A Fuzzy Optimization Technique for the Prediction of Coronary Heart Disease Using Decision Tree

Persi Pamela. I [1], Gayathri. P [2] and N. Jaisankar [3]
M.Tech Student [1], Assistant Professor (Senior) [2] and Professor [3]
School of Computing Science and Engineering, VIT University, Vellore – 632014, Tamil Nadu, India.
[1] persipamela@yahoo.com, [2] pgayathri@vit.ac.in, [3] njaisankar@vit.ac.in

***Abstract* - Data mining along with soft computing techniques helps to unravel hidden relationships and diagnose diseases efficiently even with uncertainties and inaccuracies. Coronary Heart Disease (CHD) is a killer disease leading to heart attack and sudden deaths. Since the diagnosis involves vague symptoms and tedious procedures, diagnosis is usually time-consuming and false diagnosis may occur. A fuzzy system is one of the soft computing methodologies is proposed in this paper along with a data mining technique for efficient diagnosis of coronary heart disease. Though the database has 76 attributes, only 14 attributes are found to be efficient for CHD diagnosis as per all the published experiments and doctors' opinion. So only the essential attributes are taken from the heart disease database. From these attributes crisp rules are obtained by employing CART decision tree algorithm, which are then applied to the fuzzy system. A Particle Swarm Optimization (PSO) technique is applied for the optimization of the fuzzy membership functions where the parameters of the membership functions are altered to new positions. The result interpreted from the fuzzy system predicts the prevalence of coronary heart disease and also the system's accuracy was found to be good.**

Keywords - Coronary Heart Disease, CART, Particle Swarm Optimization, Fuzzy System

I. INTRODUCTION

Coronary Heart Disease (CHD) is caused due to the blockage of arteries because of the fat deposition in the arteries supplying blood to the heart, thus restricting the blood flow. This results in myocardial infarction and sudden heart attack. The disease is becoming pandemic and affects even the younger generation. In countries like Australia and UK, heart disease is one of the highest causes of mortality [1].The symptoms are usually vague and correlated with other diseases. The diagnosis procedures are also time-consuming and prone to errors. This leads to adverse effects and sudden deaths. Proper treatment may not be given at the right time due to the false diagnosis. Patients may not accurately explain their difficulty or doctors at times may misinterpret the diagnosis. A clearer understanding of the disease including the risk factors and the attributes leading to the disease is important.

Data mining techniques with respect to medical diagnosis reveals hidden relationships and important patterns which are not disclosed in traditional diagnosis procedures [2]. Soft computing techniques like fuzzy systems perform well with inaccurate and fuzzy data while providing accurate results [3]. The applications of data mining and soft computing techniques in medical field is increasing day by day and helps in diagnosis of many complicated diseases including heart diseases. Knowledge-based systems are extensively applied in the fields anywhere knowledge is leading than data and entails logic in the reasoning to obtain new set of knowledge [4]. Fuzzy logic is an intelligent computing model which deals with imprecision and inexact data. It fuzzifies the crisp rules obtained from the decision tree algorithm and provides crisp results. The application of fuzzy logic in medical diagnosis is increasing at the present time for precise diagnosis. Decision tree plays a major role in discovering new and unseen observations through the decision rules. The Fuzzy sets when employed to medical diagnosis problems handles both the ambiguity of diagnosis and vagueness of symptoms [8].

In this paper all the existing approaches for Coronary Heart Disease (CHD) prediction have been reviewed. A fuzzy system for the prediction of coronary heart disease has been proposed along with the optimization of the fuzzy membership parameters. Particle Swarm Optimization (PSO) algorithm has been successfully applied in fuzzy systems for resolving problems which were not solved by conventional methods [7]. Due to ease of understanding and potential optimization ability, PSO is employed in solving several optimization problems [9].

II. RELATED WORK

Data mining methodologies are an efficient tool for identifying the knowledge concealed into huge medical databases [5]. Cardiovascular diseases have become the major reason for deaths in US as per the

research by American Heart Association [6] and also in many other countries like India. Coronary heart disease is also a cardiovascular disease leading to heart attack. The ability of data mining in solving medical diagnosis problems have been reported by World Health Organization (WHO) in 1997 [1].

Imran et al. proposed a comparative performance of Logistic Regression (LR), Classification and Regression Tree (CART), Multilayer Perception (MLP) Neural Networks and Self Organizing Feature Maps (SOFM) for predicting coronary heart disease and found that LR, CART and MLP performed better than SOFM but the classification accuracies were very low[10]. Artificial Immune Recognition System (AIRS) along with Principle Component Analysis (PCA) and k-Nearest Neighbor (k-NN) algorithms were used by Fatma et al, for the prediction of atherosclerosis (artery obstruction), which predicted the blockage of arteries throughout the body accurately, but not the artery obstruction of heart in particular [11]. Coronary heart disease diagnosis using Exercise Stress Testing (EST) along with neural networks was proposed by Ismail et al. but the system did not perform well with lesion localization [12]. Support Vector Machine (SVM) based heart valve disease prediction using heart sounds was proposed by Ilias et al. Though SVM is a common method of classification used in medical field, the classification accuracy was only 77% and did not in particular predict the coronary heart disease [13]. A comparative performance for feature selection using Binary Particle Swarm Optimization (BPSO) and Genetic Algorithm (GA) was proposed by Ismail et al. and found that BPSO performed better than GA [14]. The algorithms were centered only towards the attribute reduction using feature selection, not in heart disease prediction. A fuzzy-evidential hybrid inference engine for coronary heart disease prediction was proposed by Vahid et al. The fuzzy system consisted of fuzzy rule base and membership functions for diagnosis but the prediction accuracy was 91.58% [8].

A decision support system with Optimal Decision Path Finder (ODPF), which is a automatic decision making process, proposed by Chih-Lin Chi et al. for heart disease prediction [15]. Though cost savings were found in this algorithm, the prediction accuracy was only 55%. Association rule mining using multi resolution image parameterization was proposed by Matjaz et al for coronary artery disease diagnosis. The algorithm works with the scintigraphic images of heart and used with image processing. The accuracy technique involved was less than 90% and the diagnostic efficiency was very less [16]. Image processing and machine learning technique for evaluation of medical images was proposed by Luka et al. This also included the scintigraphic images and parameterization techniques but the classification accuracy was low though the diagnostic power was increased [17].

Dursan et al. proposed an analytic approach comparing SVM, Decision Trees (DT) like c5, CART and Neural Networks. The sensitivity analysis techniques were applied and process showed that SVM performed better than the other two with only 88% accuracy [5]. A fuzzy expert system approach for coronary heart disease prediction was proposed by Debabrata et al. [18]. The fuzzy rule base and knowledge base were formulated separately for the analysis. The entire set of processes indicated that fuzzy system performed better than ANN, ID3 and CART, but the accuracy of fuzzy system was only 84%. P.K. Anooj proposed a clinical decision support system using weighted fuzzy rules and fuzzy rule-based DSS.

The weighted procedure included along with fuzzy rules was an added advantage but the accuracy was 67.75% [19]. Using ANN as feature selection method for Ischemic heart disease prediction was proposed by Rajeswari et al., for reducing the number of features showed better performance [20]. The attribute reduction using back propagation algorithm was excellent with 89% accuracy only. Domain-driven decision support system for mining novelty rules from heart disease dataset [21] was proposed by Y. Sebastian et al. The rules were not properly evaluated and had many limitations such as large number of attributes and less accuracy. A feature selection method from ECG using CART for the prediction of myocardial infarction was proposed by Hui Yang et al [22]. The Electro Cardio Gram (ECG) and Vector Cardio Gram (VCG) features were selected and feature selection was made. The entire process included many complex procedures.

III. PROPOSED WORK

The proposed system deals with only the essential attributes taken from the Cleveland and Switzerland heart disease database from UCI machine learning repository. The proposed system architecture is given in fig. 1. Out of 76 attributes, 12 attributes are found to be essential for the diagnosis as mentioned in all published experiments. The attributes include age, sex, blood pressure, cholesterol, maximum heart rate, chest pain type, old peak, slope, thallium scan, and fasting blood sugar, rest ECG, exercise angina. With these 12 attributes the accuracy of the system was 93.27% [23]. Taking systolic and diastolic blood pressures into the diagnostic process the accuracy improves to 94.4%. So, 14 attributes are utilized in the CART decision tree algorithm for obtaining the crisp if-then-else rules. The output of the decision tree provides only the rules with these attributes and does not need any pruning mechanism to prune the unnecessary branches of the decision tree, this saves time and also appropriate rules can be obtained. These crisp rules are fuzzified in the fuzzy inference system through the triangular membership functions. Fuzzification is essential since a degree of membership is given for each member of the set. The parameters of these membership functions need to be tuned or optimized for a

better diagnosis. A Particle Swarm Optimization (PSO) method is employed, which locates the optimal results by iteratively improving an appropriate solution. With the optimized membership functions, the fuzzy system predicts the results more accurately.

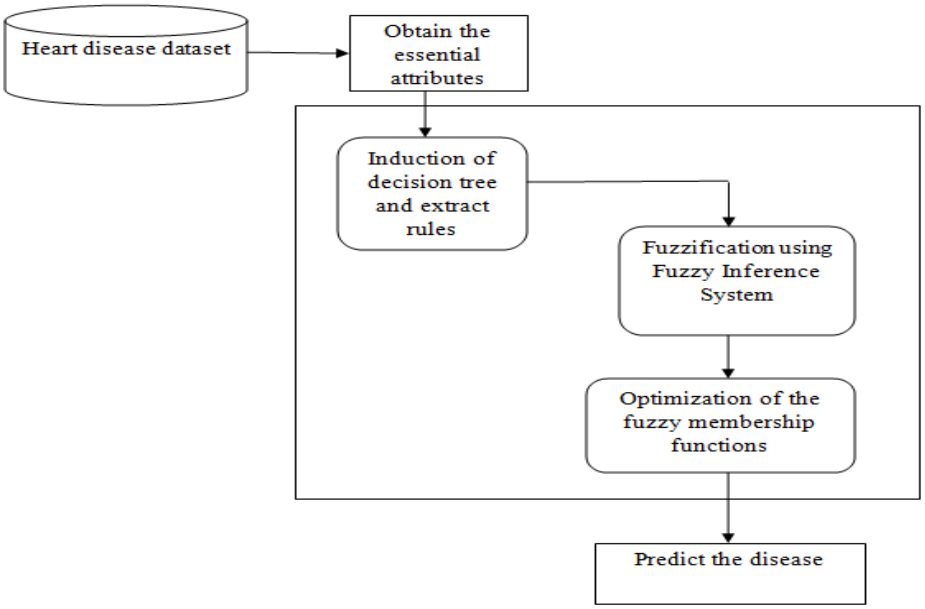

Fig. 1. System Architecture

A. CART Decision Tree Algorithm

Decision trees are a simple yet powerful method for numerous variable analyses. They provide exclusive capabilities to enhancement, accompaniment and replacement for

- Conventional statistical forms of analysis (multiple linear regression).
- A range of data mining methodologies and techniques (neural networks).
- Lately developed multidimensional structure of analysis.

Decision trees are generated by algorithms that provide different means of splitting a database into branch-like sections. These sections generated an inverted decision tree with the root node at the top followed by branches. CART is a decision tree algorithm applicable for both numerical and categorical data. It performs classification for categorical data and regression for numerical data. CART is found to be efficient since it deals with data which are not complete, data with different data types both in the input and predicted features and thus provides rules which are human-readable. CART has been previously used in many bio-medical applications for detection of cancer and heart diseases [24]. CART needs a measure of impurity to split a node, where and when to split that node. The default splitting criteria used in CART is the Gini diversity index. The measure should be at a greatest when a node is evenly separated among all classes and should be the least when the node contains only one class.

The Gini impurity measure is given as

$$i(t)=1-S$$

where S is the impurity criteria given as $\sum p^2(j|t)$, for j = 1 to k, k denotes the number of classes and p(j|t) denotes the probability of class j in node t. The decision tree yields crisp rules along with the decision tree.

B. Fuzzy System

Fuzzy system is one of the soft computing methodologies used to solve problems dealing with inaccurate and imprecise data but gives accurate results. It performs fuzzification using the fuzzy inference engine and knowledge base and finally defuzzification which gives the crisp output. The architecture of the fuzzy system is given in fig. 2 where the input is given to the fuzzifier, with the knowledge base; the inference engine fuzzifies the input. Finally defuzzification is performed to obtain the crisp result.

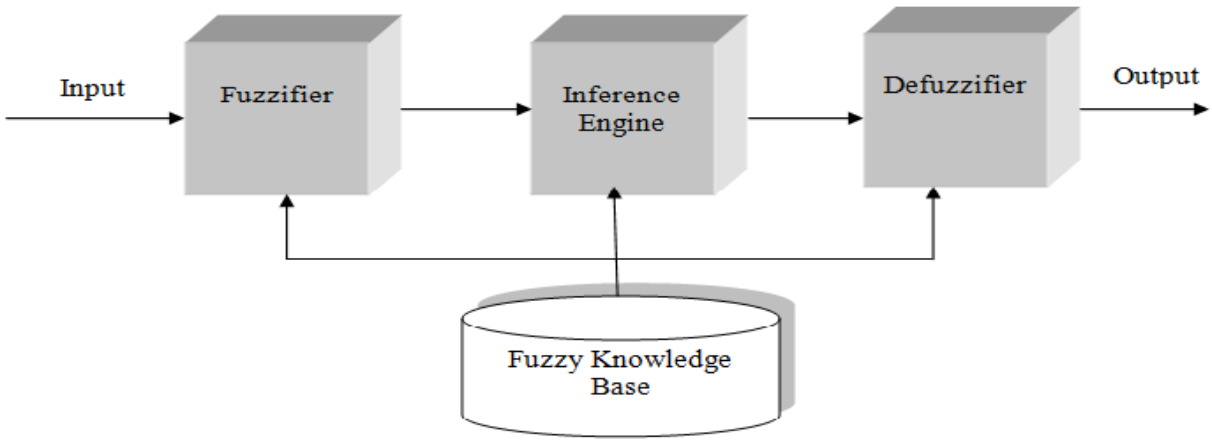

Fig. 2. Architecture of the fuzzy system

- Fuzzification - Changes the crisp input to a linguistic variable with the membership functions gathered in the fuzzy knowledge base.
- Fuzzy Inference Engine – With the help of If-Then type fuzzy rules, changes the fuzzy input into the fuzzy output.
- Defuzzification - Changes the fuzzy output of the inference engine to crisp using membership functions equivalent to those utilized by the fuzzifier.

Mamdani fuzzy inference system is used and for generating the membership functions, triangular membership functions have been employed, defuzzification is performed through Centroid of is method which is applicable to fuzzy sets of any shape.

C. Particle Swarm Optimization (PSO)

Particle Swarm Optimization is a population based optimization algorithm motivated by the communal activities of bird flocking and fish schooling. Though PSO shares similarities with genetic algorithms, it has some variations including the avoidance of genetic operators like mutation and cross-over. When compared to genetic algorithms, PSO has only less parameters to adjust and much easier to implement.

The population is initialized with a group of random particles. Every particle has two values associated with it, a personal best (pbest) and a global best (gbest) values. The best fitness value achieved by the particle is the pbest and the best value obtained in the overall population is the gbest. With these values, the velocities of the particles are calculated and thus the positions are updated. The process continues until the maximum number of iterations is reached. The optimization process takes place with the values of cognitive acceleration, social acceleration, values of the velocities at the beginning and end of the optimization process. An objective function is created for optimizing the parameters of the fuzzy membership functions. The values of the fuzzy membership function parameters are changed to new positions after optimization.

IV. IMPLEMENTATION DETAILS

The implementation was carried out in MATLAB 10. The CART decision tree algorithm is implemented, followed by the fuzzy systems and the optimization algorithm. The optimized membership functions predict the prevalence of coronary heart disease much accurately than without optimization. The implementation of the decision tree algorithm, fuzzy systems and the optimization algorithm are given below.

A. Implementation of CART decision tree

From the dataset obtained, only the essential attributes as mentioned before are utilized in the decision tree algorithm. With these attributes, crisp rules are generated along with the decision tree from which the rules can be interpreted. Gini impurity index is the default splitting criterion; the value should be at its maximum when a node is uniformly divided amongst all the classes. Splitting is performed until the terminal nodes have extremely small number of cases. The 14 attributes from the dataset are applied to the CART decision tree, the proportion of each node is calculated using the Gini index and the tree is splitted accordingly. Pruning of the nodes can be done if necessary to prune off the unwanted nodes from the decision tree. The implementation is done in MATLAB; the built-in functions are available for the generation of classification and regression trees. The obtained decision tree is easy to interpret and understand; the rules generated are crisp and human-readable and can be applied to the fuzzy system.

B. Implementation of Fuzzy System

The fuzzy logic toolbox available in MATLAB is utilized for generating the fuzzy system. The membership functions for the attributes are initialized with the membership function editor in the fuzzy editor. Triangular membership functions are employed since it is easy to understand and widely accepted for many applications. The crisp rules from the decision tree are included in the rule editor, which forms the fuzzy rule base. With the membership functions and the rule base, the rule viewer in the fuzzy editor is used for displaying the output. The rule viewer provides the defuzzified output, which is easy to interpret whether the patient is affected with coronary heart disease or not.

C. Implementation of Particle Swarm Optimization Algorithm

The parameters of the fuzzy system should be optimized in the particle swarm optimization algorithm. The values of the triangular membership functions are adjusted to new positions after optimization. An objective function is initialized first with these membership function parameters, which has to be optimized. The details of the algorithm are given above, where the particles are initialized first and the best positions of the particles are updated on every iteration. The objective function is optimized where the new values for the membership functions are obtained. The triangular membership functions are moved either right or left, accordingly the output in the rule viewer is changed to new values. With the optimized membership functions, the fuzzy system provides a more optimized and accurate results. There may be changes in the prediction results for some records which affects the performance of the system developed. The prediction results are accurate than the unoptimized output.

V. RESULTS AND DISCUSSION

The output of the CART decision tree algorithm generates the decision tree along with the rules. The rules are generated as in fig. 3. The obtained rules also predict the prevalence of coronary heart disease. 16 rules are obtained from the decision tree which is then applied to the rule editor of the fuzzy system. Fig. 4 depicts the decision tree obtained from which the rules can be easily interpreted where the leaf nodes display the outcome, whether the patient is affected with coronary heart disease or the status is normal. The options available in the fuzzy editor are utilized for generating the triangular membership functions, which are widely used for many applications. Fig. 5 show the membership function for 'BP_Systole' before optimization where the four fuzzy sets for low, medium, high and very high are depicted. Fig. 6 depicts the membership function for 'BP_Systole' after optimization where the four fuzzy sets are altered to new positions. The variations in the triangular membership functions before and after optimization are shown clearly. Similarly, for the other input membership functions optimization is done and the membership functions are moved to new positions. The rule viewer in the fuzzy editor is used for displaying the output. The rule viewer for one of the test data before and after optimization is shown in fig. 7 and fig. 8 respectively. The fuzzy rule viewer displays the input attribute values of the patients and the output value is the defuzzified result. The output value before optimization shows no prevalence of coronary heart disease, whereas after optimization, the value differs and shows the prevalence of coronary heart disease.

```
t =

Decision tree for classification
 1  if BP<139 then node 2 elseif BP>=139 then node 3 else normal
 2  if thalium<6.5 then node 4 elseif thalium>=6.5 then node 5 else normal
 3  if BP<157 then node 6 elseif BP>=157 then node 7 else CHD
 4  if chol<318.5 then node 8 elseif chol>=318.5 then node 9 else normal
 5  if BP<117.5 then node 10 elseif BP>=117.5 then node 11 else normal
 6  if chest_pain_type<1.5 then node 12 elseif chest_pain_type>=1.5 then node 13 else CHI
 7  if max_hrt_rate<162.5 then node 14 elseif max_hrt_rate>=162.5 then node 15 else CHD
 8  if chest_pain_type<1.5 then node 16 elseif chest_pain_type>=1.5 then node 17 else no:
 9  class = CHD
10  if fbs<89 then node 18 elseif fbs>=89 then node 19 else normal
11  if max_hrt_rate<162 then node 20 elseif max_hrt_rate>=162 then node 21 else CHD
12  class = normal
13  if old peak<1.25 then node 22 elseif old peak>=1.25 then node 23 else CHD
14  class = CHD
15  class = normal
16  class = normal
17  if age<75.5 then node 24 elseif age>=75.5 then node 25 else normal
18  class = normal
19  class = normal
20  if BP_diastole<72.5 then node 26 elseif BP_diastole>=72.5 then node 27 else CHD
```

Fig. 3. A part of rules from decision tree

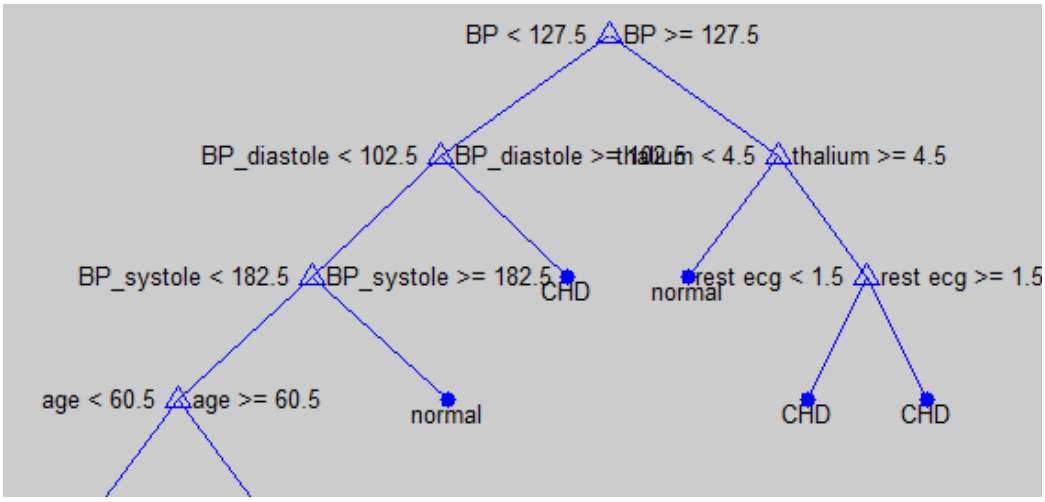

Fig. 4. A part of decision tree obtained

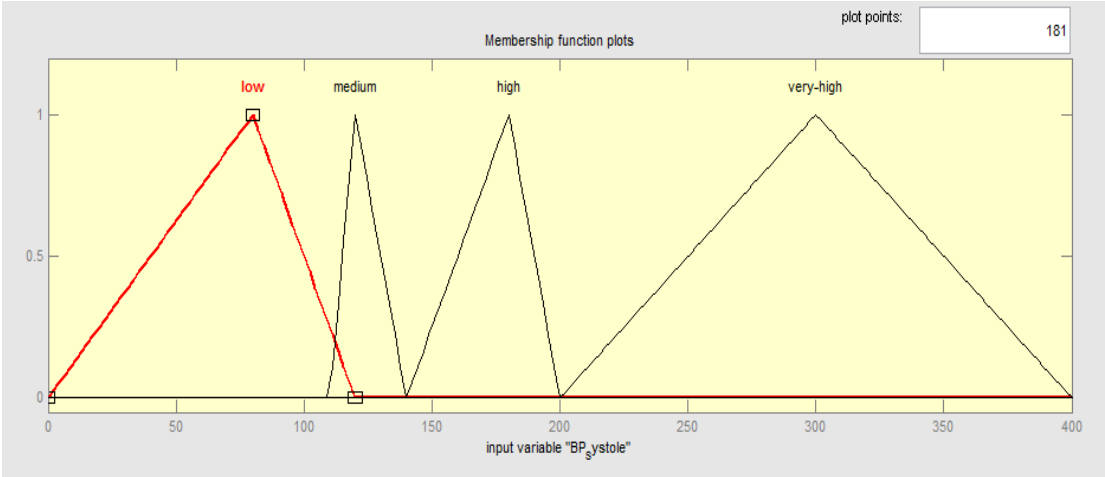

Fig. 5. Membership function for 'BP_Systole' before optimization

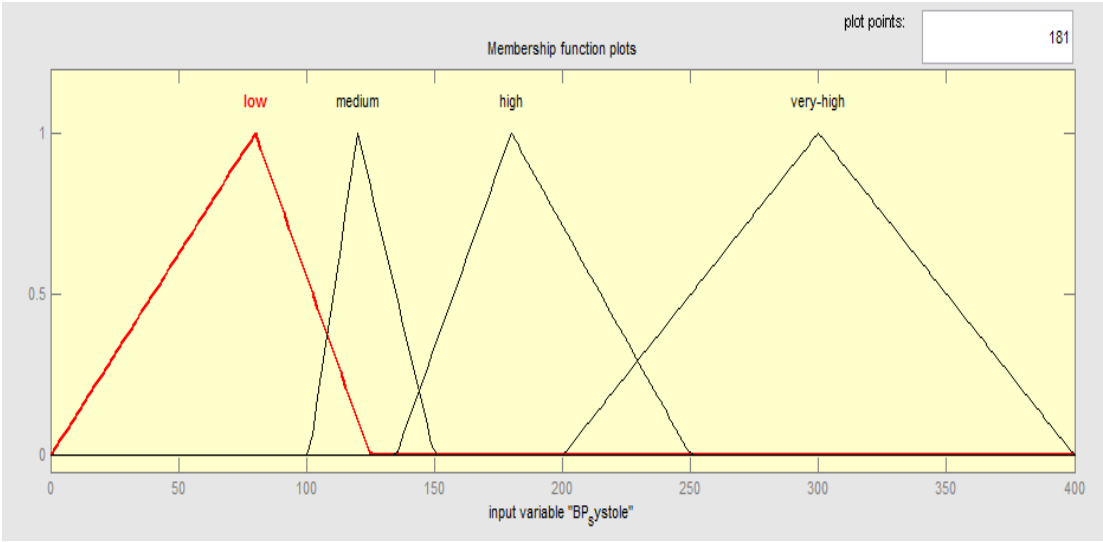

Fig. 6. Membership function for 'BP_Systole' after optimization

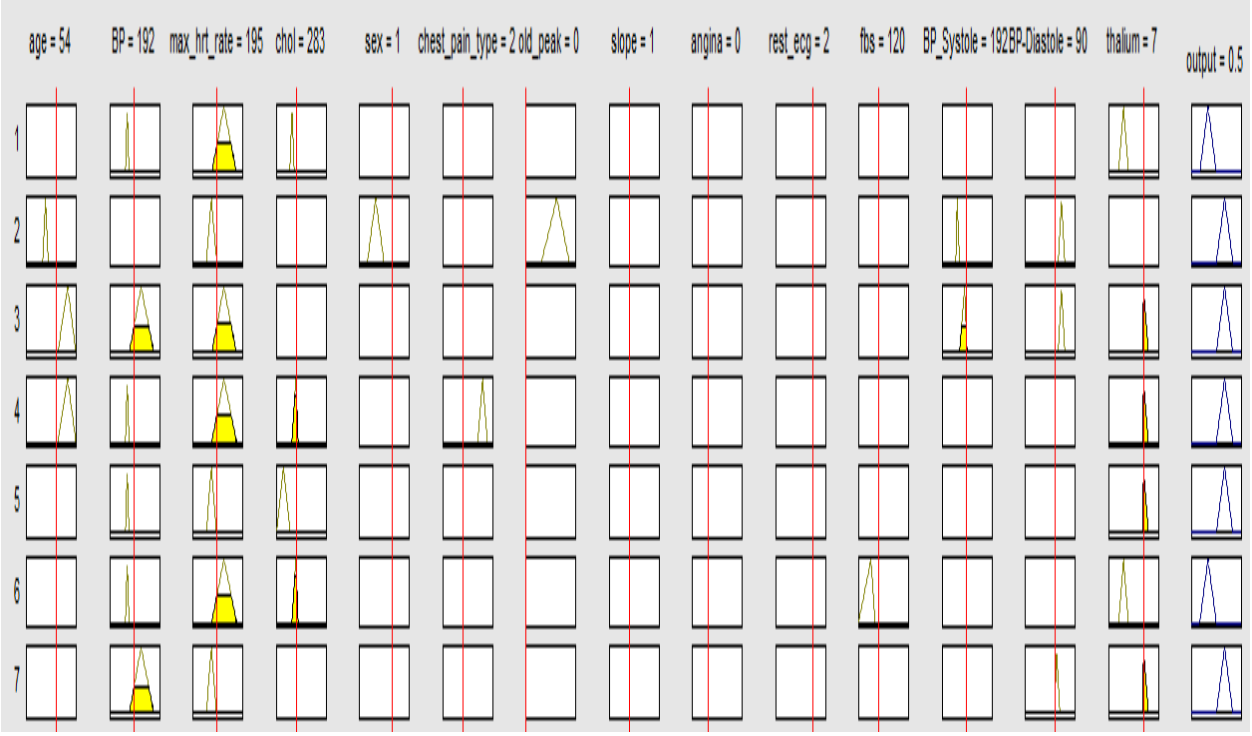

Fig. 7. Rule viewer for one of the test data before optimization

Fig. 8. Rule viewer for one of the test data after optimization

VI. PERFORMANCE ANALYSIS

For analyzing the performance of the proposed system, out of 250 instances from the Cleveland database, 90 instances were used as the test data. Another database from Switzerland hospital was taken and out of 122 instances, 50 instances were used for testing. A confusion matrix was obtained with the true positives, true negatives, false positives and false negatives. With these values sensitivity, specificity and accuracy of the system is calculated. The accuracy of the system with Cleveland database was 92.2% with specificity 90% and sensitivity 95%, before optimization and after optimization, the accuracy was 94.4% with specificity 92% sensitivity 97.5% as shown in fig. 9. While, for Switzerland data the accuracy was 86% before optimization with specificity 91.6% and sensitivity 84%. The accuracy increased to 94% after optimization with specificity 92.5% and sensitivity 95.6% as shown in fig. 10.

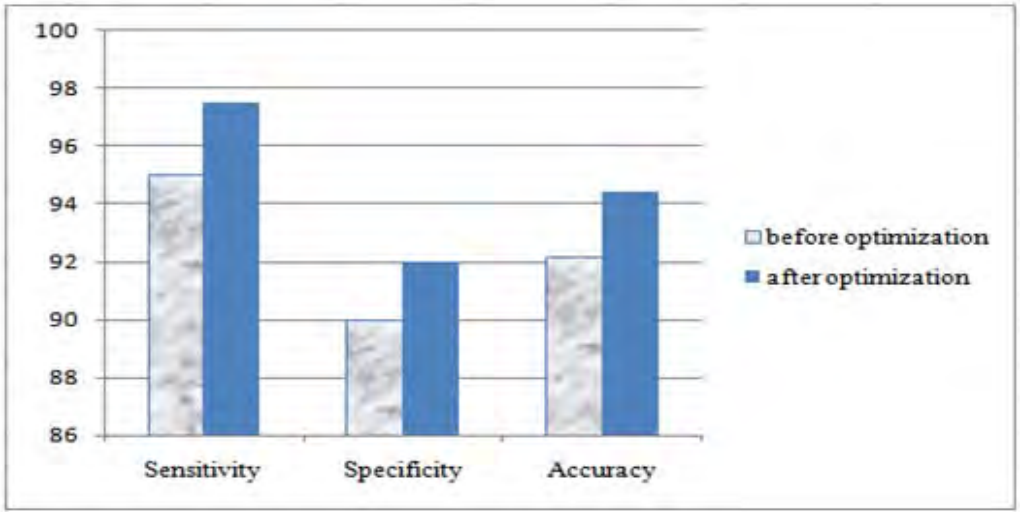

Fig. 9. Performance of Cleveland database

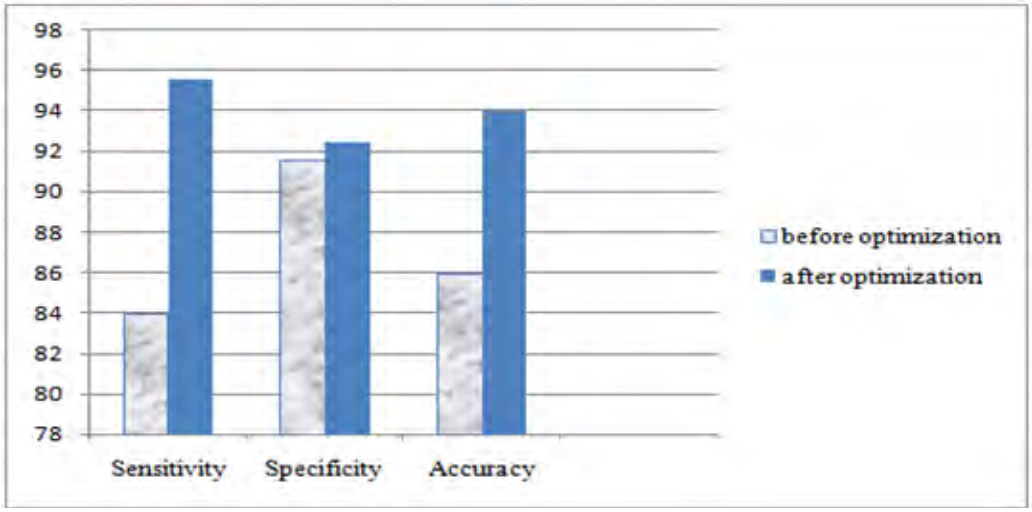

Fig. 10. Performance of Switzerland database

VII. CONCLUSION AND FUTURE ENHANCEMENT

The accuracy of the proposed system is found to be good for both Cleveland and Switzerland databases when compared to that of the existing work. Selection and application of only the essential attributes greatly influences the performance of the system. With the prediction of coronary heart disease, early treatment can be given at the right time which avoids the risk of heart attacks. Since the diagnosis involves simple procedures and is easy to obtain the required results, the proposed system is found to be efficient than the other existing systems.

However, the performance of the proposed work can be enhanced by including few additional attributes and checked for accuracy. This should be done along with detailed survey and doctors' opinion. As for the proposed system, only benchmark databases have been used, in future real-time databases can also be applied and checked for results. The optimization is performed for the fuzzy system, however with other soft computing methodologies like neural networks; this optimization technique could be applied in future.

REFERENCES

[1] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen, *"Association rule mining to detect factors which contribute to heart disease in males and females"*, Expert Systems with Applications, 40 (2013), pg: 1086–1093.
[2] Krzysztof J. Cios, G. William Moore, *"Uniqueness of medical data mining"*, Artificial Intelligence in Medicine, 26 (2002), pg: 1–24.
[3] Ahmet Yardimci, *"Soft computing in medicine"*, Applied soft Computing, 9 (2009), pg: 1029-1043.
[4] Babita Pandey, R.B.Mishra, *"Knowledge and intelligent computing system in medicine"*, Computers in Biology and Medicine, 39 (2009), pg: 215 – 230.
[5] Dursun Delen, Asil Oztekin, Leman Tomak, *"An analytic approach to better understanding and management of coronary surgeries"*, Decision Support Systems, 52 (2012), pg: 698-705.
[6] Heart disease and stroke statistics, *"Heart disease and stroke statistics update"*, American heart association, available at *http://www.americanheart.org*.

[7] Hassan M. Elragal, *"Using swarm intelligence for improving accuracy of fuzzy classifiers"*, International Journal of Electrical and Computer Engineering, 5:2 2010.
[8] Vahid Khatibi, Gholam Ali Montazer, *"A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment"*, Expert Systems with Applications, 37 (2010), pg: 8536-8542.
[9] Dong-ping ian, Nai-qian Li, *"Fuzzy particle swarm optimization algorithm"*, International Joint Conference on Artificial Intelligence, 2009.
[10] Imran Kurt, Mevlet Ture, A. Turhan Kuram, *"Comparing performances of logistic regression, classification and regression tree and neural networks for predicting coronary artery disease"*, Expert Systems with Applications, 34 (2008), pg: 366-374.
[11] Fatma Latifoglu, Kemal Polat, Sadik Kara, Salih Gunes, *"Medical diagnosis of atherosclerosis from Carotid Artery Doppler Signals using PCA, k-NN based weighted pre-processing and Artificial Immune Recognition System (AIRS)"*, Journal of Biomedical Informatics, 41 (2008), pg: 15-23.
[12] Ismail Babaoglu, Omer Kaan Baykan, Nazif Aygul, Kurtulus Ozdemir, *"Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization"*, Expert Systems with Applications, 36 (2009), pg: 2562-2566.
[13] Ilias Maglogiannis, Euripidis Loukis, Elias Zafiropoulos, Antonis Stasis, *"Support vectors machine-based identification of heart valve diseases using heart sounds"*, Computer methods and programs in biomedicine, 95 (2009), pg: 47-61.
[14] Ismail Babaoglu, Oguz Findik, Erkan Ulker, *"A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine"*, Expert Systems with Applications, 37 (2010), pg: 3177-3183.
[15] Chih-Lin Chi, W. Nick Street, David A. Katz, *"A decision support system for cost-effective diagnosis"*, Artificial Intelligence in Medicine, 50 (2010), pg: 149-161.
[16] Matjaz Kukar, Igor Kononenko, Ciril Groselj, *"Modern parameterization and explanation techniques in diagnostic decision support system. A case study in diagnostics of coronary artery disease"*, Artificial Intelligence in Medicine, 52 (2011), pg: 77-90.
[17] Luka Sajn, Matjaz Kukar, *"Image processing and machine learning for fully automated probabilistic evaluation of medical images"*, Computer Methods and Programs in Biomedicine, 104 (2011), pg: e75-e86.
[18] Debabrata Pal, K.M. Mandana, Sarbajit Pal, Debranjan Sarkar, Chandan Chakraborty, *"Fuzzy expert system approach for coronary artery disease screening using clinical parameters"*, Knowledge-Based Systems, (2012).
[19] 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, 24(2012), pg: 27-40.
[20] K. Rajeswari, Dr. V. Vaithiyanathan, Dr. T. R. Neelakantan, *"Feature selection in Ischemic heart disease identification using feed forward neural networks"*, Procedia Engineering, 41 (2012), pg: 1818-1823.
[21] Y. Sebastian, Patrick H. H. Then, *"Domain-driven KDD for mining functionally novel rules and linking disjoint medical hypotheses"*, Knowledge-Based Systems, 24 (2011), pg: 609-620.
[22] Hui Yang, Satish T. S. Bukkapatnam, Trung Le, Ranga Komanduri, *"Identification of myocardial infarction using spatio-temporal heart dynamics"*, Medical Engineering and Physics, 34 (2012), pg: 485-497.
[23] S. Muthukaruppan , M.J. Er ," *A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease"*, Expert Systems with Applications, 39 (2012), pg: 11657–11665.
[24] *Link: http://www.researchmethods.org/CARTIntroTutorial.pdf.*