



PREDICTION OF RISK SCORE FOR HEART DISEASE IN INDIA USING MACHINE INTELLIGENCE

*Dr. P. Amirtharaj, K. Rajeswari, Dr. V. Vaithiyanathan

*Associate Professor, Department of Cardiothoracic Surgery, Madras Medical College, Chennai.

Abstract.

Machine Intelligence is used for Medical Data Mining. Medical Data Mining is finding interesting information from a large collection of Medical data. Our proposed paper tries to find useful information from heart disease data set. Twenty five to Thirty percent of deaths in the most developed countries is due to Coronary Heart Disease (CHD) or Ischemic Heart Disease (EHD). India is in a risk of developing more death due to CHD. Hence a Decision Support System (DSS) is proposed to identify a risk score for predicting the Heart risk of a Patient in the subsequent Years. This will help the Patients in taking precautionary steps like following a Balanced Diet and medication which in turn may increase the life time of patient. The features for prediction are selected after considering Indian conditions from literature and based on the Expert knowledge from Doctors. Framingham Risk score which has five attributes may be used for comparison. Our proposed system has nineteen features. A discussion to reduce the number of features using Genetic algorithms is made. The system is a theoretical study which proposes implementation of Machine Intelligence algorithms.

Keywords: Decision Support System (DSS), Heart disease, Machine Intelligence, Ischemic Heart Disease(EHD), Coronary Heart Disease(CHD), Risk Score.

1. Introduction

Indians are more likely to have types of heart disease [1] that lead to worse outcomes like ischemic heart disease — a condition characterized by reduced blood supply to the heart. The WHO has drawn attention to the fact that CHD is our modern "epidemic". Myocardial Infarction is specific to CHD. Others are not and may cause confusion during diagnosis [2]. Rose [3] calculates the "incubation period" of CHD may be 10 years or more. Machine Intelligence is used for Medical Data Mining where large collection of Medical data is mined for interesting pattern.

1.1 India and Coronary Heart Disease

A large body of data exists on the occurrence of CHD in hospital patients.

However, there are only two studies one in Chandigarh, screening patients over the age 30, by a 12-lead ECG and the other in Haryana where the prevalence is 65.4 and 47.8 per 1000 males and females respectively in urban population [4]. This proposed study is carried out in Chennai, Madras Medical College for all the age group. The pattern of CHD in India has been reported to be as follows:

- CHD appears a decade earlier compared with the age incidence in developed countries. The peak period is attained between 51-60 years.
- Males are affected more than females.
- Hypertension and Diabetes account for about 40 percent of all cases. Heavy smoking is responsible aetologically in a good number of cases [5].

The aetiology of CHD is multifactorial. Apart from the obvious one like increasing age and male sex, studies have shown several important risk factors, which make the occurrence of the Disease more probable as below.

1. Smoking

Smoking has been identified as a major risk factor [6, 7]. The degree of risk of developing CHD is directly related with the number of cigarettes smoked per day [8].

2. Hypertension

Many Investigators feel that Systolic Blood Pressure is a better predictor than Diastolic Blood Pressure. However, both components are significant risk factors. The risk role of mild Hypertension is accepted [9].

3. Serum Cholesterol

The 14-years experience of the Seven Countries study [10] showed that serum cholesterol concentration is an important risk factor for the incidence of CHD levels perhaps 220 mg/dl or more. This supports the notion of a "threshold level" of cholesterol.

4. Other risk factors

- Diabetes [11], Sedentary life style [12], Type A individuals [13], are the important risk factors.
- A family history of CHD is known to increase the risk of premature death. Higher alcohol intake, defined as 75 g or more per day is an independent risk factor for CHD [9].

2. Risk Assessment Based on Clinical Conditions

Absolute Risk is divided into three classes namely, high; intermediate and low in quantitative terms as in table 2.1.

Table 2.1. Risk Classification

Risk Classification	10 Year Absolute Risk for CHD
Low	<10%
Medium	10-20%
High	>20%

Patients at High Risk require Risk lowering treatment. Patients at intermediate risk also need medical intervention to the extent that therapy is safe and successful. Patients with low risk and no risk may be encouraged by their doctors to follow Health recommendations for Prevention of CHD.

2.1 Related Work -Framingham Risk Score and Risk Factor

Framingham heart study has five features to be entered, namely Age, Total Cholesterol, Smoker/Non smoker, Systolic Blood pressure, HDL. Points are added together to predict the 2-year risk score of Myocardial Infarction. Sample score for Age is as follows [21]. Similarly scores are given to other features given as above. The scores are summed to obtain the Risk factor.

Age	Points[Male]	Points [Female]
35-39	0	0
40-44	1	1
45-49	3	2
50-54	4	3
55-59	6	4
60-64	7	5
65-69	9	6
70-74	10	7

2.2 Indian Heart Risk Predicting Features

The following 19 features are proposed to be collected and analyzed for Indian Heart risk score prediction based on extensive study and expert opinion from doctors with respect to Indian body conditions, life style and eating habits.

Age, Sex, Menopause details, Height, Weight, Body Mass Index[BMI], Waist Circumference, Systolic Blood pressure, Diastolic Blood pressure, Diabetes, Thyroid, Total Cholesterol, LDL Cholesterol, HDL Cholesterol, Smoking habits, Family History, Genetic factors, Type A personality, Sleeping Disturbance

A total of 125 patients record are collected from the Cardio Thoracic Department in Madras Medical College. Based on the data collected, a Risk score has to be predicted. Points will be given after analyses of data based on the importance of every feature. The statistical description of the collected dataset is given in Table 2.3.

Table 2.3. Descriptive Statistics of data collected for Male [N=61] and Female [N=49]

Variables	Min [Male]	Min [Female]	Max [Male]	Max [Female]	Mean [Male]	Mean [Female]	SD [Male]	SD [Female]
Age	15	15	65	60	42.34	32.51	14.608	12.762
Height	143	140	178	170	163.31	152.29	6.862	6.400
Weight	29	30	81	80	55.62	45.47	11.715	10.190
Body Mass Index	10.65	11.02	29.75	29.38	20.4308	16.7013	4.30299	3.74291
Waist Circumference	57	55	105	100	77.30	70.14	12.434	10.634
Systolic Pressure	90.00	90.00	170.00	140.00	124.3279	114.6939	14.87360	9.81114
Diastolic Pressure	70.00	70.00	100.00	100.00	79.4426	75.1020	6.99649	7.39438

3. Machine Learning

Nowadays many systems in health care domain are of multi-purpose type which support various complex tasks along-with diagnosis and combine more than one AI technique [17]. Medical care could be enhanced and costs could be reduced by means of an automated medical diagnosis system [15]. The importance of the role of data mining techniques in providing better patient care and effective diagnostic capabilities by finding patterns and extracting knowledge increases with the increase in the volume of stored data [16].

3.1 Determination of High Impact Patterns and Their Optimal Values Using Genetic Algorithm

In the case of medical data, a significant fraction of association patterns are not useful because they are medically irrelevant, and contain redundant information or simply describe well-known facts [18]. So, these mined association patterns cannot be a best match for designing the employed reasoning technique, Fuzzy logic. Also, decision making under uncertainty can be viewed as a deliberation: determining appropriate solutions to maximize the expected gain. In the proposed system, we employ Genetic Algorithm (GA) to determine the most high impact association pattern for heart disease diagnosis and the optimal values corresponding to each of the considered attributes. It can be considered as a mathematical optimization algorithm whose workings are metaphorically inspired by neo-Darwinian evolution [20] and is based on the evolutionary algorithm that models biological process to optimize highly complex cost functions by allowing a population composed of many individuals to evolve under

The fitness function used in the proposed system is aimed at determining the optimal (frequency) values corresponding to each attribute. It can be defined as a function of actual and reference value of each of the attribute. The fitness value that decides the survival of each chromosome is given as follows [19]. Av represents the actual value of the attribute and Nv represents the reference value of the attribute.

$$fitness = \sum_{i=1}^m \frac{1}{m} (freq(A_{v(i)}) * (A_{v(i)} - N_{v(i)}))$$

Among the nineteen features, more important features will be selected using Genetic Algorithm and a Risk factor can be made by summing the Risk score's of the various features. It is anticipated that data mining could help in the identification of risk subgroups of subjects for developing future events and it might be a decisive factor for the selection of therapy, i.e., angioplasty or surgery [14].

4. Conclusion

In this paper, we have proposed a Clinical Decision Support System (CDSS) for reliable heart disease risk score prediction using Machine Learning techniques. The system is designed for Indian Population. Coronary Heart Disease or Ischemic Heart Disease can be handled successfully if more research is encouraged in this area. As this approach focuses on Present Risk analysis, future work may be directed for Life time Risk analysis.

5. References

1. <http://tim.esofindia.indiatim.es.com/city/de/hi/T/Ol-campaign-against-heart-disease-a-success/articleshow/7193581.cms>
2. Pedoe, H.T (1982). In Epidemiology of Diseases, D.I. Miller and R.T.D. Farer (eds), Blackwell, Oxford.
3. Rose, G. (1982) European Heart J.3, Suppl B. 18.
4. Dewan, B.D. et al (1974), Indian heart j., 26:68.
5. Sinha, B.C (1970). Jr. Ind. Med. Assoc, 55: 171.
6. Slone, D. et al (1978). N. Eng. J. Med. 298:1273.
7. Sharper A.G. et. Al (1981). Brit. Med. J. 283:179
8. Bain, C. et. Al (1978). Lancet, 1:1087.
9. WHO (1985). Primary Prevention of CHD EURO Rep and Studio 98. Copenhagen Keys, A. 91980). Seven Countries : a multivariate analysis of death and CHD, Harvard University Press, Cambridge, M.A.
10. WHO (1985). Tech. Rep. Ser., 727.
11. Miller, N.E. et.al (1979) Lancet, 1:111.
12. Jenkins, C.D. et al (1974). N. Eng. J. Med, 290:1271.
13. Minas A. Karaolis,, Joseph A. Moutiris, Demetra Hadjipanayi, and Constantinos S. Pattichis (2010), Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees. IEEE Transactions on Information Technology in Biomedicine, VOL. 14, NO. 3.
14. Latha, P. and Subramanian, R. (2008) 'Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm', International Journal of Biological and Medical Sciences, Vol. 3, No. 3.
15. Wasan, S. K., Bhatnagar, V. and Kaur, H. (2006) 'The Impact of Data Mining Techniques on Medical Diagnostics', Data Science Journal, Vol. 5, pp. 119-126.
16. Shahina B., Ahmed, M. U. and Funk, P. (2009) 'Case-based systems in health sciences: a case study in the field of stress', WSEAS Transactions on Systems, Vol. 8, No. 3, pp. 344-354.
17. Carlos O. (2004) 'Improving Heart Disease Prediction Using Constrained Association Rules', Seminar Presentation at University of Tokyo.
18. Goldberg, D. E. (1989) 'Genetic algorithms', New York: Addison Wesley Longman, Inc.

Shanthi, D., Sahoo, G. and Saravanan, N. (2008) 'Input Feature Selection using Hybrid Neuro-Genetic Approach in the Diagnosis of Stroke Disease', International Journal of Computer Science and Network Security, Vol. 8, No. 12, pp. 99-107.

D'Agostino, Russell MW, Huse DM et al. 'Primary and subsequent coronary risk appraisal: new results from the Framingham Study', American Heart Journal 2000.

