PERDICTION OF CORONARY HEART DISEASE USING DECISION TREE

SOWMYA N.

13PCS014

A Project Report Submitted to

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore-641043

In Partial Fulfillment of the Requirements for the Master's Degree in Computer Science

March, 2015

PERDICTION OF CORONARY HEART DISEASE USING DECISION TREE

SOWMYA N.

13PCS014

A Project Report Submitted to

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore-641043

In Partial Fulfillment of the Requirements for the Master's Degree in Computer Science

March, 2015

Signature of the Supervisor **Signature of the Head of the Department**

Signature of External Examiner

ACKNOWLEDGEMENT

ACKNOWLEDGEMENT

I would like to express my sincere thanks to God Almighty, for his constant love and grace that he has showered upon me.

I am very grateful to **Dr. T. S. K. MeenakshiSundaram, M.A., M.Phil., Ph.D., Chancellor,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for his support and encouragement during the course of my study.

I heartily thank **Dr. (Mrs). SheelaRamachandran M.Sc., P.G. Dip., Ph.D., Vice Chancellor,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for extending all resources that facilitated the conduct of the present study.

I express my humble gratitude to **Dr. (Mrs) A.Venmathi M.Sc., M.Phil., Ph.D., Registrar,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all facilities necessary for the study.

I obliges the support of **Dr. (Mrs). Saroja Parabhakaran**, Director, Hall of Residence, Sri Avinashiligam Trust Hostel, Coimbatore for her heart full blessings and support.

I am also thankful to **Dr. (Mrs). A. Parvathi, M.Sc., Dip.Ed. M.Phil., Ph.D., Dean, Faculty of Science,** for granting the facility required.

I wish to place on record my deep sense of gratitude to **Dr.(Mrs).G.Padmavathi M.Sc., M.Phil., Ph.D., Professor and Head,** Department of Computer Science, for providing all the facilities to complete the project.

I owe great deal of gratitude to my esteemed guide and my project Coordinator **Dr.(Mrs).R.Vijayabhanu, MCA., M.Phil., Ph.D., Assistant Professor,** Department of Computer Science, for imparting the tremendous assistance and well timed support for triumph of my project.

I have great pleasure in expressing my deep sense of gratitude to all teaching and non-teaching staff members who stood behind the screen in making of project.

I would extend my hearty thanks to one and all that helped me directly or indirectly for successful completion of my project.

ast yet importantly, I would like to thank my parents, my brother and all my well-wishers for their kind inspiration.

SYNOPSIS

SYNPOSIS

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. An Artificial Bee Colony Optimization (ABC) 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.

CONTENTS

CONTENTS

INTRODUCTION

1. INTRODUCTION

Today's world of preset data anthology and futuristic database provides us with a copious amount of information in various e-formats. Detection and prediction with certain knowledge has become effective through data analytics. Extraction of acquaintance and preprocessing of missing attributes increase the process of such system. Splitting of huge database into branches with the classifiers such as Decision tree which enhance the process of prediction. The fuzziness of the system should be impassive and optimized to gain better accuracy of the prophecy. The significance of the prediction in the health check domain is rapidly escalating now -a -days. One of such medical forecast that prevents heart attacks and sudden deaths is the prediction of Coronary Heart Disease.

1. 1. CORONARY HEART DISEASE (CHD)

According to aetiology, **Atherosclerosis** is alleged to be the most assassinating disease in the majority of developed and developing countries like India. Atherosclerosis is a medical terminology used to describe the ruptures of the arteries of the heart muscles by causing blood clots or plaque. The **plaque** is made up of fat, cholesterol, calcium and other substances which build upon the walls of the blood vessels. The arteries are responsible for supplying the oxygen rich blood to the heart muscles. When the blood clot grows huge enough, it blocks the flow of oxygen affluent blood to the heart muscles absolutely as shown in Fig 1. This causes angina or heart attacks or even to sudden death. The angina is the chest pain or uneasiness due to the lack of oxygen in the blood [31].

The **Coronary Heart Disease (CHD)** is the damage in the interior of the coronary arteries leading to heart attacks and Arrhythmias. The Arrhythmias is the problem even in adolescent people with respective to the rate or rhythm of their heartbeat [29]. The early prediction of such tedious disease can reduce the mortality rate. These types of prophecy nowadays are quite impressive in the world of robotic technology. The following inspection provides us with better way to end with such prediction proposals. The disease is becoming pandemic and affects even the younger generation. 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 [28].

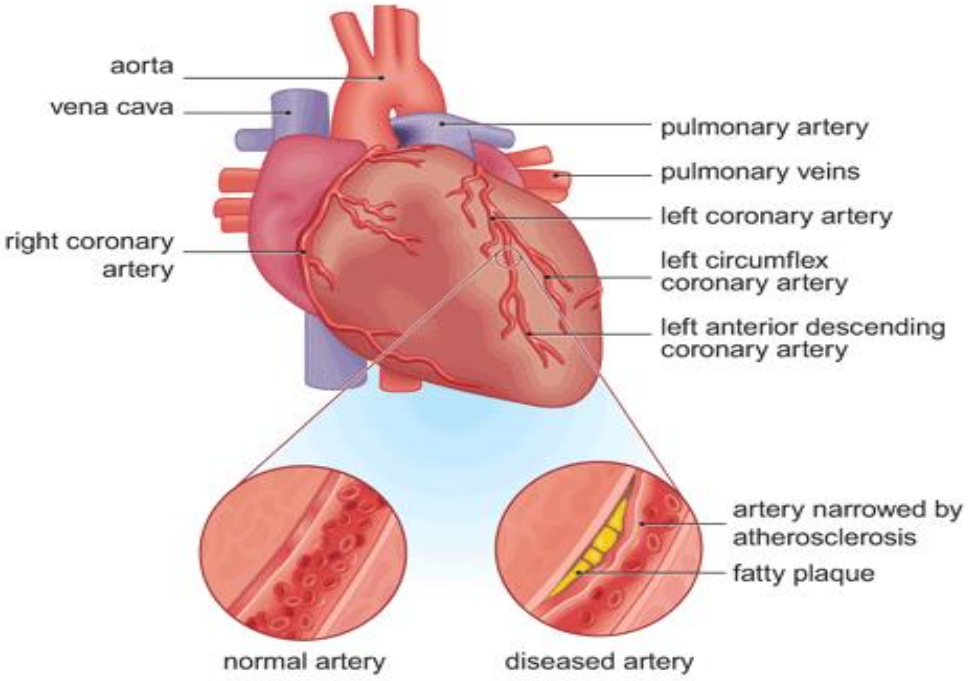

Fig 1. A Coronary artery affected by Atherosclerosis

1.2. DATA MINING

The digital world of today produces a copious source of data known as big data. This big data consists of non-trivial, hidden, previously unknown and potentially valuable data [19]. Such data can be used effectively for the prediction of future trends. This can be done through data mining processes. The veiled patterns and relationships can be retrieved knowledge based on mining. This effective tool is used to identify the acquaintance based on convinced criteria. Thus, it is sometimes known as data or knowledge discovery.

The ultimate goal of data archeology is to extract useful data and analyze them with different perceptions and gather them into useful information. There are large database of information that has been stored in various electronic forms which may consist of curtailed,

noisy and inconsistent data. The data mining methodologies are used to harvest the embedded information which is used as a cause of knowledge for decision building.

The electronic data are preprocessed by the data or pattern analysis to construct the predictive modules. These modules are rooted with the algorithms such as K-means, SVM, KNN, CART, Naïve Bayes, etc to envisage the indispensable knowledge or information for decision making.

1.3. DECISION TREE

A decision tree is a widely used data mining classifier, which incorporates both nominal and numerical data. Being uttered as a recursive partition of the instance space, the decision trees use certain discrete function of the input attributes. According to a Survey on Decision Tree Algorithm For Classification by Brijain R Patel *et al.* (2014) [5], to extort models from a large data set there are two forms of data analysis namely classification and prediction. Such analysis can be effectively worn to foresee for prospecting data trends.

The decision trees are the well known paradigm to depiction any discrete value classifier that is proficient of handling datasets that may have error and missing values. Consequently, these trendy approaches are used to predict the accuracy of CHD in a choice of related studies. Widely used heart disease datasets in decision tree research consists of 303 numbers of instances and 75 numbers of attributes.

'The tree complexity has a decisive effect on its accuracy' was the statement by *Breiman et al. (1984)* [4]. This tree complexity of the decision tree is clearly achieved by using stopping criteria and the pruning methodologies. Thus, the decision tree inducers provide exclusive potential to boost the conventional statistical forms of analysis.

The decision tree inducers are the algorithms that involuntarily construct a decision tree from a specified dataset. The primary objective is to obtain the optimal decision tree thereby minimizing the generalization error. The decision tree inducers can be reflected on either of top-down or bottom-up approaches. The greedy algorithm is considered to the indispensable learning approach that proceeds with the recursive top down approach of decision

tree structure. The decision tree algorithm has experienced a lot in the world of data mining. These inducers algorithms such as CART, C4.5, and C5 are largely used in the predictions.

1.4. FUZZY SYSTEM

A fuzzy system is considered to be the conservatory of the traditional fuzzy mathematics. The fundamental of the fuzzy mathematics are laid by the fuzzy sets and the fuzzy logic. The multi-valued logic that allows transitional values to be defined between conventional evaluations like 0/1, true/false, yes/no, high/low, etc are known as Fuzzy Logic. The fuzzy set is considered to be the basic notion of the fuzzy system. Membership functions of fuzzy sets can be distinct in any integer of ways as long as they follow the rules of the description of a fuzzy set [1].

The fuzzy logic is said to be the superset of Boolean logic that has been unmitigated to grip the concept of the partial certainty values between 'completely true' and 'completely false'. The fuzzy system logic recognizes further than simple true and false values. The expertise considers fuzzy logic as "a constitution of knowledge depiction appropriate for notions that cannot be defined accurately, but which depend upon their context".

The Classical set or the Crisp set contains the objects that can convince accurate properties of membership. The Crisp membership functions have values of either one or zero. But the fuzzy set contains the vague properties of membership in correspondence to their objects. Fuzzy is said to determine "possibility" rather than "probability". The impetus of the fuzzy logic is to alleviate difficulties in developing and analyzing complex systems encountered by the conventional mathematical utensils.

Fuzzy Logic Process

The Fuzzy logic process is a progression of computing, reasoning and modeling with the fuzzy familiarity. Despite the fact that the massiveness of the information we incorporate each day with fuzzy, most of the actions or decisions implemented by humans or machines are crisp or binary. Fuzzy logic provides a substitute way to represent linguistic and subjective attributes of the real world in computing [3]. It is able to be applied to control systems and other applications in order to improve the efficiency and simplicity of the design process.

1.5. OPTIMIZATION

The analytical methods that are used to stumble on the optimum solution or unimpeded maxima or minima of constant and differentiable function are said to be the Classical optimization techniques. The formula behind these techniques is executed iteratively by comparing diverse solutions in order to acquire the expected optimal result. There are two discrete types of optimization algorithms generally used

- **Deterministic Algorithms** - Specific rules for moving one solution to other.
- **Stochastic Algorithms** - Probabilistic translation rules for gaining popularity due to certain properties. One of such algorithm is the Swarm Intelligent (SI).

Swarm Intelligent Optimization

The synthetic intelligence which is based on the collective performance of decentralized and self-organized systems is known as Swarm Intelligent (SI). The SI is a loosely structured collection of interacting agents which can be distinguished, communicated and/or interrelated with each other. Since the agents can be easily added or removed without influencing the composition of the system, it is measured to be flexible and can be adapted in new situations.

Swarm Intelligence indicates a recent computational and behavioral metaphor for solving distributed problems that originally took its inspiration from the biological examples provided by social insects (ants, termites, bees, wasps) and by swarming, flocking, herding behaviors invertebrates. Any attempt to design algorithms or distributed problem-solving devices inspired by the collective behavior of social insects and other animal societies. Main algorithmic frameworks based on the notion of Swarm Intelligence: Collective Intelligence, Particle Swarm Optimization, Ant Colony Optimization Computational complexity, NP-hardness and the need of (meta) heuristics Some popular meta heuristics for combinatorial optimization tasks.

This manuscript deals with prophecy of the Coronary Heart Disease by preprocessing of the heart disease database to salvage the efficient attributes and classifying of the colossal database into prediction nodes. The nodes are evaluated for eliminating the vagueness of the system and are optimized to obtain the best accuracy.

REVIEW OF LITERATURE

1. RELATED WORK

Data mining methodologies are an efficient tool for identifying the knowledge concealed into huge medical databases [9]. Cardiovascular diseases have become the major reason for deaths in US as per the research by American Heart Association [11] 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 [18]. 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[15].

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 [10]. 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 [16]. 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 [14].

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 [17]. 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% [26]. 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 [6].

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 [21].

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 [20]. 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 [9].

A fuzzy expert system approach for coronary heart disease prediction was proposed by Debabrata *et al.* [7]. 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% [2].

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 [24]. 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 [25] 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 [13]. 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. The following table 1 illustrates the summary of above discussed literature survey.

S.No	Paper Reference No	Author of the paper	Year Publication	Methodology used	No. of attributes / dataset	Attained accuracy (Approximately) %
1	[1]	Persi Pamela *et al.*	2013	CART decision Tree; Fuzzy System ;PSO	14	94%
2	[3]	Markos etal.	2008	Decision Tree; Fuzzy modeling & Optimization	19	73.4%
3	[4]	Kantesh *et al.*	2014	Fuzzy reasoning	6	80%
4	[5]	K Cinetha *et al.*	2014	Fuzzy logic; Decision Tree with Clustering	1230 (Training data)	97.67%
5	[6]	Debabrata *et al.*	2012	CAD Screening Expert System; Fuzzy System	7	84.20%
6	[7]	S Muthukarupp-an *et al.*	2012	Decision Tree; PSO; Fuzzy expert System	13	93.27%
7	[8]	Dursan *et al.*	2012	SVM; Decision Tree; Neural Networks	19	87.74%
8	[9]	Ilias *et al.*	2009	SVM	198 (Heart sound signal)	77%
9	[10]	Chih-Lin Chi *et al.*	2010	Decision Support System with Optimal Decision path finder	49	50%
10	[11]	Rajeshwari *et al.*	2012	ANN ; Back Propagation	12	89.4%

Table 1

EXISTING SYSTEM

2. EXISTING SYSYEM

The proposed system deals with only the essential attributes taken from the Cleveland and Switzerland heart disease database from UCI machine learning repository. 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 and exercise angina. With these 12 attributes the accuracy of the system was 93.27% [22]. 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. An Artificial Bee Colony Optimization (ABC) 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.

3.1. CART Decision Tree Algorithm

Decision trees are a simple yet prevailing method for numerous variable analyses. Breiman *et al.*, (1984) [4] has projected the classification algorithm called the Classification and regression tree (CART) for constructing binary trees in which each internal node precisely has two retiring edges [32]. The CART algorithm has been also termed as Hierarchical Optimal Discriminate Analysis (HODA) that enables the users by providing the prior probability distribution. The cost-complexity Pruning and Gini index are used to prune the tree obtained from the CART algorithm and as the impurity measure for selecting attribute respectively. 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 , 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 algorithm yields crisp rules along with the decision tree. This algorithm makes use of both categorical and numeric variables either to construct classification or regression trees and thus it is a non-parametric decision tree learning technique.

3.2. 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. With the knowledge base; the inference engine fuzzifies the input. Finally defuzzification is performed to obtain the crisp result.

- **Fuzzification -** Converts the crisp input to a linguistic variable using the membership functions stored in the fuzzy data base.

The conversion of real inputs to fuzzy set values is the preliminary of the fuzzy system. In the real world, hardware and manuals generates crisp data, but these data are subject to investigational errors. By establishing the fact base of the fuzzy system we identify the input and output of the system. The IF THEN rules are coiled and uses unprocessed data to develop a membership function.

- **Fuzzy inference system**

The Fuzzy rules are based on fuzzy premises and fuzzy consequences. Truth value for the premise of each rule is computed, and applied to the conclusion part of each rule. This results in one fuzzy subset to be assigned to each output variable for each rule. There are two inference methods/inference rules: **MIN** and **PRODUCT**.

- **Defuzzification-** Convert the fuzzy value obtained from composition into a “crisp” value.

This process is often intricate since the fuzzy set might not interpret directly into a crisp value. But it considered being obligatory, since controllers of substantial systems require discrete signals. The conversion of a fuzzy quantity to a precise quantity, just as fuzzification is the conversion of a precise quantity to a fuzzy quantity. The output of a fuzzy process can be the logical union of two or more fuzzy membership functions defined on the universe of discourse of the output variable.

Defuzzification approach is intended at producing a non-fuzzy control action. The crisp value of the output variable is computed by finding the variable value of the center of gravity of the membership function for the fuzzy value. There different defuzzifying methods that used most commonly, one of them is Centroid of area (COA).

Centroid of area (COA)

COA finds the point where a vertical line would slice the aggregate set into two equal masses. Centroid defuzzification method finds a point representing the centre of gravity of the fuzzy set on its interval.

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.

3.3. Particle Swarm Optimization (PSO)

Particle Swarm Optimization is a population based optimization algorithm motivated by the communal activities of bird flocking and fish schooling [12]. 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 [8]. 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.

PROPOSED SYSTEM

3. PROPOSEDSYSTEM

The following fig 2 shows the flow diagram of proposed system. The automated prediction system induces the crisp rule from the decision tree, which are evaluated for the vagueness through the Fuzzy Inference System and predicts the presences of Coronary Heart Disease and optimizes its prediction accuracy.

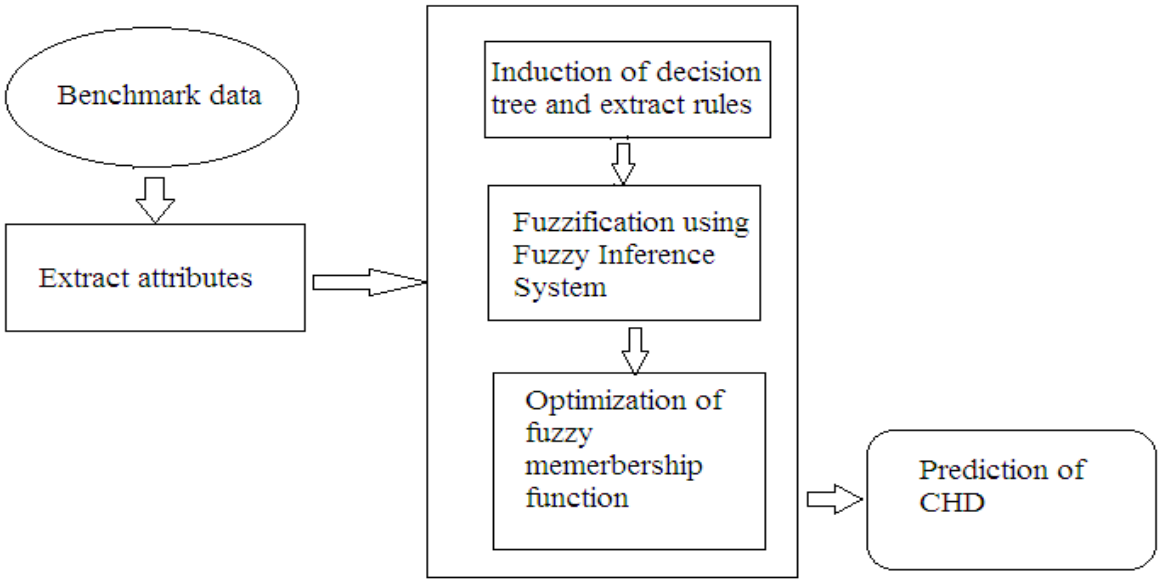

Fig 2. Flow diagram of Proposed System

3.1. METHODOLOGY

The fig 3 depicts the flow of the methodology of the proposed system as follows

3.1.1. Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues by preparing raw data for further processing and effectively select the features for the diagnosis.

3.1.2. Classification

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. From the attributes crisp rules are obtained by employing CART decision tree algorithm, which are then applied to the fuzzy system.

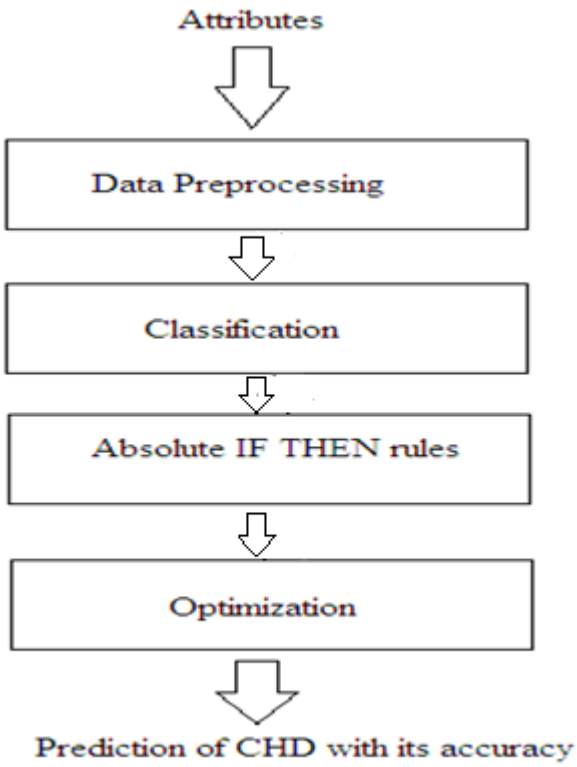

Fig 3. Flow of the methodology of the proposed system

3.1.3. Absolute IF THEN rules

Fuzzy IF THEN rules are rules whose antecedents, consequences or both are fuzzy rather than crisp. The benefit of having fuzzy antecedents is to provide a basis for an interpolation mechanism. In this view a fuzzy rule is defined by means of a conjunction rather than in terms of a multiple-valued logic implication. These crisp rules are fuzzified in the fuzzy inference system through the **triangular membership functions**.

3.1.4. **Optimization**

Artificial Bee Colony Optimization (ABC) as an optimization tool provides a population-based search procedure in which individuals called foods positions are modified by the artificial bees with time and the bee's aim is to discover the places of food sources with high nectar amount and finally the one with the highest nectar. This technique is applied for the optimization of the fuzzy membership functions where the parameters of the membership functions are altered to new positions.

4.2. A**RTIFICIAL BEE COLONY OPTIMIZATION (ABC)**

The optimization process of the proposed system has been carried out by using A**rtificial Bee Colony Optimization (ABC)** that has been anticipated by Karaboga in 2005. ABC is developed based on inspecting the behaviors of real bees on finding nectar and sharing the information of food sources to the bees in the hive. It has been worn for solving multidimensional and multimodal optimization problems. This algorithm contains three groups of bees:

- The **Employed Bee** (50%): It stays on a food source and provides the neighborhood of the source in its memory.
- The **Onlooker Bee** (50%): It gets the information of food sources from the employed bees in the hive and select one of the food source to gathers the nectar.
- The **Scout** (5-10%): The employed bee whose food source has been exhausted becomes a scout.

Scouts are the colony's explorers. The number of employed bees is equal to the number of food source and the food source position is equal to the possible solution to the problem. The amounts of nectar of a food source are identical to the quality of the solution. If the fitness value of new one is higher than that of the previous one, the bee forgets the old one and memorizes the new position. This is called as greedy selection. Then the employed bee whose food source has been exhausted becomes a scout bee to search for the further food sources once again

4. 3. ABC ALGORITHM

In ABC, the solutions represent the food sources and the nectar quantity of the food sources corresponds to the fitness of the associated solution. The number of the employed and the onlooker bees is same, and this number is equal to the number of food sources. Employed bees whose solutions cannot be improved through a predetermined number of trials, specified by the user of the ABC algorithm and called **limit**, become scouts and their solutions are abandoned.

The general scheme of the ABC algorithm is as follows:

- Bee Initialization Phase
- Set the Loop Employed Bee Phase
- Onlooker Bee Phase
- Scout Bee Phase
- Memorize the best solution found so far
- Until the loop is terminated

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.

IMPLEMENTATION OF PREDICTION SYSTEM

5. IMPLEMENTATION OF PREDICTION SYSTEM

The implementation was carried out in **MATLAB 7. 11**. 0.584. 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.

5. 1. 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 spitted 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.

5.2. 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.

5. 3. 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.

5 4. Implementation of Artificial Bee Colony Optimization Algorithm

In the initialization phase, the control parameters are set, such as colony size, iteration number (bee travel time), working to onlooker bee rate. In the next phase, the attributes is given as an input to the prediction system a mean value is obtained by using nearest neighbor method. Further when the working bees are initialized, the bee optimization loop is set. Then the random node is assigned for the bee to start, and then by computing the probabilities. The bees will work and draw the possible value to obtain the accuracy and will memorize the best solution found so far using the greedy selection strategy. Finally the bees become scout bees and the number of working bees is updated, that is the employed bee which is exhausted becomes the scout bee again. The optimization loop is terminated when the numbers of iterations are completed and the best result is obtained. The scout bees then again start to search for the new optimization value by which best prediction accuracy can be obtained.

RESULTS AND DISCUSSION

6. RESULTS AND DISCUSSION

The output of the CART decision tree algorithm generates the decision tree along with the rules. The rules are generated which is shown 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 shows the membership function for the attributes where the four fuzzy sets for low, medium, high and very high are depicted. Fig. 6 depicts the mean values for the optimization where the four fuzzy sets are altered to new positions. 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 after optimization is shown in fig. 7. 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 with accuracy is shown in fig 8.

COMPARISON OF PSO WITH ABC AND PERFORMANCE ANALYSIS

7. COMPARISON OF PSO WITH ABC AND PERFORMANCE ANALYSIS

By comparing the Swarm Intelligent algorithm, the proposed algorithm has obtained the best accuracy than the existing techniques. The PSO optimization technique that has been used in the proposed system has the following disadvantages of weakness regarding local search, slow convergence rate and may get trapped in local minima for hard optimization problems.

The following fig 4 provides the comparison of prediction accuracy obtained by both optimization techniques. While the proposed algorithm of ABC has the advantages of having strength in both local and global searches. PSO algorithm has obtained about 91.45 % of accuracy while ABC has obtained 96.73% of prediction accuracy of CHD.

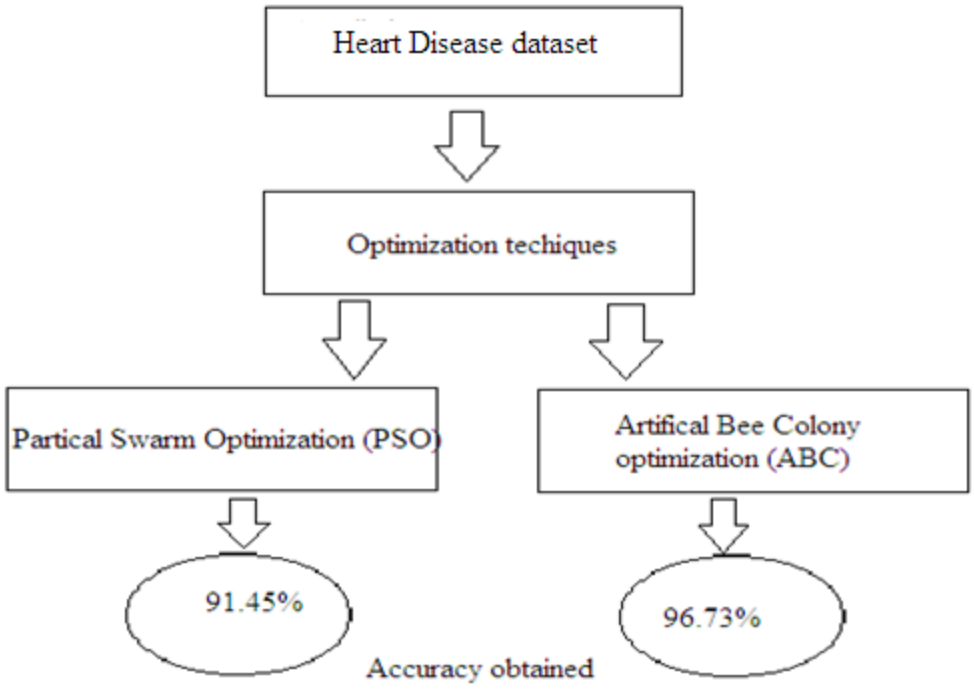

Fig 4. Comparison of PSO with ABC

CONCLUSION

8. CONCLUSION

The accuracy of the proposed system is found to be good for Cleveland databases [27] 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.

SCOPE FOR FUTURE ENHANCEMENT

9. SCOPE FOR FUTURE ENHANCEMENT

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.

BIBLIOGRAPHY

BIBLIOGRAPHY

JOURNAL REFERENCES

[1] Ahmet Yardimci, (2009), *"Soft computing in medicine",* Applied soft Computing, Pp: 1029-1043.

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

[3] Babita Pandey, R.B.Mishra, (2009), *"Knowledge and intelligent computing system in medicine"*, Computers in Biology and Medicine, Pp: 215 – 230.

[4] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984), Classification and regression trees. elmont, Calif.: Wadsworth.

[5] Brijain R Patel, Kaushik K Rana, (2014), *"Use of Renyi Entropy Calculation Method for ID3 Algorithm for Decision tree Generation in Data Mining",*International Journal of Advance Research in Computer Science and Management Studies Volume 2, Issue 5, Pp.30-34

[6] Chih-Lin Chi, W. Nick Street, David A. Katz, (2010), *"A decision support system for cost-effective diagnosis"*, Artificial Intelligence in Medicine, Pp: 149-161.

[7] Debabrata Pal, K.M. Mandana, Sarbajit Pal, Debranjan Sarkar, Chandan Chakraborty, (2012), *"Fuzzy expert system approach for coronary artery disease screening using clinical parameters"*, Knowledge-Based Systems.

[8] Dong-ping Tian , Nai-qian Li, (2009), "*Fuzzy Particle Swarm Optimization Algorithm*", Proceedings of the 2009 International Joint Conference on Artificial Intelligence, Pp.263-267.

[9] Dursun Delen, Asil Oztekin, Leman Tomak, (2012), *"An analytic approach to better understanding and management of coronary surgeries",* Decision Support Systems, Pp: 698-705.

[10] Fatma Latifoglu, Kemal Polat, Sadik Kara, Salih Gunes, (2008), *"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, Pp: 15-23.

[11] Heart disease and stroke statistics, *"Heart disease and stroke statistics update"*, American heart association, available at *http://www.americanheart.org*.

[12] Hassan M. Elragal, (2010), *"Using swarm intelligence for improving accuracy of fuzzy classifiers"*, International Journal of Electrical and Computer Engineering.

[13] Hui Yang, Satish T. S. Bukkapatnam, Trung Le, Ranga Komanduri, (2012), *"Identification of myocardial infarction using spatio-temporal heart dynamics"*, Medical Engineering and Physics, Pp: 485-497.

[14] Ilias Maglogiannis, Euripidis Loukis, Elias Zafiropoulos, Antonis Stasis, (2009), *"Support vectors machine-based identification of heart valve diseases using heart sounds"*, Computer methods and programs in biomedicine, Pp: 47-61.

[15] Imran Kurt, Mevlet Ture, A. Turhan Kuram, (2008), "*Comparing performances of logistic regression, classification and regression tree and neural networks for predicting coronary artery disease"*, Expert Systems with Applications, Pp: 366-374.

[16] Ismail Babaoglu, Omer Kaan Baykan, Nazif Aygul, Kurtulus Ozdemir, (2009), *"Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization"*, Expert Systems with Applications, Pp: 2562-2566.

[17] Ismail Babaoglu, Oguz Findik, Erkan Ulker, (2010), *"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, Pp: 3177-3183.

[18] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen, (2013), *"Association rule mining to detect factors which contribute to heart disease in males and females"*, Expert Systems with Applications, Pp: 1086–1093.

[19] Krzysztof J. Cios, G. William Moore, (2002), *"Uniqueness of medical data mining"*, Artificial Intelligence in Medicine, Pp: 1–24.

[20] Luka Sajn, Matjaz Kukar, (2011), *"Image processing and machine learning for fully automated probabilistic evaluation of medical images"*, Computer Methods and Programs in Biomedicine, Pp: 75-86.

[21] Matjaz Kukar, Igor Kononenko, Ciril Groselj, (2011), *"Modern parameterization and explanation techniques in diagnostic decision support system. A case study in diagnostics of coronary artery disease"*, Artificial Intelligence in Medicine, Pp: 77-90.

[22] Muthukaruppan, S., M.J. Er , (2012), "*A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease"*, Expert Systems with Applications, Pp: 11657–11665.

[23] Persi Pamela. I, Gayathri.P and N. Jaisanker , (2013), "*A Fuzzy Optimization Technique for the Prediction of Coronary Heart Disease Using Decision Tree",* International Journal of Engineering and Technology (IJET), Pp: 2506-2514.

[24] Rajeswari, K., Dr. Vaithiyanathan,V., Dr. Neelakantan, T. R. , (2012), *"Feature selection in Ischemic heart disease identification using feed forward neural networks"*, Procedia Engineering, Pp: 1818-1823.

[25] Sebastian, Y., Patrick H. H. Then, (2011), *"Domain-driven KDD for mining functionally novel rules and linking disjoint medical hypotheses"*, Knowledge-Based Systems, Pp: 609-620.

[26] Vahid Khatibi, Gholam Ali Montazer, (2010), *"A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment"*, Expert Systems with Applications, Pp: 8536-8542.

WEB REFERENCES

[27] www. archive.ics.uci.edu/ml/support/Heart+Disease

[28] www.nhlbi.nih.gov/health/health-topics/topics/cad

[29] www.webmd.com/heart-disease/guide/heart-disease-coronary-artery-disease

[30] www.heartfoundation.org.au/your-heart/cardiovascular-conditions/Pages/coronary-heart-disease.aspx

[31] www.bhf.org.uk/heart-health/conditions/coronary-heart-disease

[32] www.researchmethods.org/CARTIntroTutorial.pdf.

APPENDIX

APPENDIX

SCREEN SHOTS

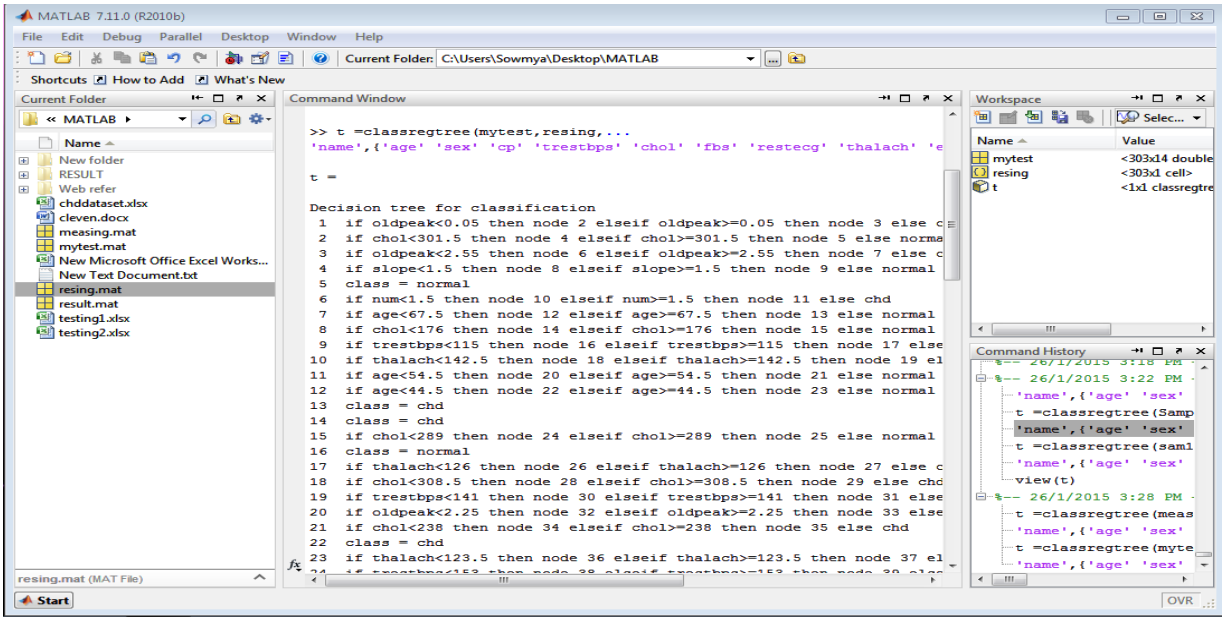

Fig 3. A part of rules from decision tree

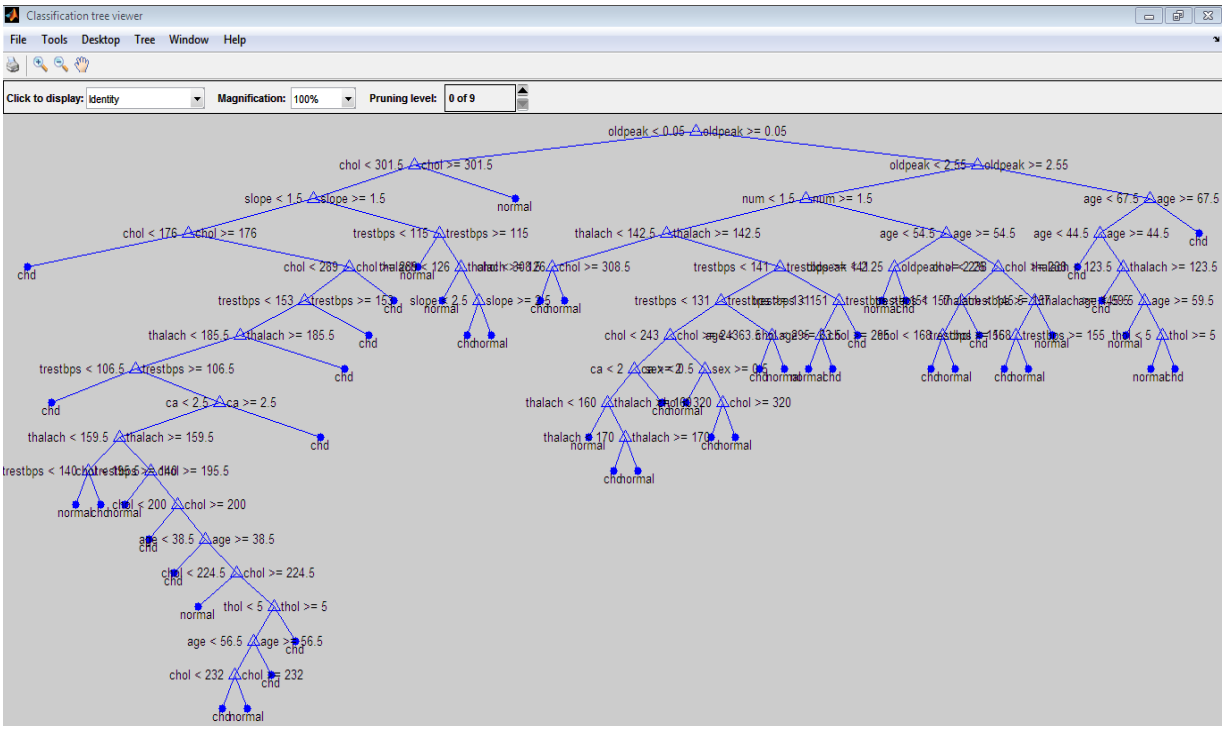

Fig 4. CART Decision tree obtained

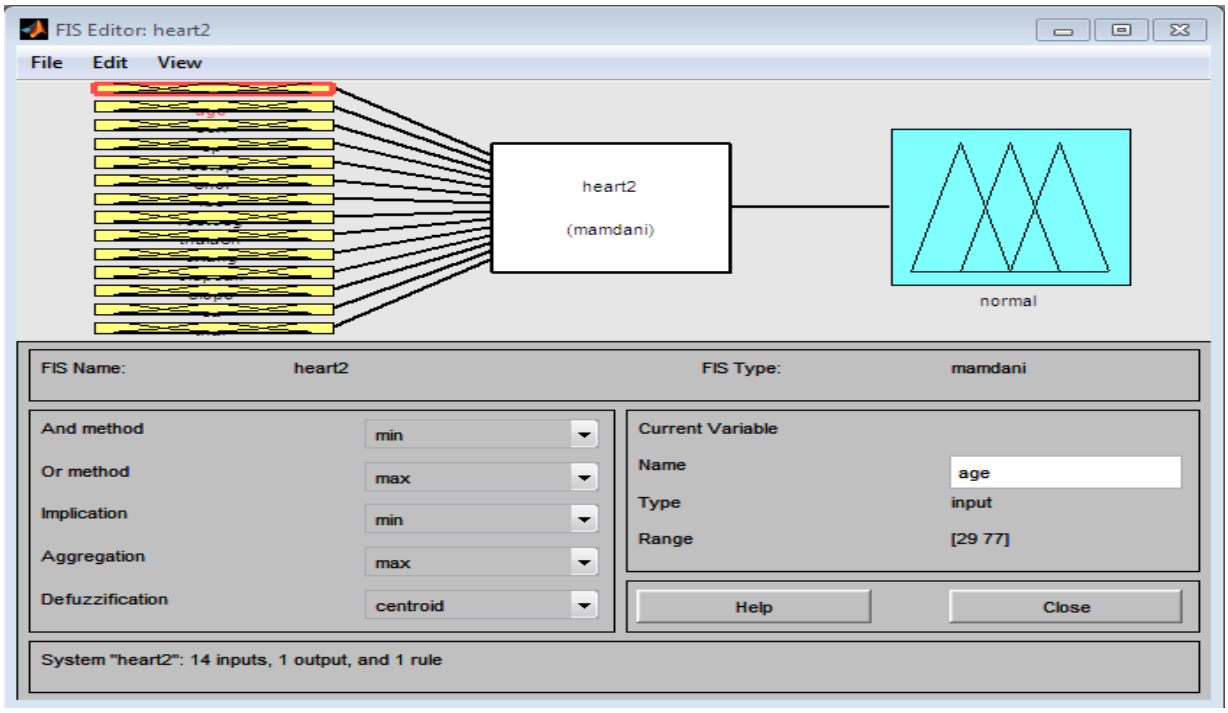

Fig 5.Fuzzy Inference System

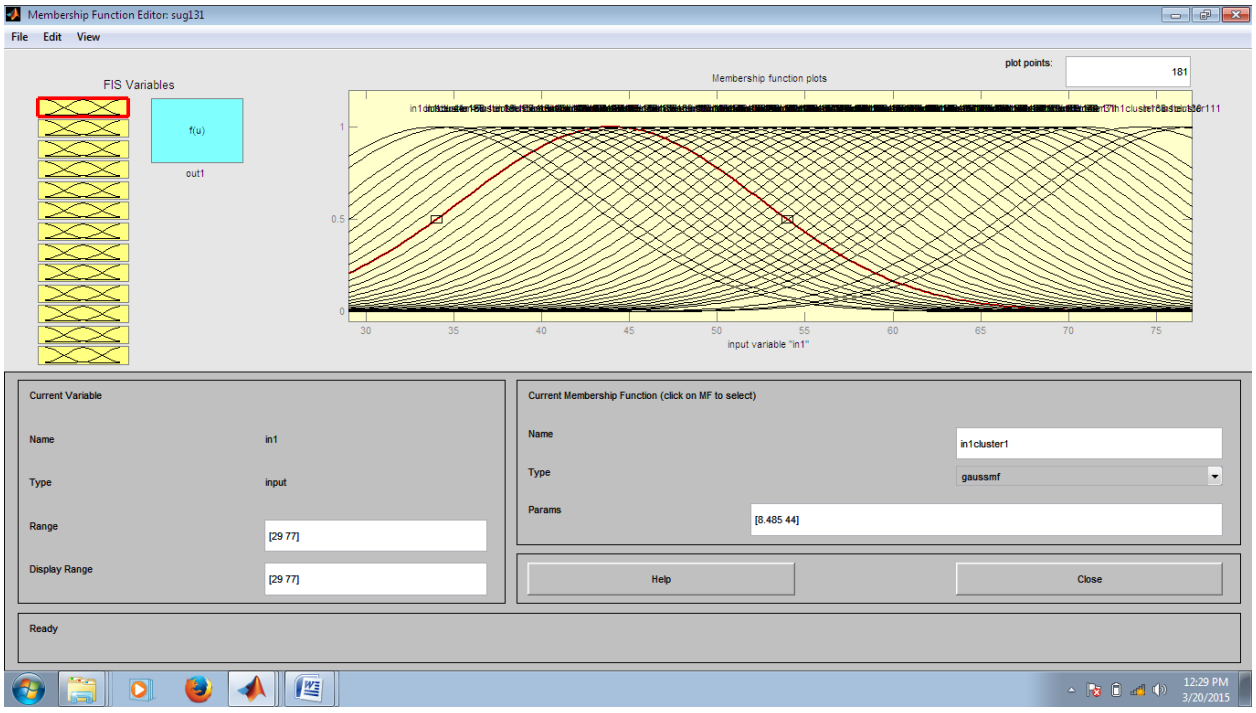

Fig 6. Membership function for attributes

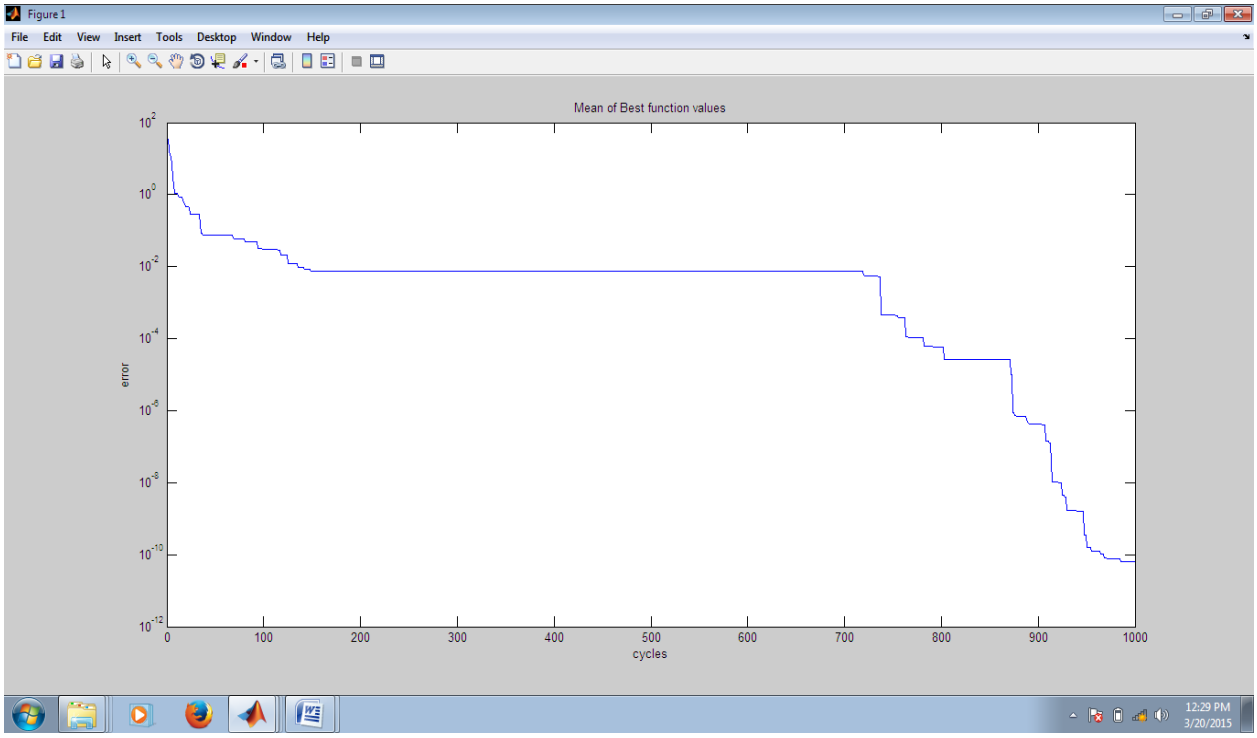

Fig 7. Mean chart used for optimization

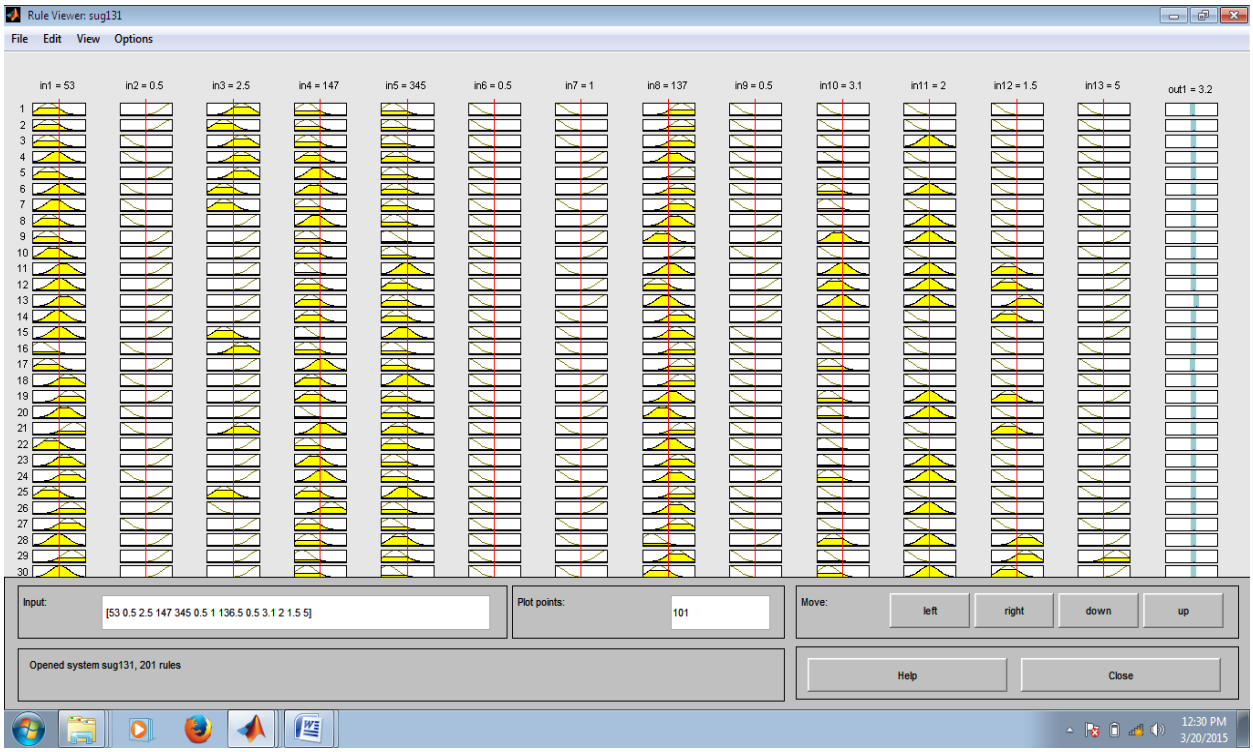

Fig 8. Rule viewer for one of the test data after optimization

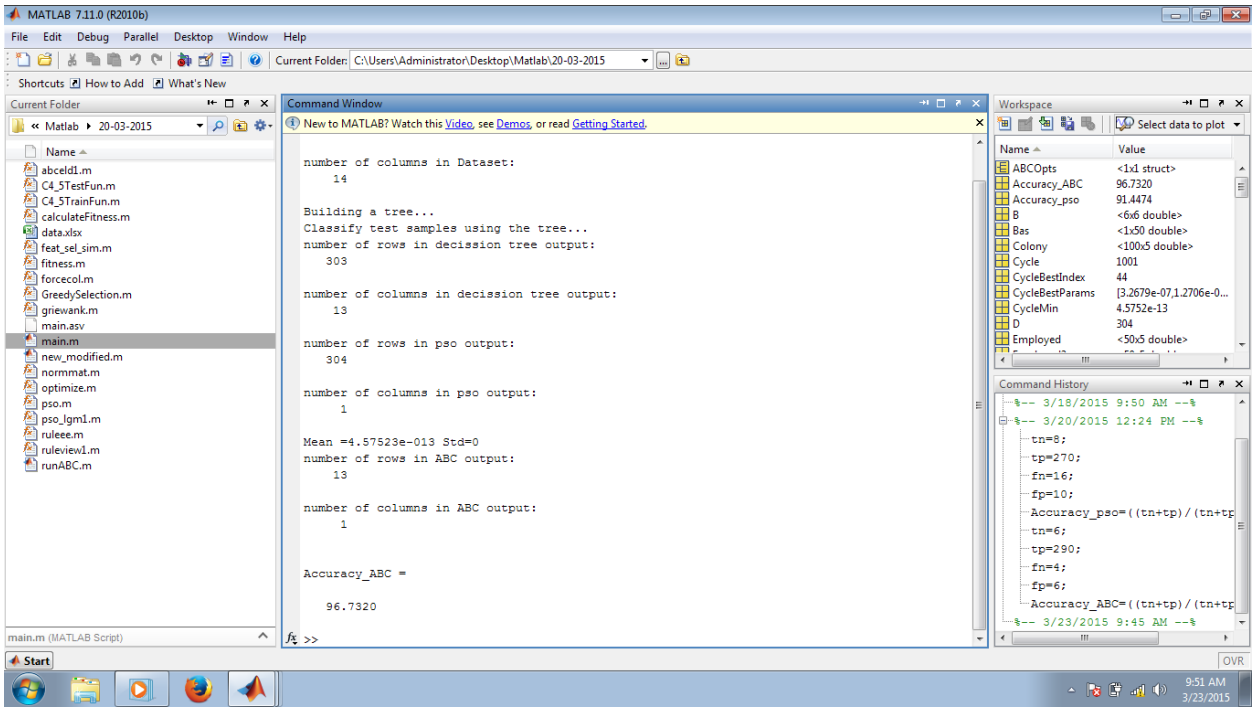

Fig 9. Prediction accuracy obtained through ABC

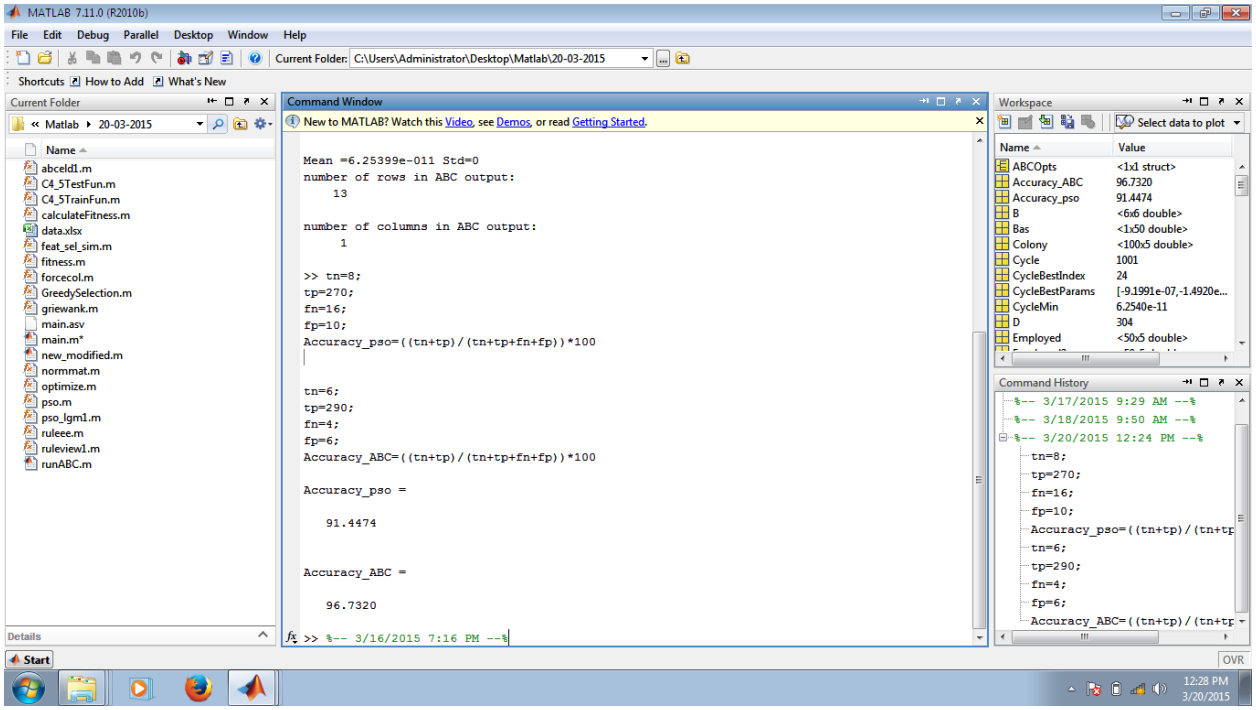

Fig 10. Comparison of PSO with ABC