

Detection of Cardiac Disease using Data Mining Classification Techniques

Abdul Aziz

Faculty of Computer Science & IT
Superior University, Lahore, Pakistan

Aziz Ur Rehman

Faculty of Computer Science & IT
Superior University Lahore, Pakistan

Abstract—Cardiac Disease (CD) is one of the major causes of death. An important task is to identify the Cardiac disease very minutely and precisely. Generally medical diagnostic errors are dangerous and costly. Worldwide they are leading to deaths. Data mining techniques are very important to minimize the diagnostic errors as well as to improve the patient's safety. Data mining techniques are very effective in designing a medical support system and enrich ability to determine the unseen patterns and associations in clinical data. In this paper, the application of classification technique, decision tree for the detection of heart disease have been introduced. Classification tree uses many factors including age, blood sugar and blood pressure; it can detect the probability of patients fallen in CD by using fewer diagnostic tests which save time and money.

Keywords—Cardiac disease; classification technique; decision tree; knowledge discovery

I. INTRODUCTION

In rapidly growing world, as the time moves individuals need to carry on with a very deluxe life, consequently they work like a gadget with a specific end goal to get a lot of cash and carry on with a casual life. But they overlook to take care of themselves. Their whole way of life is changing as their foods are changing. In this sort of life style they get tension and have blood pressure and sugar issues. It moves towards a major threat, namely, heart disease, an utmost vital organ having a deep influence in all body parts of an individual. Heart is the only organ in human body that really works hard [1]. Cardiac disease is the major cause of mortality in the worldwide over few decades [2]. Various factors exists which are supportive in detection of coronary illness, for example hypertension, smoking, high cholesterol, family history, obesity, blood sugar etc. [3].

In most cases, identification of disease is usually done by the medical specialist abilities and on the basis of current test result. To diagnose an illness is so crucial task that needs high skills and much experience [4]. The main focus is in detection of cardiac disease by using data mining techniques. The enthusiasm to my study is the approximation provided by W.H.O. According to W.H.O by year 2030 just about 23.6 million individuals will kick the bucket due to coronary illness [5].

Thus to reduce this risk the detection of coronary illness should be performed. Current year's medical sector is producing large amount of data related to diagnosis disease, patients, hospital resource and medical devices, etc. [6]. This

data is the main source for effective analysis of data and from this, extract key information that motivates the healthcare community. The detection of heart disease using Data Mining techniques provides us better result. To extract and discover unseen patterns related to heart disease form the existing coronary illness database classification tree which plays an important role. Data Mining focuses machine learning, statistical analysis and databank technology [7]. It assists the medical practitioner and analyst to mark intelligent medical decision which outmoded support system cannot. Some risky elements for CD are: obesity, family history, level of cholesterol, inactiveness and hypertension [8].

According to a survey about 50 percent of victims have no indications till heart attack arises. Analysis of many factors are done to investigate the heart disease, generally physicians make conclusion by assessing current result of the patients tests but it depends upon doctor experience and abilities.

II. LITERATURE REVIEW

In this section, current literature to diagnose the heart disease by using various data mining methods and tools have been discussed. Among of them a few researches that supported my work are discussed here. M Kumari diagnoses the heart disease by applying a data mining classifier that is Decision Tree. The research scholar analyzes the presentation of this algorithm on various factors that is accuracy rate, sensitivity and error rate. He concludes that the accuracy of Decision Tree is 79 percent [9].

Research demonstrates that by using data mining techniques in health industry then this industry would be in healthier position to fulfill their long term as well as short term goals [10]. By using biomedical mining algorithms heart disease is predicted, the author used classification technique that constructed on supervised machine learning procedures. The author use the decision tree that has error rate 0.2775 and having accuracy of 79.05% [11].

Decision tree algorithms have been applied for classification in various application areas that is production, medicine manufacturing and monetary analysis [12]. N. Kausar and S Planiappan perform a comparison between decision tree and naïve Bayes algorithms. They used UCI data set for risk prediction and stated that decision tree gave high accuracy then naïve Bayes that is 96.4% accuracy [13]. Dr. P Alli, Jenzi and Paryanka offered a new system that depends on mining algorithms to detect the cardiovascular disease. They gathered various patterns to estimate the CD. It happened to them that

decision tree was very cool to fathom and had a better accuracy rate for detecting heart disease [14]. Meenu and Kawaljeet show in his research that Bhatla applied three classifiers such as Decision Tree, Naïve Bayes and neural network for likelihood of CD. Their examination shows that neural network have extraordinary correctness in neural network and after it Decision Tree outer performed over other data mining algorithms [15].

K. Kaurand Lalit shows in their research that he performs many experiments with KNN, Naïve Bayes and Decision Tree. Among all over them DT (Decision Tree) have very high performance of accuracy. Afterwards pre-processing the data correctness of Naïve Bayes and Decision Tree have been enhanced, they use Tanagra tool to classify their data [16]. The study was carried out via C4.5, Decision Tree for identification of stroke disease. The uppermost percentage of this algorithm was 75%, 65% and 75% for Myocardial Infarction disease [17]. Most of the above work was done with WEKA tool.

III. RESEARCH METHODOLOGY

A. Decision Tree

Decision Tree is pondering the most famous technique for diagnosis the cardiac disease. To build a decision tree by using accessible data which can pact with the glitches related to different research areas is very important [18]. Corresponding to the flow chart in which each non leaf node shows a test on a specific attribute and each branch shows the result of that test and each leaf node need a class tag. Root node is the upper most node of the decision tree [19].

The utmost usage of decision tree is in processes research analysis for computing conditional probability. Few advantages of Decision Tree are easily understandable, perform well in huge dataset, simply interpret and robust as well as it knobs both categorical and numerical data [20]. In this the structures that convey supreme information are carefully chosen for classification while other features are put off, by this means computational efficacy is enhanced [21].

B. Data Source

These experiments are being carried out for the detection of heart disease using Decision Tree algorithm. The data set is taken form University of California Irvin (UCI) Cleveland Data set and there are total 52 instances from which only 8 attributes are taken for experiments work such as age, chest pain, blood pressure, blood sugar that achieved, angina electro cardiogram. SPSS Clementine 12.0 has been used for [22] calculation and analysis of Data due to its efficiency in finding patterns, analysis and having ability of good prediction.

C. Data Set

Selection of data sets is very important because all the experiments and results are based on the data sets. It has been tried to choose the latest, accurate and clean data set so that best results could be obtained. An extra care has been taken in this regards. A total of 210 instances are taken from patient database of Cheema Heart Complex and Cleveland hospital database.

TABLE. I. DATASET AND ITS DESCRIPTION

Attribute	Description
Age	30 to 50 = 1 , 51 to above = 2
Chest pain	1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
Blood Pressure	Normal (80 to 120) =1 High (above 120) =2
Blood Sugar	False = 0, True = 1
ECG	0: normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: showing probable or definite left ventricular hypertrophy
Heart Rate	Normal (60 to 100) =1 High (above 100) = 2
Angina	Yes = 1, No = 0
Class Attribute	
Disease	0: <50% diameter narrowing (No Disease) 1: > 50% diameter narrowing (Heart Disease)

From 210 only 8 attributes are selected for experiments. The data set contained 117 patients without heart disease and 92 patients with heart disease. We identify a diagnosis class having value 0 with no heart disease and value 1 with heart disease. Table 1 shows the selected attributes and their description.

D. Proposed Model

A new model is planned which gives finest result and perfections over previous models. In this section full framework has been discussed as shown in Fig. 1.

The first step in model is the selection of data that is the data source. After sourcing field option is used and a type field is selected that allows field metadata to be determined and controlled. And then the modeling phase occurs, the algorithm C5.0 is selected to constructs a predictive Decision Tree or rule set depends on your own choice and nature of data. After executing the predicted model, performance analysis is performed as shown in Fig. 2. Here performance of algorithm can be evaluated.

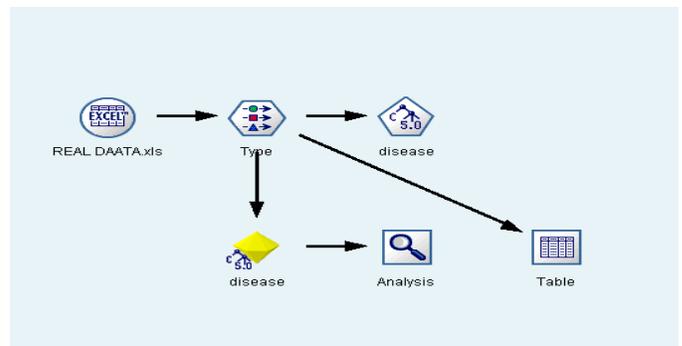


Fig. 1. Classification model.

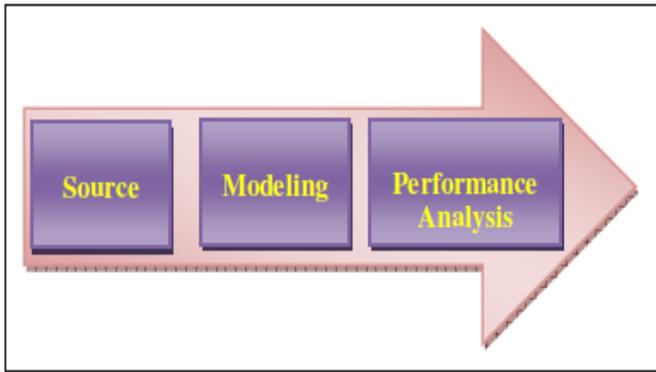


Fig. 2. Block diagram of proposed model.

E. Experiments and Results

Experiments were made using SPSS Clementine tool. Data Set of 210 Patients with 8 medical attributes is used. All attributes are in discrete form and resolved the discrepancies among them. Decision Tree performs best having a good estimation probability of 79.9% by using 8 attributes. That is shown in Fig. 3. There are total 8 attributes from which 7 are numeric and the last taking two values 0 and 1 (0 mean negative and 1 mean positive disease) is my class attribute.

The tree diagram shows all the results of predicted disease. Fig. 4 illustrates that a class attribute (disease) has two child nodes (exercise angina and chest pain) and maximum 2 depth tree. In class attribute Disease (1 presence of heart disease & 0 shows the absence of heart disease), initially 92 persons are found to be infected by heart disease (out of a dataset of 209 patients records), rest of them took further tests/observations like patients having angina during exercise.

For the sake of this, add exercise angina attribute which have two nodes with value 1 and 0 (presence or absence). In node 1, there is probability of 83% having the heart disease while rest 17% are not the victims of heart disease. In node 2 out of 137 people, 23% individuals i.e. 32 people would be infected by heart disease based on the test of chest pain, those who have chest pain type (typical angina and asymptomatic angina: typical angina and non-anginal pain) are 52 percent and 8 percent respectively. The remaining 105 people are not infected by the heart disease (see Fig. 4 and 5).

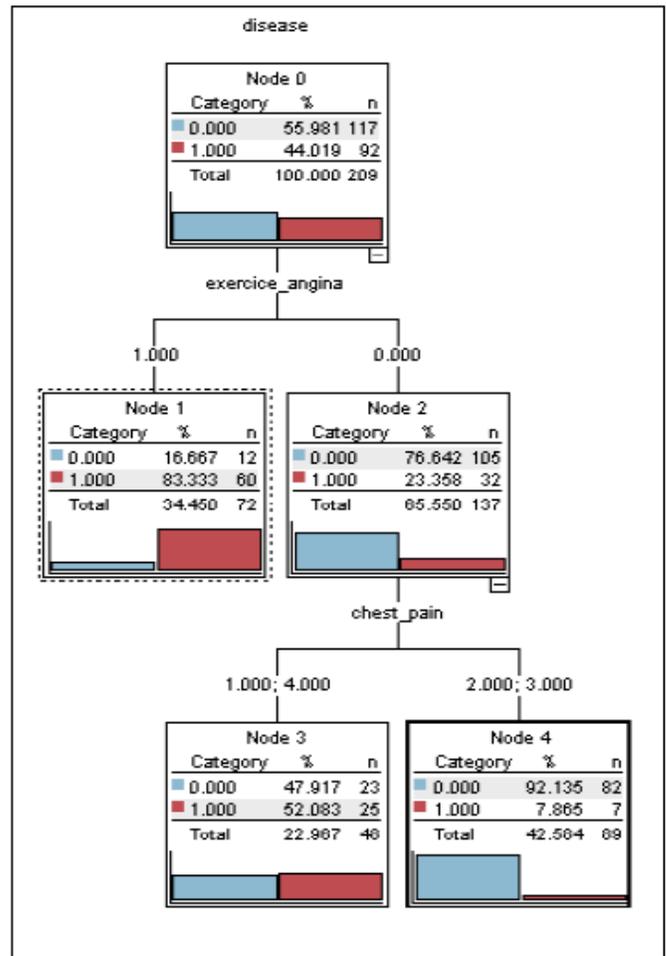


Fig. 4. Model of decision tree.

From the clinical point of view it is a common practice to carry out all the tests whenever a patient attends the clinic with chest pain. Usually, it takes much time to reach at the conclusion whether the patient has a heart disease or simply he suffers from muscular pain. Moreover, in addition to long decision time it is also very costly to patients. With the help of classification tree, numbers of diagnostic tests are reduced which also helps to reduce cost significantly.

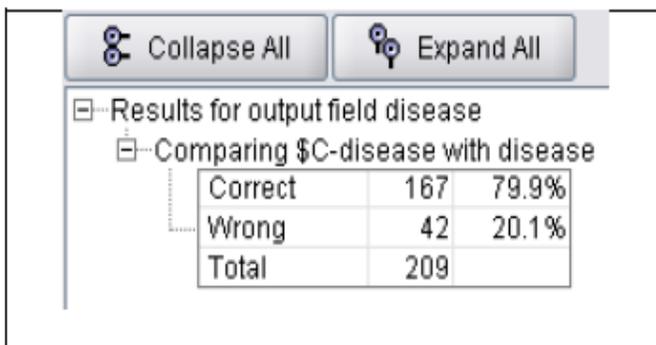


Fig. 3. Results and accuracy.

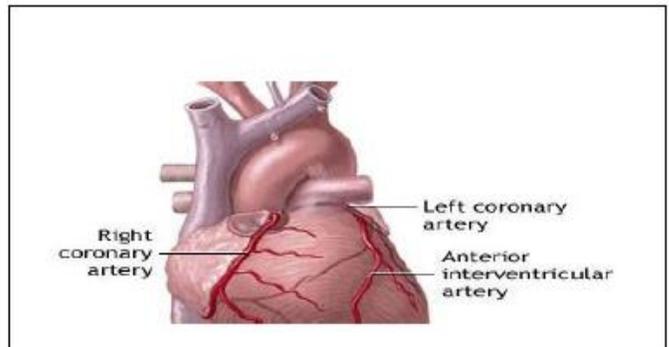


Fig. 5. Heart disease that leads to stroke.

IV. CONCLUSION

The most widely used technique of Data Mining in healthcare sector is the classification. The extensive classification method used for the prediction of heart disease is the decision tree that is used in this research. Sometimes poor observations lead towards death. All practitioners are not so expert to diagnose the heart disease with minimal number of tests. The main purpose of this research is to diagnose the heart patients more precisely and more accurately with minimum number of tests (reduction of attributes). This research plays a vital role in the cost reduction of treatment, diagnose disease and additional enhancement of the medical studies. The purposed research work can further be boosted and expended for the prediction of various types of heart diseases.

ACKNOWLEDGMENT

The authors thank and acknowledge the support and assistance of Cheema Heart Complex (CHC), Gujranwala, Pakistan. In addition to providing the useful cardiac data they also provided the technical guidance for the success of this research project. We also express gratitude for the commitment, expertise, and actions of members of the CHC. The findings and conclusions in this research paper are those of the authors and do not necessarily represent the official position of the CHC for Disease Control and Prevention.

REFERENCES

- [1] Kuldip Singh, Singh G. Alterations in Some Oxidative Stress Markers in Diabetic Nephropathy. *Journal of Cardiovascular Disease Research*. 2017; 8(1):24-7.
- [2] Nagre SW. Mobile Left Atrial Mass – Clot or Left Atrial Myxoma. *Journal of Cardiovascular Disease Research*. 2017; 8(1):31-4.
- [3] Krishnamurthy VT, Venkatesh SA. Negative Pressure Pulmonary Oedema after Sedation in a Patient Undergoing Pacemaker Implantation. *Journal of Cardiovascular Disease Research*. 2017; 8(1):28-30.
- [4] Paryad E, Balasi LR, Kazemnejad E, Booraki S. Predictors of Illness Perception in Patients Undergoing Coronary Artery Bypass Surgery. *Journal of Cardiovascular Disease Research*. 2017; 8(1):16-8.
- [5] Berenson GS, Srinivasan SR, Hunter SM, Nicklas TA, Freedman DS, et al. (1989) Risk factors in early life as predictors of adult heart disease: The Bogalusa heart study. *Am J Med Sci* 298:141-151.
- [6] Dhuper S, Buddhe S, Patel S (2013) Managing cardiovascular risk in overweight children and adolescents. *PaediatrDrugs* 15:181-190.
- [7] Berenson GS, Srinivasan SR, Bao W, Newman WP III, Tracy RE, et al. (1998) Association between multiple cardiovascular risk factors and atherosclerosis in children and young adults. The Bogalusa Heart Study. *N Engl J Med* 338:1650-1656.
- [8] Newman WP, Freedman DS, Voors AW, Gard PD, Srinivasan SR, et al. (1986) Relation of serum lipoprotein levels and systolic blood pressure to early atherosclerosis. The Bogalusa Heart Study. *New Eng J Med* 314:138-144.
- [9] National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents (2004). The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents; *Paediatrics* 114:555-76.
- [10] Dinarević S, Mulaosmanović V (2005) Primary prevention of Hypertension in Sarajevo Children: Role of Adiposity. 29th UMEMPS Congress Union of Middle Eastern and Mediterranean Paediatric Societies, pp. 154-156.
- [11] Dinarević S, Mesihović H, Simeunović S, Zulić I (1994) Dyslipoproteinaemia in Children with Heart Disease. *Intercontinental Cardiol* 3:126-9.
- [12] Kapur G, Ahmed M, Pan C, Mitsnefes M, Chiang M, et al. (2010) Secondary hypertension in overweight and stage 1 hypertensive children: A Midwest Paediatric Nephrology Consortium report. *J Clin Hypertens* 12:34-39.
- [13] Dinarević S, Hasanbegović S (2010) Problem of obesity in children and youth in Canton Sarajevo. *Pediatr Res* 68:1091.
- [14] WHO Europe (2007) The challenge of obesity in the WHO European Region and the strategies for response. Copenhagen: WHO Regional Office for Europe.
- [15] Daniels SR, Arnett DK, Eckel RH, Gidding SS, Hayman LL, et al. (2005) Overweight in children and adolescents: Pathophysiology, consequences, prevention and treatment. *Circulation* 111:1999-2012.
- [16] Berenson GS, Wittingly WA, Tracy RE, Newman WP, Srinivasan SR, et al. (1992) Atherosclerosis of the aorta and coronary arteries and cardiovascular risk factors in persons aged 6 to 30 years and studied at necropsy (The Bogalusa Heart Study). *Am J Cardiol* 70:851-858.
- [17] McNiece KL, Gupta-Malhotra M, Samuels J, Bell C, Garcia K, et al. (2007) National High Blood Pressure Education Program Working Group: Left ventricular hypertrophy in hypertensive adolescents: Analysis of risk by 2004 National High Blood Pressure Education Program Working Group staging criteria. *Hypertension* 50:392-5.
- [18] Torrance B, McGuire KA, Lewanczuk R, McGavock J (2007) Overweight, physical activity and high blood pressure in children: a review of the literature. *Vasc Health Risk Manag* 3:139-149.
- [19] Jiber H, Blitti MC, Bouarhroum A. Acute type B Aortic Dissection Complicated by Acute Limb Ischemia: Case Report. *Journal of Cardiovascular Disease Research*. 2016;7(2):97-9.
- [20] Muhammad Subhi Al- Batah "Testing the probability of Heart Disease using Classification and Regression Tree" *Annual Research & Review in Biology* 4 (11): 1713-1725,(2014).
- [21] Setty HS, Hebbal VP, Channabasappa YM, Jadhav S, Ravindranath KS, Patil SS, et al. Assessment of RV function following Percutaneous Transvenous Mitral Commissurotomy (PTMC) for rheumatic mitral stenosis. *Journal of Cardiovascular Disease Research*. 2016; 7(2):58-63.
- [22] Patnaik L, Pattanaik S, Sahu T, Panda BK. Awareness of symptoms and risk factors of Myocardial Infarction among adults seeking health care from a rural hospital of India. *Journal of Cardiovascular Disease Research*. 2016;7(2):83-5.