Computation Methods for the Diagnosis and Prognosis of Heart Disease

Deepthi S
Dept. of Computer Science & Engineering
Mar Baselios College of Engineering & Technology
Thiruvananthapuram, India

Aswathy Ravikumar
Dept. of Computer Science & Engineering
Mar Baselios College of Engineering & Technology
Thiruvananthapuram, India

ABSTRACT
Medical data mining is the application of machine learning to highly voluminous medical database. The mining process aims at exploring the patterns and relationships within the medical data that could be exploited for the diagnosis and prognosis of the disease. Since medical databases are humongous in nature, feature selection is done using pre-processing, in order to reduce the dimensionality without compromising the accuracy, by removing immaterial data present in the database. Health care organizations are challenged to provide high quality and cost-effective care to patients. Heart disease prognosis is considered to be one of the tiresome and complicated tasks in medical field. Heart disease has given primary importance because of its exponential rise in recent years. It has been estimated that, by the year 2016, most important cause of mortality in India will be due to heart diseases. Hence an efficient heart disease prediction system is crucial. Different heart disease prediction systems have been introduced for the prognosis of heart disease. Through this paper, we propose an amalgamation of different algorithms such as Support Vector Machine, Random Forest and Naive Bayes. For feature extraction and dimensionality reduction, Principle Component Analysis is used along with Firefly algorithm for the optimization of output.

Keywords
Heart disease prediction, Support Vector Machine, Random Forest, Naive Bayes, Principle component analysis, Firefly Algorithm.

1. INTRODUCTION
Medical data mining is a technique used to extract features from humongous amount of medical data present in the medical database[2]. Since medical data are highly voluminous, there should be some efficient mechanism to extract the data. The extracted data is extremely useful for recognition, risk analysis and prognosis of disease

Heart diseases are leading cause of death in India and worldwide[2]. Various types of heart disease are in prevalence. Symptoms and effects of these diseases varies. Heart is also called as cardio. So diseases related to heart are generally termed as Cardio Vascular Diseases (CVD). Some examples of Cardio Vascular Diseases includes Coronary Heart Disease(CHD), Congenital Heart Disease(CHD), Ischemic, Coronary Artery Disease(CAD), Atrial Fabrication, Heart Attacks, Cardiomyopathy, Arthrosclerosis, Arrhythmias, Angina pectoris, and Myocarditis[7][8][10]. These diseases affects people in low and middle income developing countries like India as compared to high income developed countries. WHO has announced India as global CAD capital [9]. Past records show that heart diseases are the major reason for death in India because of rapid urbanization and industrialization. According to the California-based CADI, by the year 2015, India will be having 62 billion people with heart disease [10]. People get affected at their younger age itself because of their unbalanced diet and unhealthy daily routine. In rural areas, 30% of the deaths are due to CVD. Medical diagnosis at the earliest is an essential yet complicated venture that needs to be accomplished efficiently and accurately. The automation of diagnosis of disease is preferable. Automation can be accomplished using different Artificial Intelligence (AI) methods. Most powerful technique in AI is machine learning. Machine Learning (ML) systems earns information automatically from the data present and creates some generalized models based on the extracted data. ML uses various efficient algorithms for this purpose. Nowadays ML plays an important role in the field of medicine for the diagnosis and prognosis of various diseases.

In the past few decades, medical data mining have played a remarkable role in the field of heart disease research. In order to find medical data which are hidden inside the database, so as to differentiate between healthy and disease affected individuals, existing clinical data plays a powerful role for classification and prediction of heart disease. Machine learning and statistics are two approaches which have been applied for the prediction of heart disease based on the prevailing clinical data.

In India the number of individuals with heart disease is growing in an exponential pace. Hence there needs to be a decision support system for the prediction of heart disease. The main intention of this paper is to provide a model for the prognosis of heart disease which would help the physicians to take an appropriate decision at accurate time.

This paper is as arranged as follows: Next section provides a brief idea about the structure and function of heart. Overviews of works related to heart disease are summarized in section 3. In section 4, the importance of extracted features is explained briefly. Proposed algorithm is presented in section 5. Section 6 provides the comparison of existing methods and finally section 7 concludes this paper.

2. STRUCTURE AND FUNCTION OF HUMAN HEART
Blood Circulation in human body is carried virtue of human heart, which is about the size of a clenched human fist. Human heart acts as a pump which drives the blood to flow in the Human Circulatory System. Due to the pumping activity of Heart, it sucks the deoxygenated blood from different parts

of human body via veins and pumps back the fresh oxygen-rich blood from Lungs to different parts of the human body through arteries. Human lungs are the responsible organs for supplying sufficient amount of oxygen to the blood.

2.1 Structure of Human Heart

Human Heart is divided into four chambers; two in left unit and the other two in right unit which are separated by a partition called Septum. Each unit has an upper chamber called Atrium and a lower chamber called Ventricle. Left Atrium sits on top of the left Ventricle and the right Atrium sits on top of the right Ventricle.

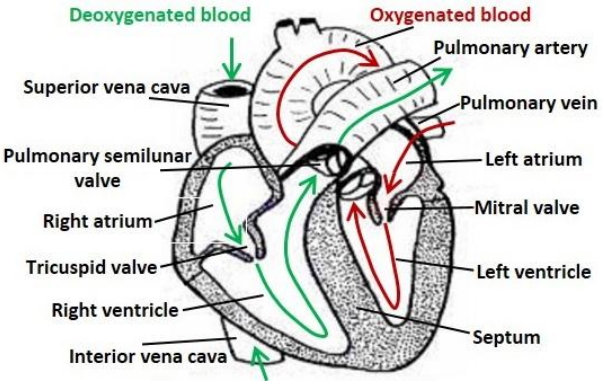

Fig 1. Structure and schematic functioning of human heart

The four chambers of the heart are connected by valves and are attached to major veins or arteries that either bring blood into or carry blood away from the heart. Heart consists of four valves connecting the atriums and ventricles [18]. They consist of flaps which control the direction of flow of blood inside Heart; the open to allow the forward flow and close to prevent the backward flow; through one valve flow can happen in only one direction. In between Right Atrium and the Right Ventricle there is a Tricuspid valve. Pulmonary valve connects right ventricle and pulmonary artery. Mitral valve is between the left atrium and the left ventricle whereas aortic valve connects the left ventricle and the aorta.

2.2 Function of Human heart

The two largest veins in the human body, the superior and inferior vena cava, bring the deoxygenated or oxygen-poor blood from different parts of the body to the heart into the right atrium. This impure blood flows through the tricuspid valve into the right ventricle. Flaps of tricuspid valve gets closed once the blood pass through to prevent it from flowing back to the right atrium. Once the right ventricle gets filled with blood, it contracts to pump the blood through the pulmonary valve into the pulmonary artery hence to the lungs[19]. Pulmonary valve closes to prevent the back-flow of blood into right ventricle. Pulmonary artery carries the blood into both the lungs through its branches. Blood gets freshen with respired oxygen at the lungs and expels carbon dioxide.

Pulmonary veins bring the oxygen-rich blood from the lungs to the heart into the left Atrium. Through the mitral valve oxygen-rich blood flows from left Atrium to left Ventricle and the valve closes to prevent back-flow. The left ventricle, the most muscular chamber of human heart, contracts to pump blood into the Aorta through the aortic valve which closes after the flow to prevent backward flow of blood into the left ventricle. Aorta branches to arteries, arterioles and capillaries to supply oxygen-ring blood to the entire human body.

Human Heart is controlled by an electric impulse system traveling through the heart. Frequency of heart pumping is being controlled by the heart's built in control system called the natural Pacemaker called the Sinus node, which is located at the top of the right atrium. The signals sent by the Sinus node travel through the heart muscle tissues to actuate the contracting and relaxing of Atria and Ventricles[20].

3. RELATED WORKS

The term Cardio Vascular Diseases (CVD) covers various types of heart diseases that affect the human heart. Different types of heart diseases are there which leads to the rise in rate of mortality. There are various correlated factors that results in the development of heart diseases. This feature makes the diagnosis of Heart Disease a multi-layered problem which may guide to incorrect conjecture associated with uncertain effects. Hence data mining has emerged into medical care of heart disease as an intelligent prediction tool to assist physicians to refine the prediction accuracy and diminish the event of heart disease. In this section, comparison of different prediction tools developed in the last few years are compared.

3.1 Intelligent Heart Disease Predication System (IHDPS, 2008)

IHDPS model in[4] aims to extract relationships and hidden patterns associated with heart disease from a historical medical database. IHDPS can estimate the probability of individuals getting HD using attributes such as age, sex, blood pressure and chest pain. IHDPS is a hybrid system that exploits three data mining techniques for prediction: Naive Bayes (NB), Decision Tress (DT) and Artificial Neural Network (ANN). For model creation, prediction, training and content access, Data Mining Extension (DME) query language is used. Left Chart and Classification Matrix are used in order to evaluate the prediction [4].

3.2 Intelligent and Effective Heart Attack Prediction System (IEHAPS, 2009)

IEHAPS aims to squeeze patterns that are relevant to heart attack from the HD database through consumption of Different predictive methods: K-means, Maximum Frequent Item set Algorithm (MAFIA) and ANN. In IEHAPS approach, clustering is performed on the preprocessed data warehouse using K means algorithm with K=2. This will generate two clusters; first one consists of data that are relevant to heart attack. Next, the frequent patterns most relevant to HD diagnosis are mined the. MAFIA integrates multiple algorithmic ideas (such as Apriority and FP Tree) to mine association rules from the clustered dataset. After mining frequent patterns, patterns having weightage greater than a fixed threshold are picked to support the prediction of heart attack. Eventually ANN is trained with the chosen patterns in order to predict heart attack. IEHAPS makes use of feed-forward ANN and MLPNN (Multi- Layer Perceptron Neural Network) with back-propagation (BP) algorithm[17].

3.3 Ischemia Prediction using ANFIS (2010)

An advanced algorithm called Adaptive Neuro Fuzzy Interference System (ANFIS) is used to predict Ischemia diseases by using Electrocardiogram (ECG) signals. Pre-processing of ECG signal has been implemented in order to suppress the noise content. Evaluation of validity of prediction accuracy is done using Root Mean Square Error

criterion. Then ANFIS classifier classify as normal or Ischemia beaten. Performance is evaluated with the help of criteria such as Sensitivity and Specificity [12].

3.4 Classification Ensemble Optimization Using Genetic Algorithm (CEO-GA, 2011)

Support Vector Machine (SVM) algorithm with three different types of kernel functions is used for the base classification. Different kernels functions used are linear kernel, polynomial kernel and radial basis kernel. In order to optimize the results from the ensemble of base classifiers, genetic algorithms are used. The classifier is evaluated based on the values produced for accuracy, specificity and sensitivity. Four different datasets are used for this purpose. This ensemble technique has shown much improvement in terms of classification accuracy [13].

3.5 A classifier based on Binary Particle Optimization and Support Vector Machine (BPSO-KNN-SVM, 2012)

BPSO is a computer-aided prediction system for heart valve diseases. It makes use of Support Vector Machine (SVM) and Binary Particle Swarm Optimization along with K-nearest neighbor (KNN) and leave-one-out cross-validation. This system uses heart sound signals as dataset heart sound signals. This approach is initiated by an algorithm based on binary particle swarm optimization in order to choose the most weighted features. Optimization is followed by the implementation of SVM to classify the heart signals into two end results: healthy or unhealthy having a heart valve disease[14].

3.6 Predictive Risk Assessment of Arthrosclerosis (PRAA, 2013)

Atherosclerosis is a Coronary heart disease (CHD) caused by hardening of artery walls due to cholesterol. In Predictive Risk Assessment of Arthrosclerosis (PRAA) approach, an imputation algorithm and particle swarm optimization (PSO) is used to predict the risk factors associated. The strength of PRAA is the PSO search which identifies the physical inactivity as one of the risk factors for the onset of atherosclerosis other than already known factors. PSO is used for feature subset selection. This improves the accuracy. Using this approach decision tree is generated which can predict the disease with an accuracy of 99.73% which is very much higher than machine learning techniques employed earlier[15].

3.7 Artificial Neural Network using hybrid algorithm for optimization (ANN-GSO-ABC, 2014)

In ANN-GSO-ABC method, to strengthen the training process of the artificial neural network to predict the heart disease effectively, hybrid algorithm which is an amalgamation of Group Search Optimization (GSO) and Artificial Bee Colony (ABC) is used. To begin with, an initial population that has number of members are generated. Training of neural network is done by assigning weights to each member in the population. In order to identify a perfect member to train the neural network, operation using hybrid algorithm are performed. Then each member is assigned to the neural network and fitness for each member is determined. Then each member is assigned to perform different hybrid operation. After performing corresponding operation by each member, we gain a new set of members. This process will be repeated until we gain a sensible member for producer operation. Then we select the weight values of the producer for the training of neural network in order to predict the heart diseases [16].

4. FEATURE SELECTION

In this paper, we use heart disease dataset from University of California Irvine (UCI) machine learning repository [21]. 14 clinical features are extracted from the database.

Table 1. Clinical features, their description and value

Feature	Description	Value
Age	Instance age in years	Any
Sex	Instance gender	(1,0)
Cp	Chest pain type	(1-4)
Trestbps (mmHg)	Resting blood pressure	Any
Fbs	Fasting blood sugar	(1,0)
Restecg	Resting ECG result	(1,0)
Thalach	Maximum heart rate achieved	Any
Oldpeak	ST depression induced by exercise relative to rest	Any
Slope	Slope of peak exercise ST segment	(1,2,3)
Ca	No. of vessels coloured	(1,2,3)
Thal	Type of defect	(3,6,7)
Class	Diagnosis of HD	(1,0)
Chol(mg/dl)	Serum cholestrol	Any
Exang	Exercise induced angina	(1,0)

Table 1 display 14 attributes which is to be used in this system, including 8 symbolic and 6 numeric: age (age in years), sex (male, female). Male is indicated by 1 and female is indicated by 0, Chest pain types are typical angina, atypical angina, non-angina pain and asymptomatic and their corresponding values are 1, 2, 3 and 4. Trestbps(resting blood pressure in mm Hg), cholesterol(serum cholesterol in mg/dl), fasting blood sugar < 12mg/dl(true or false), resting electrocardiographic results(normal, having ST-T wave abnormality, showing probable or definite left ventricular hypertrophy by Estes' criteria), max heart rate, exercise induced angina(true(1) or false(0)), oldpeak (ST depression induced by exercise relative to rest), slope(up, flat, down) represented by values (1,2,3), number of vessels colored by fluoroscopy(0-3), thal(normal(3), fixed defect(6), reversible defect(7)), and class(healthy(1), with heart-disease(0)) [3]. Any in table 1 represent any integer value.

5. RECOMMENTATION

A new method is suggested in this paper, which is the ensemble of heterogeneous classifiers for the classification purpose and finally the results are optimized using Firefly Algorithm (FA). At the initial stage, heart disease dataset can be chosen from University of California Irvine (UCI) machine learning repository. Then the dataset should be devoted for preprocessing in order to remove noise present in the dataset. Pre-processing can also be employed for conversion of

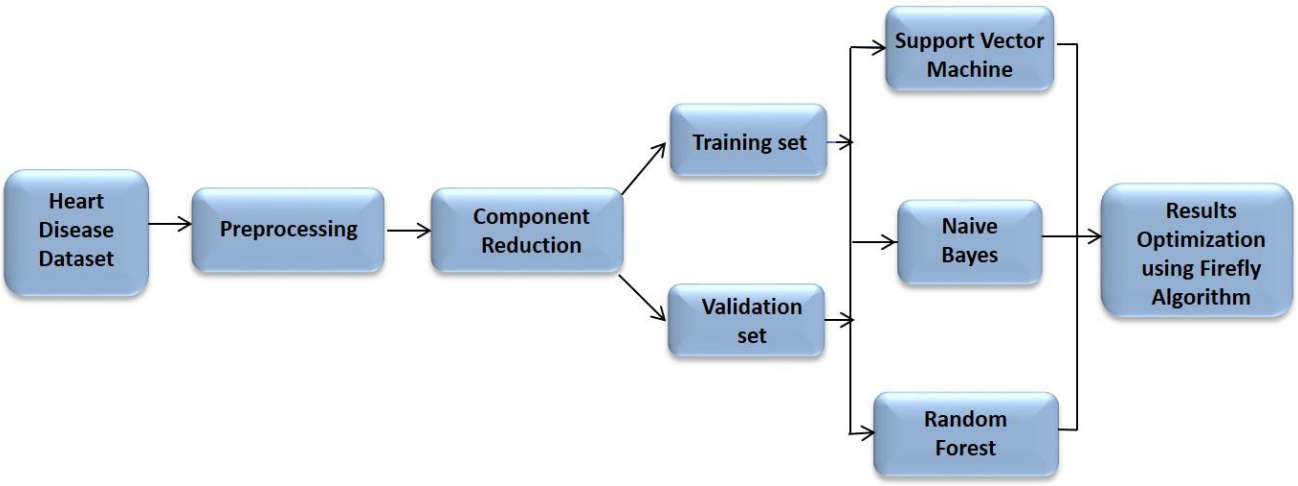

Fig 2: Amalgamation of three heterogeneous base classifiers; Support Vector Machine, Random Forest and Naive Bayes

alphanumeric numbers to integers. After preprocessing, the dataset should be treated for dimensionality reduction. Dimensionality reduction can be performed efficiently using Principle Component Analysis (PCA). PCA is a statistical procedure that uses orthogonal transformation for converting set of correlated variable to set of uncorrelated variables termed as principle component. Then dataset will be categorized into two as training set and validation set.

In the classification phase, amalgamation of three heterogeneous base classifiers like Support Vector Machine (SVM), Random Forest (RF) and Naive Bayes (NB) are shown in Fig 2. The data given for training and validation will be given to the base classifiers such as SVM, NB and RF. The outcome of these classifiers will be ensemble and given to Firefly Algorithm for result optimization.

Ensemble of classifier is recommended because the combination of decisions of individual classifiers makes a final decision more accurate. If one classifier disagree with the decision, ensemble will be advantageous since the other classifier will come in effect.

6. COMPARISON

IHDPS is developed using three basic algorithms (DT, NB, ANN) to predict the heart disease. IHDPS is a hybrid approach which used for the prediction of heart disease in general. Advantages of IHDPS are reliability, user-friendliness, and extensibility.

IEHAPS is a hybrid approach that is used for the prediction of heart attacks. The Data mining techniques used in this system are K-means, MAFIA and ANN. It combines new methods for the extraction of association rules and for calculation of frequently occurring patterns. Drawbacks of IEHAPS are the need of expensive computation, heavy pre-processing and ratability.

ANFIS is used for the prognosis of Ischemia. ANFIS is a hybrid model consists of Adaptive Neural Network and Fuzzy Inference system. It uses ECG signals as the dataset. Performance can be measured using the parameters such as sensitivity, specificity etc.

CEO-GA is an ensemble system that ensembles homogeneous classifier for heart disease classification and then results are optimized by using Genetic algorithm. Support Vector Machine algorithm with different kernel functions is used as base classifier [13]. Since CEO-GA makes of ensemble of different SVM kernel functions, the accuracy of decision will be high. This system is used for the prediction of heart diseases in general.

Table 2. Summary of various prediction systems

HD prediction system	HD type	Mode	Data Mining technique
IHDPS	HD in general	Hybrid	DT,NB,ANN
IEHAPS	Heart Attack	Hybrid	K-means, MAFIA,ANN
ANFIS	Ischemia	Hybrid	ANN, FIS
CEO-GA	HD in general	ensemble	SVM
BPSO-KNN-SVM	Heart Valve Diseases	Hybrid	BPSO, KNN,SVM
PRAA	Arthrosclerosis	Hybrid	PSO, imputation algorithm
ANN-GSO-ABC	HD in general	Hybrid	ANN,GSO, ABC

BPSO-KNN-SVM introduces a computer-aided diagnosis system of the heart disease using binary particle swarm optimization and support vector machine, along with K-nearest neighbor and with leave-one-out cross-validation. This system is used for the prediction of Heart Valve Diseases in specific.

PRAA is a novel approach which is used for the prediction of risk factor based on an in-built imputation algorithm and Particle Swarm Optimization technique. PRAA is used for the diagnosis of arthrosclerosis in particular. This has an accuracy of about 99.73%.

ANN-GSO-ABC is a hybrid algorithm, that uses neural network for the training purpose and optimization of result is performed using Group Search Optimization (GSO) and Artificial Bee Colony (ABC). Accuracy of this technique is very high because of the result optimization.

In Table 2, the comparison of seven different prediction systems is performed.

7. CONCLUSION

In this paper, eight different heart disease prediction systems developed during the last decade are compared. Each one varies in their prediction capability, data mining techniques used for classification and training, type of heart disease diagnosed, mode of design, attributes extracted etc.

Finally a novel approach for heart disease prediction is suggested. The propose method ensemble three base classifiers for training the dataset. The classifiers used are SVM, NB and RF. After training the results obtained will be optimized using an efficient optimization algorithm, Firefly.

8. REFERENCES

[1] Ethem Alpaydin, "Introduction to Machine Learning," 2nd ed., 2012.

[2] M.Akhil Jabbar, B.L Deekshatulu, Priti Chandra , "Heart Disease Prediction using Lazy Associative Classification,"IEEE, pp. 40-46, 2010.

[3] I. S. Jacobs and C. P. Bean, "HDPS: Heart Disease Prediction System," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Intelligent Heart Disease Prediction System using Data Mining Techniques,"pp. 108-115, 2008.

[5] R. Nicole, "Association Rule Discovery with the Train and Test Approach for Heart Disease prediction," IEEE transaction on Information Technology in Biomedicine, pp.334-343, 2008.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "An Empirical Study on Prediction of Heart Disease using Classification Data Mining Techniques," International Conference on Advances in Engineering, science and management(ICAESM), pp. 514-518, 2012.

[7] http://www.webmd.com/heart-disease/guide/heart-disease-symptoms-types

[8] http://www.healthline.com/health/heart-disease/types Reddy KS, Yusuf., Emerging epidemic of CVD in developing countries. Circulation 1998:596-601

[9] K Raj Mohan,Ilango Paramasivam, Subhashini Sathya Narayan, "Prediction and Diagnosis of Cardio Vascular Disease–A Critical Survey," World Congress on Computing and Communication Technologies, pp.246-251, 2014.

[10] http://www.preservearticles.com/201105216892/essay-on-the-structure-and-functions-of-heart-for-students.

[11] A.Emam, H.Tonekabonipour, M.Teshnelab, M.Aliyari Shoorehdeli, "Ischemia prediction using ANFIS," pp. 4041-4044, International conference on System Mans and Cybernetics, 2010.

[12] Benish Fida, Muhammad Nazir, Nawazish Naveed, Sheeraz Akram, "Heart Disease Classification Ensemble Optimization Using Genetic Algorithm," International Conference on Multitopic Conference, pp. 19-24, 2011.

[13] Mona Nagy Elbedwehy, Hossam M. Zawbaa, Neveen Ghali, and Aboul Ella Hassanien, "Detection of Heart Disease using Binary Particle Swarm Optimization," Federated Conference on Computer Science and Information Systems, pp. 177–182, 2012.

[14] V. Sree Hari Rao and M. Naresh Kumar, "Novel Approaches for Predicting Risk Factors of Atherosclerosis," IEEE Journal Of Biomedical And Health Informatics, Vol. 17, no. 1, 2013.

[15] B. Srinivasa Rao , K. Nageswara Rao, S. P. Setty, "An Approach for Heart Disease Detection by Enhancing Training Phase of Neural Network Using Hybrid Algorithm", IEEE International Advance Computing Conference (IACC), pp.1211-1220, 2013.

[16] Eman AbuKhousa and Piers Campbell, "Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems," International Conference on Innovations in Information Technology (IIT), pp.267-272, 2012.

[17] http://www.livescience.com/34655-human-heart.html

[18] https://www.cardiosmart.org/Heart-Basics/How-the-Heart-Works

[19] http://www.chop.edu/service/cardiac-center/heart-conditions/how-the-normal-heart-works.html

[20] UCI Machine Learning Repository. Arlington: The Association; 2006[updated 1996 Dec 3; cited 2011Feb2].Available from: http://archive.ics.uci.edu/ml/datasets/Heart+Disease