# Identification of Breast Cancer Using Ensemble of Support Vector Machine and Decision Tree with Reduced Feature Subset

**H.S. Hota**

***Abstract*** ***Breast cancer is very common disease found in woman in which breast masses are increases abnormally .A recent survey in united kingdom proved that breast cancer is not only a problem of young woman but it is also a problem of old age woman those who have crossed the age of sixty or even seventy. An early identification and then prevention with proper medication of breast cancer can save life of human being. A robust and efficient breast cancer identification system is necessary for this purpose. Statistical technique like support vector machine and data mining technique like decision tree are widely used by the researcher since last few years. These techniques proved their ability to efficiently diagnose breast cancer problem. In this research work an ensemble model based on above two techniques are explored with special reference to feature selection. A rank based feature selection technique reduces features one by one based on its rank of breast cancer data ,downloaded from UCI repository site. An ensemble of support vector machine and C5.0 decision tree technique with reduced subset of only five features produced high accuracy of 92.59%.***

***Keywords*** ***Decision Tree (DT), C5.0, Support Vector Machine (SVM),Feature Selection (FS).***

## 1 INTRODUCTION

Identification of any disease like breast cancer with the help of expert system has many benefits[10] over any experienced doctor. Expert system never forget what it has learnt ,there is no depreciation or loss of knowledge which are acquired, and so on. Due to all these benefits, an expert system is an important need in current health care diagnosis process and for this a framework with suitable techniques is required to be designed. Decision tree techniques are used by many authors [2][3][8][9][12][13] due to its simplicity and pattern extraction capabilities .Results obtained by all of them are quit satisfactory . on the other hand statistical technique like support vector machine has also proved its capability in health care domain. Ramana et al. have utilized SVM for liver disease diagnosis.Four SVM modeling techniques including Multi level SVM (MLSVM) proposed by Zhong W. and et al.[15] are implemented and evaluated , performances are also compared on three large and complex real health care data sets. There is a wide range of literature in which statistical techniques are combined with data mining and soft computing techniques. However decision tree technique and SVM have their own weakness but these weaknesses can be avoided and their strength can be captured by integrating these two techniques as an ensemble model. Literature proves that ensemble model in health care diagnosis process is more efficient than others.

**Manuscript received February, 2014.**
**H.S. Hota**, Guru Ghasidas Central University,(C.G.),India

This research work in intended to develop a model based on ensemble of SVM and C5.0 for identification of breast cancer, which is very commonly found in woman in the young age as well as in old age. Many decision tree (DT) based algorithms like C4.5,Classification and Regression Tree (CART),Quick, Unbiased and Efficient Statistical Tree (QUEST) and Chi-Squared Automation Interaction Detection (CHAID) are tried as individual as well as all possible combinations of ensemble and results are compared in terms of accuracy ,sensitivity and specificity. Less number of features may be beneficial in order to develop a health care identification model since it will require less number of pathological and other tests. A feature selection techniques [8][9] can find out relevant features from medical data and removes unwanted or irrelevant features to improve the overall accuracy of the model. A rank based feature selection technique is applied in this research work to reduce features form breast cancer data downloaded from UCI repository site with 33 features in all .After reducing features one by one a feature sub set with only five features are selected and supplied to ensemble of SVM and C5.0 model which produces 92.59%accuracy .Experimental works for developing and testing the models are carried out using Clementine data mining toll under windows environment.

## 2 RESEARCH METHODOLOGY

Decision tree (DT) (Han et al., 2011) is a most popular and powerful classification technique where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The top most nodes in a tree are the root nodes. DT is so popular because construction of DT classifiers does not require any domain knowledge or parameter setting and, therefore, is appropriate for exploratory knowledge discovery. Various decision tree based techniques are widely accepted and applied on health care diagnosis process. On the other hand some statistical technique like SVM is also used as classifier for health care diagnosis.

Interactive Dichotomiser 3 (ID3) and C4.5 are the two very popular DT algorithms proposed by Quinlan [6]. C4.5 is a successor of ID3 in terms of removing the problem of bias over attributes of higher values in ID3,both algorithms split the tree based on features available in the data set using information gain ratio. After constructing DT with training samples, testing samples can be used to check the performance of classifier. C4.5 is again improved and a new technique as C5.0 is suggested by Quinlan [6] .On the other hand, classification and regression tree (CART) [1] and Chi-Squared Automation Interaction Detection (CHAID) are other classifications algorithms which are based on DT induction [4].CART method uses recursive partitioning to



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

split the training records into segments with similar output field values using Gini index. While CHAID DT is constructed by partitioning the data set into two or more data subsets, based on the values of one of the non-class attributes, the number of subsets in a partition can range from two up to the number of distinct values of the splitting attribute. In this regard, CHAID differs from CART, which always forms binary splits (two subgroups). In a nut shell, we can conclude that ID3 and C4.5 uses Entropy based information gain, CART uses the Gini index and CHAID uses the chi-squared test for splitting and constructing DT.

### *2.1 C5.0*

This is a decision tree based classifier developed by Ross Quinlan [16] 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.

### *2.2 Support Vector Machine (SVM)*

SVM is based on statistical learning theory of Vapnik [5].SVM is a robust classification and regression technique 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 [4] 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.

### *2.3 Ensemble Model*

An ensemble model is a combination of two or more models to avoid the drawbacks of individual and to achieve high accuracy. The two models are combined by using high confidential wins scheme[11] 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.

### *2.4 Feature Selection*

Feature subset selection[17] is an important problem in knowledge discovery, not only for the insight gained from determining relevant modeling variables, but also for the improved understandability, scalability, and possibly, accuracy of the resulting models. In the feature selection, the main goal is to find a feature subset that produces higher classification accuracy. Feature selection technique with feature ranking is applied to select best feature subset. The simple feature selection procedure is based on evaluate of classification power of individual features, then ranking such evaluated features, and eventually selecting the first best features. A criteria applied to an individual feature could be of either of the open-loop or closed-loop type. This also relies on an assumption that the final selection criterion can be expressed as a sum or product of the criteria evaluated for each feature independently. We can expect that a single feature alone have a low classification power. However, this feature, when put together with others, may exhibit substantial classification powers.

## 3 BREAST CANCER IDENTIFICATION THROUGH ENSEMBLE MODEL

Breast cancer is a very common disease commonly found in women, the abnormal growth of cells in the breast is the main cause of breast cancer. A proper and prior diagnosis of this disease is essential for medications.

### *3.1 Breast Cancer Data*

To develop ensemble model breast cancer data set available in UCI repository site [7] is considered. The data set has 198 patterns with 34 features, out of which 32 are real valued features and one is identification and another one is output with 151 nonrecurring and 47 recurrent patterns. Patterns are categorized either as recurrent or nonrecurring . Patterns are divided into three different size of training and testing part to check the efficiency of model in terms of partition size. In each partition larger size is used for training the models and smaller size is used for testing of models.

### *3.2 Model Measurement*

Effectiveness of classification model developed by many authors [2][3][8][9][11][12][13][14] have been tested using three well known measures as shown in equation 1 ,2 and 3. These measures are defined by true positive (TP), true negative (TN), false positive (FP) and false negative (FN) under positive (P) and negative (N) cases .

$$\text{Accuracy} = (TP + TN) / (P + N) * 100 \quad (1)$$

$$\text{Sensitivity /Recall} = TP / (TP + FN) * 100 \quad (2)$$

$$\text{Specificity} = TN / (TN + FP) * 100 \quad (3)$$

These measures can be calculated using above equations after obtaining confusion matrix in terms of TP,TN,FP,FN. Accuracy measures [8][9] clearly reflect the classification ability of the model but may produce bias evolution results; this is why models can be verified with some other measures like sensitivity and specificity.

### *3.3 Ensemble of SVM and C5.0*

In this work many models with the combination of all the individual models are tried and finally an ensemble model with combination of SVM and C5.0 is selected, because this model produces highest accuracy among all the ensemble models as well as individual models. Figure 1 show the pictorial representation of combining two models to develop ensemble model as C5.0 and SVM. Training and testing data are supplied respectively for model development and model testing for all individual as well as ensemble model. Patient pathological and other test data are supplied to classify it as recurrent or nonrecurrent.



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

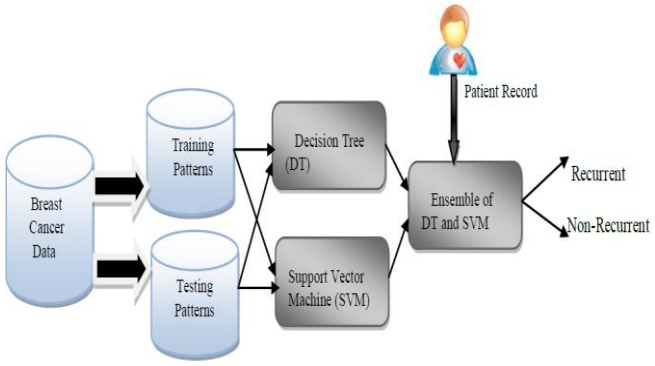


Figure 1: Ensemble Model for breast cancer identification

## 4 EXPERIMENTAL STUDIES

Experimental work is carried out using Clementine data mining tool. This tool is very popular among the researchers to be utilized for classification ,predication and association rule mining.

### *4.1 Model Development*

Figure 2 show the flow diagram designed for development of ensemble of SVM and C5.0 as breast cancer classifier. Data source is provided to the flow stream as wpbc-breast available in form of excel file with all its features. Data node is then connected through filter node, which filtered ID feature from the data set , as there is no role of this feature in classification process. Filter node is then attached with type node which decides the input features as well as output features of the data set. Data partition node used to generate partitions as decided by the developer as training and testing parts. In this study data are partitioned by the ratio 75:25% ,80:20% and 90:10% for training and testing respectively. These partitioned data are supplied to individual decision tree techniques :CART, CHAID,QUEST,C5.0 as well as statistical technique SVM as shown in figure by associating respective nodes with partition node. Two generated nuggets: C5.0 and SVM are combined and an ensemble model is generated, all these models are analyzed through analysis node. Obtained results are presented in table 2 which show that ensemble of SVM and C5.0 with 90-10 partition produces 88.89% accuracy in case of testing patterns.

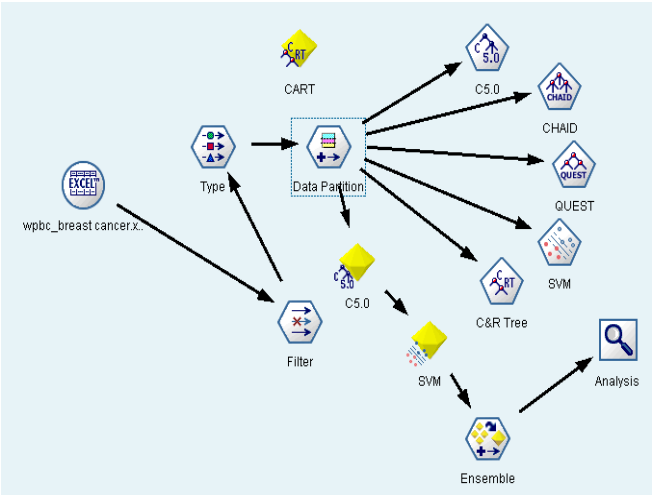


Figure 2: Model developed using SPSS Clementine data mining tool

Table 2 : Accuracy of various individual and ensemble models with different partition size

| Technique | 75-25% Partition | 80-20% Partition | 90-10% Partition |
|---|---|---|---|
| CART | 80.0 | 81.25 | 70.37 |
| C5.0 | 80.0 | 77.08 | 81.48 |
| CHAID | 70.91 | 72.92 | 77.78 |
| QUEST | 72.73 | 83.33 | 81.48 |
| SVM | 78.18 | 77.08 | 81.48 |
| CART+SVM | 83.64 | **85.42** | 85.19 |
| C5.0+SVM | 83.64 | 81.25 | **88.89** |

### *4.2 Rank Base Feature Selection*

Applying feature selection technique is beneficial in terms of accuracy produced by the models in health care diagnosis process. A DSS with reduced set of features can improve the performance of the system while on the other hand it is also beneficial for the patient in terms of number of tests required to collect. A rank based feature selection technique was applied in this research work to extract irrelevant features from the data set. After applying feature selection technique on breast cancer data set, ranking of the features are obtained and are reduced. Features of data set are extracted one by one starting from lowest rank to obtain new feature subsets as shown in table 3.

It can be clearly observed and analyzed that applying feature selection technique on breast cancer data set has many advantages in terms of various pathological tests required to diagnose the disease. Developed ensemble model is performing well with 92.52 % accuracy ,100 % sensitivity and 60 % specificity (As calculated using equations 1,2 and 3) with only five features. However model is not well suited in terms of specificity but two out of three measures are producing better results. Comparative results are also shown in form of bar chart in figure 2.

Table 3 :Performance of ensemble models with reduced feature subsets

| Number of feature | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| 32 (Excluding ID) | 88.89 | 95.45 | 60 |
| 31 | 88.89 | 95.45 | 60 |
| 29 | 88.89 | 95.45 | 60 |
| 27 | 85.19 | 90.90 | 60 |
| 25 | 81.48 | 81.81 | 80 |
| 23 | 81.48 | 81.81 | 80 |
| 21 | 85.19 | 90.90 | 60 |
| 19 | 85.19 | 90.90 | 80 |
| 17 | 85.19 | 90.90 | 60 |
| 15 | 88.89 | 90.90 | 80 |
| 13 | 81.48 | 81.81 | 80 |
| 11 | 85.19 | 90.90 | 60 |
| 9 | 85.19 | 90.90 | 60 |
| 7 | 85.19 | 90.90 | 60 |
| 5 | **92.59** | **100** | **60** |
| 3 | 81.48 | 81.81 | 80 |



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

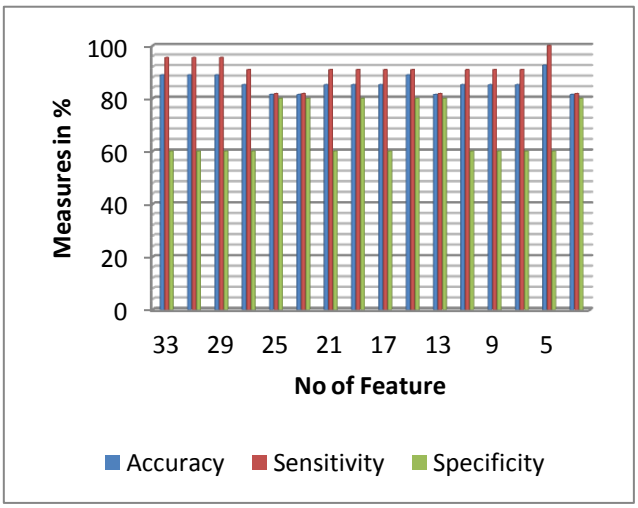


Figure 2: A comparative bar chart showing different measures for ensemble of C5.0 and SVM

## 5 CONCLUSION

Identification of any disease like breast cancer is a challenging task and requires lots of experience ,same can be done with the help of expert system developed through decision tree and statistical technique. This paper presents utilization of these techniques to develop ensemble model. Experimental work is performed through Clementine software. An ensemble model as SVM and C5.0 is developed, further ranking based feature selection technique was applied to reduce feature. A desired model with only five features is producing 92.52% accuracy, 100 % sensitivity and 60 % specificity. In future, an integrated decision support system (DSS) will be developed for identification of many life threatening diseases to assist physicians as well as to medical students.

## ACKNOWLEDGEMENT:

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

## REFERENCES

[1] Breimen.R. & Anand.T. The process of knowledge discovery in databases:A *human centered approach. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds),Advances in Knowledge Discovery and Data Mining.Cambridge:MIT press,1996.*

[2] Kurt, Imran,Ture, Mevlut, and Kurum, A. Comparing performance of logistic regression,classification and regression tree and neural network for predicting coronary artery disease. Expert Systems with Application34(1),366-374,2008.

[3] Gupta, S., Kumar, D., and Sharma, A.(2011).Performance analysis of various data mining classification techniques on health care data. International Journal of Computer Science and Information Technology,3(4),155-169,2011.

[4] Jiawei Han, Kamber Micheline. Data mining: Concepts and Techniques, Morgan Kaufmann Publisher,2009.

[5] Vapnik, V.: The nature of statistical learning theory. Springer, New York ,1995.

[6] Quinlan.J.R. (1993).C4.5:Programs for machine learning (1st edition), San Francisco, Morgan Kaufmann Publishers,1993.

[7] UCI (2014).Web source:http://archive.ics.uci.edu/ml/datasets.html,last accessed on Jan 2014.

[8] Dinesh K. Sharma, Hari, S.Hota ,"Development of rule base system using intelligent techniques to diagnose life threatening diseases ",Proceeding published in review of business and technology research ,Vol. 9 ,No. 1 ,Pp 14-19,2013.

[9] Hari S.Hota , "Data mining techniques for effective and intelligent health care predictive model " Proceeding published in review of business and technology research ,Vol. 8 ,No. 1 ,Pp 143-149,2012.

[10] Ali K.,Ayturk K.,Ugur Y. "Expert system based on neuro-fuzzy rules for diagnosis breast cancer ",Expert system with applications Vol. 38,Pp 5719-5726,2011.

[11] Elsayad, A. M. (2010). Predicting the severity of breast masses with ensemble of Bayesian classifiers. Journal ofComputer Science, 6(5), 576-584,2010.

[12] Bendi V.R., Prasad M. S. Babu and Venkateswarlu N. B, "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 506-516,2012.

[13] Bendi V. R., Prasad M. S. Babu and Venkateswarlu N. B. "A Critical Study of Selected ClassificationAlgorithms for Liver Disease Diagnosis", International Journal of Database Management Systems (IJDMS), Vol.3,No.2, PP 101-114,2011.

[14] Bendi V. R ,"A Critical Evaluation of Bayesian Classifier for Liver Diagnosis using Bagging and BoostingMethods", International Journal of Engineering Sciences and Technology (IJEST) ,Vol.3,No. 4 PP 3422-3426,2011.

[15] Wei Z. ,Rick C.,Jieyue H. "Clinical charge profiles prediction for patients with chronic disease using multi-level support vector machine" Expert systems with application ,Vol. 39,Pp. 1474-1483,2012.

[16] Web source www.rulequest.com/see5-info.html, last accessed on Jan 2014.

[17] Wang, J. (2003). Data Mining: opportunities and challenge, Idea Group, USA.



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*