Oncology
and Radiotherapy ©
Vol.16 Iss.4:1-6 • Research Article

Early stage breast cancer detection using ensemble approach of random forest classifier algorithm

Dumpala Shanthi
Department of Computer Science and Engineering, Sreyas Institute of Engineering and Technology, India

SUMMARY

Breast Cancer is one of the most deadly diseases in the world and is commonly seen in women. Based on the severity, breast cancer is classified into two types. One is Benign type of breast cancer, which can be detected at early stages and can be cured with the help of medication. Other is Malignant type of breast cancer, which shows severe affect and might lead to death. To detect breast cancer at early stages, wide variety of algorithm techniques are used such as Navie Bayes, Convolution Neural Network, KNN, adaptive voting ensemble machine learning algorithm and so on. Most latest algorithm that is under practice is adaptive voting ensemble machine learning algorithm. In this algorithm, Wisconsin Breast Cancer dataset and CNN algorithm is used to classify images and for object detection. But the major drawback of ensemble machine learning algorithm is lack of accuracy. It is proved that Neutral Network works more effective on humans mostly in analysing data and to perform pre-diagnosis without medical knowledge. In this paper, we propose Random Forest Classifier algorithm to achieve more accuracy.

Key words: breast cancer, random forest, diagnosis

Address for correspondence:

Dumpala Shanthi, Department of Computer Science And Engineering, Sreyas Institute of Engineering and Technology, India, email: dshanthi01@gmail.com

Word count: 3803 **Table:** 04 **Figures:** 08 **References:** 09

Received:- 19 February, 2021, Manuscript No. OAR-21-26506

Editor assigned:- 23 February, 2021, PreQC No. OAR-21-26506(PQ)

Reviewed:- 10 March, 2021, QC No. OAR-21-26506(Q)

Revised:- 20 December, 2021, Manuscript No. OAR-21-26506(R)

Published:- 29 April, 2022 , Invoice No. J- 26506

INTRODUCTION

The most scary disease in this world is cancer and Breast Cancer is the more scarier among the women. Numerous expire due to malignant growth. Distinguishing the malignant growth physically takes lot of time and it is hard for the doctor to detect. So distinguishing the malignant growth through different programmed indicative methods is very vital. There is different strategy and calculation accessible for recognizing disease, for example, Support Vector Machine, Naïve Bayes, KNN and Convolution Neural System is the most recent calculation in Deep learning that is moreover utilized for classification. CNN and deep learning calculation for the most part utilized for pictures and article recognition. In this paper we use UCI open database for training and testing reason in which two classes of Tumour are accessible, one is benign tumour and the other is threatening in which benign Tumour is non-carcinogenic and the dangerous is a malignant growth Tumour. Numerous scientists are yet performing research for identifying disease in the beginning period. The early stage malignant growth is costly to complete its treatment and numerous catalysts are yet attempting to building up an appropriate analysis framework for identifying the tumour. So the treatment can be begun earlier before and the rate for goals may increase

Existing method

In the early stages of research breast cancer detection is based on low energy X-ray mammography was in practice. Later Magnetic Resonance images, Ultra Sound images are also preferred. In pre-processing stage various methods such as binarization, Image thinning, Image gray scale extending, discrete wavelets, real valued or complex valued continuous filters, fuzzy filters were applied by many researchers. Similarly, transformation techniques were used to extract physical features or textural features of an image. Transformation techniques Euclidean Distance Transform, Fourier Transform, Discrete Wavelet Transform were widely used. There are several classifiers to in process obtain the decision of finding the presence or absence of the cancer cells. The classification results vary based on the research work to find the whether it is benign or malignant. Some researchers provide the stages of cancer. But most of the research works was based on single classifier. In general, the majority voting method was also applied to find the optimized result. Parameter tuning is not needed all the time after tuning individual classifier. A weighted voting method is another approach to find the optimal solution.

Disadvantages

- Lack of accuracy when compared to neural networks which works more effective on humans in data analysis and diagnosis.
- By using number of filters we are detecting the cancer which increases the cost of the system and also expert person need to be there to detect the cancer

Random forest classifier algorithm

In recent days, highly developed machine learning techniques had been used in a wide range to detect breast cancer at early stages. While performing diagnosis, the data gathered from the patient, data analysis and decisions taken by experts plays vital role to acquire better results. Diagnosis performed using different algorithm techniques can avoid human errors in analysis data and cancer detection at early stages. Proposed method Random Forest Classifier algorithm had around 99.7% accuracy in classifying the data and to obtain better results. Also the proposed method can be used to analyse other breast cancer problems which has high rate of training data. In our proposed method we used Wisconsin Breast Cancer Data set (WBC-DD). This is widely used by researchers who use different algorithm techniques to cure breast cancer, as it is used to compare our system with other references related to the same issue.

The rest of the paper is organized as follows: Section 2 includes Literature survey to propose method Section 3 provides the overview of proposed method Random Forest Classifier Algorithm Section 4 presents architecture and control flow of proposed method. Section 5 presents results obtained by using the proposed method to diagnose breast cancer. Section 6 includes the future scope of proposed method and concludes the paper work. Finally, Section 7 includes references.

Literature survey

Literature survey includes researches or work conducted in the past, that is related to breast cancer detection. A lot of research work has been done to diagnosis breast cancer at early stages with maximum accuracy. Below mentioned few techniques that are in practice [1].

In this paper, author explained about advantages of Positron Emission Tomography (PET) algorithm when compared to other algorithms like CT imagining and X-ray imaging. Computed Tomography (CT), also known as Computed Axial Tomography (CAT), is one of the scanning techniques that uses both x-ray and computer technologies. It generates 2D image that represents the cross sectional part of the body. This can be converted to three dimensional image using special algorithms. Whereas X-rays are the high frequency electromagnetic waves, which are transmitted through the body and creates a black and white images. Generally, hard particles appears as white and spaces as black in colour. Coming to PET imaging, Positron Emission Tomography (PET) detects fast growing tissues in the body based on image functional processes. Positron emits isotope which has short life span and contains organic substances like glucose. This organic substance creates F18-fluorodeoxyglucose, which determines metabolic utilization. With the help of PET, one can detect rapidly growing tissues like tumor and other infections in the body. PET images can be used to compare with CT scans to determine co-relation. This process can be performed on the same device without disturbing patient. It also reduces image reconstruction. In addition to this, author also explained future scope and improvements that can be made to get better accuracy [2].

In this paper, author explained unique method MR molecular imaging to diagnose breast cancer. One of the clinically relevant targets is the tyrosine kinase Her-2/neu receptor, which has a significant role in staging and treating breast cancer. In this paper, receptors were imaged in a panel of breast cancer cells expressing different numbers of the receptors on the cell membrane. Commercially available streptavidin-conjugated super-paramagnetic Nano particles were used as targeted MR contrast agent. These targeted MR contrast agents were directed to receptors pre-labelled with a bio-tinylated monoclonal antibody and generated strong T2MR contrast in Her-2/neu-expressing cells. The contrast observed in MR images was proportional to the severity level of Her-2/neu receptors determined independently with FACS analysis. In these experiments, iron oxide Nano-particles were attached to the cell surface and were not internalized into the cells, which is a major advantage for in vivo applications of the method [3].

In this paper, author proposed an automatic diagnosis system to detect breast cancer based on Association Rules (AR) and Neural Network (NN). In this method, Association Rules (AR) is used to decrease the breast cancer data base field. The size of the breast cancer database field was reduced from nine to four. Whereas, Neural Network (NN) is used for intelligent classification. During validation, to verify the classification accuracy and system performance, a 3-fold cross validation method was applied to the Wisconsin breast cancer database and achieved 90%. This paper demonstrated that rapid automatic diagnostic results can be achieved by using AR+NN method and can be applicable for other diseases [4].

In this paper, author implemented an algorithm to diagnosis breast cancer based on neuro-fuzzy rules. Mammography is one of most effective and widely used techniques used in today's world to detect breast cancer. But due to low precision rate and unnecessary biopsies, the classifier accuracy of mammography is approximately around 70%. The main objective of this paper is to develop strong algorithm using fuzzy rules in order to detect breast cancer with more accuracy. Artificial Intelligence technologies had been used in order to extract these fuzzy rules. To perform neuro-fuzzy classification NEFCLASS tool is used. This method uses Breast Imaging Reporting and Data Base System (BI-RADS), mass shape, and mass margin attributes. The predictive rate of this rule base is 75% positive 93% on negative side. Overall, on an average, approximately 70% rate of unnecessary biopsy in the diagnosis process is taken into consideration and perform diagnosis [5].

In this paper, author used rough set supporting vector machine (RS_SVM) classifier to detect breast cancer with more classifier accuracy. In this method (RS_SVM), RS is a reduction algorithm

where it clears all the redundant features. Then to improve the classifier accuracy, SVM algorithm is used. This method uses Wisconsin Breast Cancer Dataset (WBCD) to calculate the efficiency, accuracy and sensitivity. Some researches proved that RS_SVM method, is capable to achieve high classifier accuracy and also it can provide information regarding five different informative features which reduces human efforts to diagnose breast cancer.

Here are few more experiments that were performed in the past and got succeeded in attaining more accuracy. In Kemal Polat and Salih Güne [6] using Least Square Support Vector Machine (LS-SVM) classifier algorithm conducted diagnosis on breast cancer and obtained 98.53%. Classification accuracy. In Sahan et al. [7] a new hybrid method based on fuzzy artificial immune system and KNN algorithm was used and the obtained accuracy was 99.14%. A. Marciano et al. [8] proposed a method named AMMLP based on the biological meta-plasticity property of neurons and Shannon's information theory and obtained total classification accuracy of 99.26%. All the above mentioned observations were tested using Wisconsin Breast Cancer Dataset which is quite different from WBCDD and WBCPD. Below mentioned few researches conducted based on WBCDD and WBCPD. In T. T. Mu and A. K. Nandi [9] a research conducted based on WBCDD, by including both radial basis function networks and self-organizing maps to attained 98% accuracy. A three-stage algorithm was proposed. In first stage, a non-linear transformation method is used to extend classification related information from the original data attribute values for a small data set. In the second stage, optimal subset of all the features are obtained by applying Principle Component Analysis (PCA) on the newly transformed, which eventually are used as inputs to support vector machine and obtained 96.35% accuracy.

METHODOLOGY

Overview of proposed method

Random Forest Classifier algorithm consists of multiple decision trees which are developed based on randomness. Decision tree will be in binary format and also known as binary tree. Previously, there used to be method called "CART" to develop these binary trees. In this method, a binary partition splits the tree into homogeneous nodes, which eventually forms daughter nodes. These daughter nodes are fed or developed by the parent tree. In general, a parent tree represents the sample of original data.

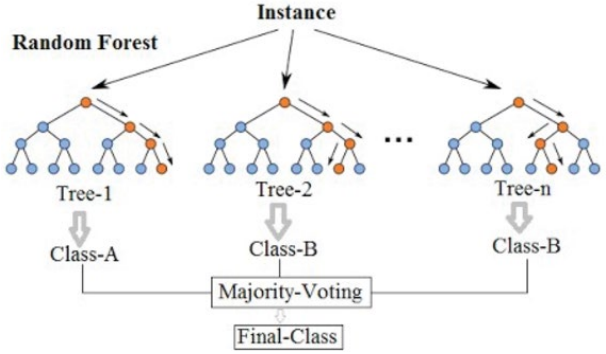

Fig. 1. Classification of random forest

RF tree differs from CART, as RF uses two stage randomization procedures. In first stage, RF develops tree as similar to CART using sample of original data. In second stage, instead of splitting tree using all the inputs, RF selects random variables at each node and uses only selected variables to find out the best split in the whole tree. Few researchers conclude that RF tree is more efficient in low data sample applications. Whereas, if the tree is deeply grown or if the original data sample is huge, RF tree yields low bias and eventually reduces variance.

Architecture of random forest classification

Random tree is grown as explained below:

- Initially, the training set and testing set are differentiated from the original data set.
- In the next step, new data set named "in bag" will be formed from the training set using bootstrap method as shown below. In general, the in bag data set and training data set will contain same number of samples. But the only difference between these two is, in bag data set contains the duplicates or replacements of training data set. This method is referred as "boot strapping".
- Another data set named "Out of Bag (OOB) will be derived from the training data set. Based on Boot strapping technique, OOB data set contains one-third of the training data set samples as shown in the above figure. Out of Bag data set is also known as" Left-Over data".
- Random variables are selected at each node and uses only selected variables to find out the best split in the whole tree.

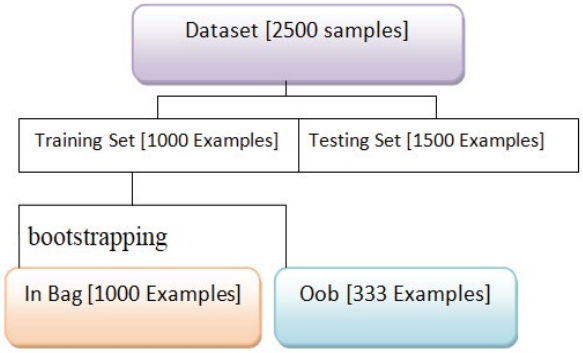

Fig. 2. RF Tree development

The above procedure will be repeated until all the possible trees are derived from the parent tree. Later, when tree development is completed, Out of Bag (OOB) samples are used to verify each individual tree and are applicable throughout forest. Out of Bag error will be estimated based on average misclassification of all the trees. Performance of the machine and the variable weights can be derived based on the error estimates. Below figure explains random forest classifier architecture for multiple applications (Figure 1).

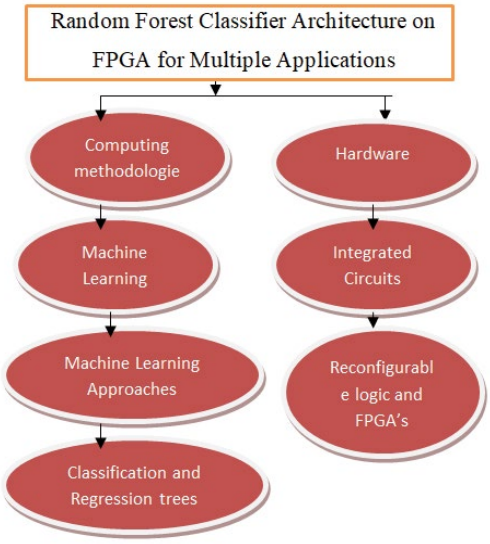

Fig. 3. Random forest classifier architecture for multiple applications

The proposed method can be explained in two phases to improve accuracy. In the first phase, as mentioned above Random Forest classifier algorithm will be trained first and fed to training set to test. This testing procedure is followed to select random variables from the whole tree and allocate rankings to each node. The feature of allocating ranks to each node will be performed based on Bayesian probability. Bayesian probability, allocates ranking to each node and arrange all nodes in ascending order. Node with least ranking will be eliminated first and the elimination process continues by comparing classification accuracy before and after eliminating the node. Overall, in the first phase, set of randomly selected nodes will be generated. Output of the first phase will be fed again to train the classifier in order to improve the classifier accuracy. N-fold cross validation is one of approach where the initial data will be randomly partition into N equal sets with same size and apply RF learning algorithm N number of times. Every time, one of the set in N set will be considered as Test set and model will be trained on remaining sets (i.e N-1). Average of errors in each set will be calculated and one set with least error will be selected and learn model parameters.

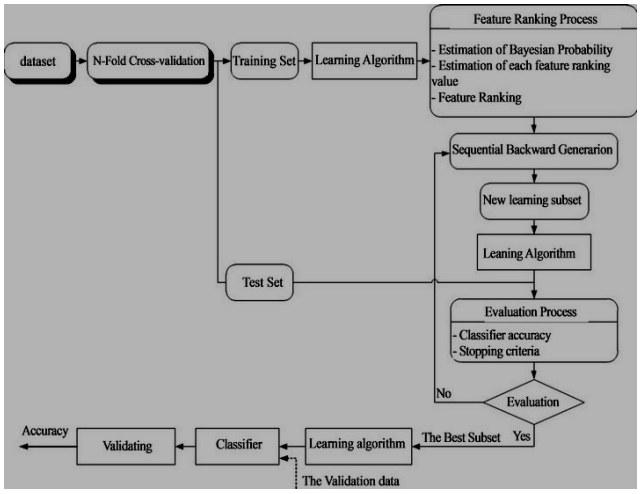

Fig. 4. Architecture of random forest classifier

The grouping of data in the Decision Tree is based on the values of attributes of the given data. A Decision Tree is formed from the pre-classified data. The division into classes is set upon the features that best divides the info. The data items are split according to the values of these features. This process is applied to each split subset of the data items recursively. The process terminates as for as all the data items in current subset belong to the same class.

REP Tree

In REP Tree, decision/regression tree is made with information gain because the splitting criterion and reduced error pruning is employed to prune it. It sorts values only for numeric attributes once. The method of fractional instances is used to handle missing values with C4.5. REP Tree is a fast Decision Tree learner.

Random tree

A random tree is a tree constructed randomly from a set of n possible trees having K random features at each node. Or we can say that trees have a "uniform" distribution. Random trees are often generated efficiently and therefore the combination of huge sets of random trees generally results in accurate models. There has been an extensive research in the recent years over Random trees in the field of machine Learning.

Random forests

Random Forest developed by Leo Breiman is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process. Prediction is formed by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Each tree is grown as described in: By Sampling N randomly, If the amount of cases within the training set is N but with replacement, from the first data. This sample is going to be used because the training set for growing the tree. For M number of input variables, the variable m is selected.

RESULTS AND DISCUSSION

In this section we present experimental results of the proposed method through fifty times of trials. The proposed method for breast cancer diagnosis and prognosis were implemented by using the R program language version 2.1.5.2 with RF package. The both of datasets were divided randomly into training set and validation set in the ratio of 1 to 1. (Table 1) The parameters of the proposed method in this experiment were determined as follows: Number of trees in RF: 25; Number of remaining features: 15-the proposed method will be stop if number of remaining features in dataset is greater or equal number of remaining features. Random Forest classifier algorithm was applied and tested 50 times on both WBCDD and WBCPD.

Figures 2-4 represent the performances of random forest classifier method on training set and test set of WBCDD and WBCPD.

(Table 2 and 3) shows number of original nodes and number of nodes were selected. The result indicates that the feature subsets

Tab.1. Results of 50 trials on WBCDD and WBCPD

Dataset	Mean of classification accuracy (%)	Sd (%)	Min (%)	Max (%)
WBCDD	99.82	0.39	98.24	100
WBCPD	99.7	0.78	96.87	100

Tab. 2. Comparison of classifier accuracy of WBCDD and WBCPD

Dataset	Classification accuracy (%)		No. of features selected		Min (%)		Max (%)
	Original Method		Proposed Method		Original Method		Proposed Method
WBCDD	57		99.82		30		18.36
WBCPD	70.4		99.7		33		17.06

Tab. 3. Time-consumption (second) of the proposed method (50 trails)

Dataset	Mean of classification accuracy (%)	SD (% (%)	Min (%)	Max (%)
WBCDD	3.52	2.3	1.81	13.05
WBCPD	2.65	0.62	1.67	4.27

Tab. 4. The sensitivity and specificity of the proposed method

Dataset	Sensitivity (%)	Specificity (%)	AUC
WBCDD	99.83	99.72	99.78
WBCPD	99.97	99.91	99.84

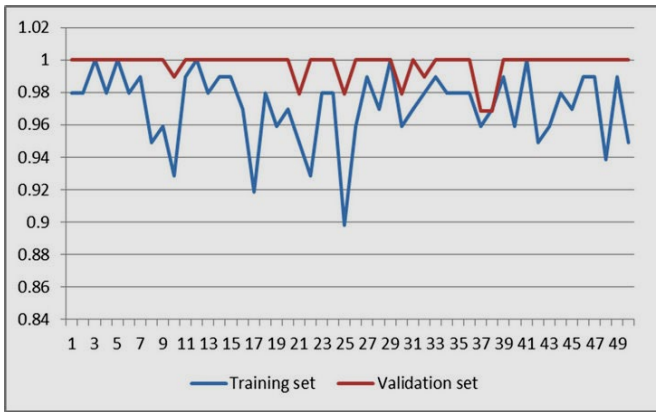

Fig. 5. Result of 50 trials of Random Forest Classifier method on WBCDD

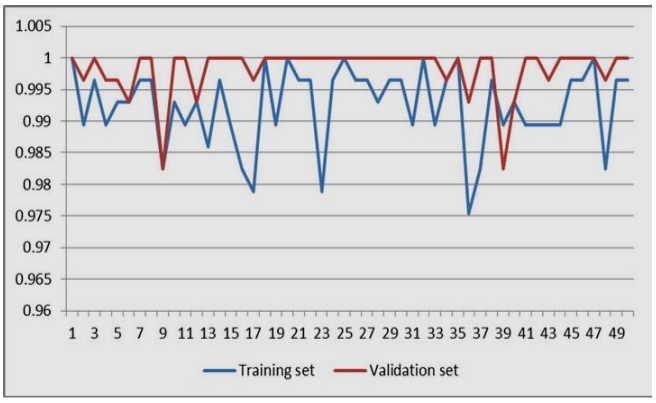

Fig. 6. Result of 50 trials of Random Forest Classifier method on WBCPD

selected by the proposed approach have a better classification performance than that produced by the original RF.

Carried out on a laptop computer with the central processing unit Intel Core 2 Duo 2.13 GHz so that time consuming is not a challenge of the proposed method. The sensitivity and specificity of the proposed method are presented in Table 4. Figures 5 and 6 show ROC curve those are built based on sensitivity (Table 4).

The results indicate that the proposed method is a reliable diagnostic tool for breast cancer diagnosis and prognostic. As we can see from the results, our method obtains the highest classification accuracy so far.(Figure 7 and 8).

Results when the previous experiment- Ensemble Voting Algorithm is used. The accuracy obtained is 92%. The proposed method Random Forest classifier has the accuracy of 96%. We give test and train data to the algorithm. It splits the data to both test and train. Initially, takes trained data and the system will learn and try to predict through test data. This way, Random Forest Algorithm gives accuracy higher which is 96%.

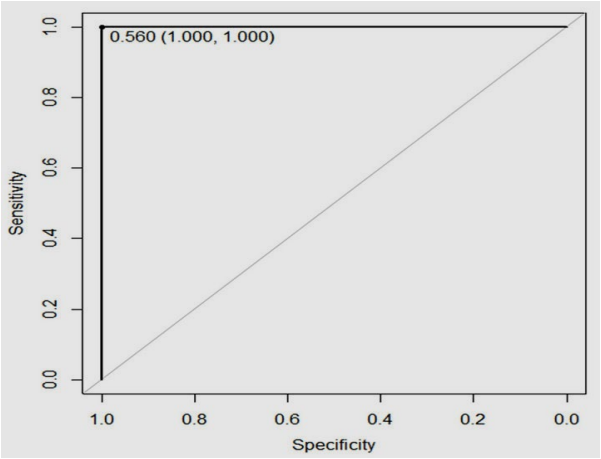

Fig. 7. ROC curve of proposed method on WBCDD

FUTURE SCOPE AND CONCLUSION

There are few other possible improvements that can be made to Random Forest Classifier algorithm in order to improve more accuracy. The below proposed improvements are beyond

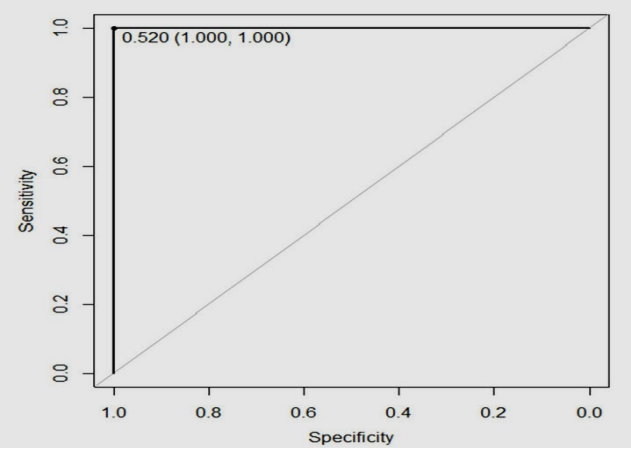

Fig. 8. ROC curve of proposed method on WBCPD

the scope of this paper and are mentioned under future scope section.

- In the proposed method, Out of Box (OOB) error analysis is performed to provide rankings and calculate the strength of the tree. Instead of OOB, margin of the forest can be considered to calculate the weight of the tree. This can be useful to improve classifier accuracy.
- In our research of Diversity based dynamic pruning, we have used bootstrap datasets for finding diversity. Instead, OOB datasets can be used to find diversity. It would be interesting to compare and see the results of the two approaches.
- Research can be extended to design precise stopping criterion for Diversity based dynamic pruning approach. Q statistics can be used as a design parameter.

REFERENCES

1. Avril N, Rose C, Schelling M, Doce J, Kuhn W, et al. Breast imaging withpositron emission tomography and fluorine-18 fluorodeoxy- glucose: use and limitations. J Clin Oncol. 2000;18:3495-3502.
2. Weissleder R, Mahmood U. Molecular imaging. Radiol. 2001;219:316-333.
3. Ince MC, Karabatak M. An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl. 2009;36:3465-3469.
4. Keles A, Keles A and Yavuz U. Expert system based on neuro-fuzzy rules for diagnosis breast cancer. Expert Syst Appl. 2011;38:5719-5726.
5. Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Syst Appl. 2011;38:9014-9022.
6. Sahana S, Polat K, Kodaz H, Günes S. A new hybridmethod based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Comput Biol Med. 2007;377;415-423.
7. Marcano-Cedeño A, Quintanilla-Dominguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst Appl. 2011;38:9573-9579.
8. Mu TT, Nandi AK. Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier. J Franklin Inst. 2007;344:285-311.
9. Li DC, Liu CW, Hu SC. A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artif Intell Med. 2011;52:45-52.