**J**ournal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 14, Issue. 2, No. 2: 2023
ISSN : **1906-9685**


# PREDICTION OF CORONARY HEART DISEASE RISK FOR OCCUPATIONAL DRIVERS USING DECISION TREE ANALYSIS

**Dr.S.Akila** Assistant Professor & Head, Department of Computer Science, Vellalar College for Women,Thindal, Erode-638012,Tamilnadu, India.
**Dr.K.Rajendran** Associate Professor,Department of Electronics,LRG Government Arts College for Women,Tirupur-641604,Tamilnadu, India.

## Abstract

Coronary Heart Disease (CHD) is a common disease which is closely linked with lifestyle related behaviors. Correct and in time diagnosis is very important to prevent from death. The most accurate Angiography diagnosis method has side effects and it is costly. The existing studies have used wide range of computational methods and tools for analyzing medical data. In the present study effective data mining methods have been applied to develop CHD prediction model in a cost effective manner. The aim of the study is to develop a rapid and automated prediction of CHD risk by analyzing the physical and biochemical factors using C4.5 algorithm with 10-fold cross validation. This method provides early intervention of CHD and reduces the risk for occupational drivers. In this work, information gain plays a vital role in determining the root and nodes of the decision tree. The proposed method has yielded good accuracy of 98.66%. In order to evaluate the performance of the proposed method sensitivity and specificity are analyzed which helps to reduce further invasive CHD risk examination cost for the individuals.

**Keywords:** Coronary Heart Disease; Ischemic heart disease; Risk Prediction; Decision Tree Classification; Receiver Operating Characteristics; 10-fold cross validation

## Introduction

Coronary Heart Disease (CHD) is a serious medical emergency in which the supply of blood to an area of the heart is inadequate. This inadequacy is due to the development of plaques within the walls of one or more coronary arteries resulting in narrowing of lumen of coronary arteries (Muthukaruppan and Er, 2012). Under this circumstance the supply of the oxygenated blood to the heart decreases leading to myocardial ischemia and subsequently myocardial infarction (MI) or sudden cardiac death. The majority of CHD is caused by risk factors that can be controlled, treated or modified, such as high blood pressure, blood sugar, cholesterol, obesity and lack of physical activity (Susan et al.,2012; Katzmarzyk et al., 2006; Song et al., 2007; Grundy et al., 2004). The most useful way to reduce deaths due to CHD is early diagnosis and treatment. The combination of symptoms with medical reports to diagnose the CHD risk requires much experience and knowledge. Nowadays, Medical databases have huge amount of health care data, which contains hidden information. However, there is a task of designing the effective analysis tool to discover the hidden relationships in data for the prediction of CHD.

Knowledge that is hidden inside the huge amount of data can be discovered using data mining (Roohallah et al., 2013). It can reveal the patterns and relationships using statistical analysis, pattern recognition and machine learning techniques. Data mining is used in various applications such as marketing, banking, insurance, crime detection, agriculture, medicine, privacy preservation, etc. (Jiawei et al., 2012). Data mining is a popular predicting tool for medical reports stored in the massive database which seeks to identify and exploit patterns and relationship among large number of variables.

In medical diagnosis, attribute selection plays a vital role to identify the disease which in turn leads to a successful performance of the analysis in a reliable and effective manner. The data mining tool combines expert's knowledge with advanced techniques to analyze the hidden relationships among data from different perspective and summarizes its features to diagnose the CHD. In order to accomplish this task, the proposed work acts as an aid for the domain experts to deeply involve them into the process of knowledge discovery by acquiring domain knowledge and using it to focus the analysis as well as to filter the findings.

Data mining uses data analysis method with sophisticated algorithms in order to discover unknown patterns. Such algorithm include decision tree that have been extensively used in medicine (Karaolis et al., 2010). The objective of the paper is to develop a data mining technique based on decision tree classification using C4.5 algorithm for the assessment of CHD risk. In order to improve the classification accuracy the cross validation method is used (Podgorelec et al., 2002). In this model easily available low cost clinical and medical test data has been taken for analysis. The key idea of this work is early identification of the CHD risk which helps to reduce the CHD events. Further medical investigation can be carried out only for the CHD risk identified persons to avoid the financial burden.

## Methods

I. Data Description

To implement the proposed algorithm occupational drive's master health checkup data was collected from Institute of Road and Transport Perundurai Medical College and Hospital. The professional drivers in the transportation industry are at higher risk due to irregular diet and sedentary behavior. They experience high psychological demands and low physical activity at work which has been associated with Ischemic heart disease (IHD) (Bigert et al., 2003; Ragland et al., 1998). Furthermore, drivers have long working hours, shift work, exposure to nitrous oxide, carbon monoxide and traffic noise exposure which have also been linked with IHD (Claire et al., 2013).

Each individual is described by a set of nineteen attributes that includes screening of both clinical and biochemical data. The medical data records are transformed into operational format and a database is created. Database cleaning is done by removing the irrelevant attributes related to this study and duplicate records from the dataset. The missing values for the attributes are filled based on the data estimation and 375 instances are taken for this study. The numerical and categorical variables are converted into binary data based on the cutoff points. A unique number has been assigned to each patient for identification in the database for further reference.

After rigorous assessment only six predominant attributes out of nineteen attributes has been taken for this analysis. The predominant attributes includes three biophysical parameters Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Body mass index (BMI) and three blood chemical parameters Fasting blood sugar (FBS), Post prandial blood sugar (PPBS) and Triglycerides (TG). The threshold for all these attributes are BMI>=25kg/$m^2$, BP>140/90 mmHg, FBS>=126mg/dl, PPBS>=200mg/dl, TG>=170mg/dl, fixed based on the standards provided by World health organization (WHO).

II. Decision tree Induction

Decision tree algorithm is an induction learning algorithm which has the advantages of simplicity, transparency and ability to extract decision rules which are most commonly used for classification to predict the group in which a case belongs to. There are many decision tree algorithms like ID3, Classification and Regression trees (CART), Chi-squared Automatic Interaction Detection (CHAID), C4.5 and C5.0 (Quinlan, 1993; Breiman et al., 1984). For this study C4.5 algorithm is used since it is a top down divide and conquer strategy that partitions a given set of instances into smaller and smaller subsets in step with the growth of the tree.

The algorithm C4.5 that has been applied in the present paper for the investigation of CHD risk is described as follows. The CHD dataset is selected as an input to the algorithm for analysis. The attribute selection is carried out using information gain and gain ratio that provide a ranking for each attribute. Based on the ranking the algorithm identifies the most significant independent variable and

sets it as a root node, which is followed by the next best variables (Worachartcheewan et al., 2010). The C4.5 algorithm finds the variable-threshold for each leaf node that maximizes the homogeneity and splits the input observation into two or more subgroups (Delen et al., 2005). Based on the splitting criterion, branches are grown for each leaf node until the tree is completed.

II. A. *Information Gain*

Information gain is based on Claude Shannon's information theory (Shannon, 1948). The C4.5 algorithm uses information gain as its attribute selection measure. In the information gain, the information of the dataset is the expected information required to classify an instance. Attribute's information of the dataset is the new amount of information needed to classify an instance of the dataset after partitioning by that attribute. Thus to identify the splitting attribute of the decision tree, one must calculate the information gain for each attribute and select the attribute that maximizes the information gain. The information gain for each attribute is calculated using the formula.

$$\text{Information Gain}\ (\text{A}) = \text{Info}(D) - \text{Info}_{\text{A}}(D) \quad (1)$$

where A is the attribute investigated

$$Info(D) = -\sum_{i=1}^{k} p_i \log_2(p_i) \quad (2)$$

where

$p_i$ – probability (class $i$ in dataset $D$)

$k$ – number of class values

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j) \quad (3)$$

where

$|D_j|$ - number of observations with attribute values $j$ in dataset $D$

$|D|$ - total number of observations in dataset $D$

$D_j$ - sub dataset of D that contains attribute values j

$v$ - all attribute values

II. B. *Gain Ratio*

The information gain prefers to select attributes with large number of distinct values which results in large number of partitions which may lead to a useless classification. To overcome the effect of the bias in information gain, a variant known as gain ratio was introduced by the C4.5 algorithm. The gain ratio adjusts the information gain for each attribute by applying a kind of normalization to each attribute by using the formula

$$GainRatio(A) = \frac{InformationGain(A)}{SplitInfo_A(D)} \quad (4)$$

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \quad (5)$$

**Results**

In this study the classes are taken as positive instances when CHD=No, while it is negative instances when CHD=Yes. P is the number of positive instances i.e. 287 and N is the number of negative instances i.e. 88. For each instance, a comparison is made with the classifier's predicted class label and the actual class label.

The following four additional terms used to compute the evaluation measures

1. *True Positive* (TP): Refers to the positive instances (CHD=No) that were correctly labeled by the classifier as CHD=No.

2. *True Negative* (TN): Refers to the negative instances (CHD=Yes) that were correctly labeled by the classifier as CHD=Yes.

3. *False Positive* (FP): These are negative instances (CHD=Yes) that were incorrectly labeled by the classifier as positive i.e. CHD=No.

4. *False Negative* (FN): These are positive instances (CHD=No) that were incorrectly labeled by the classifier as negative i.e. CHD=Yes.

**Table 1:** Confusion Matrix

| Class predicted | No CHD risk | CHD risk |
|---|---|---|
| No CHD risk | 285 | 2 |
| CHD risk | 3 | 85 |

The confusion matrix is a useful tool for analyzing the C4.5 classifier by recognizing the instances of different classes. TP and TN are used to know when the classifier is getting the correct data while FP and FN are used to know when the classifier is getting wrong data. The confusion matrix for this study is shown in Table 1. It contains predicted classification information done by the C4.5 classification system as TP= 285, TN=85, FP=3, FN=2. From the confusion matrix the accuracy of the classifier is calculated using the formula

$$Accuracy = \frac{TP + TN}{P + N} \quad (6)$$

$$= \frac{285 + 85}{287 + 88} = 98.66\%$$

The error rate or misclassification rate of the classifier calculated using the formula

$$ErrorRate = \frac{FP + FN}{P + N} \quad (7)$$

$$= \frac{2 + 3}{287 + 88} = 1.33\%$$

The accuracy rate of 98.66% may not be acceptable as the classifier could be correctly labeling only the CHD=Yes instances and misclassifying all the CHD=No tuples. In order to find how well a classifier finds the positive instances and negative instances, sensitivity and specificity measures are used. The sensitivity is calculated using the formula

$$Sensitivity = \frac{TP}{P} \quad (8)$$

$$= \frac{285}{287} = 99.30\%$$

The specificity is calculated using the formula

$$Specificity = \frac{TN}{N} \quad (9)$$

$$= \frac{85}{88} = 96.59\%$$

**Table 2:** Performance summary of CHD risk using decision tree analysis

| S.No. | Performance Estimator | Values (%) |
|---|---|---|
| 1. | Sensitivity | 98.95 |
| 2. | Specificity | 97.70 |
| 3. | Accuracy | 98.66 |
| 4. | Positive predictive value | 99.30 |
| 5. | Negative predictive value | 96.59 |

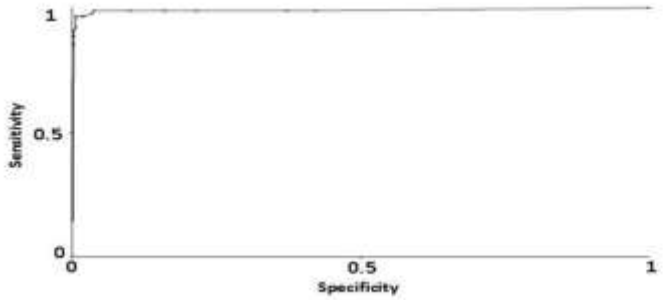


**Fig. 1:** ROC graph for Specificity versus Sensitivity has a larger area under the curve which shows higher performance of the algorithm

The above said performance measure values are tabulated in Table 2. From the measures it is observed that the classifier has high accuracy and it has accurately classified the positive instances i.e. sensitivity, but its ability to find the negative instances i.e. specificity is comparatively low. The comparison between sensitivity and specificity is plotted using Receiver Operating Characteristics (ROC) curve in Figure 1. The positive predictive value in Table 2 shows the probability that the patients with CHD=No truly have no CHD risk and the negative predictive value shows the probability that the patients with CHD=Yes truly have CHD risk.

The ROC graph has long been used in signal detection theory to depict trade-offs between true positive rate (sensitivity) and false positive rate (specificity) (Mukar et al., 1999). In this work an ROC curve allows one to visualize the trade-offs between the rates at which the classifier can accurately recognize positive instance versus the rate at which it mistakenly identifies negative instances as positive. Similarly the dataset were evaluated by the classifier on the above said performance measures on percentage split and supplied train and test set method but the accuracy is 77.66%. But the most significant improvement was obtained only by using 10-fold cross validation method. The 10-fold cross validation method has been used in C4.5 algorithm to build the decision tree. The fold versus accuracy graph is shown in Figure 2.

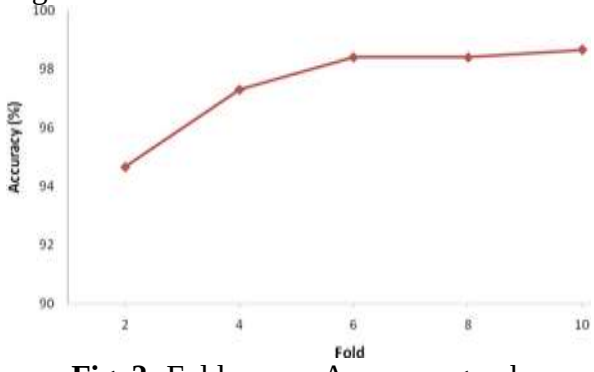


**Fig. 2:** Fold versus Accuracy graph

This shows that the accuracy increases rapidly up to six folds and it becomes stable. In order to fix the predominant attribute the information gain and gain ratio for the attributes are computed by using the equation (2) as

$$Info(D) = \left(-\frac{88}{375}\log_2\frac{88}{375}\right) - \left(\frac{287}{375}\log_2\frac{287}{375}\right) = 0.7860$$

Computing equation (3) as

$$Info(TG) = \frac{154}{375} I(r_1, n_1) + \frac{221}{375} I(r_2, n_2) \qquad (10)$$

From the above equation information gain of instances with high TG (i.e., 154) among these CHD=Yes ($r_1$=75) and CHD = No ($n_1$=79).

$$I(r_1, n_1) = \left( -\frac{75}{154}\log_2\frac{75}{154} \right) - \left( \frac{79}{154}\log_2\frac{79}{154} \right) = 0.9995$$

Similarly the information gain of instances with normal TG is also calculated and its value is 0.3226 and substituted in equation (10).

$$Info(TG) = \frac{154}{375}(0.9995) + \frac{221}{375}(0.3226) = 0.6005$$

Substituting the above values in equation (1) we get

$$\text{Information Gain } (TG) = 0.7860 - 0.6005 = 0.1854$$

The above calculated information gain value is substituted in equation (4) we get the gain ratio of TG as

$$\text{Gain Ratio } (TG) = 0.1272$$

Similarly information gain and gain ratio are calculated for other attributes and tabulated as shown in Table 3.

**Table 3:** Information gain and Gain ratio for all the attributes calculated using the formula (1) & (4) for selecting the root node of the tree

| S.No. | Attributes | Information Gain | Gain Ratio |
|---|---|---|---|
| 1. | TG | 0.184 | 0.236 |
| 2. | BMI | 0.138 | 0.175 |
| 3. | DBP | 0.104 | 0.133 |
| 4. | PPBS | 0.100 | 0.127 |
| 5. | SBP | 0.099 | 0.126 |
| 6. | FBS | 0.079 | 0.102 |

The information gain and gain ratio of the attribute TG is higher than other attributes so the decision tree selects TG as its root node. The process of selecting the next node to the root node is done by selecting TG=High as the target and information gain is calculated for the remaining attributes. The information gain of BMI is found to be higher than other attributes and is selected as the next node to the root node in the tree. This process continues for each leaf node until every attribute has been included along this path through the tree. The final decision tree learned by C4.5 algorithm for 375 instances is shown in Figure 3. In the figure, classified terminal leaf nodes have two numbers in parenthesis where the former denotes the number of correctly classified samples and latter denotes the number of incorrectly classified samples respectively.

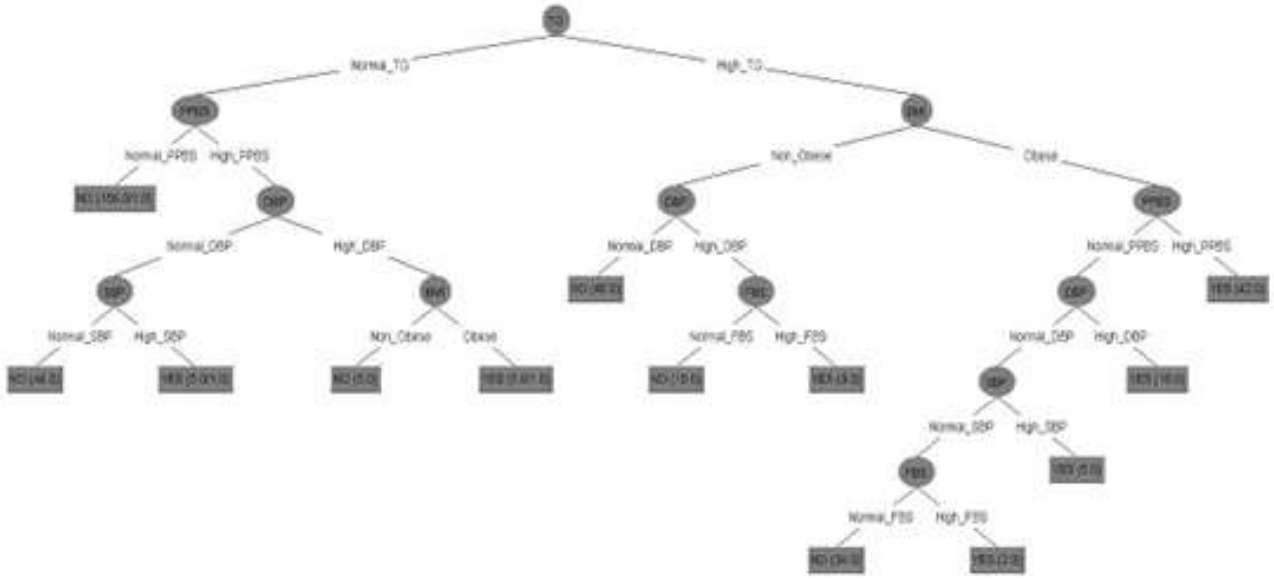

**Fig. 3:** Decision tree for CHD risk prediction shows the leaf node with YES for CHD risk and NO for No CHD risk

**Discussion**

The aim of the present study is to characterize a population of occupational drivers with regard to clinical and chemical parameters and to find out the coronary heart disease risk. The most significant results were reached using the decision tree C4.5 algorithm and 10-fold cross validation. The predictive power of medical data mainly depends on the specificity and sensitivity than accuracy. The clinicians especially wanted to see if it is possible to increase the specificity of the predictive process without affecting the sensitivity too much which may lead to the danger of number of patients who actually has the CHD risk, but omitted without any examination. In this work due to higher sensitivity (99.3%) the percentage of misclassification is very less which would minimize the no of patients going for further invasive investigations and also shorten the waiting time of the truly ill patients.

The above said two misclassifications are very less in this work which shows the high effective classification of C4.5 algorithm along with 10-fold cross validation method. Out of 375 patients 285 patients are identified as no CHD risk which reduces the further invasive CHD risk examination cost for the individuals.

Considering the correctly classified instances, 77% of the instances need not go for further analysis as they have no CHD risk and only 23% has to undergo for further analysis. This shows the improved diagnostics performance of the proposed work with very few predominant attributes. The misclassified 1.33% of instances can be corrected by further analyzing the instances by applying hybrid method by taking the output of the decision tree as the input to soft computing technique for further analysis.

**References**

Muthukaruppan S, Er MJ (2012) A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert systems with applications 39:11657-11665

Susan van Dieren, Andre PK, John C et al (2012) Effects of blood pressure lowering on cardiovascular outcomes in different cardiovascular risk groups among participants with type 2 diabetes. Diabetes research and clinical practice 98:83-90

Katzmarzyk PT, Janssen I, Ross R, Church TS, Blair SN (2006) The importance of waist circumference in the definition of metabolic syndrome: prospective analyses of mortality in men. Diabetes Care 29:404-409

Song Y, Manson JE, Meigs JB, Ridker PM, Buring JE, Liu S (2007) Comparison of usefulness of body mass index versus metabolic risk factors in predicting 10-year risk of cardiovascular events in women. Am J Cardiol 100:1654-1658

Grundy SM, Cleeman JI, Merz CNB, Brewer HB Jr, Clark LT, Hunninghake DB, Pasternak RC, Smith SC Jr, Stone NJ (2004) Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. Circulation 110:227-239

Roohallah A, Jafar H, Mohammad JH, Hoda M, Reihane B, Asma G, Behdad B, Zahra AS (2013) A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine 111:52-61

Jiawei H, Micheline K, Jian P (2012) Data mining concepts and techniques, 3rd edn. Morgan Kaufmann Publishers

Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS (2010) Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE transactions on information technology in biomedicine 14:559-566

Podgorelec V, Kokol P, Stiglic B, Rozman I (2002) Decision trees: an overview and their use in medicine. Journal of medical systems 26: 445-463

Bigert C, Gustavsson P, Hallqvist J, Hogstedt C, Lewne M, Plato N, Reuterwall C, Scheele P (2003) Myocardial infarction among professional driver. Epidemiology 14:333–339

Ragland DR, Krause N, Greiner BA, Fisher JM (1998) Studies of health outcomes in transit operators: policy implications of the current scientific database. Journal of Occupational Health Psychology 3:172-187

Claire HQ, Jen MN, Troy HP (2013) Does occupational driving increase the risk of cardiovascular disease in people with diabetes? Diabetes research and clinical practice 99: e9-e11

World Health Organization (2007) Prevention of Cardiovascular Disease Pocket Guidelines for Assessment and Management of Cardiovascular Risk. WHO Press

Quinlan JR (1993) C4.5: Programs for machine learning, Morgan Kaufmann publishers

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Taylor & Francis

Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V (2010) Identification of metabolic syndrome using decision tree analysis. Diabetes research and clinical practice 90: e15-e18

Delen D, Walker G, Kadam A (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 34:113-127

Shannon CE (1948) A Mathematical Theory of Communication. Bell System Technical Journal 27:379-423, 623-656.

Mukar M, Kononenko I, Groselj C, Kralj K, Fettich J (1999) Analyzing and improving the diagnosis of ishaemic heart disease with machine learning. Artificial Intelligence in Medicine 16: 25-50