HEART DISEASES PREDICTION USING MACHINE LEARNING TECHNIQUES

[1]Dr.S. Jabeen Begum, [2]Saravanan S , [3]Vaasuki A.S, [4]Ahamed Shihabudeen M.I ,[5]Dr.V. Latha Jothi

[1]Professor and Head,[2,3,4]students,[5]Professor

Computer Science and Engineering

Velalar College of Engineering and Technology,Erode-638012.Tamilnadu,INDIA.

Email:[1]sjabeenbegum@gmail.com, [5]lathajothi.s@gmail.com

Contact No:+91 9894651159,+91 9842723701

Abstract *Prediction of CHD is difficult within the place of clinical record analysis. The quantity of data within the healthcare enterprise is huge. It is difficult to discover coronary heart disease because of several risk factors along with high blood pressure, diabetes, cholesterol and anomalous pulse rate. Data mining with classification plays particular role within the prediction of CHD and data investigation. In present paper, expecting the result by usage of the algorithm named particle SVM is introduced . By this algorithm, disorder are anticipated primarily based on the pulse rate, age, sex, blood pressure. Neural networks and kmeans and KNN are also used to prediction of heart disease. One of the most useful algorithm for prediction is unsupervised Learning in Machine Learning is k-Means Clustering. So we are imposing a heart disease prediction system using data mining techniques and K-means method.*

It facilitates in predicting the heart disease using numerous attributes.

Keywords: *Data Mining, Machine Learning, Heart Disease, Prediction, Classification.*

1.INTRODUCTION

Data mining has end up a fundamental methodology for computing applications in medical informatics. Progressing of data mining applications are manifested inside the region of information management in health informatics, epidemiology, healthcare organizations, affected patient care and monitoring systems, assistive technology, large-scale image analysis to automatic identification and information extraction of unknown classes. Various algorithms associated with data mining have considerably helped to recognise medical data more clearly, with the aid of distinguishing pathological data from normal data, is helpful in supporting decision making and then visualization and identification of hidden complicated relationships between diagnostic capabilities of various patient groups. Coronary Heart Disease is a major cause of disability in adults and unusual purpose of demise in countries like Europe, United States, South Asia, etc., It has been predicted that all the countries of the world might be affected due to this disease by the year 2020. Coronary Heart Disease refers to the failure of coronary circulation to supply needed circulation to cardiac muscle and its surrounding tissue. This restricts the supply of blood and oxygen to the heart, particularly at some stage in exertion when the myocardial

metabolic needs are increased. As the degree of coronary artery disorder progresses, there can be complete obstruction of the lumen of the coronary artery, critically affecting the flow of oxygen carrying blood to the myocardium. Individuals with these degree of coronary artery disease generally have suffered from more myocardial infarctions i.e heart attacks, and may have signs and symptoms of chronic coronaryischemia, including symptoms of Heart at rest and flash pulmonary edemae.

2. LITERATURE SURVEY

Data mining plays a vital role in identification and prediction of disease by various sorts of metabolic syndrome thus the various sorts of diseases can be discovered.

The development of a data mining model is mainly concerned with the Random Forest classification algorithm.[1] Random forest algorithm runs efficiently on large databases and has the capability of handling thousands of input variables. In their study classification trees, have certain advantage over logistic regression models. The developed model will have the functionalities such as predicting the occurrence of various events related to each patient record prevention of risk factors such as high blood pressure, sugar, cholesterol level is associated with CHD and to improve the overall prediction accuracy using random forest classification algorithm.

Backpropagation Neural Network (multi layered Feed Forward Neural Network) algorithm [2] for prediction of heart disease, blood pressure and sugar. The benchmark set is used in this work are the signs, symptoms and the results of physical evaluation of a patient. The dataset of Benchmark datasets has total 166 records are used for train a neural network on different categories of neonatal diseases . The author used 13 medical attribute for heart disease prediction and produced accuracy of 75% with higher stability.

An application of Artificial Neural Networks (ANN) is the prediction of patient coronary heart disease function. [3] Multilayer perceptron (MLP) which is a type of ANN architecture were used. In this paper Models that are evaluated according to accuracy, sensitivity, and specificity measures. Each method is evaluated by using 6-fold cross validation algorithm. It results the optimised ANN system achieved a greater accuracy level.

The ability of Fuzzy neural network model[4] to predict the coronary heart disease by evaluating individulas based on knowledge of their biomarker, risk habits and demographic profiles.It was calculated in terms of percentage accuracies and compared with the prediction performance of Logistic regression .for the prediction of coronary heart disease they taken the samples namely body mass index, systolic blood pressure, total cholesterol level, and age. Tenfold cross-validation were applied.

Machine Learning plays a important role in diagnosing a heart disease. Some of the machines learning techniques that are used such as [5] decision trees, neural networks, Naïve Bayes classification, genetic algorithms, regression and support vector machines. The decision tree algorithm is used for extracting rules in predicting heart disease. A graphical user based interface is used to input the patient data and predict whether the patient is suffering from heart disease or not by using Weighted Association rule based Classification.

Heart disease is actually complex process, as it requires depth knowledge and rich experience. In general, the prediction of heart disease is the traditional way of examining the medical report such as ECG, MRI, Blood Pressure, Sugar level, Stress tests by a medical practitioner. Now days, a numerous volume of medical data is available on medical industry and it acts as a great source for predicting useful and hidden facts in all medical problems. These facts, helps the practitioners to make accurate predictions. [6] The novel method of Artificial Neural Network concepts has also been contributing themselves in yielding highest prediction accuracy level over medical data.

Heart Attacks are the major cause of mortality in the world today, particularly in India. The need to predict the heart disease is necessity for improving the country's healthcare sector. Accurate and quick prediction of the heart disease mainly depends on Electrocardiogram data and clinical dataset. These dataset must be fed to a non linear disease prediction model. [7]This non linear heart function monitoring module can monitor the arrhythmias such as tachycardia, bradycardia, myocardial

infarction, atrial, ventricular fibrillation, atrial ventricular flutters and PVC's. An effective method to analyse the clinical data and ECG data, so as to train the Artificial Neural Network to accurately diagnose the heart and predict abnormalities if any.

The computational intelligence techniques is used for the detection of heart disease[12]. By comparing six well known classifiers of Cleveland dataset are used. It performs by feature selection process and compared with computational intelligence based feature selection mechanism. It shows to predict and improve the accuracy level.

Heart disease investigate the risk of sickness is mainly contribute to males and female. It uses three rule generation algorithm named apriori, predictive apirori and tertices. The prediction process is depend on gender. It compares, male and female ECG[13], either normal or hyper and slope. By comparing it gives the healthy status for both genders.

Heart disease experiences the association rule in medical dataset[14] for the prediction of heart disease It uses the attribute of chest pain to predict either it is normal or abnormal.It conclude the result by females more chonce of free from coronary heart disease.

In the world, large amount of people sufferes from heart disease. Artificial neural network, backpropagation algorithm is used. Neural network[15] uses 13 attributes to feature selection and backpropagation used to find the presence and absence of heart disease.

Heart disease are increasingly day by day due to hereditary. It uses the parameter such as blood pressure, cholesterol and pluse rate ranges to compare [16] predict the results. They uses classification techniques such as Naïve bayes, KNN, decision tree, neural network. It predicts the greater accuracy.

Heart disease prediction is achieved through decision support system. The performance become harder when the dataset contain missing value, they use Probabilistic Principal Component Analysis (PPCA)[17] to deal the problems of missing values. It uses medical test of various patients as input and proposed a result. In proposed methodology extracts high impact features by applying Probabilistic Principal Component Analysis(PPCA). The use of Radical Basis Function(RBF) and Support Vector Machine (SVM) to classify the heart prediction and normal patients.

Cancer and Cardiovascular disease provides thousands of problems to human health over a decade.It can be predicted by data mining [8] and several other techniques.In this the dataset from cardiovascular patients have been collected by UCI laboratory by using several algorithm such as decision tree,naïve bayes. The hybrid method achieves 86.8% accuracy.

In Multiple Criteria Decision Making problems(MCDM)[9] has several disadvantage for determining the best alternative. The TOPSIS and MCDM are applied with completely different dataset and different methods such as intuitionistic fuzzy TOPSIS. It is applied on IFS (Intuitionistic fuzzy dataset).In this they propose two methods are standard derivation(SD) and Preference Solution Index(PSI). They are compared and the accuracy is 87.1%.

Over the last decade various methods and tools been discovered for predicting heart disease. In this they uses an SAS software 91.3 for predicting (CVD)[10]. Neural networks are the center of this system. This method is achieved by combining the obtained propabilities from the methods. By these they achieved the accuracy of about 80.95 to 89.91%.

Yet many information is produced in association of medical industry but those informations are not used wisely. The successful analysis methods[11] are not there to find link in the health care. This paper gives the detailed techniques of knowledge abstraction using data mining. They have been used several methods such as Naïve bayes and decision tree for analysing medical datasets. They achieved an overall accuracy of about 83.9%.

The wireless sensor networks (WNS) acts as a core data that collects data from the smart cities. The securities are the key problems to this services. To get rid of this problem we used a hierarchical framework called Usage Control (UCON)[18]. In addition to that they use a dynamic adaptive chance discovery to discover unknown attacks. To design and construct such mechanism , a unified framework is applied in which low level attacks are being controlled in the base station. The software defined network and network function virtualization techs are used to perform attack migration when the attacks are detected. A controlled prototype was created to examine the consumption of resources and attack rates. The results are demonstrated in the means of feasibility and efficiency.

3.EXISTING SYSTEM

It is quite difficult to identify the heart disease. It is because of several risk factors such as cholesterol, blood pressure , diabetics , and high pulse rate etc,. The heart disease is complex in nature and the disease must be carefully handled. If it fails this may affect the heart and thereby causes serious infection. The aim of the medical science and data mining are used for predicting metabolic syndrome. The usage of data mining methods in the healthcare industry has been proved to take less time for the prediction of results with accurate results. This method uses effective prediction crossed with the mutation. For experimental validation, they use the well known Cleveland dataset which is collected from a UCI machine learning repository. HRFLM makes use of ANN and back propagation algorithm along with 13 clinical features as the input. The obtained results are comparatively analyzed against traditional method. The feature selection plays a prominent role in the prediction of heart disease. ANN with back propagation is proposed is used for prediction of the disease. The results obtained from the application of ANN are highly accurate and very precise heart disease.

4.PROPOSED SYSTEM

By using K-Means clustering and K-NN algorithms are used to predict the heart disease. Kmeans clustering is one among the simplest and popular machine learning algorithms. To process the learning data, the K-means algorithm in data mining starts with a primary group of randomly selected centroids, that are used as the beginning points for every cluster, and then plays iterative (repetitive) calculations to optimize the positions of the centroids. K-Nearest Neighbour is considered as one of the most basic yet essential classification algorithms in Machine Learning.

4.1 PREPROCESSING DATA

The process of rule generation advances in two stages. During the first stage, the SVM model is built using training data During each fold, this model is utilized for predicting the class labels .The rules are evaluated on the remaining 10% of test data for determining the accuracy, precision, recall and F-measure. In addition, ruleset size and mean rule length are also calculated for each fold of cross-validation.

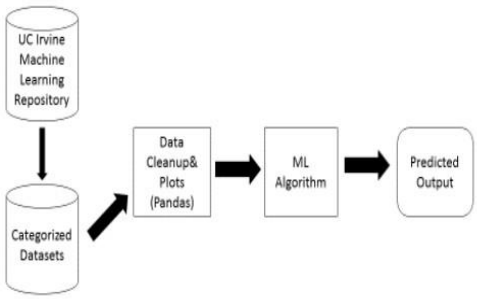

Figure 1: methodology used to construct the model

4.2 TRAINING

The training set is different from test set. In this study, we used this method to verity the universal applicability of the methods. In k-fold cross validation method, the whole dataset is used to train and test the classifier to Heart Stoke.

From among the 13 attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining 11 attributes are considered as important because they contain vital clinical records. Clinical records are useful for diagnosis and learning the severity of heart disease. As previously mentioned in this experiment, several (ML) techniques are used namely, NB,SVM,DT and KNN and K-means. This experiment was repeated with all the ML techniques using all 13 attributes.

Table 1: dataset attributes detailed information

Attribute	**Description**	**type**
Age	Paitents age in years	Numeric
Sex	Paitents gender(male represented as 1,female represented a 0	Nominal
Cp	Types of Chest Pain categorized into 4 values: 0.Typival angina 1.Atypical angina 2.Non typical angina 3.asymptomatic	Nominal
Trestbps	Level of blood pressure at resting mode(in mm/Hg at the time od admitting in the hospital)	Numeric
Chol	Serum cholesterol in mg/dl	Numeric
FBS	Blood sugar level >120 mg/dl; represented as 1 in case of sugar level is high and 0 in case of sugar level is low	Nominal

Resting	Results of Electrocardiogram is represented as 1. Normal state is 0 2. Abnormal state is 1	Nominal
Thali	The accomplishment of the maximum rate of the Heart	Numeric
Exang	Angina induced by exercise (0 as No, 1 as yes)	Nominal
Oldpeak	Exercise induced by ST depression	Numeric
Slope	ST segment measured in terms of sloping 1.unsloping 2. Flat 3.downsloping.	Nominal
Ca	Fluoroscopy coloured vessels numbered from 1 to 4	Numeric
Thal	Status of the Heart	Nominal
Num	Heart disease diagnosis	Nominal

4.3 CLASSIFICATION MODELLING

The clustering of datasets is done on the basis of the variables and criteria of Decision Tree (DT) features. Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the above results based on their low rate of error. The performance is further optimized by choosing the DT cluster with a high rate of error and extraction of its corresponding classifier features. The performance of the classifier is evaluated for error optimization on this data set.

1)DECISION TREES

For training samples of data, the trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top down recursive divide and conquer(DAC) approach.

2) SUPPORT VECTOR MACHINE

Let the training samples having dataset Data =$\{yi, xi\}; i = 1,2,...,n$ where xi ∈Rn represent the ith vector and $yi \in Rn$ represent the target item. The linear SVM finds the optimal hyperplane of the form $f(x) = wTx + b$ where w is a dimensional coefficient vector and b is a offset. This is done by solving the subsequent optimization problem:

$$\boldsymbol{Minw, b, \xi i\ 1\ 2\ w2 + C\ n\ X\ i = 1\ \xi i}$$

$$\boldsymbol{s.t.\ yiwTxi + b \geq 1 - \xi i, \xi i \geq 0, \forall i\ \in \{1,2, \dots, m\}}$$

4.4 K MEANS CLUSTRING

k-means is one of the simplest algorithms is used for solving the well known clustering problem. The procedure follows a simple given dataset through a certain number of clusters The main idea is to define k centers, one for each cluster. These centers should give a different result because of different

location centres. The next step is to take each point belonging from a given data set and associate it to the nearest center. Where there is no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data of set points and the nearest new center. As a result of this loop that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

4.5 K-NEAREST NEIGHBOR (KNN)

Our proposed approach KNN algorithm to improve the classification accuracy of heart disease data set. By this algorithm used to search as a goodness measure to prune redundant and irrelevant attributes, and to rank the attributes which contribute more towards classification. It contains Least ranked attributes that are removed, and classification algorithm is built based on evaluated attributes. This classifier is trained and it helps to classify heart disease data set as either healthy or sick. Our proposed algorithm consists of two parts.

1.First part deals with evaluating the attributes by using genetic search.

2.Second part deals with building classifier and measuring accuracy of the classifier Proposed algorithm.

Step 1: load the data set
Step 2: Apply K-Means on the data set
Step 3: attributes are ranked based on their value
Step 4: selects the subset of higher ranked attributes
Step 5: Apply (KNN) on the subset of attributes that maximizes classification accuracy
Step 6: calculates accuracy of the classifier, which measures the ability of the classifier to correctly classify unknown sample.

5. PERFORMANCE MEASURES

Several standard performance metrics such as accuracy, precision and error in classification have been considered for the computation of performance efficacy oft his model.

Accuracy in the current context would mean the percentage of instances correctly predicting from among all the available instances. Precision is defined as the percentage of corrective prediction In this approach, the classification accuracy rates for the datasets were measured. For example, in the classification The True Positive rate (TP) and True Negative rate (TN) are correct classifications. A False Positive (FP) occurs when the outcome is incorrectly predicted it shown as positive but it is actually negative. A False Negative (FN) occurs when the outcome is incorrectly predicted as negative but it is actually positive.

prediction		disease	
		+	–
test	+	True positive (TP)	False positive (FP)
	–	False negative	True negative

		(FN)	(TN)

1.Accuracy - It refers to the total number of records that are correctly classified by the classifier. Accuracy = (TP+TN) / (TP+FP+FN+TN)

2. Classification error - This refers to the misclassified datasets from the correctly classified records.

3. True Positive Rate (TP) : It corresponds to the number of positive examples that have been correctly predicted by the classification model.

4. False Positive Rate (FP) : It corresponds to the number of negative examples that have been wrongly predicted by the classification model.

5.Precision : is the fraction of retrieved instances that are relevant.

Precision = (TP) / (TP+FP)

6.Recall : Is the fraction of revelent instance that are retrieved.

Recall = (TP) / (TP+FN)

Since the class label prediction is of multi-class, the result on the test set will be displayed. Each matrix element shows the number of test cases for which the actual class is the row and the predicted class is the column.

Table2: Performance Analysis

performance measure	**prediction levels**
Accuracy(%)	90.3
True positive rate	0.630
False positive rate	0.254
precision(%)	0.575
Recall (%)	0.630
classification error (%)	35.44

6.RESULT AND DISCUSSIONS

Support vector model algorithm runs efficiently on large databases and has the capability of handling thousands of input variables. It generates an internal unbiased estimate of the generalization error as the building progresses and has an effective method for estimating the missing data and maintains the accuracy when large proportion of the data are missing. The tree forest that has been generated can be saved in order to make comparative study about the features of the attributes**.**

At first the suitable attributes are taken from the training dataset by the proposed approach and then they are tested and trained to find the prediction

To measure the effectiveness of the approach experiments have been conducted using the UCI machine learning dataset consisting of 13 attributes. The attributes involved are age, sex, chest pain type, serum cholesterol, fasting blood sugar, resting electro cardio graphic results, maximum heart rate achieved, exercise induced angina, ST depression, the major vessels colored by fluoroscopy, and thal..

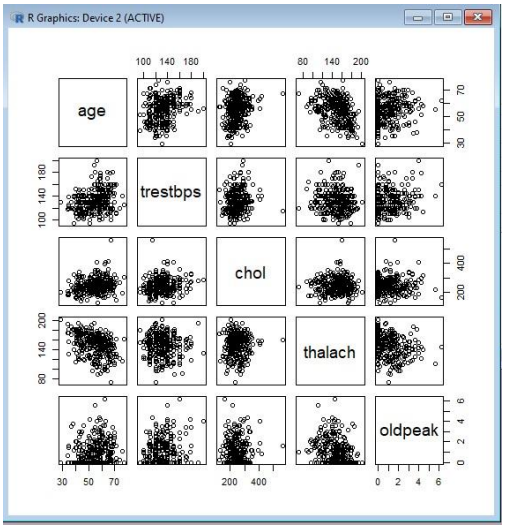

Figure1:Clustring the dataset with 5 attributes

Categories the attribute based on 5 attributes are 1.Age factor is used.
2.Age factor is compared with sugar level to predict the disease.
3.Age factor is compared with cholesterol level to predict the disease.
4.Age factor is compared with thalach for prediction.
5.Age factor is compared with oldpeak for prediction.

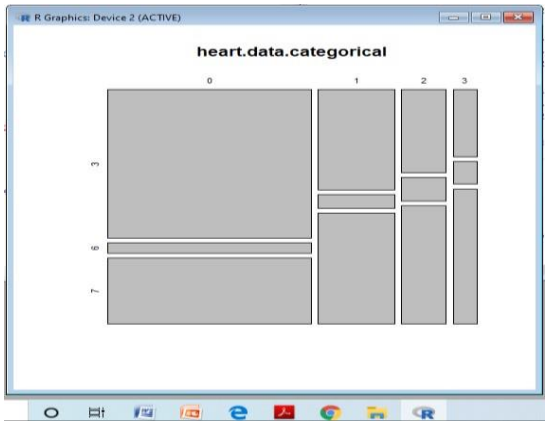

Figure2:comparing the attributes

After Clustring these 5 attributes are compared and predict whether the male ,female among these 5 attributes.
Comparison is based on chest pain and chi squared value. Where,0,1,2,3 are chest pain categories,

7,6,3 are based on chi square value.

Using two types of comparision method,
1.comparing with 6 attributes.
2.comparing with 100 attributes.

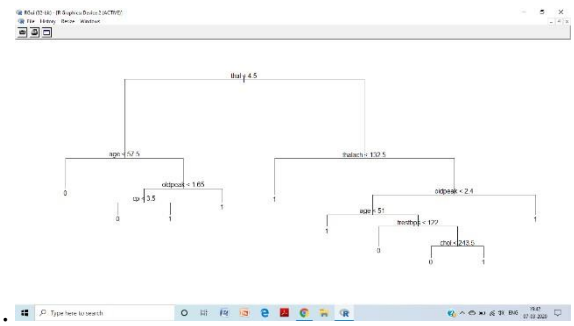

Figure3: Bilateral tree method for comparing 6 attributes

Thal is taken and compared with age, sex, cp, oldpeak, chol, thresbp.

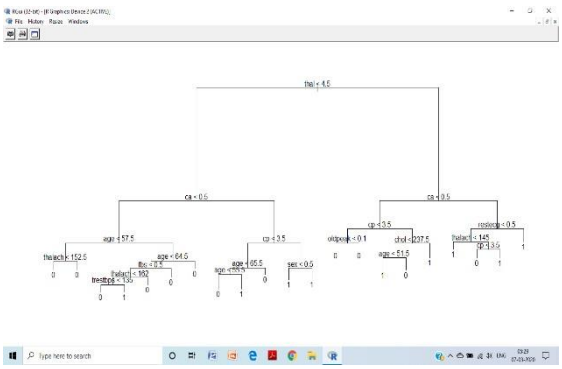

Figure4: Bilateral tree method for comparing 10 attributes

Taken thal for comparing other attributes. Their comparion is between ca, cp, age, sex, thalach, oldpeak, chol, ECG, thresbps in form of Bilateral tree.

7.CONCLUSION

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques helps to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of theoretical approaches and simulations. In proposed hybrid KNN approach is used.

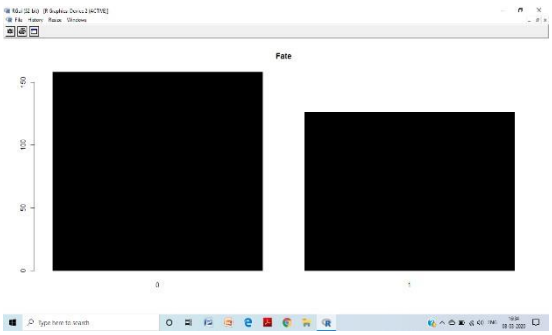

Figure5:Gender comparison

Based on comparison, male has more number of attaining heart disease than female.

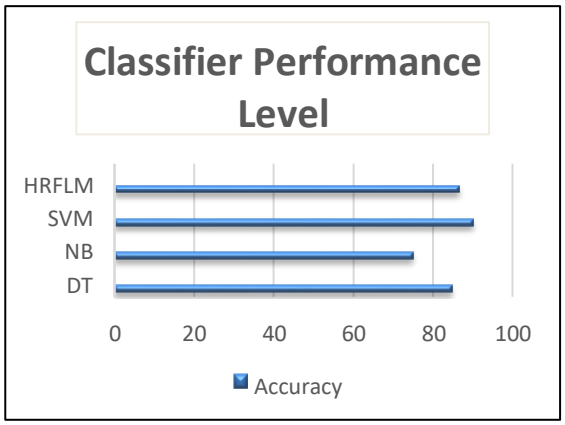

Figure6: Comparing Performance of Accuracy level

8. REFERENCES

[1] A. S. Abdullah and R. R. Rajalaxmi, ``A data mining model for predicting the coronary heart disease using random forest classifier," in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, April. 2012, pp. 2225.

[2] N. Al-milli, ``Backpropogation neural network for prediction of heart disease," J. Theor. Appl.Inf. Technol., volume. 56, no. 1, pp. 131135, 2013.

[3] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ``Analysis of neural networks based heart disease prediction system," in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, July. 2018, pp. 233239.

[4] P. K. Anooj, ``Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," J. King Saud Univ.-Comput. Inf. Sci., vol. 24, no. 1, pp. 2740, January. 2012. doi: 10.1016/j.jksuci.2011.09.002.

[5] M Durairaj ,V Revathi , "Prediction of Heart disease using Back Propagation MLP algorithm" August 2015.

[6] M Nagaraj ,C Lutimath, Chethan , Basavaraj S Pol, "Predicition of heart disease using machine learning" September 2019.

[7] D. K. Ravish,Nayana R Shenoy " Heart function monitoring,prediction and prevention of heart attacks: using Artificial Neural Network" October 2015.

[8] H. A. Esfahani, M. Ghazanfari, "Cardiovascular disease detection using a new ensemble classifier", *Proc*. IEEE 4th Int. Conf. Knowl.-Based Eng. Innov. *(*KBEI), pp. 1011-1014, December. 2017.

[9] F. Dammak, L. Baccour, A. M. Alimi, "The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains", Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), vol. 9, pp. 1-8, August. 2015.

[10] R. Das, I. Turkoglu, A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles", Expert Syst. Appl., volume. 36, no. 4, pp. 7675-7680, May 2009.

[11] N. K. S. Banu, S. Swamy, "Prediction of heart disease at early stage using data mining and big data analytics: A survey", Proc. Int. Conf. Elect. Electron. Commun. Comput. Optim. Techn. (ICEECCOT), pp. 256-261, December. 2016.

[12] J. Nahar, T. Imam, K. S. Tickle, Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach", Expert Syst. Appl., volume. 40, no. 1, pp. 96-104, 2013.

[13] J. Nahar, T. Imam, K. S. Tickle, Y.-P. P. Chen, "Association rule mining to detect factors which contribute to heart disease in males and females", Expert Syst. Appl., volume. 40, no. 4, pp. 1086-1093, 2013.

[15] T. Karayılan, Ö. Kılıç, "Prediction of heart disease using neural network", Proc. Int. Conf. Comput. Sci. Eng. (UBMK), pp. 719-723, October. 2017.

[16] J. Thomas, R. T. Princy, "Human heart disease prediction system using data mining techniques", Proc. Int. Conf. Circuit Power Comput. Technol. *(ICCPCT)*, pp. 1-5, March. 2016

[17] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, S. A. Hussain, "Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis", *Phys.* A Stat. Mech. Appl., volume. 482, pp. 796-807, 2017.

[18] J. Wu, K. Ota, M. Dong, C. Li, "A hierarchical security framework for defending against sophisticated attacks on wireless sensor networks in smart cities", IEEE Access, volume. 4, pp. 416424, 2016.