of the samples in node t, assigned to the right node t_R and the left node t_L, respectively. The task now reduces to one of adopting, from the set of candidate questions, the one that performs the split, leading to the highest decrease in impurity according to Eq. (3).

3.1.3 Stop-splitting rule

A simple stop-splitting rule has been adopted, when the maximum value of $\Delta I(t)$, over all possible splits, is less than a threshold T; then, splitting is stopped. Other alternatives are to stop splitting either when the cardinality of the subset X_t is small enough or when X_t is pure, in the sense that all points in it belong to a single class [82]. A critical factor in designing a decision tree is its size: it must be large enough, but not too large; otherwise it tends to learn the particular details of the training set and exhibits poor generalization performance. Experience has shown that the use of a threshold value, for the impurity decrease as stop-splitting rule, does not always lead to optimum tree size. Many times, it stops the tree growing either too early or too late. The most commonly used approach is to grow a tree up to a large size first and then prune its nodes according to a pruning criterion [62]. Tree size has a significant importance to the present study since it is dealing with a two-class problem. Trees too large or too small will incorrectly represent the feature vectors.

3.1.4 Class assignment rule

Once splitting is stopped, a node is declared to be a leaf, and a class label ω_j is given using the majority rule:

$$j = \arg\max_i P(\omega_i|t) \quad (4)$$

In other words, a leaf t of the tree is assigned to the class where the majority of the vectors X_t belong to.

3.2 Boosted decision tree (BDT)

The normal recursive learning procedure for the tree-structured classifier is to split the source set of the parent node into subsets for child nodes based on the parent node's classification test. The potential problem of a tree-structured classifier is that the node of a tree-structured classifier loses distribution information from the entire data set and is very susceptible to over-fitting. The addition of boosting to a decision tree as a means to improve prediction accuracy is known as adaptive boosting and was proposed by Freund and Schapire [32, 33], Quinlan [69, 70] and [4]. Adaptive boosting is based on a learning algorithm of a decision tree classifier over a repeated series of trials: $t = 1, \ldots, T$. One possible approach is to select a best weight and tree structure from the distribution of weights over the training set.

For a training set $(x_i, y_i) \cdots (x_m, y_m)$, x_i belongs to X and y_i belongs to label set Y. This generates the weak hypothesis $h_t(i)$: $X \rightarrow \{-1, +1\}$, as $D_t(i)$ is the weight distribution on training instance i at trial t.

The error of the hypothesis is given as

$$\varepsilon_t = Pr_{i\sim D_t}[h_t(x_i) \neq y_i] = \sum_{i:h_t(x_i)\neq y_i} D_t(i) \quad (5)$$

where the $Pr_{i\sim D_t}[.]$ is the probability with respect to the distribution $D_t(i)$ when the weak learner was trained. The parameter of weight will be chosen as follows:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \quad (6)$$

where α_t increases when ε_t decreases. After updating $D_t(i)$, the final hypothesis H measures the confidence in the boosting prediction and is given as

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \quad (7)$$

The final hypothesis H is a majority vote in $t = 1, \ldots, T$, where α_t is the weight of h_t.

Many classifiers are constructed from a single training data set for boosting. Each classifier is constructed to form a SDT structure or a rule set using the training data. New classifications are based on votes from many classifiers, while the predicted and final classes are decided from the votes. The first step of this boosting procedure is to build a SDT structure or a rule set from the training data. This classifier will usually contribute to the errors for some cases in the data. The first decision tree structure generates the wrong class for some cases in the training data. Next, the second classifier is constructed with greater attention to correct classification. The second classifier will consequently be different from the first classifier. The third classifier construction step is comparatively even more focused, although it also will make mistakes in some cases. By setting the boost trial number in advance, the boosting process continues iteratively by updating $D_t(i)$. The final step of the boosting process is stopped when the most recent classifier is either extremely accurate or inaccurate.

3.3 Decision tree forest (DTF)

Random forests (RF) is one of the most successful ensemble learning techniques which have been proven to be very popular and powerful techniques in the pattern recognition and machine learning for high-dimensional classification and skewed problems [11, 59]. A drawback associated with tree classifiers is their high variance. In practice, it is common for a small change in the training data set to result in a very different tree. The reason for this lies in the hierarchical nature of the tree classifiers. An error that occurs in a node close to the root of the tree propagates all the way to the leaves. In order to make tree classification more stable, a decision forest methodology has been invented. The methodology was initially proposed by Ho [39, 40], Amit and Geman [2] and later by Breiman [11], in an integrated form (as "RF"). A decision forest is an ensemble of decision trees. It can be seen as one classifier which contains several classification methods or one method but various parameters of work. A new input vector is classified by each individual tree of the forest. Each tree yields a certain classification result. The decision forest chooses the classification which has the most votes over all the trees in the forest. The RF methodology contains Breiman's "bagging" idea and Ho's "random selection features". Bagging, which stands for "'bootstrap aggregation", is a type of ensemble learning introduced by Breiman [10], in order to improve the accuracy of a weak classifier by creating a set of classifiers. In this method, each classifier's training set is generated by randomly drawing N examples, with replacement, with N, the size of the original training set. The learning system generates a classifier from the sample and aggregates all the classifiers generated from the different trial to form the final classifier. To classify an instance, every classifier records a vote for the class to which it belongs and the instance is labeled as a member of the class with the most votes. In case that more than one class jointly receives the maximum number of votes, then, the winner is selected at random. Every tree in the ensemble is grown on an independently drawn bootstrap replica of input data. Observations not included in this replica are "out-of-bag" for this tree [10]. The prediction error of the bagged ensemble is estimated by computing predictions for each tree on its out-of-bag observations, averaging these predictions over the entire ensemble for each observation and then comparing the predicted out-of-bag response with the true value at this observation. Bagging works by reducing variance of an unbiased base learner, such as a decision tree. This technique tends to improve the predictive power of the ensemble, as the random selection of features reduces the correlation between trees in the ensemble. The RF algorithm is summarized as shown in Fig. 2.

4 Subjects and methods

In this comparative study, the medical data related to breast cancer is considered. This database was obtained from the university of Wisconsin hospital, Madison from Dr. William H. Wolberg [57, 58, 85]. Wisconsin Breast Cancer database can be used to predict the severity (benign or malignant) of a mammographic mass lesion from and the patient attributes. It consists of 9 input real variables, 2 output classes and 699 cases, of which 458 are diagnosed as benign (class 1) and the remaining 241 are known to be malignant (class 2) that have been identified on full field digital mammograms. After the 16 instances are removed from the data set due to missing values, there are 683 instances, of which 444 are benign and the remaining 239 are diagnosed as malignant. The nine measurements taken from fine needle aspirates from human breast tissues correspond to cytological characteristics of benign or of malignant sample. There are totally 10 attributes (1 class and 9 numeric features) detailed in Table 1. Each of these nine attributes of the fine needle aspirates was graded 1–10 at the time of sample collection, with 1 being the closest to benign and 10 the most anaplastic.

4.1 Data analysis

The classification process starts by obtaining a data set (input–output data pairs) and dividing it into a training set and validation data set. The commercially available software package, namely decision tree and regression (DTREG) [77] was used to implement decision trees to predict the class proportions. The ideal split would divide a group into two child groups in such a way so that all of the rows in the left child have the same value on the target variable and all of the rows in the right group have the same target value—but different from the left group. Such a perfect split is possible only if the rows in the node being split have only two possible values on the target variable. Unfortunately, perfect splits do not occur often, so it is necessary to evaluate and compare the quality of imperfect splits [77]. Various criteria have been proposed for evaluating splits, but they all have the same basic goal which is to favor homogeneity within each child node and heterogeneity between the child nodes. The heterogeneity or

Table 1 Wisconsin breast cancer data description of attributes

Attribute number	Attribute description	Values of attributes	Mean	SD
1	Clump thickness	1–10	4.42	2.82
2	Uniformity of cell size	1–10	3.13	3.05
3	Uniformity of cell shape	1–10	3.20	2.97
4	Marginal adhesion	1–10	2.80	2.86
5	Single epithelial cell size	1–10	3.21	2.21
6	Bare nuclei	1–10	3.46	3.64
7	Bland chromatin	1–10	3.43	2.44
8	Normal nucleoli	1–10	2.87	3.05
9	Mitoses	1–10	1.59	1.71

N 699 observations, 241 malignant and 458 benign

dispersion of target categories within a node is called the "node impurity". The goal of splitting is to produce child nodes with minimum impurity. The impurity of every node is calculated by examining the distribution of categories of the target variable for the rows in the group. A "pure" node, where all rows have the same value of the target variable, has an impurity value of zero. When a potential split is evaluated, the probability weighted average of the impurities of the two child nodes is subtracted from the impurity of the parent node. This reduction in impurity is called the improvement of the split. The split with the greatest improvement is the one used [77]. DTREG provides two methods for evaluating the quality of splits when building classification trees: (1) Gini and (2) entropy. Only one method is provided when building regression trees and that is minimum variance within nodes. The minimum variance/least squares criteria are essential the same criteria used by traditional, numeric regression analysis (i.e., line and function fitting). One of the classic problems in building decision trees is the question of how large a tree to build. Early programs such as AID (automatic interaction detection) used stopping criteria such as the improvement in splits to decide when to stop. This is known as forward pruning. But analysis of trees generated by these programs showed that they often were not of the optimal size. DTREG does not use its stopping criteria as the primary means for deciding how large a tree should be. Instead, it uses relaxed stopping criteria and builds an overly large tree. It then analyzes the tree and prunes it back to the optimal size. This is known as backward pruning. Backward pruning requires significantly more calculations than forward pruning, but the optimal tree sizes are much more accurately calculated [77]. This pruning process is usually done with a fresh "test data set" by subsequently removing branches of the large tree getting simpler and simpler trees and finally the "optimal" tree size. However, this tree may be only marginally better than an even smaller one with just a slightly larger error value [77]. In particular, for relatively small data sets ($n < 1{,}000$; [18], V-fold cross-validation [9, 19] is an approved and often applied pruning procedure (e.g. [17, 84]. Thereby, the data are randomly divided into more or less equal portions. Using the very effective $V = 10$ cross-validation [77], repeatedly nine out of these ten portions are used as "training data" to generate the model while the tenth is used as test data (validation data) to evaluate the model repeated until all test data sets have been used. In DT modeling approach, tenfold cross-validation was used and pruned back the trees to minimum cross-validated relative error [23]. Further, we searched if certain variables were masked by primary splitters using percentage variable importance. These scores were given as relative values to the most important variable scaled to 100 %. Obviously, the feature that is selected as a splitter earlier in the tree is more important than others in the decision-making process. Score "0" is considered as the least important one and can be rejected from the regression tree. Score "100" is considered as the highest score, which indicates that this feature is the most significant one.

5 Results and discussion

In this section, the performance and the comparison of the three DT classifiers presented in Sect. 3 are demonstrated. The simulations were performed by using an Intel (R) Core (TM) i3 CPU 530–2.93 GHz personal computer and a Microsoft Windows 7 64-bit operating system.

5.1 Performance analysis

The performance of each DT classifier was evaluated by using performance indices such as sensitivity, specificity, PPV, NPV, accuracy and F measure. Some of the main formulations are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100\,\% \tag{8}$$

$$\text{Senstivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\,\% \tag{9}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \times 100\,\% \tag{10}$$

PPV is the proportion of positive test results that are true positives (such as correct diagnoses). It is a critical measure of the performance of a diagnostic method, as it reflects the probability that a positive test reflects the underlying condition being tested for. The NPV is a summary statistic used to describe the performance of a diagnostic testing procedure. It is defined as the proportion of subjects with a negative test result who are correctly diagnosed. A high NPV for a given test means that when the test yields a negative result, it is most likely correct in its assessment.

$$\mathrm{PPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\,\% \tag{11}$$

$$\mathrm{NPR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \times 100\,\% \tag{12}$$

In Eqs. (8)–(12), TP is the number of true positives (benign breast tumor); FN is the number of false negatives (malignant breast tumor); TN is the number of true negatives; and FP is the number of false positives. They are defined as a confusion matrix. Receiver operating characteristic, ROC, curves are also used to evaluate the performance of a diagnostic test [45, 67]. This method consists of a lot of information for comprehensibility and improving classifiers' performance. The ROC curve plots the true-positive rate as a function of the false-positive rate. It is parameterized by the probability threshold values. The true-positive rate represents the fraction of positive cases that are correctly classified by the model. The false-positive rate represents the fraction of negative cases that are incorrectly classified as positive. Therefore, it provides a trade-off between sensitivity and specificity. The advantages of ROC analysis are the robust description of the network's predictive ability and an easy way to change the existence network based on differential cost of misclassification and varying prior probabilities of class occurrences. However, it requires visual inspection because the best classifiers are hard to recognize when the curves are mixed.

5.2 Training phase of classifiers

5.2.1 Single decision tree (SDT) analysis

Ten test trees are built for SDT model using the reduced data sets with the unused 10 % in each case then run through each test tree and the classification error for that tree computed. SDT parameters are summarized in Table 2. The Gini criterion is used here for evaluating the quality of splits. This default rule often works well across a broad range of problems. Gini has a tendency to generate trees that include some rather small nodes highly concentrated with the class of interest. Figure 1 shows an example of the SDT created for breast cancer classes using the training data set. As an illustration, the operation at node 2 is described in some detail. At this node, the splitting rule is "uniformity of cell size ≤2.5". Out of 683 training samples at node 2, 418 training samples satisfy this rule, that is, $N = 418$ at node 2. The misclassification accuracy at this node is 2.87 % as shown in Fig. 2.

Once the tree has been built, the records in the learning data set can be run through the tree to see how well the tree fits the data. The rate of classification errors measured when running the learning data set through a tree built using that data set is known as the "resubstitution cost" for the tree [77] (it is called resubstitution because the same data is rerun through the tree). For the learning data set, the accuracy of the fit always improves (resubstitution cost decreases) as the tree is grown larger. It is always possible to grow a sufficiently large tree to provide 100 % accuracy in predicting the learning data set. In an extreme case, the

Table 2 Training parameters of SDT model

Parameter type	Value
Minimum rows in a node	5
Minimum tree levels	10
Maximum splitting levels	10
Splitting algorithm	Gini
Maximum categories for continuous predictors	200
Tree pruning and validation method	*k*-fold cross-validation
Number of cross-validation folds	10
Tree pruning criterion	Minimum cost complexity

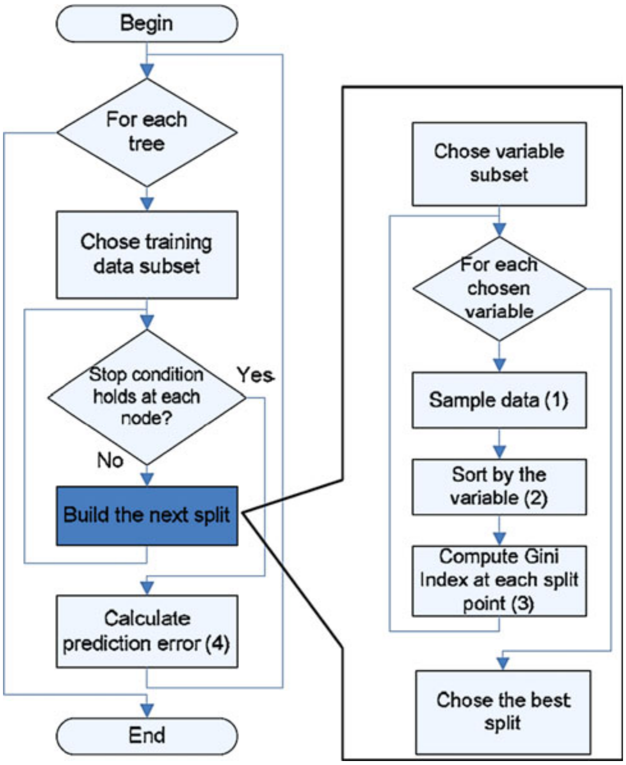


Fig. 1 Random forest algorithm *flow chart*

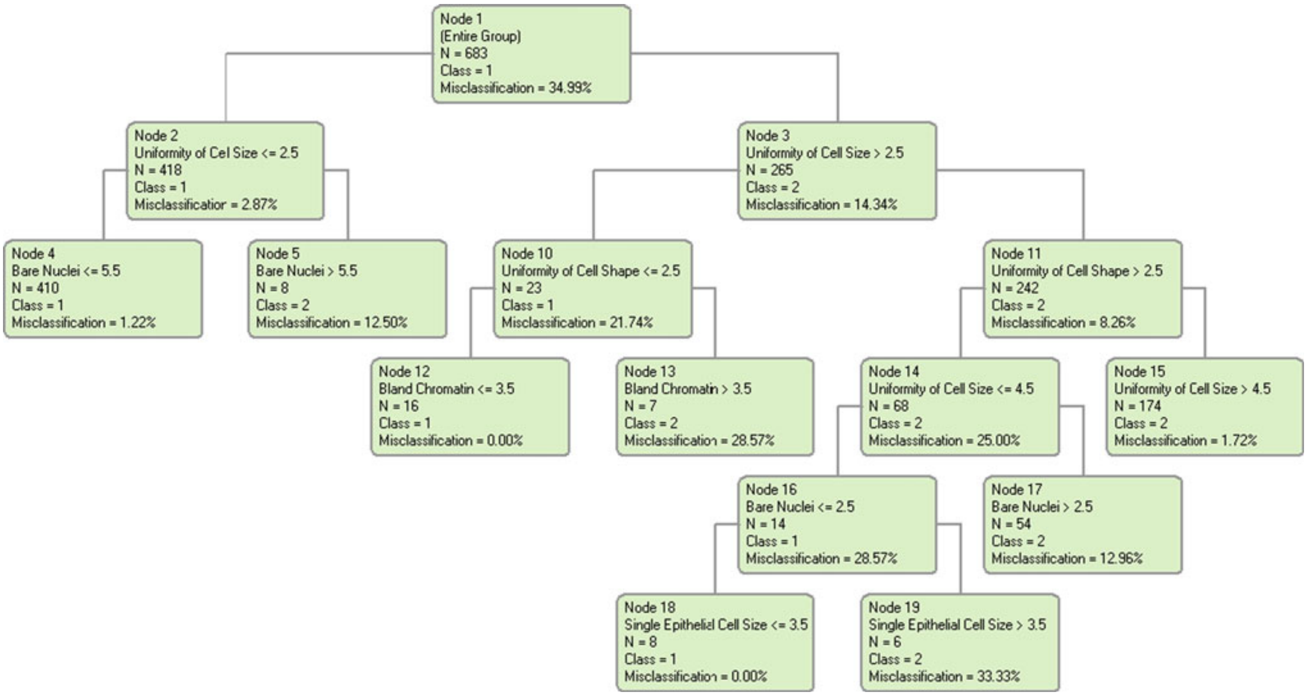

Node 1
(Entire Group)
N = 683
Class = 1
Misclassification = 34.99%
Node 2
Uniformity of Cel Size <= 2.5
N = 418
Class = 1
Misclassificatior = 2.87%
Node 3
Uniformity of Cell Size > 2.5
N = 265
Class = 2
Misclassification = 14.34%
Node 4
Bare Nuclei <= 5.5
N = 410
Class = 1
Misclassification = 1.22%
Node 5
Bare Nuclei > 5.5
N = 8
Class = 2
Misclassification = 12.50%
Node 10
Uniformity of Cell Shape <= 2.5
N = 23
Class = 1
Misclassification = 21.74%
Node 11
Uniformity of Cell Shape > 2.5
N = 242
Class = 2
Misclassification = 8.26%
Node 12
Bland Chromatin <= 3.5
N = 16
Class = 1
Misclassification = 0.00%
Node 13
Bland Chromatin > 3.5
N = 7
Class = 2
Misclassification = 28.57%
Node 14
Uniformity of Cell Size <= 4.5
N = 68
Class = 2
Misclassification = 25.00%
Node 15
Uniformity of Cell Size > 4.5
N = 174
Class = 2
Misclassification = 1.72%
Node 16
Bare Nuclei <= 2.5
N = 14
Class = 1
Misclassification = 28.57%
Node 17
Bare Nuclei > 2.5
N = 54
Class = 2
Misclassification = 12.96%
Node 18
Single Epithelial Cell Size <= 3.5
N = 8
Class = 1
Misclassification = 0.00%
Node 19
Single Epithelial Cell Size > 3.5
N = 6
Class = 2
Misclassification = 33.33%

Fig. 2 Single decision tree model for breast cancer diagnosis

tree might be grown so large that every row of the learning data set ended up in its own terminal node. Obviously, with such a tree, an exactly correct value of the target value for every row could be predicted. The primary goal of the pruning process is to generate the optimal size tree that can be generalized to other data beyond the learning data set. As shown in Fig. 3, the maximum depth of the tree is 8 nodes with relative error value of 0.1212 and a standard error of 0.0119. For single-tree models, the model size is the number of terminal nodes in the tree. Therefore, the tree will be pruned from 9 to 8 nodes.

5.2.2 Boosted decision tree (BDT) analysis

BDT models often have a degree of accuracy that cannot be obtained using a large, single-tree model. They can handle hundreds or thousands of potential predictor variables. The number of terminal nodes in a tree is equal to 2^k where k is the number of levels [77]. Because many trees contribute to the model generated by BDT, usually, it is not necessary for individual trees to be very large. The depth should be at least as large as the number of variable interactions. BDT parameters are summarized in Table 3.

To avoid overfitting problems during modeling process, k-fold cross-validation was used for better reliability of test results [31]. In k-fold cross-validation, the original sample is randomly partitioned into k subsamples. A single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated k times (the “folds”), with each of the k subsamples used exactly once as the validation data. The average of the k results gives the validation accuracy of the algorithm [23]. The advantages of k-fold cross-validation are that the impact of data dependency is minimized and the reliability of the results can be improved [73]. Research has shown [35] that the predictive accuracy of a BDT can be improved by apply a weighting coefficient that is <1 ($0 < v < 1$) to each tree as the series is constructed. This coefficient is called the “shrinkage factor”. The effect is to retard the learning rate of the series, so the series has to be longer to compensate for the shrinkage but its accuracy is better [77]. Tests have shown that small shrinkage factors in the range of 0.1 yield dramatic improvements over BDT series built with no shrinkage ($v = 1$). The trade-off in using a small shrinkage factor is that the BDT series is longer and the computational time increases. In this study, shrinking factor was adjusted to 0.05. BDT series are less prone to problems with over-fitting than SDT models, but they can benefit from validation and pruning to the optimal size to minimize the error on a test data set. In the case of a BDT series, “pruning” consists of truncating the series to the optimal number of trees. The software [77] will not prune the series to a length shorter than the specified value; therefore, the specified minimum number of trees was set at 10 trees. Some BDT series have erratic behavior with small numbers of trees. Sometimes, the error rate is very low with series consisting of one or two trees; then, the

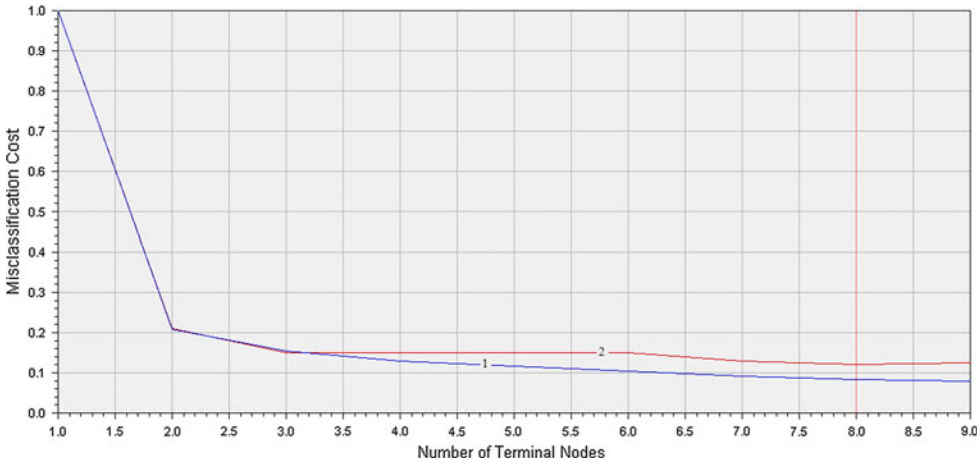


Fig. 3 SDT optimal model size during training (*blue line* or *line no. 1*) and validation phases (*red line* or *line no. 2*) (color figure online)

Table 3 Training parameters of BDT model

Parameter type	Value
Maximum trees in tree boost series	400
Maximum splitting levels	5
Minimum size node to split	10
Maximum categories for continuous predictors	200
Shrinkage factor	0.05
Tree pruning and validation method	*k*-fold cross-validation
Number of cross-validation folds	10
Tree pruning criterion	Minimum absolute error

error rate jumps up and gradually declines. In cases like this, the short series is unreliable, and it is undesirable to prune to that length even if the minimum error occurs with one or two trees. By specifying the minimum number of trees in the series, we can guarantee that pruning will not truncate the series below a specified length. To allow pruning the series to a smaller number of trees than the minimum validation point, pruning was allowed to increase the error by up to 10 %.

As shown in Fig. 4, the BDT series were pruned to 348 trees with minimum error of 0.0293. The minimum error of BDT with the training data occurred with 373 trees (0.0263) while the minimum error with the test data occurred with 353 trees (0.0263). Fluctuations in the error rate were smoothed out by averaging the misclassification rates for neighboring tree series sizes. Sometimes, the error rate fluctuates as the tree size increases, and an anomalous minimum "spike" may occur in a region where the surrounding error rates are much higher. This happens more often when using random row-holdback validation than when using *V*-fold cross-validation which tends to average out error rate values [77]. Therefore, the minimum point was smoothed by 5 trees.

5.2.3 Decision tree forest (DTF)

DTF models provided greater predictive accuracy than single-tree models, but they have the disadvantage that cannot be visualized; DTF models are more of a "black box". Generally, the larger a DTF is the more accurate prediction. There are two types of size controls available (1) the number of trees in the forest and (2) the size of each individual tree. Parameters of DTF are summarized in Table 4. Specify the maximum number of levels (depth) that each tree in the forest may be grown to. Some research indicates that it is best to grow very large trees, so the maximum levels should be set large, and the minimum node size control would limit the size of the trees. Therefore, maximum tree levels were adjusted at 50 trees. Surrogate splitters were used to handle missing values to compute the association between the primary splitter selected for a node and all other predictors including predictors not considered as candidates for the split. If the value of the primary predictor variable is missing for a row, the software will use the best surrogate splitter whose value is known for the row.

The classification results of the training phase that obtained from the three SDT and BDT classifiers are displayed in Table 5 by using a confusion matrix. RF tree never reports results on training data. When trees are grown out to their maximal size, we can readily expect near perfect classification on training data but this result is not useful for model assessment. The primary RF performance measures were based on OOB data.

In a confusion matrix, each cell contains the raw number of exemplars classified for the corresponding combination of desired and actual classifier outputs where TP and TN are the number of samples which are correctly identified as positives or negatives by the classifier in the test set, respectively, and FN and FP represent the numbers of samples corresponding to those cases as they are

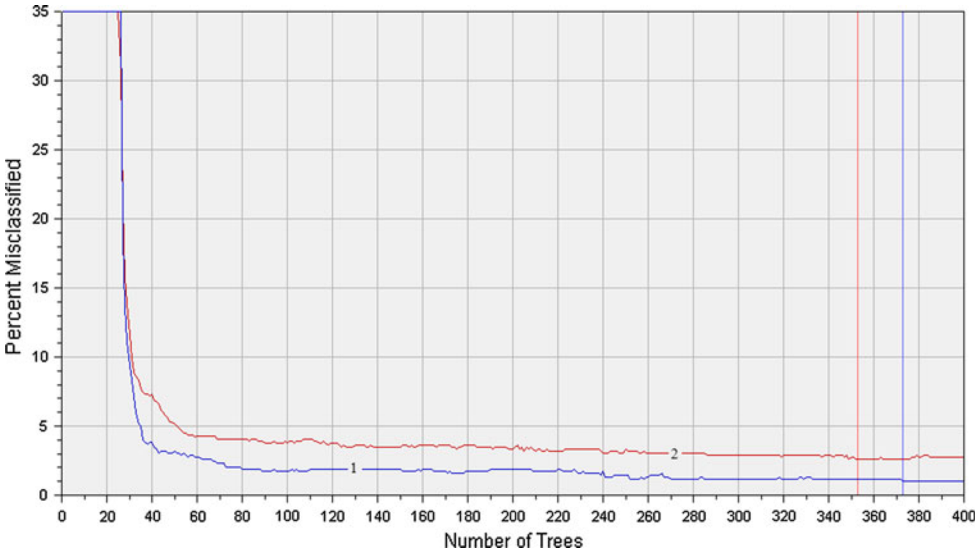


Fig. 4 BDT optimal model size during training (*blue line* or *line no. 1*) and validation phases (*red line* or *line no. 2*) (color figure online)

Table 4 DTF model parameters

Parameter type	Value
Maximum trees in decision tree forest	200
Maximum splitting levels	50
Minimum size node to split	5
Maximum categories for continuous predictors	200
Handle missing values	Surrogate splitters
Tree validation method	Out-of-bag (OOB)

Table 5 Classification comparison of SDT and BDT classifiers during training phase for breast cancer diagnosis

Category	SDT		BDT	
	Benign	Malignant	Benign	Malignant
Benign	429 (TP)	15 (FN)	437 (TP)	7 (FN)
Malignant	5 (FP)	234 (TN)	1 (FP)	238 (TN)

“Benign” and “malignant” in the column headings indicate histologic findings

TP true-positive, *TN* true-negative, *FN* false-negative, *FP* false-positive

mistakenly classified as benign or malignant, respectively. Considering imbalanced positive and negative samples in the data sets, another appropriate quantity for evaluating the classification accuracy of imbalanced positive and negative samples is the Matthews correlation coefficient (MCC), which is given as follows [89]:

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FN}\cdot\mathrm{FP}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})}} \tag{13}$$

Obviously, the scope of the MCC is within the range of [−1, 1]. The larger the MCC value, the better the classifier performance.

The performance analysis of the training phase by SDT and BDT classifiers is given in Table 6 and also represented graphically in Fig. 5. The results illustrated the overall accuracies of SDT and BDT achieved 97.07 % with 429 correct classifications and 98.83 % with 437 correct classifications, respectively. The results indicated that BDT performed better than SDT for all performance indices. Value of ROC and MCC for BDT achieved 0.99971 and 0.9746, respectively, which is superior SDT classifier.

5.3 Validation phase of classifiers

Once the model structure and parameters have been identified, it is necessary to validate the quality of the resulting model. In principle, the model validation should not only validate the accuracy of the model, but also verify whether the model can be easily interpreted to give a better understanding of the modeled process. It is therefore important to combine data-driven validation, aiming at checking the accuracy and robustness of the model, with more subjective validation, concerning the interpretability of the model. There will usually be a challenge between flexibility and interpretability, the outcome of which will depend on their relative importance for a given application. While it is evident that numerous cross-validation methods exist, the choice of the suitable cross-validation method to be employed in the DT is based on a trade-off between maximizing method accuracy, stability and minimizing the operation time. In this research, tenfold cross-validation method is adopted for SDT and BDT because of its accuracy and possible implementation. For DTF, cross-

Table 6 Performance indices for training phase of SDT and BDT classifiers

Performance index	SDT	BDT
Accuracy	97.07 %	98.83 %
Sensitivity	96.62 %	98.42 %
Specificity	97.91 %	99.58 %
Geometric mean of sensitivity and specificity	97.26 %	99.00 %
Positive predictive value (PPV)	98.85 %	99.77 %
Negative predictive value (NPV)	93.98 %	97.14 %
Geometric mean of PPV and NPV	96.38 %	98.45 %
Precision	98.85 %	99.77 %
Recall	96.62 %	98.42 %
MCC	0.9367	0.9746
F measure	0.9772	0.9909
Area under ROC curve (AUC)	0.9827	0.999708

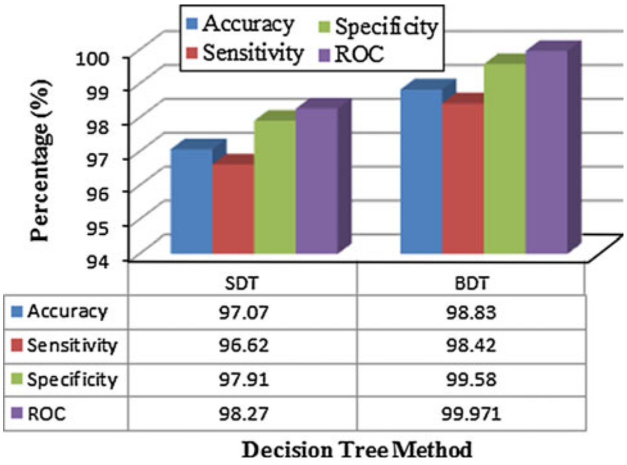


Fig. 5 Performance comparison of SDT and BDT classifiers during training phase

validation is unnecessary as it generates an internal unbiased estimate of the generalization error (test error) as the forest building progresses. The classification and performance results of the validation phase by SDT, BDT and DTF are given in Tables 7 and 8.

It can be noted from Tables 7 and 8 that the overall accuracies of DTF during validation phase achieved 97.51 %, which is superior to SDT (95.75 %) and BDT (97.07 %). Value of ROC and MCC for DTF achieved 0.99382 and 0.9462, respectively. Performance indices comparisons are shown in Fig. 6.

ROC curves for detecting benign breast tumors using DTF classifier during validation phase are shown in Fig. 7. The generalization error of DTF was calculated by computing the out-of-bag (OOB) rows for each tree through the tree. The error rates for all of the trees in the forest were then averaged to give the overall generalization error rate for the entire forest. There are several advantages to this method of computing generalization error: (1) all of the rows are used to construct the model, and none have to be held back as a separate test set, (2) the testing is fast because only one forest has to be constructed (as compared to *V*-fold cross-validation where additional trees have to be constructed). Furthermore, the BDT model showed the highest sensitivity than other methods. This means that the boosting approach is also useful to alleviate instability, which is a significant limitation of the SDT approach.

Moreover, the BDT provides additional information, that is, confidence values of overall alternatives, the structure of each tree, and attribute usage of each factor, which helps the practitioner comprehend the decision-making process. This feature is valuable for supporting the practitioner in making decisions intuitively as the number of alternatives increase in the formwork method selection process. Therefore, these results indicate that BDT has the dual advantages of boosting and the DT technique.

Besides classification accuracy, the amounts of times needed for classifier construction and for classification are also an important factor for consideration. In this regard, computational expenses were compared between all methods of DT (see Table 8). Results showed that the analysis time of BDT was much longer than the other methods. SDT was significantly faster than other types and closely followed by DTF.

The focus category analysis provides information about the “focus category” of the target variable. The impurity of the focus category is the percentage of the rows predicted to be the focus category which are actually some other category. In other words, it is the percent of the misclassified cases predicted to be the focus category. If every case that is predicted to be the focus category is actually the

Table 7 Classification results of DT classifiers during the validation phase

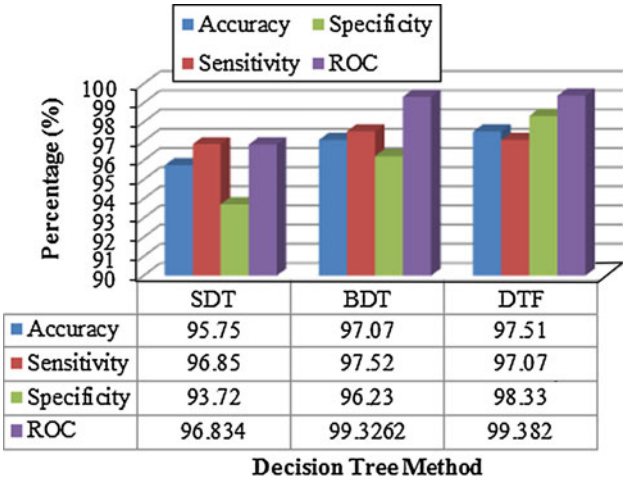
Category	SDT		BDT		DTF	
	Benign	Malignant	Benign	Malignant	Benign	Malignant
Benign	430 (TP)	14 (FN)	433 (TP)	11 (FN)	431 (TP)	13 (FN)
Malignant	15 (FP)	224 (TN)	9 (FP)	230 (TN)	4 (FP)	235 (TN)

“Benign” and “malignant” in the column headings indicate histologic findings

TP true-positive, *TN* true-negative, *FN* false-negative, *FP* false-positive

Table 8 Performance indices for validation phase of DT classifiers

Performance index	SDT	BDT	DTF
Accuracy	95.75 %	97.07 %	97.51 %
Sensitivity	96.85 %	97.52 %	97.07 %
Specificity	93.72 %	96.23 %	98.33 %
Geometric mean of sensitivity and specificity	95.27 %	96.88 %	97.70 %
Positive predictive value (PPV)	96.63 %	97.96 %	99.08 %
Negative predictive value (NPV)	94.12 %	95.44 %	94.76 %
Geometric mean of PPV and NPV	95.37 %	96.69 %	96.90 %
Precision	96.63 %	97.96 %	99.08 %
Recall	96.85 %	97.52 %	97.07 %
MCC	0.9066	0.9358	0.9462
F measure	0.9674	0.9774	0.9807
Area under ROC curve (AUC)	0.96839	0.993262	0.99382
Time	00.00.19	00.05.95	00.00.32



	SDT	BDT	DTF
🟦 Accuracy	95.75	97.07	97.51
🟥 Sensitivity	96.85	97.52	97.07
🟩 Specificity	93.72	96.23	98.33
🟪 ROC	96.834	99.3262	99.382

Fig. 6 Performance comparison of DT classifiers during validation phase

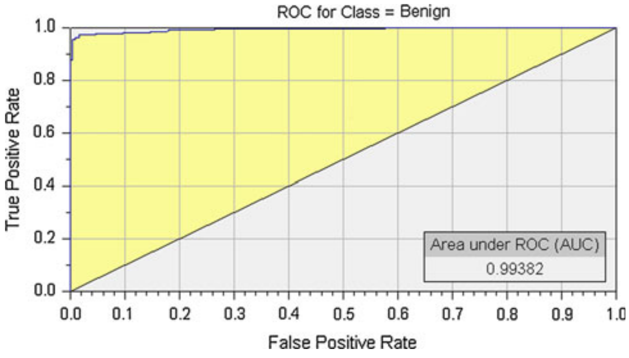


Fig. 7 ROC *curves* for detecting benign breast tumors using DTF classifier during validation phase

focus category, then the impurity is 0.0. Loss of the focus category is the percentage of actual focus category cases which are misclassified as some other category. If every case of the focus category is correctly predicted to be the focus category, then the loss is 0.0. This report shows how the impurity and loss for the focus category change with varying model sizes. The impurity values of the focus category of SDT during training and validation phases are summarized in Table 9 and also represented graphically in Fig. 8. As noted in Table 9, the full tree has 8 nodes. The minimum impurity during training phase (4.66 %) occurred with 6 nodes while in the validation phase, the minimum impurity was 5.76 % and occurred with 8 nodes (see italicized values in Table 9). The minimum loss in the training and validation phases (2.09 and 5.42 %) occurred with 8 and 4 nodes, respectively.

For BDT and DTF models, the model size is the number of trees in the model. The impurity values of the focus category of BDT during training and validation phases are summarized in Table 10. The minimum impurity during training and validation phases was 0.0 % and occurred with 1 tree. The minimum loss in the training and validation phases was 0.42 and 3.39 %, respectively, and occurred with 349 trees. The minimum loss of BDT is also represented graphically in Fig. 9.

5.4 Feature importance by DT classifiers

The software offers three methods for computing the importance of features [77]: (1) *Use split information* by adding up the improvement in classification gained by each split that used the predictor. This was the same method used to compute the importance for SDT and BDT classifiers. Generally, this method produced good results and was calculated quickly; (2) *Type 1 margins*, in this method, the misclassification rate for the model is calculated first using the actual data values for all predictors. Then, for each predictor, it randomly permutes (rearranges) the values of the predictor and computes the misclassification rate

Table 9 Focus impurity and loss values of SDT classifier

Nodes	Training		Validation	
	Impurity %	Loss %	Impurity %	Loss %
8	6.02	*2.09*	*5.76*	6.25
7	5.35	3.77	5.81	7.08
6	*4.66*	5.86	8.04	6.67
4	8.48	4.18	9.10	*5.42*
3	8.26	7.11	8.17	6.67
2	14.34	5.02	13.68	5.89

Table 10 Focus impurity and loss values of BDT classifier

Trees	Training		Validation	
	Impurity %	Loss %	Impurity %	Loss %
1	0.0	100	0.0	100
349	2.86	0.42	3.96	3.39

for the model using the permuted values. The difference between the misclassification rate with the correctly ordered values and the misclassification rate for the permuted values is used as the measure of importance of the feature. This method of calculating variable importance often is more accurate than calculating the importance from split information, but it takes much longer to compute because of the time required to permute the rows for each predictor; (3) *Type 1 + 2 margins*, in this method, the importance using type 1 margins is computed as described above. It then examines each data row and determines how many trees in the forest correctly voted for the row with the original data minus the number of trees that correctly voted for the row using the permuted data. The two measures of importance are then averaged. This is usually the most accurate measure of importance, but it is also the slowest to compute.

The importance of features calculated by DT classifiers is summarized in Table 11 and represented graphically in Fig. 10. As usual, all importance scores are rescaled to have values between 0 and 100. It is clear from the results that mitosis (feature 9) is irrelevant for each class using all three methods of DT classifiers. Uniformity of cell size (feature 2) is very important feature for each class using SDT and BDT classifiers while bare nuclei (feature 6) is the most relevant feature for each class using DTF classifier (see italicized and bold values in Table 11) followed by uniformity of cell size (feature 2). SDT classifiers reduced the number of features to five features.

6 Conclusion

The prediction and classification of breast cancer has been a challenging research problem for many researchers. Early detection of breast cancer will help to increase the chance of survival since the early treatment can be decided for the patients who suffer this disease. To attain the best solution in a specific problem, several techniques must be tested. The goal of the classification is to distinguish between the cancerous (malignant) and non-cancerous (benign) tumors. In this study, SDT, BDT and DTF classifiers were compared as decision support tools for automatic detection of breast cancer. The performance of the proposed structure is evaluated in terms of sensitivity, specificity, accuracy and ROC. The results revealed that the overall accuracies of SDT and BDT in the training phase achieved 97.07 % with

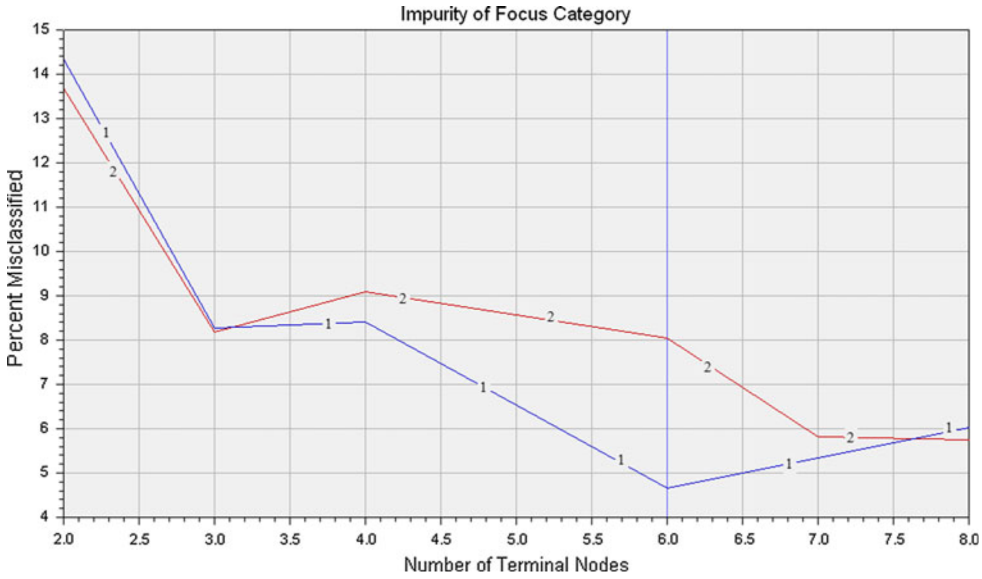


Fig. 8 Impurity during training (*blue line* or *line no. 1*) and validation phases (*red line* or *line no. 2*) of SDT classifier (color figure online)

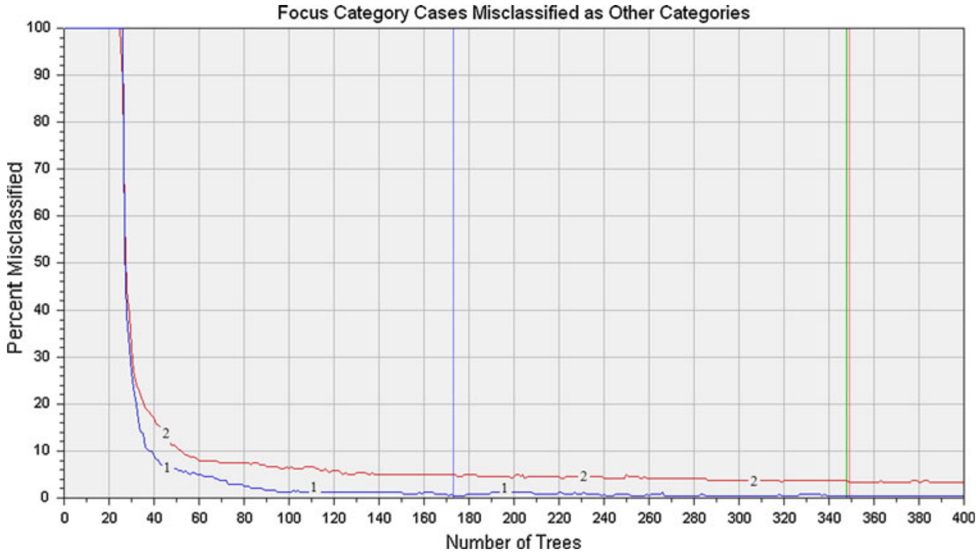
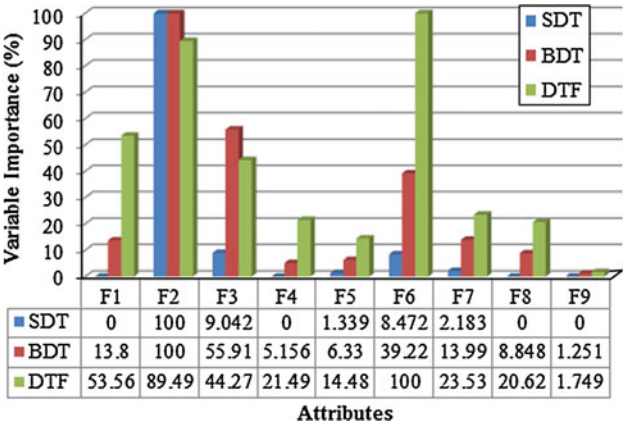


Fig. 9 Focus category loss during training (*blue line* or *line no. 1*) and validation phases (*red line* or *line no. 2*) of BDT classifier (color figure online)

Table 11 Importance of features by DT classifiers

Features	Attribute description	SDT (%)	BDT (%)	DTF (%)
F1	Clump thickness	0	13.797	53.556
F2	Uniformity of cell size	*100*	*100*	89.493
F3	Uniformity of cell shape	9.042	55.906	44.268
F4	Marginal adhesion	0	5.156	21.490
F5	Single epithelial cell size	1.339	6.330	14.481
F6	Bare nuclei	8.472	39.219	*100*
F7	Bland chromatin	2.183	13.992	23.529
F8	Normal nucleoli	0	8.848	20.621
F9	Mitoses	**0**	**1.251**	**1.749**



	F1	F2	F3	F4	F5	F6	F7	F8	F9
■SDT	0	100	9.042	0	1.339	8.472	2.183	0	0
■BDT	13.8	100	55.91	5.156	6.33	39.22	13.99	8.848	1.251
■DTF	53.56	89.49	44.27	21.49	14.48	100	23.53	20.62	1.749

Fig. 10 Feature importance by DT classifiers

429 correct classifications and 98.83 % with 437 correct classifications, respectively. BDT performed better than SDT for all performance indices than SDT. Value of ROC and MCC for BDT achieved 0.99971 and 0.9746, respectively, which is superior SDT classifier. During validation phase, DTF achieved 97.51 %, which was superior to SDT (95.75 %) and BDT (97.07 %) classifiers. Value of ROC and MCC for DTF achieved 0.99382 and 0.9462, respectively. DTF is an ensemble method that combines the predictions of many individual tree models (the base classifiers) to provide a prediction that tends to be more accurate than any of the individual classifiers' predictions. BDT showed the best performance in terms of sensitivity, and SDT was the best only considering speed. The experimental results proved that DTF techniques provided satisfactory results for the classification task of breast cancer. The results showed that the proposed system is a useful tool to be used by clinicians that may further assist the selection of appropriate adjuvant treatments for the individual patient As a result, for the CADx system developers, the use of proposed DT methods strategy might be recommended either for gene expression data sets or for ordinary medical data sets. Future work can also include integrating the tree structure into the Markov random field-based relabeling system by using the classification confidence of each test sample to improve the overall accuracy.

Acknowledgments I would like to highly appreciate and gratefully acknowledge Phillip H. Sherrod, software developer and consultant on predictive modeling, for his support and consultation during modeling process.

References

1. Alsabti K, Ranka S, Singh V (1998) CLOUDS: a decision tree classifier for large datasets. In: Proceedings of the fourth international conference on knowledge discovery and data mining (KDD-98), August 27–31. AAAI Press, New York City, NY, USA, pp 2–8

2. Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. Neural Comput 9(7):1545–1588
3. Ankerst M, Elsen C, Ester M, Kriegel HP (1999) Visual classification: an interactive approach to decision tree construction. In: Proceedings of international conference on knowledge discovery and data mining (KDD '99), San Diego, CA, USA
4. Arditi D, Pulket T (2005) Predicting the outcome of construction litigation using boosted decision trees. J Comput Civil Eng 19(4):387–393
5. Balakumaran T, Vennila ILA, Shankar GC (2010) Microcalcification detection in digital mammograms using novel filter bank. Procedia Comput Sci 2:272–282
6. Bick U, Diekmann F (2007) Digital mammography: what do we and what don't we know? Eur Radiol 17(8):1931–1942
7. Boyle P, Levin B (2008) World cancer report 2008. International Agency for Research on Cancer, Lyon
8. Bradford JP, Kunz C, Kohavi R et al (1998) Pruning decision trees with misclassification costs. In: Proceedings of the 10th European conference on machine learning, Chemnitz, Germany, pp 131–136, April 21–23, 1998
9. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth & Brooks, CA
10. Breiman L (1994) Bagging predictors. Technical report 421. Department of Statistics, University of California, Berkeley
11. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
12. Brown DE (2008) Introduction to data mining for medical informatics. Clin Lab Med 28(1):9–35
13. Burrell HC, Sibbering DM, Wilson AR et al (1996) Screening Interval breast cancers: mammographic features and prognostic factors. Radiology 199(3):811–817
14. Burrell HC, Pinder SE, Wilson AR et al (1996) The positive predictive value of mammographic signs: a review of 425 non-palpable breast lesions. Clin Radiol 51(4):277–281
15. Clark LA, Pregibon D (1992) Tree-based models. In: Chambers JM, Hastie TJ (eds) Statistical models (chap 9). S. Chapman & Hall, New York, pp 377–420
16. Christoyianni I, Koutras A, Dermatas E, Kokkinakis G (2002) Computer aided diagnosis of breast cancer in digitized mammograms. Comput Med Imaging Graph 26(5):309–319
17. Cummings MP, Segal MR (2004) Few amino acid positions in rpoB are associated with most of the rifampin resistance in Mycobacterium tuberculosis. BMC Bioinform 5:137–143
18. De'ath G (2007) Boosted trees for ecological modeling and prediction. Ecology 88(1):243–251
19. De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81(11):3178–3192
20. Delen D, Walker G, Kadam A (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 34(2):113–127
21. Dershaw DD (2006) Status of mammography after the digital mammography imaging screening trial: digital versus film. Breast J 12(2):99–102
22. DeSantis C, Siegel R, Bandi P, Jemal A (2011) Breast cancer statistics. CA Cancer J Clin 61(6):409–418
23. Diamantidis NA, Karlis D, Giakoumakis EA (2000) Unsupervised stratification of cross-validation for accuracy estimation. Artif Intell 116(1–2):1–16
24. Dietterich TG (1990) Machine learning. Annu Rev Comput Sci 4(1):255–306
25. Doi K, MacMahon H, Katsuragawa S et al (1999) Computer-aided diagnosis in radiology: potential and pitfalls. Eur J Radiol 31(2):97–109
26. Doi K (2007) Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph 31(4–5):198–211
27. Elatar I (2002) Cancer registration, NCI Egypt 2001. National Cancer Institute, Cairo. http://www.nci.edu.eg/Journal/nci2001%20.pdf, accessed 26 May 2012
28. Endo A, Shibata T, Tanaka H (2008) Comparison of seven algorithms to predict breast cancer survival. Biomed Soft Comput Hum Sci 13(2):11–16
29. Fan CY, Changb PC, Linb JJ, Hsieh JC (2011) A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl Soft Comput 11(1):632–644
30. Ferri U, Flach PA, Hernandez-Orallo J (2003) Improving the AUC of probabilistic estimation trees. In: Lecture notes in artificial intelligence, vol 2837, pp 121–132
31. Francois D, Rossi F, Wertz V, Verleysen M (2007) Resampling methods for parameter-free and robust feature selection with mutual information. Neurocomputing 70(7–9):1276–1288
32. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Proceedings of the 13th international conference on artificial intelligence: machine learning. International Machine Learning Society, pp 148–156
33. Freund Y, Schapire RE (1999) A short introduction to boosting. J Jpn Soc Artif Intell 14(5):148–156
34. Friedman JH, Kohavi R, Yun Y (1996). Lazy decision trees. In: Proceedings of the 13th national conference on artificial intelligence and eighth innovative applications of artificial intelligence conference, vol 1. AAAI Press/The MIT Press, AAAI 96, IAAI 96, August 4–8, 1996, pp 717–724
35. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Statist 29(5):1189–1232
36. Fulton T, Kasif S, Salzberg S, Waltz D (1996) Local induction of decision trees: towards interactive data mining. In: Proceedings of the second international conference on knowledge discovery and data mining, Portland, OR, USA, pp 14–19
37. Garofalakis M, Hyun D, Rastogi R, Shim K (2000). Efficient algorithms for constructing decision trees with constraints. In: Proceedings of the 6th ACM SIGKDD international conference on knowledge discovery and data mining, Boston, MA, USA, pp 335–339
38. Hambly NM, McNicholas MM, Phelan N, Hargaden GC, O'Doherty A, Flanagan FL (2009) Comparison of digital mammography and screen-film mammography in breast cancer screening: a review in the Irish breast screening program. Am J Roentgenol 193(4):1010–1018
39. Ho T (1995) Random decision forest. In: 3rd international conference on document analysis and recognition, Montreal, Canada, August 14–18, 1995, pp 278–282
40. Ho T (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844
41. Houssami N, Given-Wilson R, Ciatto S (2009) Early detection of breast cancer: overview of the evidence on computer-aided detection in mammography screening. J Med Imaging Radiat Oncol 53(2):171–176
42. Ibrahim NA, Kudus A, Daud I, Abu Bakar MR (2008) Decision tree for competing risks survival probability in breast cancer study. Proc World Acad Sci Eng Technol 38:15–19
43. Islam SR, Aziz SM (2012) Mammography is the most effective method of breast cancer screening. Mymensingh Med J 21(2):366–371
44. Kallergi M (1998) Digital mammography: from theory to practice. Cancer Control 5(1):72–79
45. Kerekes J (2008) Receiver operating characteristic curve confidence intervals and regions. IEEE Geosci Remote Sens Lett 5(2):251–255
46. Kuo WJ, Chang RF, Chen DR, Lee CC (2001) Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. Breast Cancer Res Treat 66(1):51–57

47. Jemal A, Murray T, Samuels A, Ghafoor A, Ward E, Thun MJ (2003) Cancer statistics. CA Cancer J Clin 53:5–26
48. Jerez-Aragonés JM, Gómez-Ruiz JA, Ramos-Jiménez G, Muñoz-Pérez 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
49. Lavrac N (1999) Selected techniques for data mining in medicine. Artif Intell Med 16(1):3–23
50. Laya MB, Larson EB, Taplin SH, White E (1996) Effect of estrogen replacement therapy on the specificity and sensitivity of screening mammography. J Natl Cancer Inst 88(10):643–649
51. Lee MY, Yang CS (2010) Entropy-based feature extraction and decision tree induction for breast cancer diagnosis with standardized thermograph images. Comput Methods Program Biomed 100(1):269–282
52. Lewin JM, D'Orsi CJ, Hendrick RE, Moss LJ, Isaacs PK, Karellas A, Cutter GR (2002) Clinical comparison of full-field digital mammography and screen-film mammography for detection of breast cancer. Am J Roentgenol 179(3):671–677
53. Li H, Giger ML, Yuan Y, Chen W, Horsch K, Lan L, Jamieson AR, Sennett CA, Jansen SA (2008) Evaluation of computer-aided diagnosis on a large clinical full-field digital mammographic dataset. Acad Radiol 15(11):1437–1445
54. Lim TS, Loh WY, Shih YS (1998) An empirical comparison of decision trees and other classification methods. Technical report 979. Department of Statistics, University of Wisconsin
55. Llora X, Garrell JM (2001) Evolution of decision trees. In: Proceedings of the 4th Catalan conference on artificial intelligence (CCIA '2001). ACIA Press
56. Locasale JW, Cantley LC (2010) Altered metabolism in cancer. BMC Biol 88:88
57. Mangasarian OL, Wolberg WH (1990) Cancer diagnosis via linear programming. SIAM News 23(5):1–18
58. Mangasarian OL, Setiono R, Wolberg WH (1990) Pattern recognition via linear programming: theory and application to medical diagnosis. In: Coleman TF, Li Y (eds) Large-scale numerical optimization. SIAM, Philadelphia, pp 22–30
59. Meinshausen N (2006) Quantile regression forests. J Mach Learn Res 7:983–999
60. Mehta M, Agrawal R, Rissanen J (1996) SLIQ: a fast scalable classifier for data mining. In: Proceedings of the 5th international conference on extending database technology, Avignon, France, March 25–29, pp 18–32
61. McAree B, O'Donnell ME, Spence A et al (2010) Breast cancer in women under 40 years of age: a series of 57 cases from Northern Ireland. Breast 19(2):97–104
62. Mingers J (1989) An empirical comparison of selection measures for decision tree induction. Mach Learn 3(4):319–342
63. Muller S (1997) Full-field digital mammography designed as a complete system. Eur J Radiol 31(1):25–34
64. NHS breast screening programmes: annual review 2011. ISBN: 978-1-84463-079-0. http://www.cancerscreening.nhs.uk/breastscreen/
65. Noble M, Bruening W, Uhl S, Schoelles K (2009) Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. Arch Gynecol Obstet 279(6):881–890
66. Omar S, Khaled H, Gaafar R et al (2003) Breast cancer in Egypt: a review of disease presentation and detection strategies. East Mediterr Health J 9(3):448–463
67. Park SH, Goo JM, Jo CH (2004) Receiver operating characteristic (ROC) curve: practical review for radiologists. Korean J Radiol 5(1):11–18
68. Pryke M (2012) Effect of population-based screening on breast cancer mortality. Lancet 379(9823):1297–1298
69. Quinlan JR (1993) C4. 5: programs for machine learning. Morgan Kaufmann, San Mateo
70. Quinlan JR (2003) Data mining tools See5 and C5.0. RuleQuest Research, Austria. http://www.rulequest.com/see5-info.html
71. Richards G, Rayward-Smith VJ, Sönksen PH, Carey S, Weng C (2001) Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med 22(3):215–231
72. Russell S, Norvig P (2002) Artificial intelligence: a modern approach. Prentice-Hall, NJ
73. Salzberg SL (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Discov 1(3):317–327
74. Shah AJ, Wang J, Yamada T, Fajardo LL (2003) Digital mammography: a review of technical development and clinical applications. Clin Breast Cancer 4(1):63–70
75. Shapiro S, Strax P, Venet L (1966) Evaluation of periodic breast cancer screening with mammography. Methodology and early observations. JAMA 195(9):731–738
76. Shanthi S, Bhaskaran VM (2011) Intuitionistic fuzzy C-means and decision tree approach for breast cancer detection and classification. Eur J Sci Res 66(3):345–351
77. Sherrod PH (2012) DTREG predictive modeling software. www.dtreg.com, accessed 16 Sep 2012
78. Shiraishi A (2008) Current state of digital mammography. Breast Cancer 15(3):194–199
79. Sinclair N, Littenberg B, Geller B, Muss H (2011) Accuracy of screening mammography in older women. Am J Roentgenol 197(5):1268–1273
80. Skaane P (2009) Studies comparing screen-film mammography and full-field digital mammography in breast cancer screening: updated review. Acta Radiol 50(1):3–14
81. Štajduhar I, Dalbelo-Bašic′ B (2012) Uncensoring censored data for machine learning: a likelihood-based approach. Expert Syst Appl 39(1):7226–7234
82. Theodoridis S, Koutroumbas K (2006) Pattern recognition, 3rd edn. Academic Press, San Diego
83. Ture M, Tokatli F, Kurt I (2009) Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Syst Appl 36(1):2017–2026
84. Tyler RM, Brady DC, Targett TE (2009) Temporal and spatial dynamics of diel—cycling hypoxia in estuarine tributaries. Estuaries Coasts 32:123–145
85. UCI (2012) Machine learning repository. http://archive.ics.uci.edu/ml/index.html, accessed 16 Sep 2012
86. Van Ongeval Ch (2007) Digital mammography for screening and diagnosis of breast cancer: an overview. JBR BTR 90(3):163–166
87. Vinnicombe S, Pinto Pereira SM, McCormack VA et al (2009) Full-field digital versus screen-film mammography: comparison within the UK breast screening program and systematic review of published data. Radiology 251(2):347–358
88. Wilkinson JE (2011) Effect of mammography on breast cancer mortality. Am Fam Physician 84(11):1225–1227
89. Yuan Q, Cai C, Xiao H et al (2007) Diagnosis of breast tumours and evaluation of prognostic risk by using machine learning approaches. Commun Comput Inf Sci 2:1250–1260. doi:10.1007/978-3-540-74282