Diagnosis of Breast Cancer using Intelligent Techniques

H.S.Hota

Abstract- Breast cancer is a serious and life threatening disease due to its invasive and infiltrative character and is very commonly found in woman .An abnormal growth of cells in breast is the main cause of breast cancer actually this abnormal growth of cells can be of two types benign (Non-Cancerous) and malignant (Cancerous) ,these types must be diagnosed clearly for proper meditation and for proper treatment. A physician with full of experience and knowledge can deal complex problem in the breast cancer diagnosis process to identify disease but modern medical diagnosis system is totally based on data obtained through clinical and/or other test ,most of the decision related to a patient to find out disease is taken based on these data Better classification of a disease is a very crucial and challenging job , a small error can cause the problem because it is directly related to the life of a human being. In this research work ,various intelligent techniques including supervised Artificial Neural Network (ANN) ,unsupervised Artificial Neural Network ,Statistical and decision tree based have been applied to classify data related to breast cancer health care obtained from UCI repository site. The various individual models developed are tested and combined together to form ensemble model .Experimental works were done using MATLAB and SPSS Clementine software obtained results shows that ensemble model is better than individual models accuracy obtained in case of ensemble model is ,which is higher than all individual models ,however counter propogation network (CPN) is a competitive model among all other individual models and accuracy of this model is very near to that is obtained in case of I ensemble model. In order to reduce dimensionality of breast cancer data set a ranking based feature selection technique is applied with best ensemble model ,experimental result show that model has less accuracy with less number of features .Models are also analyzed in terms of other error measures like sensitivity and specificity.

Keywords: Decision Tree (DT),Supervised and Unsupervised Artificial Neural Network (ANN), Breast Cancer . Support Vector Machine (SVM).

I. INTRODUCTION

Proper diagnosis of any human disease accurately and efficiently is a challenging task for the people involved in health care organization and provides a strong base for further treatment and medication ,on the basis of this diagnosis only a physician can suggest for proper treatment on the other hand a wrong diagnosis may create problem in treatment process . There is a wide variety of health care systems around the world in earlier days it is all based on human expert ,an expert physician can diagnose a problem in better way. In order to provide this facility to all the physician an intelligent decision support system (DSS) is must. A decision support system can help the physicians to diagnose the disease correctly Health care data are collected with the help of various tests and based on these data physician can take appropriate decision about a disease.Data mining techniques provides us facility to design and develop a predictive model. A predictive model is nothing but a classifier which identifies patient having any disease or as a normal patient.

Many others have worked on health care to design and develop classification models using data mining techniques. A significant work is done by bindi et.al. (2012) to analyze the common features available in the two data sets using ANOVA and MANOVA they have also used several data mining based classification techniques like Naïve Bayes,C4.5,Back Propogation ,K-NN and SVM with various feature subsets obtained with the help of ranking based algorithm to classify BUPA and Indian liver disorder data sets collected from Andhra Pradesh (AP Data Set),India. They found that AP data set has better feature to produce highest accuracy, precision, sensitivity and specificity as compare to BUPA data set for all classification algorithms. This paper extends the work of Bindi et.al to design a predictive model using above data mining algorithm with special interest of ensemble of two or more than two algorithms to improve the performance of model, because accuracy of this types of model must be very high in all respect since diagnosis of health care is related to the life of human being ,another contribution (Bendi V.R ,2011) in this area ,in which Bayesian classifier is evaluated for liver diagnosis using bagging and boosting methods .Obi J.C.and et al. have (2011) designed intelligent decision support system for identification of Alzheimer disease using neuro-fuzzy technique. Breast cancer is one of the dangerous disease very commonly found in woman, author (Alaa M.Elsayad,2010) worked on this area ,he has designed a model for predicting breast masses with ensemble of Bayesian classifier.

In this paper various data mining algorithms are explored to develop a predictive model for breast cancer .Many individual models like CART,C5.0,Bayesian Net and support vector machine have been tested on breast cancer health care data set, obtained from UCI repository (web source http://www.archive.ics.uci.edu/ml/datasets.html) site. Ensemble models are also developed and tried on the same data set and it is observed that ensemble of Bayesian Net and C5.0 is performing well .A rank based feature selection is also applied on ensemble model and results are concluded for various feature subsets.

II. BREAST CANCER DIAGNOSIS PROCESS

An implementation process of breast cancer diagnosis process is depicted in Figure 1, this overall process can be viewed as three different stages :Health Care Data ,Model Building and Model Evaluation all these are explained in more detail as below:

Manuscript received on January, 2013

H.S.Hota, Guru Ghasidas Central University ,Bilaspur (C.G) ,India

Health Care Data: The data set used for experimental purpose is downloaded from university of California of Iravin (UCI) repository site (web source http://www.archive.ics.uci.edu/ml/datasets.html). The data set has 699 instances from which 458 belongs to benign category while 241 belongs to malignant category with 11 features (attribute) ,first feature is sample code number which do not play any role in classification and the last is the class, hence there are 9 features in all, which will be used to classify the data .The detail of data set is shown in table 1.

Table 1: Breast Cancer (Wisconsin) data set

S. No.	Attribute	Value
1	Sample Code Number (ID)	Numeric
2	Clump Thickness (CT)	1 – 10
3	Uniformity of Cell Size (UCS)	1 – 10
4	Uniformity of Cell Shape (UCSh)	1 – 10
5	Marginal Adhesion (MA)	1 – 10
6	Single Epithelial Cell Size (SECS)	1 – 10
7	Bare Nuclei (BN)	1 – 10
8	Bland Chromatin (BC)	1 – 10
9	Normal Nucleoli (NN)	1 – 10
10	Mitoses (M)	1 – 10
11	Class (C)	2 for benign, 4 for malignant

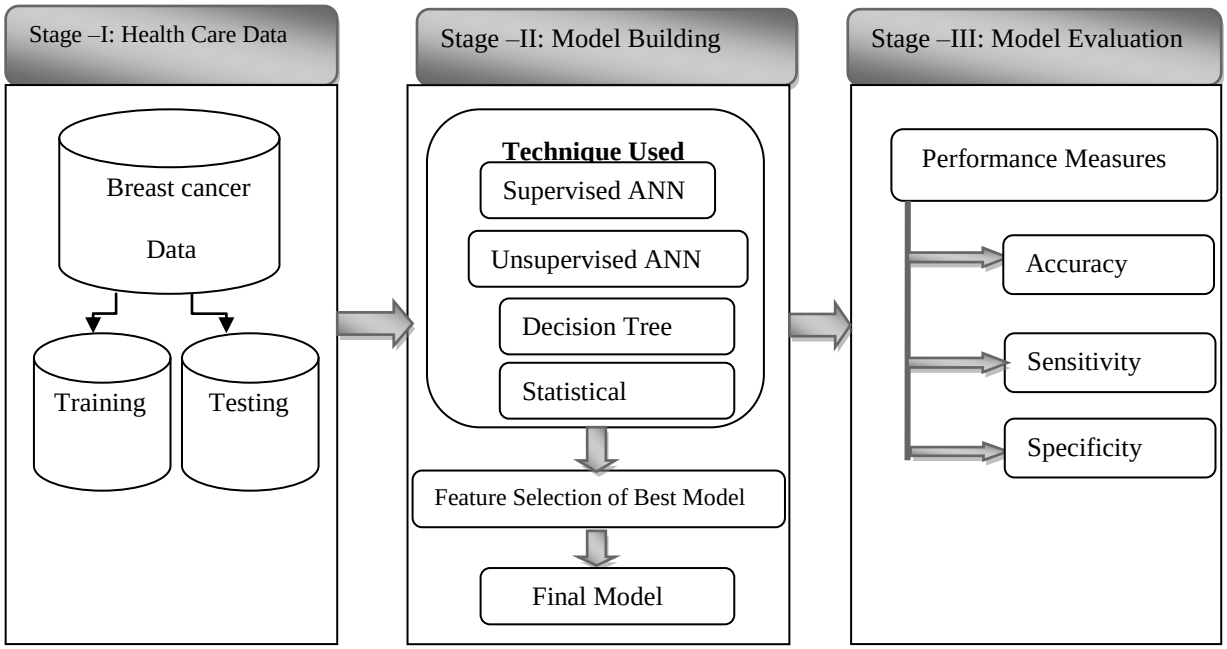


Figure 1: Implementation of Breast Cancer Diagnosis Process

Model Building: Building a model to diagnose breast cancer is basically a binary classification problem which is one of the important application of data mining. A developed model can assist a physician to take proper decision regarding a problem. Various classification algorithms are used to build classifiers, these classifiers are divided into three different categories: (1)Artificial Neural Network technique (2)Decision tree technique and (3) Statistical technique ,these techniques are explained in more detail as below:

(1)Artificial Neural Network Technique:

An Artificial neural network may be defined as an information-processing model that is inspired by the way biological nervous system, such as brain, processes information (Shivanandan & Deepa, 2011). An ANN is composedof a large number of highly interconnected processing elements (neurons), which helps to solve specific problems. Since the stock market Indices move in a nonlinear fashion, an ANN based predictive model better captures the input patterns for future forecasting precisely. The study attempts to construct a three layer ANN architecture withfour neurons in inputs and hidden layer, and one neuorn in the output to solve the complex nonlinear problems ofstock index forecasting. However, a more complex ANN architecture with more number of hidden layers helpsachieve a high level of precision in forecasting.

There are many architectures exist in neural network ,all these ANN are having their own strength and weakness ,some very well known and most suitable ANN applied for breast cancer diagnosis process are explained as below:

(A) Error Back Propogation Network (EBPN):

ANN model can be developed using two phases: (1) Training and (2) Testing. The most significant development in

Neural network technique is Error Back Propagation Algorithm (EBPA), which yields efficient results, and of course, is an extremely popular technique. This technique applies the learning algorithm to multilayer feed-forward

Networks consisting of processing elements with continuous differentiable activation functions. The algorithm provides a procedure of finding appropriate weights, for a given set of input-output pairs such that the prediction is

Precise for similar inputs. The basic concept behind this weight updated algorithm is simply the gradient-descent method, which is used in the cases of simple perceptron networks with differentiable units. This is a method where error is propagated back to the hidden unit after the weights are updated, and, finally ANN are tested for the unseen data (Testing data) to verify the prediction accuracy of the model.

It is a supervised learning method, and is a generalization of the delta rule. It requires a teacher that knows, or can calculate the desired output for any input in the training set. It is most useful for feed-forward networks (networks that have no feedback, or simply, that have no connections that loop). The term is an abbreviation for "backward propagation of errors". Back propagation requires that the activation function used by the artificial neurons (or "nodes") be differentiable. For better understanding, the back propagation learning algorithm can be divided into two phases: propagation (Forward Pass) and weight update (Backward Pass).

Phase1: Propagation- Each propagation involves the following steps:

i. Forward propagation of a training pattern's input through the neural network in order to generate the propagation's output activations.
ii. Backward propagation of the propagation's output activations through the neural network using the training pattern's target in order to generate the deltas of all output and hidden neurons.

Phase 2: Weight Update

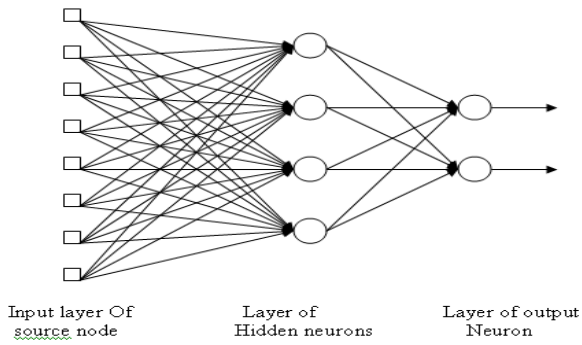
For each weight-synapse:

i. Multiply its output delta and input activation to get the gradient of the weight.
ii. Bring the weight in the opposite direction of the gradient by subtracting a ratio of it from the weight.

This ratio influences the speed and quality of learning; it is called the learning rate. The sign of the gradient of a weight indicates where the error is increasing, this is why the weight must be updated in the opposite direction.

Repeat the phase 1 and 2 until the performance of the network is good enough.

A simple architecture of Error back propogation network is shown in figure 2 with three layers :one input ,one hidden and one output layer with two neurons which will classify the data either as benign or malignant .Number of hidden neurons can be chosen in expertise manner while number of neurons in input layer should be equal to number of features present in data set in our case it is 9.



(*B) Learning Vector Quantization (LVQ*):

Learning vector quantization (LVQ) [ ] is a process of classifying the patterns, wherein each output unit represents a particular class. Here, for each class several units should be used. The output unit weight vector is called the reference vector or code book vector for the class which the unit represents. This is a special case of competitive net, which uses supervised learning methodology. During training, the output units are found to be positioned to approximate the decision surface of existing Bayesian classifier. Here the set of training patterns with known classifications is given to the network, along with an initial distribution of the reference vectors. When the training process is complete, an LVQ net is found to classify an input vector by assigning it to the same class as that of the output unit, which has as its weight vector very close to the input vector. Thus, LVQ is a classifier paradigm that adjusts the boundaries between categories to minimize existing misclassification.LVQ is used for optimal character reorganization, converting speech into phonemes and other applications as well.LVQ net may resemble KSOFM net. Unlike LVQ, KSOFM output nodes do not correspond to the known classes but rather correspond to unknown clusters that the KSOFM finds in the data autonomously.

LVQ network is used to classify health care related data i.e. cancer data. The network that is required for classification of cancer data is shown in fig 3.3.In this network there are three layers. Input layer with 9 neuron and one hidden layer which is competitive layer or kohenen layer with 9 neuron and output layer with one neuron which produces the class 1 for benign and 2 for malignant.

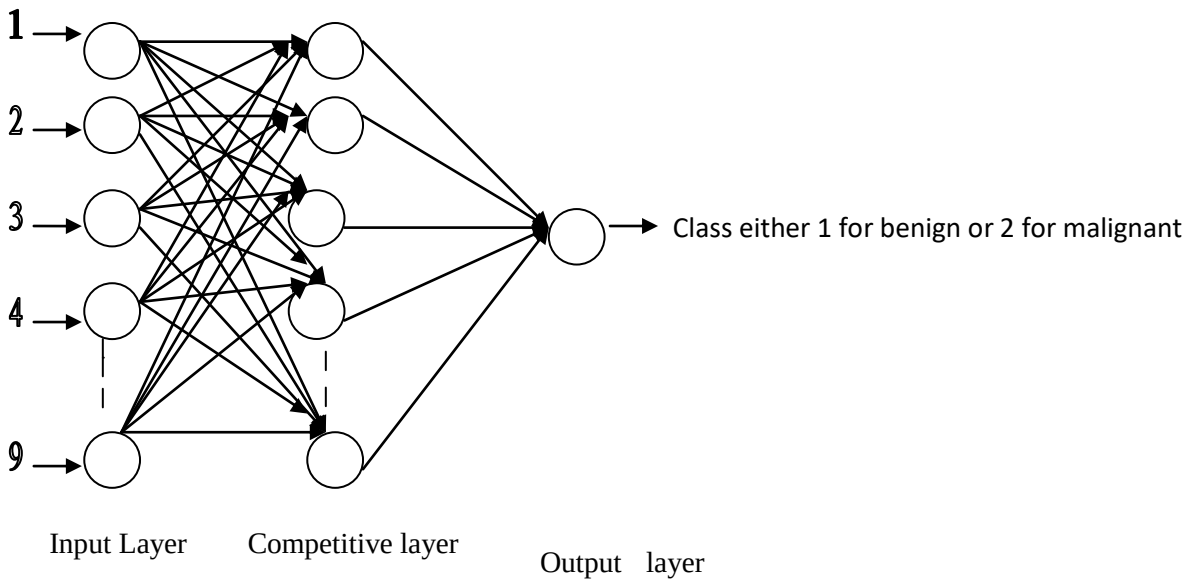
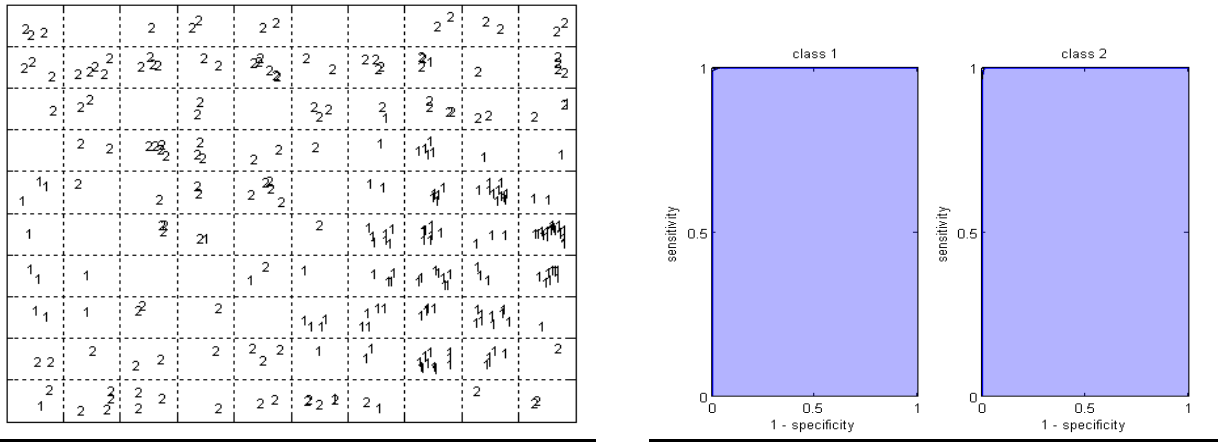


Figure :Architecture for Breast cancer data classification

(C) Counter Propogation Network (CPN):

Counterpropagation networks were proposed by Hecht Nielsen in 1987. They are multilayer networks based on the combinations of the input, output and clustering layers.The applications of counterpropagation nets are data compression, function approximation and pattern association. The counterpropagation network is basically constructed from an instar-outstar model. This model is three layer neural network that performs input-output data mapping, producing an output vector y in response to an input vector x, on the basis of competitive learning. The three layer in instar-outstar model are the input layer, the hidden (competitive) layer and the output layer. The connection between input layer and the competitive layer are the in-star structure, and the connection existing between the competitive layer and output layer are the out-star structure. The competitive layer is going to be a winner-take-all network or a Maxnet with lateral feedback connections. There exist no lateral connection within the input layer and the output layer. The connections between the layers are fully connected.

There are two stages involved in the training process of a counter propagation net. This input vector is clustered in the first stage. Originally, it is assumed that there is no topology included in the counter propagation network. However on the inclusion of linear topology ,the performance of the net can be improved. The clusters are formed using Euclidean distance method or dot product method. In the second stage of training, the weights from the cluster layer units to the output units are tuned to obtain the desired response. There are two types of counter propagation nets: (i) Full counter propagation net and (ii) forward-only counter propagation net.



(a) **(b)**

Figure : (a) Distribution of sample in kohpnen layer of CPN (b) ROC curve of CPN for two classes

2. Decision Tree :

(A)CART (Classification and Regression Tree) :CART classification algorithm which is based on decision tree induction (Jiwai H. and Micheline Kamber,2009) which is a learning of decision trees from class label training tuples . The Classification and Regression (CART) tree method uses recursive partitioning to split the training records into segments with similar output field values. The CART tree node starts by examining the input fields to find the best split, measured by the reduction in an impurity index that results from the split. CART uses Gini index splitting records measures in selecting the splitting attribute. Pruning is done in CART by using a training data set. The split defines two subgroups, each of which is subsequently split into two more subgroups, and so on, until one of the stopping criteria is triggered. All splits are binary (only two subgroups).

(B) C5.0 – This is a decision tree based classifier developed by Ross Quinlan (rulequest.com/see5-info.html,2010) and is an extension of C4.5 and ID3 decision tree algorithms .It automatically extracts classification rules in the form of decision tree from given training data .C5.0 has many benefits over C4.5 in terms of time and memory space required ,the tree generated by C5.0 is also very small as compared to C4.5 algorithm which ultimately improves the classification accuracy.

3.Statistical Technique: Most of the earlier work done by the researchers are based on statistical techniques ,some time results found using these techniques are better than the intelligent techniques .Two very well known statistical techniques are used here for breast cancer data classification ,these are :

(A) Support Vector Machine (SVM)**:** SVM is a robust classification and regression technique (Mitra & Acharya, 2004) that maximizes the predictive accuracy of a model without over fitting the training data. SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. Support vector machine (Kamber) uses a nonlinear mapping to transform the original training data into a higher dimension .Within this new dimension ,it searches for the linear optimal separating hyper plane. With an appropriate non linear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane. The SVM finds this hyper plane using support vectors and margins. The Support Vector Machines (SVM) are a general class of learning architectures, inspired by the statistical learning theory that performs structural risk minimization on a nested set structure of separating hyper planes.

(B) Bayesian Net: Bayesian net (Han, J., & Micheline, K., 2006) is statistical classifiers which can predict class membership probabilities, such as the probability that a given tuple belong to a particular class. Let X is a data sample whose class label is unknown. Let H be some hypothesis, such as that the data sample X belongs to a specified class C. For classification problems, we want to determine P (H|X), the probability that the hypothesis H holds given the observed data sample X.P (H|X) is the posterior probability, or a posteriori probability, of H conditioned on X. Bayesian classifier is very popular and applied for health care domain by many authors (Bendi V.R ,2011) (Alaa M.Elsayad,2010).

4. Ensemble Technique : An ensemble model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. The two models are combined by using high confidential wins scheme (Elsayad, et al.,2010) where weights are weighted based on the confidence value of each prediction. Then the weights are summed and the value with highest total is again selected. The confidence for the final selection is the sum of the weights for the winning values divided by the number of models included in the ensemble model. In this work many models with the combination of all the above individual models are tried and finally an ensemble model with combination of Bayesian net and C5.0 is selected, because this model produces highest accuracy among all the ensemble models as well as individual models.

Feature selection (Cios, K. J., et al., 2000) is an optimization process in which one tries to find the best feature subset, from the fixed set of the original features, according to a given processing goal and feature selection criteria, without feature transformation or construction.

There are benefits of feature selection techniques applied in health care predictive model ,because it will reduce number of test required to identify a disease and will be benefited to the patients in term of money and time consumed for tests. A rank based feature selection technique is applied here to decide the importance of features of health care data.

Model Evaluation: Based on data mining techniques as explained above all the developed models are evaluated in terms of following error measures -

Accuracy: Is a percentage of samples that are classified correctly **.**It is calculated as follows:

Accuracy = (TP + TN) / (P + N) (1)

Sensitivity :Is also known as true positive rate (TPR) which can be calculated as follows:

Sensitivity = TP/ (TP+FN) (2)

Specificity : Is also known as true negative rate (TNR). It is calculated as follows:

Specificity = TN/ (TN +FP) (3)

Where TP,TN, FP and FN are true positive ,true negative ,false positive and false negative respectively.

III. EXPERIMENTAL SETUP

Experimental work is carried out with the help of MATLAB and SPSS Clementine software ,initially

training and testing data are divided into 75% and 25% ratio respectively .Two different types of ANN based model are used for model building these are :Supervised ANN like EBPN , LVQ and unsupervised ANN like CPN .Training data are supplied to EBPN and results are obtained ,convergence curve for training EBPN with the help of MATLAB is shown in Figure armis trained with training data Data is supplied to each category of model one by one and a confusion matrix showing all the cases are obtained as depicted in table 2 for all the predictive models .Each cell of the table represents number of samples classified in that category .All the records are either classified as benign or malignant.

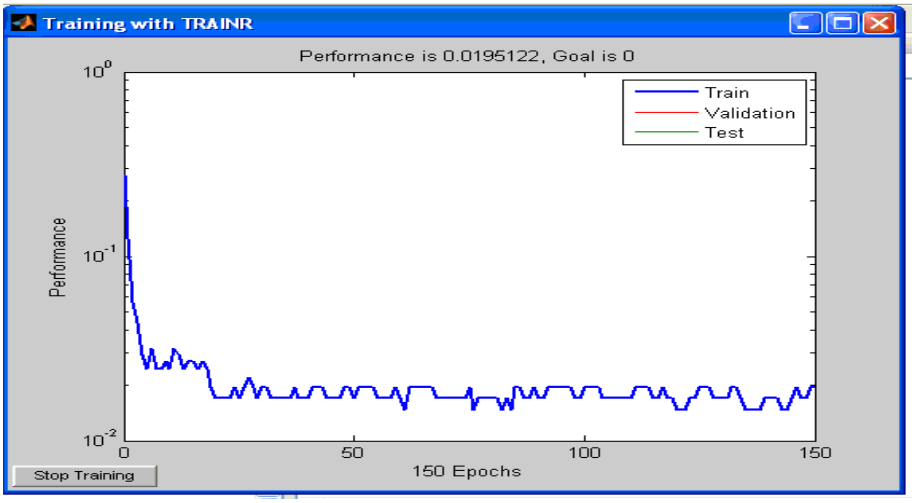


Figure 3:Convergence curve during training of EBPN

Table 2:Confusion matrix of various predictive model

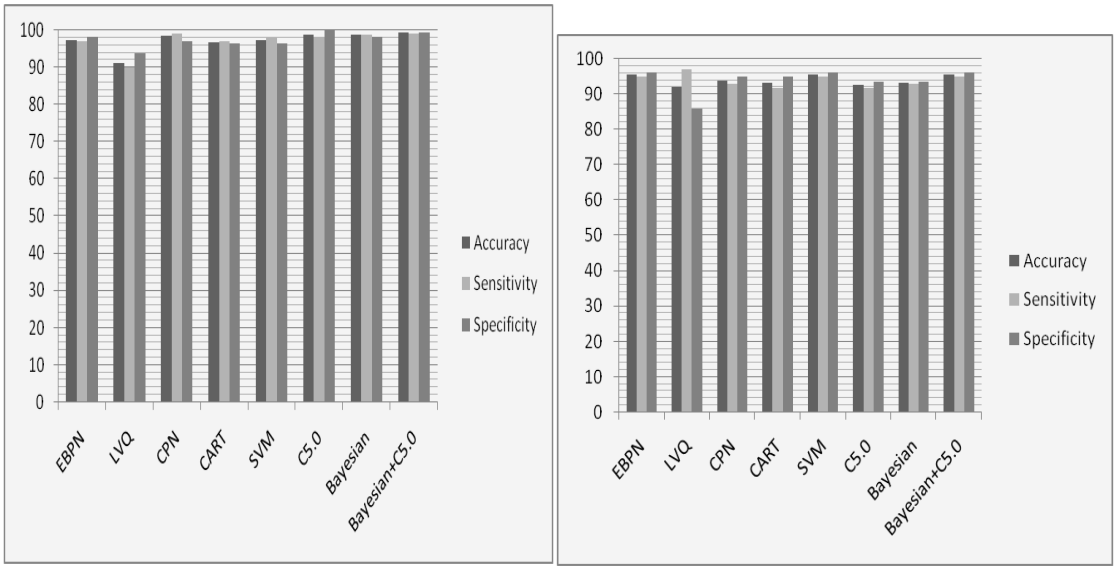
Predictive model	Target class	Training		Testing	
		Begin	Malignant	Begin	Malignant
EBPN	Begin	349	11	93	5
	Malignant	3	160	3	75
LVQ	Begin	324	36	95	3
	Malignant	10	153	11	67
CPN	Begin	357	3	91	7
	Malignant	5	158	4	74
CART	Begin	349	11	90	8
	Malignant	6	157	4	74
Bayesian Net	Begin	356	4	91	7
	Malignant	3	160	5	73
C5.0	Begin	354	6	90	8
	Malignant	1	162	5	73
SVM	Begin	352	8	93	5
	Malignant	6	157	3	75
Ensemble of Bayesian	Begin	357	3	93	5

Net and C5.0	Malignant	1	162	3	75

Further predictive models are evaluated as discussed in model evaluation section using equations 1,2 and 3 and calculated results are presented in table 3 in terms of accuracy, sensitivity and specificity. From this table it is clear that ensemble of Bayesian net and C5.0 is performing well as compared to other models. Accuracy, sensitivity and specificity in this case at training stage is 99.24%,99.16%.99.18% respectively while it is 95.45%,94.89%,96.15% respectively at testing stage.

Table 3:Error measures of various predictive models

Predictive Model	Partition	Accuracy	Sensitivity	Specificity
EBPN	Training	97.32	96.95	98.15
	Testing	95.45	94.89	96.15
LVQ	Training	91.20	90.00	93.86
	Testing	92.04	96.93	85.89
CPN	Training	98.47	99.16	96.93
	Testing	93.75	92.85	94.87
CART	Training	96.75	96.94	96.31
	Testing	93.18	91.83	94.87
SVM	Training	97.32	97.77	96.31
	Testing	95.45	94.89	96.15
C5.0	Training	98.66	98.33	99.83
	Testing	92.61	91.83	93.58
Bayesian	Training	98.66	98.88	98.15
	Testing	93.18	92.85	93.58
Ensemble of Bayesian and C5.0	Training	99.24	99.16	99.38
	Testing	95.45	94.89	96.15



(a) (b)

Figure: A comparative Bar Chart showing Error Measures of all classifiers for (a) Training and (b) Testing classifiers For selecting feature subsets a ranking based algorithm is applied and rank of the features of breast cancer data set as UCS, UCSh, BN, SDC, CT, NN, SECD, MA, M (Starting from higher rank) is

obtained, among all these features uniform cell size (UCS) has got highest rank.

A tested ensemble model is used to check for the various feature subsets after discarding feature with lowest rank one by one and a confusion matrix as shown in table 4 at training and testing stages both is prepared,

based on this table again error measures are calculated and presented in table 5 .Performance of the model is neither improving nor decreasing with feature subset in absence of feature 9 and 8.However there is an improvement in testing accuracy while discarding feature 9,8 and 7 , testing accuracy in this case is 96.59% which is 1.14% higher than the testing accuracy of data set with all 9 features .A comparative graph is also shown in figure 2 for all the measures in case of ensemble model for various feature subsets.

Table 4 :Confusion matrix for various feature subsets

No. of feature	Target class	Training		Testing	
		Benign	Malignant	Benign	Malignant
7	Benign	353	7	93	5
	Malignant	1	162	1	77
6	Benign	351	9	92	6
	Malignant	1	162	1	77
5	Benign	350	10	92	6
	Malignant	2	161	1	77
4	Benign	351	9	92	6
	Malignant	5	158	1	77
3	Begin	350	10	91	7
	Malignant	8	155	2	76

Table 5 : Error measures for various feature subsets

No. of Feature	Accuracy		Sensitivity		Specificity	
	Training	Testing	Training	Testing	Training	Testing
9	99.24	95.45	99.16	94.89	99.38	96.15
8	99.24	95.45	99.16	94.89	99.38	96.15
7	98.47	96.59	98.05	94.89	99.38	98.71
6	98.09	96.02	97.50	93.87	99.38	98.71
5	97.71	96.02	97.22	93.87	98.77	98.71
4	97.32	96.02	97.50	93.87	96.93	98.71
3	96.56	94.89	97.22	92.85	95.09	97.46

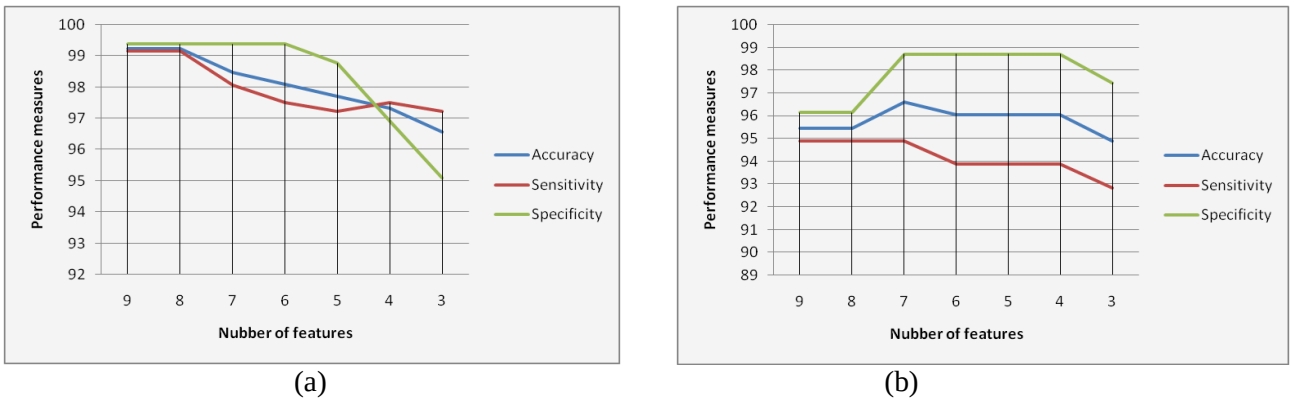


Figure 2 :A comparative graph in case of feature selection for ensemble model (a) At training stage (b) At testing stage

IV. CONCLUDING REMARKS

Health care classification is a crucial and essential task now a days for medical diagnosis. Data mining techniques provides facility to design and develop predictive model for health care classification.This paper explores various data mining techniques to design an ensemble model for classification of breast cancer related health care data. A testing accuracy of model show the efficiency of ensemble model. Models are also measured in terms of sensitivity and specificity. Feature subsets are obtained after applying ranking based feature selection

algorithm and models are tested on these data sets. Result show that ensemble model with 7 feature could be a best alternative as a health care predictive model for diagnosis of breast cancer.

Acknowledgement: This research work is supported by University Grant Commission (UGC), India under minor research project (No. F. 41-1357/2012(SR)).

REFERENCES

1. Bendi V.R., Prasad M. S. Babu andVenkateswarlu N. B.(2012), A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis, International Journal of Computer Science Issues, Vol.9. Issue 3 ,No. 2 ,PP 506-516.
2. Bendi V. R., Prasad M. S. Babu and Venkateswarlu N. B.(2011) ,A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosi, International Journal of Database Management Systems (IJDMS), Vol.3, No.2, PP 101-114.
3. Bendi V. R (2011),A Critical Evaluation of Bayesian Classifier for Liver Diagnosis using Bagging and Boosting Methods, International Journal of Engineering Sciences and Technology (IJEST) ,Vol.3,No. 4 PP 3422-3426.
4. Obi J.C. and Imainvan A.A(2011) ,Decision Support System for the Intelligent Identification of Alzheimer using Neuro Fuzzy Logic ,International Journal of Soft Computing (IJSC) ,Vol 2,No. 2 ,PP 25-38
5. Elsayad, A. M. (2010). Predicting the severity of breast masses with ensemble of Bayesian classifiers. Journal of Computer Science, 6(5), 576-584.
6. Jiawei Han, Kamber Micheline (2009). Data mining: Concepts and Techniques, Morgan Kaufmann Publisher.
7. "UCI Machine Learning Repository of machine learning database", University of California, school of Information and Computer Science, Irvine. C.A. http://www.ics.uci.edu/ last accessed on Dec 2012.
8. Simon Heykens (1999) .Neural Network comprehensive foundation " 2nd edition ,prentice hall.