HEART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS

Bhagyashree S. Shelke[1], K. C. Waghmare[2], Anil Kumar Gupta[3], Amarjeet Sharma[3]

P.G. Student, Department of Computer Engineering, PICT, Pune, Maharashtra, India.

Associate Professor, Department of Computer Engineering, PICT, Pune, Maharashtra, India.

CDAC, Pune, Maharashtra, India.

Abstract : Nowadays, health diseases are increasing rapidly specially the heart disease. So it is necessary to predict the heart disease in early stage to save the number of patients with mortality. Previous studies show that various data mining and machine learning techniques have been used for the prediction of heart disease with different numbers of parameters. Patient's different parameters are considered essential for the prediction such as blood pressure, heart rate, cholesterol, and so on. In this paper dataset used for experiment purpose is Body Sensor Network (BSN) dataset which contains different attributes. In this paper, the proposed system consists of three algorithms, Q-Learning algorithm, Naive Bayes algorithm and Heart Disease Prediction (HDP) algorithm. Q-Learning and Naive Bayes are two already existing machine learning algorithms and the proposed one is Heart Disease Prediction (HDP) algorithm. Q-Learning and Naive Bayes algorithms provide class label as an output based on the previously defined threshold in the algorithms. The output of these two algorithms used as input to the HDP algorithm. Finally HDP generates patient health report that is prediction result.

IndexTerms **- Supervised learning, Machine Learning, Heart disease, Data mining techniques, Clustering, and Classification, Prediction system**

I. INTRODUCTION

The current survey shows that mortality rate increasing in vast amounts because of heart disease. So to minimize the mortality rate intelligent heart disease prediction system is required. There are various reasons for heart disease like changing lifestyle, more stress and so on. So heart disease prediction very important needs of life. As we have studied in literature various data mining techniques have been used for the prediction of heart disease. Parameters considered for experiment are blood pressure, heart rate, cholesterol, pulse rate, and so on. Each day in large quantity medical data is generated so important knowledge extraction from this big data is challenging task. Heart is the main part of human life, if heart is properly working then human health is ok. Improper functioning of the heart will influence other parts of the human body like brain, kidney etc. [7]
So implementing the more accurate heart disease prediction system is the main objective of this paper, and finally reduces the treatment cost, which can be affordable for all classes (poor, rich) peoples. In previous studies various data mining algorithms are used, but in this paper, we have used machine learning algorithms. HDP algorithm which is proposed one finally generates the patient report i.e. prediction result.
The work in this paper is organized as shown: literature survey is given in section II, section III gives the overview about a proposed system, Results and discussion is described in section IV and section V is about final conclusion.

II. Literature Survey:

In [1] author has been developed heart disease prediction for the prediction of the heart disease. Author has been performed comparative analysis with various algorithms and found out best one. Six machine learning algorithms considered for experiment purpose, such as Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. Among all these six algorithms Logistic Regression gives good accuracy than other algorithms.
The authors in [2] have used heart disease prediction technique called Apriori Algorithm. This algorithm is implemented in WEKA and MATLAB software. MATLAB gives better accuracy than WEKA software. Apriori algorithm uses the strong rules and it is analyzed using different strong rules. It also used to find out the hidden pattern present in large scale data. To check the accuracy of algorithm support and confidence as two measures has used.
The paper [3] proposed algorithm used for the prediction is Neural Network, mainly predicts the cardiac illness using patients electronic medical data. The Dataset used for prediction is electronic health records (EHR) dataset. First of all it is converted into standard form and then used as input for the algorithm. Word Vector and One-hot encoding have used for representing patient's diagnostic events. The proposed system gives better accuracy as compared to previous implemented algorithms for heart disease prediction.
In [4] proposed a new ensemble model for heart disease prediction. This ensemble model consists of machine learning algorithms and Fast Fourier Transformation. The machine learning algorithms used in this ensemble model are, Neural Network, Support Vector Machine and Naive Bayes. The experimental results presented in this paper show that the proposed system performs better than other algorithms.
In paper [5] author proposed system for heart disease prediction using clustering algorithm, which effectively helps in heart disease prediction. A system implemented in WEKA tool. Best algorithm is decided on different parameters such as error functions and building times. According to experimental results presented in this paper three algorithms perform best, Filtered Cluster, Make Density base Cluster, and Simple K-means.

The author in [6] proposed prediction system that predicts heart failure and also count mortality rate of patients. The proposed algorithm is Improved Random Survival Forest (iRSF). The main objective of the proposed system is that to identify various predicators which help in the prediction of heart disease.

In paper [7] author proposed decision support system for heart disease prediction. Proposed system constitutes three techniques Naive Bayesian classifier, which requires a small amount of training data, Decision Tree, which requires fixed set of training data, Genetic algorithm, which uses fitness function for finding out a solution.

The work cited in [8] gives overview of classification algorithms which found out the risk level of each person for heart disease. Different parameters have used for calculating the risk level such as blood pressure, heart rate, cholesterol, pulse rate and so on. Algorithms used for the predictions are Naive Bayes, K Nearest Neighbour, Decision Tree and Neural Network. The survey given in this paper shows that accuracy of algorithms gets increased when we use number of attributes.

In [9] Naive Bayes algorithm used for heart disease prediction. The Main system consists of two parts classification and prediction. The proposed system is also called as web based questionnaire application. The user inputs passed to this system, then from the user inputs it find out the required information which helps in solving the complex queries. The proposed system in this paper reduces the treatment cost of the patients.

The author in [10] provides brief overview on various techniques used for the heart disease prediction, such as SVM, KNN, ANN, Logistic Regression and so on. From this survey author observed that there are two models used for the heart disease. One is single model and another one is hybrid model and found that hybrid model is more efficient than single model.

Table 1 gives the overview of prediction algorithms for heart disease.

Table 1. Overview of heart disease prediction algorithms

Ref. No.	Title	Technique Used	Advantage	Disadvantage
[1]	Heart Disease Prediction System using Data Mining Techniques [2016]	Decision Tree, Naive Bayes, KNN	KNN in this paper used with large size dataset and provides more accurate results	Real time dataset have not been used for experiment purpose. Comparative analysis is not performed with other classification algorithms.
[2]	Hybrid Recommendation System for Heart Disease Diagnosis based on Multiple Kernel Learning with Adaptive Neuro-Fuzzy Inference System [2017]	Multi Kernel Learning with Adaptive Neuro-Fuzzy Inference System (MKL with ANFIS)	This algorithm is analyzed and studied with many existing deep learning algorithms.	Proposed system only applicable for heart disease prediction.
[3]	A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases [2017]	Scalable IOT based algorithm	Proposed algorithm considers only significant parameters for the prediction not the all parameters.	Proposed algorithm can't be used for a continue health monitoring system.
[4]	Prediction of Cardiac Arrhythmia type using Clustering and Regression Approach (P-CA-CRA) [2017]	Clustering algorithm (DBSCAN) and Logistic Regression	Experiment has been carried out on real time dataset.	More number of attributes has been used. Attribute reduction is not performed.
[5]	Heart Disease Prediction with Data Mining Clustering Algorithms [2017]	Clustering Algorithms	Proposed system can be used as practical tool for prediction of heart disease in any medical centres.	More time consuming.
[6]	Heart Disease Prediction Using Hybrid Genetic Fuzzy Model [2015]	Genetic Algorithm, Fuzzy Logic	It is easy to build the model.	Proposed algorithm can only applicable to heart disease dataset.
[7]	A heart disease recognition embedded system with fuzzy cluster algorithm [2013]	Fuzzy Cluster Algorithm	Reduction in time due to reduction in data samples. It has used for recognition of many types of heart diseases.	Proposed system only work as assistance for the newly graduated physicians.
[8]	Heart Disease Prediction System using Naive Bayes [2013]	Naive Bayes	Proposed system can be used as training tool in many medical centres.	Proposed system is generalised system it can be only applicable for heart disease dataset.

III. Proposed System:

In the proposed system as shown in the Fig 1, first system collected the current input states from each sensor then converted it from analog to digital signals using ADC, once the conversion has done, it has received by micro-controller, and at the same time it is stored into the database. In the network module some background knowledge data is stored, which is taken from the doctors. Performed some manual pre-processing of this data and we generated some rules. This data is also called as standard data which we have been used for the experimental purpose. Then applied proposed machine Learning, Q-Learning and Naive Bayes algorithms on run time input coming from the sensors and at the same time it has checked all input values with desired threshold of generating rules. If any time values show below the minimum threshold values as well as maximum the threshold values then HDP algorithm predicts that patient has heart disease or not. Dataset used for the implementation purpose is Body Sensor Network (BSN) dataset which contains the attributes such as Blood Pressure, Cholesterol, heart rate, time-stamp so on.

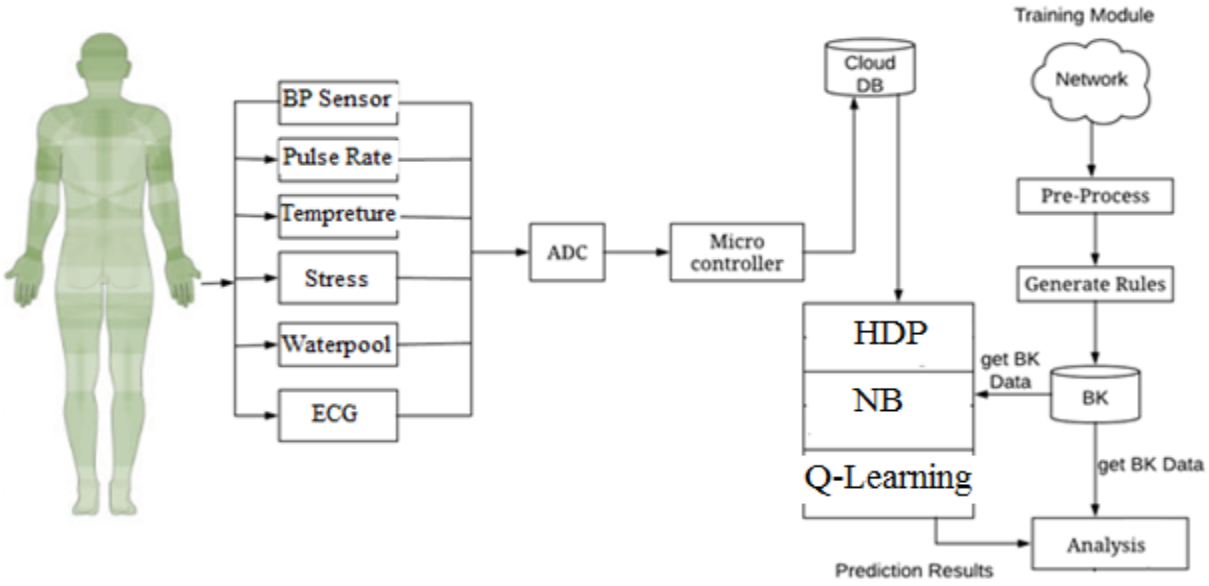

Fig 1: Proposed System Architecture

Dataset Used in the System:

Dataset used in the system named as Body Sensor Network (BSN) dataset which contains 13 clinical parameters such as, systolic blood pressure, diastolic blood pressure, heart rate, cholesterol, low cholesterol, high cholesterol, stress, sugar, QTR interval and PR interval as ECG parameters, oxygen, haemoglobin, timestamp.

IV. System Analysis:

Analysis of the system is done by using the following algorithms.
Algorithm 1 returns class labels such as High, Normal and moderate.
Algorithm 2 returns the AvgTscore for each parameter.
Algorithm 3 takes the input from algorithm 1 and algorithm 2 and finally predicts that heart disease is present or not for particular patient.

Algorithm1. Q-Learning Algorithm:

Input: In [1…..n] all input parameters which are generated by sensors, Threshold group Tmin [1…n] and Tmax [1…n] for all sensors, Desired Threshold Th.
Output: Trigger executed for output device as label.
Step 1: Read all records from database (r into DB)
Step 2: Parts [] ← Split(r)

Step 3: Calculate Cval

$$\text{Cval} = \sum_{k=0}^{n} Parts[k] \quad (1)$$

Step 4: check (Cval with Respective threshold of Tmin [1…n] and Tmax [1…n]**)**
Step 5: T← get current state with timestamp
Step 6: if(T.time > Defined Time)
Read all measure of for penalty TP and reward FN
Else continue. Total++
Step 7: calculate penalty score

$$\text{Score} = (\text{TP} * 100 / \text{Total}) \quad (2)$$

Step 8: if (score >= Th)
Generate event
End for

Algorithm2. Naive Bayes:

Input: User input file data record which contains all body parameters sensor values, Patient id pid, Timestamp T.
Output: Classified label
Step 1: Read r {all attributes} from current parameters.
Step 2: Map with train features with each sample.
Step 3: calculate average weight of train DB with same evidences

$$\text{AvgTscore} = \sum_{k=0}^{n} (sc) \quad (3)$$

Step 4: evaluate AvgTscore > threshold
Step 5: Return AvgTscore

Algorithm3. Heart Disease Prediction:

Input: Input values for all parameters HashMap <Double Value, String class> which contains the all attributes values like {BP, Heart rate, Cholesterol, Stress, Sugar, ECG, Oxygen saturation, Hemoglobin, CI} etc. policy patterns {A1, A2, An}
Output: Generate sample report for individual patient.
Step 1: for each read Hashmap

$$\text{Extracted\_Attribute[i][j]} = \sum_{i=0,j=0}^{n} (a_{[i]}, a_{[j]} \ldots\ldots a_{[n]}, a_{[n]}) \quad (4)$$

Step 2: if Extracted_Attribute[j] similar to A[1]
NormalPos = +1
MasterLits1. Add ← (NormalPos)
Step 3: if Extracted_Attribute[j] similar to A[2]
AbnPos = +1
MasterLits2. Add ← (AbnPos)
Step 4: if Extracted_Attribute [j] similar to A[n]
DenPos = +1
MasterLits3. Add ← (DenPos)
Step 5: end for
Step 6: calculate the fitness factor for all classes using below formula for all class list

$$f = \sum_{k=0}^{n} \frac{F(x)}{Sum\ F(x)} \quad (5)$$

Step 7: Weight _CurrentList[w] = $\frac{\text{MasterLits[i]}}{\text{TotalTest}}$ *100 (6)
Step 8: Sort _CurrentList[w] using desc order
Step 9: Recommend _CurrentList[0] for final class for patient profile.
Step 10: end procedure.

V. Experimental Evaluation:

A. Testing:

For the proposed system experiment is carried out on Body Sensor Network dataset (BSN). Dataset contains 13 attributes and 1000 instances. Attributes used as systolic blood pressure, diastolic blood pressure, heart-rate, cholesterol, low-cholesterol, high-cholesterol, stress, sugar, QTR interval and PR interval as ECG parameters, oxygen, haemoglobin, timestamp. 4 classes used for prediction such as Normal, Mild, Moderate and Severe.

B. Performance Analysis:

The accuracy of proposed algorithm is evaluated using the below equation (7),

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

Where,

TP is the number of true positive
TN is the number of true negative
FP is the number of false positive
FN is the number of false negative

VI. Results and Discussion:

In the implementation of proposed system we applied the Q-Learning algorithm and Naive Bayes algorithm on BSN dataset. Q-Learning algorithm gives the count of number of records falls under three classes Normal, High and Moderate. Q-Learning algorithm uses 4 parameters for the experiment purpose such as systolic blood pressure, heart rate, cholesterol, timestamp. Naive Bayes algorithm gives the classification based on the age for each individual person. Naive Bayes algorithm is performed on 11 attributes. Finally we pass the output of Q-Learning and Naive Bayes algorithms to HDP algorithm and it gives the prediction report of the patients, which predicts that whether patients has heart disease or not.

Conclusion:

We proposed a method for the prediction of heart disease called as Heart Disease Prediction (HDP) algorithm using two existing algorithms such as Q-Learning and Naive Bayes algorithms. Proposed HDP algorithm gives 89% accuracy compared with other existing algorithms. Q-Learning algorithm gives the count of number of records which falls under three categories such as High, Normal and Moderate. Naive Bayes algorithm gives the classification based on the age. It gives the classification of parameters of each individual person.

Future Scope:

We can use attribute reduction algorithm such as Principal Component Analysis (PCA) on dataset and increase the accuracy using these reduced attributes.

VII. Acknowledgement:

I take this opportunity to express my deep sense of gratitude towards Dr. Anil Kumar Gupta, Centre of Development of Advanced Computing (CDAC), Pune for giving me this splendid opportunity to select and present this topic. I would like to thank my guide Prof K. C. Waghmare, for her indispensable support, priceless suggestions and for most valuable time given when it was required. I wish to express my thanks to our Head Of Department Dr. R. B. Ingle, HOD Computer Department, PICT, for encouragement and providing me with the best facilities for work. I thank all the staff members, for their indispensable support and priceless suggestions. I thank all the people who are directly or indirectly involved in this project.

References:

[1] Ashok Kumar Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease", Neural Computing & Applications, Vol. 29(10), 2016, pp. 685-693.

[2] Mirpouya Mirmozaffari, Alireza Alinezhad, Azadeh Gilanpour, "Data Mining Apriori Algorithm for Heart Disease Prediction", Int'l Journal of Computing, Communications & Instrumentation Eng. (IJCCIE) Vol. 4, Issue 1, (2017), pp.20-23.

[3] Bo Jin, (Senior Member, IEEE), Chao Che, Zhen Liu, Shulong Zhang, "Predicting the Risk of Heart Failure with EHR Sequential Data Modeling: Heart Disease", IEEE Journal, Vol. 6, (2017), pp. 9256-9261.

[4] Ji Zhang , Raid Luaibi Lafta ,Xiaohui Tao , Yan Li ,Fulong Chen ,Yonglong Luo ,Xiaodong Zhu, "Coupling a Fast Fourier Transformation with a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment", IEEE Journal, Vol.5, 2017, pp.10674-10685.

[5] Mirpouya Mirmozaffari1, Alireza Alinezhad, and Azadeh Gilanpour. "Heart Disease Prediction with Data Mining Clustering Algorithms.", Int'l Journal of Computing, Communications & Instrumentation Eng. (IJCCIE) Vol. 4, Issue 1, (2017), pp.16-19.

[6] Fen Miao, Yun-Peng Cai , Yu-Xiao Zhang , Xiao-Mao Fan ,Ye Li , "Predictive Modeling of Hospital Mortality for Patients with Heart Failure by Using an Improved Random Survival Forest", IEEE Journal, Vol. 6, (2017), pp. 7244-7253.

[7] N Ch Sriman Narayana Iyenger, "Heart disease prediction system using data mining and hybrid intelligent techniques: A review", Int. J. Bio-Sci. Biotechnol, Vol.8, (2016), pp.139-148.

[8] Theresa Princy. R. J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT], 2016, pp.1-5.

[9] S.Indhumathi, Mr.G.Vijaybaskar, "Web Based Health Care Detection Using Naive Bayes Algorithm", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4, Issue 9, 2015, pp.35-36.

[10] G. Purusothaman, P. Krishnakumari, "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease", Indian Journal of Science and Technology, Vol 8(12), 2015, pp. 1-5.

[11] Theresa Princy. R, J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT], 2016, pp. 1-5.