A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic

Nidhi Bhatla
GNDEC, Ludhiana, India

Kiran Jyoti
GNDEC, Ludhiana, India

ABSTRACT
Cardiovascular disease is a term used to describe a variety of heart diseases, illnesses, and events that impact the heart and circulatory system. A clinician uses several sources of data and tests to make a diagnostic impression but it is not necessary that all the tests are useful for the diagnosis of a heart disease. The objective of our work is to reduce the number of attributes used in heart disease diagnosis that will automatically reduce the number of tests which are required to be taken by a patient. Our work also aims at increasing the efficiency of the proposed system. The observations illustrated that Decision Tree and Naive Bayes using fuzzy logic has outplayed over other data mining techniques.

Keywords
Cardiovascular disease; data mining; fuzzy logic; weka tool; decision tree; naive bayes; classification via clustering.

1. INTRODUCTION
WHO report Global Atlas on cardiovascular disease prevention and control states that cardiovascular disease (CVDs) are the leading causes of death and disability in the world. Although a large proportion of CVDs is preventable, they continue to rise mainly because preventive measures are inadequate.

Clinical problem solving or diagnostic reasoning is the skill that physicians use to understand a patient's complaints and then to identify a short, prioritized list of possible diagnoses that could account for those complaints. This differential diagnosis then drives the choice of diagnostic tests and possible treatments. Despite striking advances in information technology, clinical problem solving has not yet been effectively replicated by computers, making it essential that clinicians work to develop expertise in this very important skill set. Hence, more adequate systems for diagnosis of cardiovascular disease need to be developed.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. In today's era, data mining has its successful application in various fields including healthcare. On the other hand, fuzzy logic provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information.

Our work attempts to incorporate both the above mentioned techniques for the development of the proposed system and to increase its efficiency.

2. RELATED WORK
Ample number of systems such as information systems, Decision Support Systems, Image and Scan processing systems in healthcare sector has been deployed for effective diagnosis of various diseases. Our work is an endeavour to predict accurately the presence of cardiac disease with reduced number of attributes. P.K. Anooj (2012) [1] developed Clinical Decision Support System for heart disease using weighted Fuzzy Rules. E.P. Ephzibah et al (2012) [2] framed Fuzzy Rules for Heart Disease diagnosis using 6 attributes. Sulabha S. Apte et al (2012) [4] compared various data classification techniques by using 15 attributes for heart disease diagnosis. M. Anbarasi et al (2010) [5] developed an Enhanced Prediction System for heart disease with feature subset selection using Genetic Algorithm. Moreover, three classifiers Decision Tree, Naive Bayes and Classification via Clustering have been used and Decision Tree performed with good prediction probability of 99.2%. B.Patil et al (2009) [11] used Artificial Neural Network for developing heart disease prediction system. Carlos (2006) [19] compared Association Rules and Decision Trees for disease prediction. The rest of the sections are classified in the following manner. Section 3 explains the data set used. Section 4 discusses about developing the heart disease prediction system using fuzzy logic. Section 5 illustrates the classification process and outcomes. Section 6 exhibits the efficiency of the proposed system.

3. DATA SET
In our work, six attributes have been reduced to four attributes which are employed for heart disease prediction. The data of various patients is entered in the proposed system and the diagnosed results generated by the system corresponding to patients have been saved in the database. The resultant data set thus obtained is used by the classification model for calculating the efficiency of the proposed system. Attributes have been converted to categorical form for more clarity [5]. Moreover, training set method is used as the test mode.

Input Attributes:
1. Type - Chest Pain Type
2. Rbp - Resting blood pressure
3. Eia - Exercise induced angina
4. Oldpk - Old peak
5. Vsl - No. of vessels colored
6. Thal -Maximum heart rate achieved

Fig 1: Attributes list

Fig. 1 illustrates the original list of attributes and fig. 2 illustrates the reduced set of attributes.

Reduced Input Attributes:
1. CPTYPE - Chest Pain Type
2. BP - Blood Pressure
3. CA - No. of vessels colored
4. TMT – Treadmill Test

Fig 2: Reduced Attributes list

4. PROPOSED SYSTEM

The proposed system has been developed with an aim to efficiently diagnose the presence of heart disease in an individual. Fuzzy logic has been used in Matlab for this development process.

A fuzzy set is a collection of distinct elements with a varying degree of relevance or membership. The membership function takes interval values between 0 and 1. These values express the degrees with which each object is compatible with the properties or features that are distinctive to the collection. A fuzzy set is a generalization of the concept of a set whose characteristic function takes only binary values. A fuzzy inference model can be created using the properties of fuzzy set. The knowledge base of a fuzzy inference system is to link the fuzzified inputs with the associated reasoning mechanism.

Originally, 6 attributes were used for heart disease prediction but our work differs by using only 4 attributes. Patient's basic information and other attributes values are entered in the proposed system. On the basis of fuzzy rules defined, the diagnosed result is shown by the system. The diagnosed results are classified into four values viz. Normal, Low Risk, Medium Risk and High Risk. Moreover, the results are also stored in the database connected with the system for calculating the efficiency and for the record maintenance of the patients. A unique Patient ID is generated by the system during this storage process. Fig. 3 illustrates the overview of the proposed system. Fig. 4 exhibits the Knowledge Flow Layout for both the Classifiers. Fig. 5 illustrates the Decision Tree generated by graph viewer connected with J48 and Fig. 6 shows the results generated by text viewer connected with Naive Bayes.

5. CLASSIFIERS

Classification is a technique that predicts categorical class labels. It classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. Classification is a two – step process consisting of Model Construction and Model Usage. Model Construction is defined as a process of describing a set of predetermined classes whereas Model Usage is helpful for classifying future or unknown objects.

Formerly, three Classifiers Decision Tree, Classification via Clustering and Naive Bayes were used to diagnose the presence of heart disease in patients [5]. Our work intends to use Decision Tree and Naive Bayes Classifier using fuzzy logic.

Decision Tree Classifier: is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached. Our work uses J48 tree for classification.

Naive Bayes Classifier: technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, it can outperform more sophisticated classification methods.

Classification via Clustering: Clustering is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. Classification is the task of generalizing known structure to apply to new data. Hence, classification is performed based on clustering.

5.1 Examining Classifiers

Classifiers have been trained to classify the medical data into four classes viz. "Normal", "Low Risk", "Medium Risk" or "High Risk". The confusion matrix of four classes for both the Classifiers using fuzzy logic is shown in Table 1 [5]

Predicted Class

Actual Class	**C1**	**C2**	**C3**	**C4**
C1	**Valid Positives**	**Invalid Negatives**	**Invalid Negatives**	**Invalid Negatives**
C2	**Invalid Positives**	**Valid Negatives**	**Invalid Positives**	**Invalid Positives**
C3	**Invalid Positives**	**Invalid Positives**	**Valid Negatives**	**Invalid Positives**
C4	**Invalid Positives**	**Invalid Positives**	**Invalid Positives**	**Valid Negatives**

Table 1: Confusion Matrix

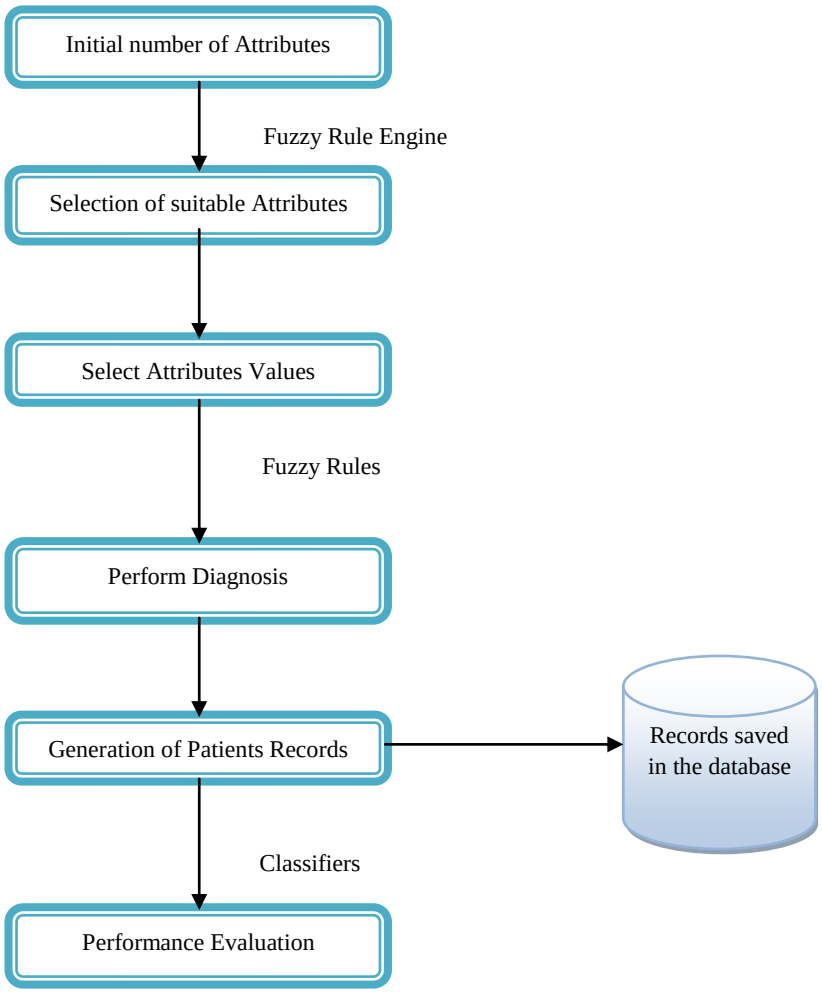

Fig 3: Overview of the Proposed System

In Table 1, Valid Positives indicate the positive results which were correctly identified by the classifier, while Valid Negatives are the negative results that were correctly identified by the classifier. Invalid Positives are the negative results that were incorrectly identified by the classifier, while Invalid Negatives are the positive results that were incorrectly identified by the classifier [5].

Table 2 depicts the Confusion Matrix for the proposed system using Classifiers and fuzzy logic. Observations exhibit that Invalid Negatives and Invalid Positives are 0 in this Confusion Matrix for various classes.

The Valid positives, Valid negatives, Invalid positives and Invalid negatives are also useful in assessing the costs and benefits associated with a classification model.

6. EXPERIMENTS AND RESULTS

Experiments were performed with Weka 3.6.2 Tool. Data set of 100 patients with 4 attributes is used. All attributes are converted into categorical form and discrepancies are resolved.

Originally, Decision Tree outperforms with good prediction probability of 99.62% by using 15 attributes. Our work exhibits more efficient results by using Decision Tree and Naive Bayes Classifier with Fuzzy Logic and reduced set of 4 attributes.

Results are shown in Table 3. It can be perceived from Table 3 that result of the proposed system outplays all other previous Classifiers.

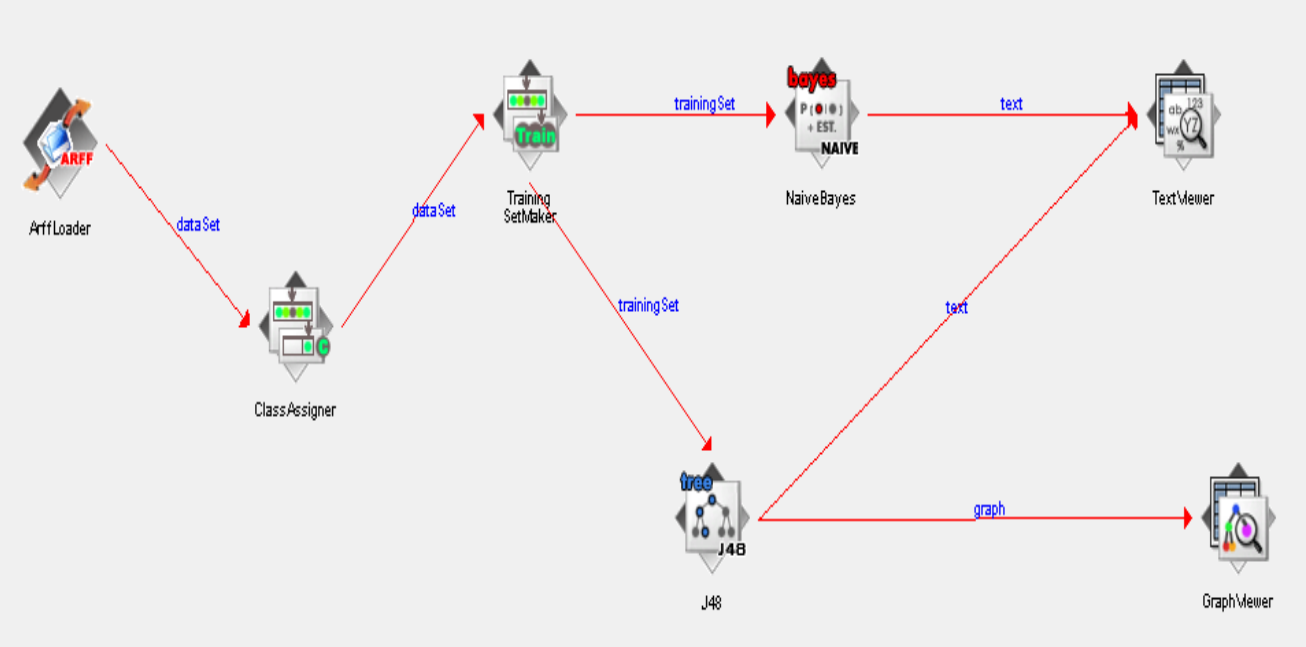

Fig 4: Knowledge Flow Layout for Decision Tree and Naive Bayes Classifiers

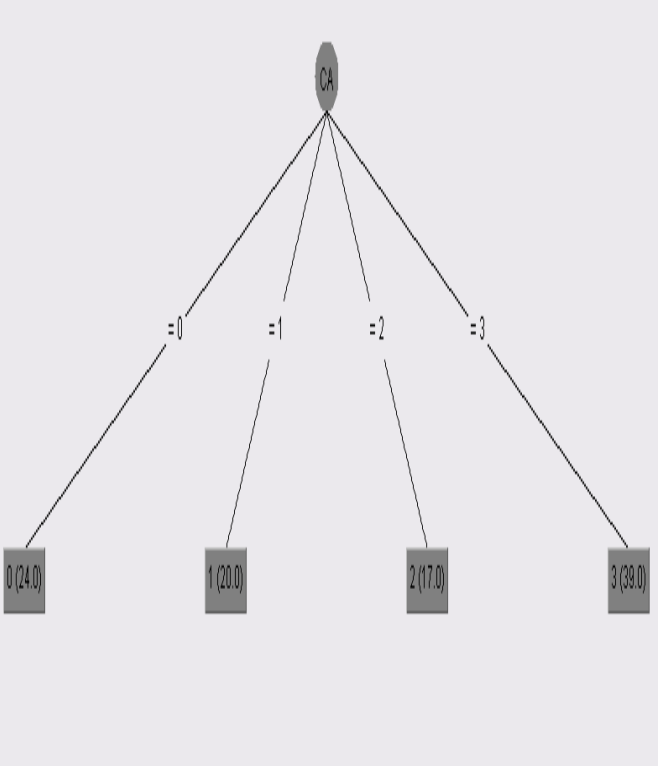

Fig. 5: Decision Tree generated by J48

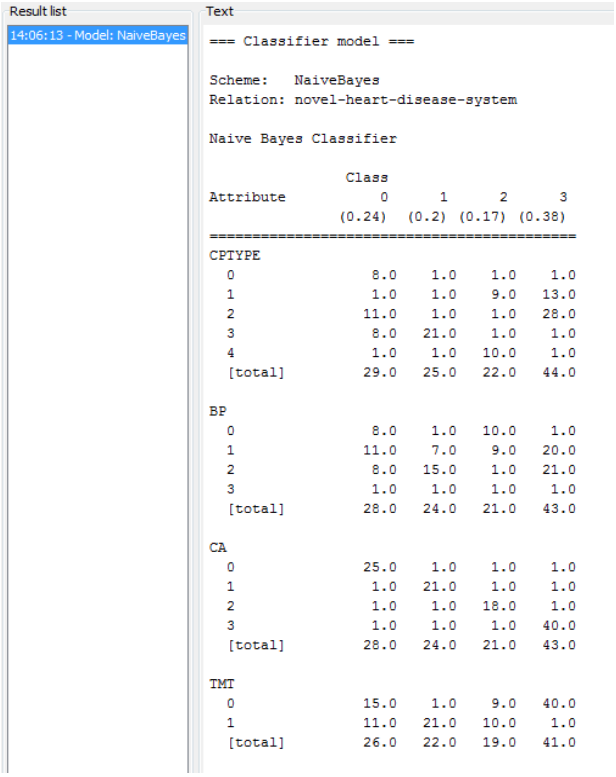

Fig. 6: Result List generated by Naive Bayes

	Healthy	Low Risk	Medium Risk	High Risk
Healthy	24	0	0	0
Low Risk	0	20	0	0
Medium Risk	0	0	17	0
High Risk	0	0	0	39

Table 2: Confusion Matrix for the Proposed System

DM Techniques	Efficiency			
	15 Attributes	13 Attributes	6 Attributes	4 Attributes (Proposed System)
Decision Tree	99.62%	96.66%	99.2%	100%
Naive Bayes	90.74%	94.44%	96.5%	100%

Table 3: Comparison Table for various Data Mining Techniques

7. CONCLUSION

The objective of our work is to predict more accurately the presence of heart disease with reduced number of attributes. Originally, six attributes were involved in predicting the heart disease. In our work, six attributes are reduced to four attributes which automatically reduces the number of tests to be taken by a patient. Subsequently, Decision Tree and Naive Bayes Classifiers using fuzzy logic are used for calculating the efficiency of the proposed system. Also, observations illustrated that mean absolute error for Decision Tree Classifier is 0 and it is 0.0072 in case of Naive Bayes Classifier.

8. REFERENCES

[1] P .K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules"; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40.

[2] E.P.Ephzibah, Dr. V. Sundarapandian, "Framing Fuzzy Rules using Support Sets for Effective Heart Disease Diagnosis"; International Journal of Fuzzy Logic Systems (IJFLS) Vol.2, No.1, February 2012.

[3] A.Sudha, P.Gayathri, N.Jaisankar, "Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability"; International Journal of Computer Applications (0975 – 8887) Volume 41– No.17, March 2012.

[4] Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques"; International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[5] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm"; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[6] E.Sivasankar, Dr.R.S.Rajesh, "Knowledge Discovery in Medical Datasets Using a Fuzzy Logic rule based Classifier"; 978-1-4244-7406-6/10/$26.00, IEEE, 2010.

[7] M.A. Saleem Durai, et. al. "Effective analysis and diagnosis of lung cancer using fuzzy rules"; International Journal of Engineering Science and Technology Vol. 2(6), 2102-2108, 2010.

[8] Mostafa Fathi Ganji, Mohammad Saniee Abadeh, "Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease"; Proceedings of ICEE 2010, May 11-13, 2010, 978-1-4244-6760-0/10/$26.00©2010 IEEE.

[9] Huang Hai, "Data Mining Based on a Compensative Fuzzy Neural Network"; International Conference On Computer Design And Applications (ICCDA), 2010.

[10] M.A.Saleem Durai, N.Ch.S.N.Iyengar, "Effective Analysis and Diagnosis of Lung Cancer Using Fuzzy Rules"; International Journal of Engineering Science and Technology, Vol. 2(6), 2010.

[11] Shantakumar B.Patil, Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network"; European Journal of Scientific Research, ISSN 1450-216X Vol.31 No.4, 2009.

[12] Rupa G. Mehta, Dipti P. Rana, Mukesh A. Zaveri, "A Novel Fuzzy Based Classification for Data Mining using

Fuzzy Discretization"; World Congress on Computer Science and Information Engineering, 2009.

[13] Markos G. Tsipouras et. al., "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling"; IEEE Transactions on Information Technology In Biomedicine, Vol. 12, No. 4, July 2008.

[14] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques"; 978-1-4244-1968-5/08/$25.00©2008 IEEE.

[15] Latha Parthiban, R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm"; International Journal of Biological and Life Sciences 3:3 2007.

[16] Niti Guru, et al. "Decision Support System for Heart Disease Diagnosis Using Neural Network"; Delhi Business Review, Vol. 8, No. 1, 2007.

[17] Kemal Polata, Salih Gunesa, Sulayman Tosunb, "Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted pre-processing"; Elsevier , Pattern recognation, 2007.

[18] Harleen Kaur, Siri Krishan Wasan, "Empirical Study on Applications of Data Mining Techniques in Healthcare"; Journal of Computer Science 2 (2): 194-200, 2006.

[19] Carlos Ordonez, "Comparing association rules and decision trees for disease prediction"; ACM, 2006.

[20] Boleslaw Szymanski, Long Han, Mark Embrechts, Alexander Ross, Karsten Sternickel, Lijuan Zhu, "Using Efficient Supanova Kernel for Heart Disease Diagnosis"; proc. ANNIE 06, intelligent engineering systems through artificial neural networks, vol. 16, pp:305-310, 2006.

[21] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease"; Springer, Vol:345, pp: 721- 727, 2006.

[22] Cleveland database: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[23] Han, J., Kamber, M, "Data Mining Concepts and Techniques"; Morgan Kaufmann Publishers, 2006.

[24] American Heart Association. Heart Disease and Stroke Statistics — 2004 Update. Dallas, Tex.: American Heart Association; 2003.

[25] Statlog database: http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart