# Empirical Study on Performance of Decision Trees (CART) and Ensemble Methods in Medical Diagnosis

Rahul Samant,
*SVKM'S NMIMS, Shirpur Campus, India;*

Srikantha Rao,
*TIMSCDR, Mumbai University, Kandivali, Mumbai, India,*

## Abstract

*This paper investigates the ability of decision trees and ensemble methods to predict the probability of occurrence of Hypertension and Diabetes in a mixed patient population. A detailed database comprising healthy, hypertensive and diabetic patients from a university hospital was used for constructing the decision trees using CART algorithm. Ensemble algorithms such as bagging and multiple versions of boosting were used to improve the performance of basic CART algorithm for building various classification models for prediction of medical diagnosis. The measure of percentage misclassification error was considered to determine the effectiveness of classifier model. Even though CART shows acceptable classification error for the given datasets, ensemble methods such as bagging still improves the performance by building multiple trees.*

## 1. Introduction

Decision tree is one of the classifying and predicting data mining techniques, belonging to inductive learning and supervised knowledge mining. It can generate easy-to-interpret If-Then decision rule, it has become the most widely applied technique among numerous classification methods [1]. Decision tree is a tree diagram based method, the node on the top of its tree structure is a root node and nodes in the bottom are leaf nodes. Target class attribute is given to each leaf node. From root node to every leaf node, there is a path made of multiple internal nodes with attributes. This path generates rule required for classifying unknown data. Moreover, most of decision tree algorithms contain two-stage task, i.e., tree building and tree pruning.

In tree building stage, a decision tree algorithm can use its unique approach (function) to select the best attribute, so as to split training data set. The final situation of this stage will be that data contained in the split training subset belong to only one certain target class. Recursion and repetition upon attribute selecting and set splitting will fulfill the construction of decision tree root node and internal nodes. On the other hand, some special data in training data set may lead to improper branch on decision tree structure, which is called over-fitting. Therefore, after building a decision tree, it has to be pruned to remove improper branches, so as to enhance decision tree model accuracy in predicting new data [2]. Among developed decision tree algorithms, the commonly used ones include ID3 [3], C4.5 [4], CART [5] and CHAID [2]. C4.5 was developed from ID3 (Iterative Dichotomiser 3) algorithm, it uses information theory and inductive learning method to construct decision tree. C4.5 improves ID3, which cannot process continuous numeric problem. CHAID algorithm is featured in using chi-square test to calculate p-value of node category in every splitting, so as to determine whether to allow decision tree to grow without pruning. CHAID cannot process continuous data, so it is not applicable to many medical issues with continuous numeric data. CART algorithm is a binary splitting method, applied in data whose attributes are continuous. Gini index is used to evaluate data discretion as basis of choosing splitting condition. The Gini index of a node is

$$1 - \sum_i p^2\,(i) \qquad (1)$$

Where, the sum is over the classes $i$ at the node, and $p(i)$ is the observed fraction of classes with class $i$ that reach the node. A node with just one class (a *pure* node) has Gini index 0; otherwise the Gini index is positive. So the Gini index is a measure of node impurity [1].

Since this study is to process medical data with multiple attributes, CART is chosen as the decision tree algorithm. The decision tree algorithm has been applied in many medical tasks.

*B. Ensemble Methods*

We have investigated the performance of the following ensemble techniques: Bagging and Boosting.

- *Bagging* (bootstrap aggregating), generates a collection of new sets by re-sampling the given training set at random and with replacement. These sets are called *bootstrap samples*. New classifiers are then trained, one for each of these new training sets. They are amalgamated via a majority vote, [3]. Bagging is probably the most widely used ensemble method [1].

- *Boosting* trains several classifiers in succession. Every next classifier is trained on the instances that have turned out more difficult for the preceding classifier. To this end all instances are assigned weights, and if an instance turns out difficult to classify, then its weight increases. In Boosting, we used *AdaBoost, LogitBoost, GentleBoost* and *RobustBoost* techniques to analyze the performance and effectiveness of each method using four different datasets.

## 2. Literature Review

Marsala et al. [6] developed the solution for the problem of the construction of fuzzy decision trees when there exists a graduality between the values of attributes and values of the class. They proposed a new measure, extended from the measure of classification ambiguity, that takes into account both discrimination power and graduality with regards to the class and validate it by presenting a Medical application.

Chen et al. [8] used an algorithm of decision trees, Chi-squared automatic interaction detection (CHAID), to build a classifier for predicting breast cancer and fibro-derma. The results demonstrate that the decision tree technique was more favorably than logistic regression in terms of rule accuracy and knowledge transparency to physicians.

Ture et al. [7] compared performances of three decision trees, four statistical algorithms, and two neural networks in order to predict the risk of essential hypertension disease. MLP and RBF—two neural networks procedures—performed better than other techniques in predicting hypertension.

Yang et al. [9] proposed a recommender system based approach based on a hybrid method using multiple classifications models for Chronic Disease Diagnosis (CDD). Multiple classifications based on decision tree algorithms were applied to build an accurate predictive model that predicts the disease risk diagnosis for the monitored cases. They reported respectable accuracy for the diagnosis system.

Pogorelc et al. [10] proposed novel features for training a machine learning classifier that classifies the user's gait into four health problems and a normal health state. Decision tree classifier was able to reach 95% of classification accuracy using 7 tags and 5 mm standard deviation of noise. Neural network outperformed it with classification accuracy over 99% using 8 tags with 0-20 mm noise.

Rao, V.S.H. at el.[11] developed an alternating decision tree method that employs boosting for generating highly accurate decision rules. The predictive models were found to be more accurate than the state-of-the-art methodologies used in the diagnosis of the Dengue Fever.

Samant et al. [12, 13, 14] developed a medical decision support system to predict diseases such as hypertension and diabetes, using three alternate structures of artificial neural networks (ANNs) and five different kernel functions of support vector machines (SVMs). ANN models reported up to 90% prediction accuracy and SVMs models showed prediction accuracy of 92%. The study was based on pathological parameters as symptom parameters to predict deceases.

## 3. Experiments

The database used for analysis in this study has been compiled as a part of an earlier study entitled Early Detection Project (EDP) conducted at the Hemorheology Laboratory of the erstwhile Inter-Disciplinary Programme in Biomedical Engineering at the School (now Department) of Biosciences and Bioengineering, Indian Institute of Technology Bombay (IITB), Mumbai, India. Spanning over a period from January 1995 to April 2005, it compiled 981 records, each with 13 parameters, which encapsulated the biochemical, hemorheological and clinical status of the individuals. We note that the Hemorheology Laboratory has pioneered the research in the field of Clinical Hemorheology by conducting the baseline hemorheological studies in the Indian population and correlating various hemorheological parameters with several disease conditions.

In all, 13 parameters were noted for each respondent. Table 1 describes the symptom (input) variables used for the present study. They include age, health indicators (e.g. systolic blood pressure (BP1), diastolic blood pressure (BP2)) and biochemical parameter like Serum Proteins (SP), Serum Albumin (SALB), Hematocrit (HCT), Serum Cholesterol (SC), Serum Triglycerides (STG), along with various hemorheological (HR) parameters (e.g.; Whole Blood Viscosity(CBV), Plasma Viscosity(CPV), using a Contraves 30 viscometer, and Red Cell Aggregation (RCA). The CART algorithm was used to build a decision tree with thirteen input variables that would yield the best classification of individuals into these diseases categories.

For inputs to the decision tree, the first 13 columns of data represent the patient's health parameters. The 14th column represented the diagnosis made by the doctor for the patient. The diagnosis is a four value parameter. It can have values such as Hypertensive, diabetic, healthy and hypertensive as well as diabetic. We have divided the original dataset into four different subsets. Dataset DS-1 is a data set, having samples of unhealthy and healthy patients. Dataset DS-2 is a dataset which stores data about hypertensive and diabetic patients. DS3 is a dataset, having diagnosis information about patients who are healthy and hypertensive. Dataset DS4 is a mixed dataset about healthy, hypertensive and diabetic patients.

**Table 1: Description of the datasets**

| Dataset | Description | Class labels |
|---|---|---|
| DS-1 | Healthy vs. NotHealthy | 2 |
| DS-2 | Healthy vs. Diabetic(DT) | 2 |
| DS-3 | Healthy vs. Hypertensive(HT) | 2 |
| DS-4 | Mixed {Healthy vs. DT vs. HT vs. HT+DT} | 4 |

## 4. Results and Discussion

The importance of features for constructing a decision tree using CART algorithm for dataset DS-1 is shown in figure 1. The algorithm used is Gini Index. All the thirteen parameters used in the study were influential on the diagnosis of the patients.

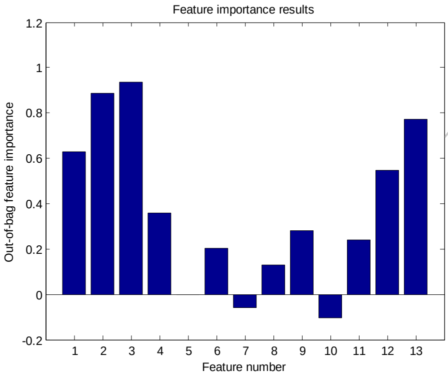


**Fig. 1 Important features listed by CART algorithm for dataset DS-1**

Table 2 lists the misclassification error for the four datasets used in our study. Then minimum error rate was for dataset DS-2 (Healthy vs. Diabetic) – 7.95% and maximum error rate is for DS-4 (mixed diagnosis) – 23.9%. This is expected as first three datasets, namely, DS-1, DS-2 and DS-3 were having only two class labels but fourth dataset, DS-4 consisted of four class labels.

**Table 2: Performance results of CART algorithm.**

| Dataset | % Misclassification error |
|---|---|
| DS-1 | 21.19 |
| DS-2 | 7.95 |
| DS-3 | 14.08 |
| DS-4 | 23.93 |

**Table 3: Classification error in different ensemble classifier methods**

| Dataset | Ensebmle Agorithm | Num. of trees | % Misclassification error |
|---|---|---|---|
| DS-4 | Bag | 42 | 1.0 |
| DS-1 | AdaBoostM1 | 50 | 10.0 |
| | LogitBoost | 95 | 5.0 |
| | GentleBoost | 95 | 2.0 |
| | RobustBoost | 80 | 3.0 |
| | Bag | 15 | 1.0 |
| DS-2 | AdaBoostM1 | 52 | 1.0 |
| | LogitBoost | 62 | 1.0 |
| | GentleBoost | 35 | 1.0 |
| | RobustBoost | 50 | 3.0 |
| | Bag | 12 | 1.0 |
| DS-3 | AdaBoostM1 | 90 | 2.0 |
| | LogitBoost | 90 | 1.0 |
| | GentleBoost | 62 | 1.0 |
| | RobustBoost | 28 | 1.5 |
| | Bag | 12 | 1.0 |

Table 3 lists the number of trees in each ensemble algorithm to reach the acceptable level of mis-classification error. Bagging technique was used with Dataset, DS-4 as it was having multiclass labels in the diagnosis.

Fig 3 showed the decision tree (Pruned to level 3) for the dataset, DS-2, consisted of healthy and diabetic patients. The decision tree showed contribution of almost all the selected parameters to the diagnosis. Fig 4,5,6 and 7 showed the cross validation error and re-substitution error plot for the CART algorithm for the different datasets used in the study. Cross validation error was calculated over 10 fold data. Re-substation error is estimate of only training error. So with more training, re-substitution error reduces. Better measure of prediction accuracy is cross-validation error. The widening gap between the cross validation error and re-substitution error is due to over-fitting.

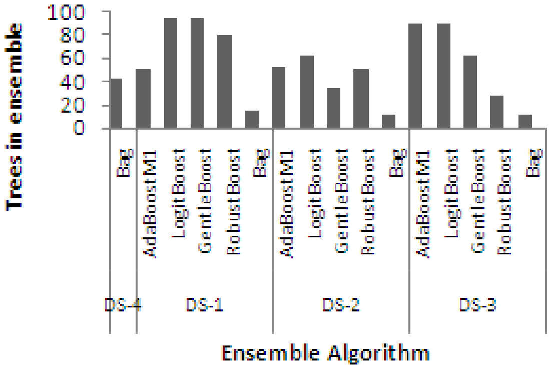
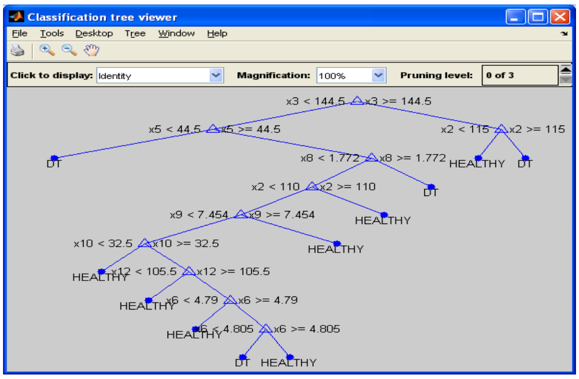


**Fig. 2 Performance of Ensemble algorithms with classification error <0.01**



Classification tree viewer

Click to display: Identity

Magnification: 100%

Pruning level: 0 of 3

**Fig. 3 Decision tree (Pruned to level 3) plotted for DS-2 (Healthy patients vs. Diabetic patients)**

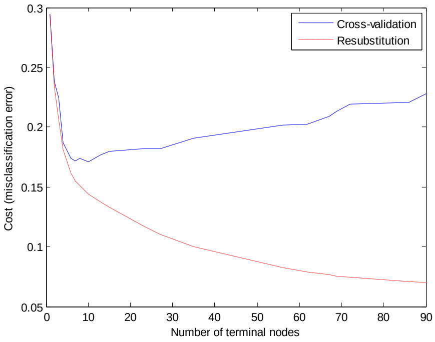


**Fig. 4 Performance curve of CART for dataset DS-4**

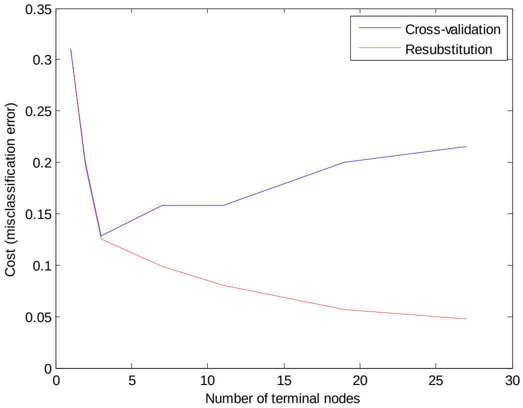


**Fig. 5 Performance curve of CART for dataset DS-1**

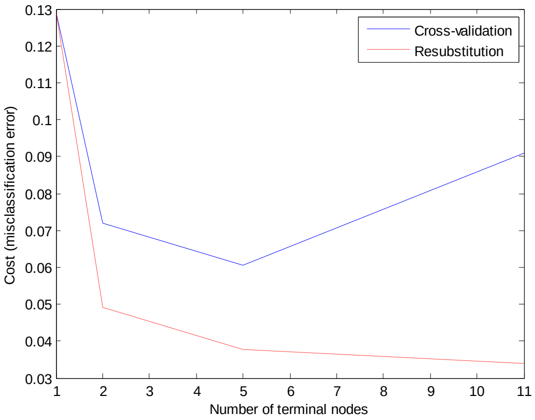


**Fig. 6 Performance curve of CART for dataset DS-2**

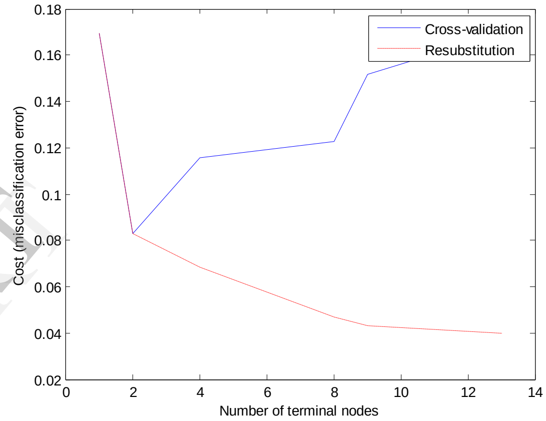


**Fig. 7 Performance curve of CART for dataset DS-3**

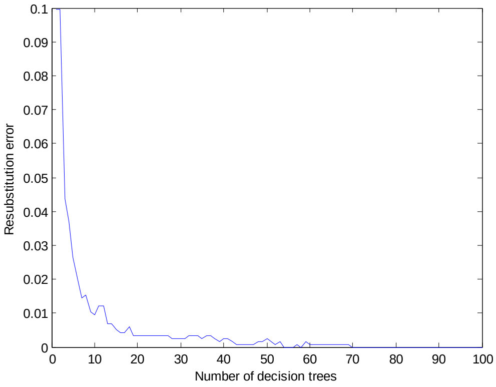


**Fig. 8 Performance of Bagging for multiclass classification ensemble for dataset DS-4**

Figure 8 plots the graph of number of decision trees built in the bagging ensemble vs. re-substitution error. Re-substitution error was almost zero when the number of decision trees reached the count of 70.

## 5. Conclusion

In this paper, we presented a novel medical decision support system built using decision trees and ensemble methods for early detection of hypertension and diabetes using pathological data. Bagging and boosting techniques in decision trees performs satisfactorily. Bagging algorithm required lesser number of trees than boosting algorithms in the ensemble to reach the minimum level of error. Among the boosting algorithms used in the study, *RobustBoost* algorithm performs the best. This paper presents promising results in detecting the occurrence of hypertension and diabetes using decision tree and ensemble methods. We can extend this work by using multi-level classifiers. In future we may extend this work by building optimized models using more sample sets.

## 6. References

[1] Han, J and M Kamber (2006). *Data Mining Concepts and Techniques*, 2nd ed., Morgan Kaufman.

[2] Quinlan, J. R. (1993). C4.5: Programs for machine learning. San Francisco: Morgan Kaufmann.

[3] Maher, P. E., & Clair, D. S. (1993). Uncertain reasoning in an ID3 machine learning framework. In Proceedings of the 2nd IEEE international conference on fuzzy systems, FUZZ-IEEE'93 (Vol. 1, pp. 7–12).

[4] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Belmont, CA: Wadsworth International Group.

[5] Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29(2), 119–127.

[6] Marsala, C. (2012), Gradual fuzzy decision trees to help medical diagnosis, IEEE International Conference on Fuzzy Systems, 2012, pp-1-6

[7] Ture, M., Kurt, I., Kurum, A. T., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications, 29(3), 583–588.

[8] Mu-Chen Chen(2006) ; Hung-Chang Liao ; Cheng-Lung Huang (2006) Predicting Breast Tumor via Mining DNA Viruses with Decision Tree , 2006. SMC '06. IEEE International Conference on Systems, Man and Cybernetics, Vol :5, 2006, pp 3585-3589

[9] Qiao Yang ; Shieh, J.S. ,(2008) A multi-agent prototype system for medical diagnosis ISKE 2008. 3rd International Conference on Intelligent System and Knowledge Engineering Vol: 1, 2008, pp-1265-1270

[10] Pogorelc, B. Gams, M. (2010), Diagnosing health problems from gait patterns of elderly, , 2010 Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBC), pp: 2238 - 2241

[11] Rao, V.S.H. ; Kumar, M.N. (2012), A New Intelligence-Based Approach for Computer-Aided Diagnosis of Dengue Fever , IEEE Transactions on Information Technology in Biomedicine, Volume: 16 , Issue: 1 , Page(s): 112 – 118

[12] Rahul Samant and Srikantha Rao.(2013) ,Evaluation of Artificial Neural Networks in Prediction of Essential Hypertension. *International Journal of Computer Applications* 81(12):34-38, November 2013. Published by Foundation of Computer Science, New York, USA

[13] Rahul Samant, Srikantha Rao,(2013) Performance of Alternate Structures of Artificial Neural Networks in Prediction of Essential Hypertension, *International Journal of Advanced Technology & Engineering Research (IJATER)*Volume 3, Issue 6, Nov. 2013 ISSN No: 2250-3536 pp:22-27

[14] Rahul Samant, Srikantha Rao,(2013) A study on Comparative Performance of SVM Classifier Models with Kernel Functions in Prediction of Hypertension, *International Journal of Computer Science and Information Technology* ,ISSN 0975 - 9646,VOLUME 4 ISSUE 6 November - December 2013, pp : 818-821