Heart Disease Prediction

Conference Paper · February 2021

CITATIONS
7

READS
21,758

1 author:

Nayab Akhtar
Fatima Jinnah Women University
5 PUBLICATIONS **14** CITATIONS

SEE PROFILE

Heart Disease Prediction

Sibgha Taqdees
Sibgha1998oct@gmail.com
Department of Software Engineering
Fatima Jinnah Women University, The Mall, Rawalpindi, Pakistan

Nayab Akhtar,
Nayabf52@gmail.com
Department of Software Engineering
Fatima Jinnah Women University, The Mall, Rawalpindi, Pakistan

Kanwal Dawood
Kanwaldaud789@gmail.com
Department of Software Engineering
Fatima Jinnah Women University, The Mall, Rawalpindi, Pakistan

Abstract: Heart disease is the major cause of deaths worldwide. To give treatment for heart disease, a lot of advanced technologies are used. In medical center it is the most common problem that many of medical persons do not have equal knowledge and expertise to treat their patient so they deduce their own decision and as a result it show poor outcome and sometime leads to death. To overcome these problems predictions of heart disease using machine learning algorithms and data mining techniques, it become easy to automatic diagnosis in hospitals as they are playing vital role in this regard. Heart disease can be predicted by performing analysis on patient's different health parameters. There are different algorithm to predict heart disease like naïve Bayes, k Nearest Neighbor (KNN), Decision tree ,Artificial Neural Network(ANN).We have used different parameters to predict heart disease. Those parameters are Age, Gender, Cerebral palsey (CP), Gender, Cerebral palsey (CP), Blood Pressure (bp), Fasting blood sugar test (fbs) etc. In our research paper, we used built in dataset .we have implement the five different techniques with same dataset to predict heart disease These implemented algorithm are Naive Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest .This paper investigates that which technique gives more accuracy in predicting heart disease based on health parameters. Experiment show that Naïve Bayes has the highest accuracy of 88%.

Keyword: Naive Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest, Heart Disease

Introduction:

Heart disease is the major cause of deaths globally. More people die annually from CVDs than from any other cause, an estimated 12 million people died from heart disease every year. Heart disease kills one person every 34 seconds in the United States.

Heart attacks are often a tragic event and are the result of blocking blood flow to the heart or brain. People at risk of heart disease may show elevated blood pressure, glucose and lipid levels as well as stress. All of these parameters can be easily measured at home by basic health facilities.

Coronary heart disease, Cardiomyopathy and Cardiovascular disease are the categories of heart disease. The word "heart disease" includes a variety of conditions that affect the heart and blood vessels and how the fluid gets into the bloodstream and circulates there in the body. Cardiovascular disease (CVD) causes many diseases, disability and death. Diagnosis of the disease is important and complex work in medicine.

Medical diagnosis is considered as crucial but difficult task to be done efficiently and effectively .The automation of this task is very helpful. Unfortunately all physicians are not experts in any subject specialists and beyond the scarcity of resources there some places. Data mining can be used to find hidden patterns and knowledge that may contribute to successful decision making. This plays a key role for healthcare professionals in making accurate decisions and providing quality services to the public. The approach provided by the health care organization to professionals who do not have more knowledge and skills is also very important. One of the main limitations of existing methods is the ability to draw accurate conclusions as needed. In our approach, we are using different data mining techniques and machine learning algorithms, Naïve Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest to predict the heart disease based on some health parameters.

Related Work:

In Paper [1], uses the data from UCI data repository. Propose heart disease prediction using KStar, J48, SMO, and Bayes Net and Multilayer perception using WEKA software. Based on performance from different factor SMO (89% of accuracy) and Bayes Net (87% of accuracy) achieves optimum performance than KStar, Multilayer perceptron and J48 techniques using k-fold cross validation. The accuracy performance achieved by those algorithms is still not satisfactory. In Paper [2] they use data from Kaggle propose application of knowledge discovering process on prediction of stroke patients based on Artificial Neural Network (ANN) and Support Vector Machine (SVM), which give accuracy of 81.82% and 80.38% for ANN and SVM respectively for training data set and 85.9% and 84.26% for Artificial Neural Network (ANN) and Support Vector Machine (SVM) in test dataset respectively. Paper [3] use data from UCI repository and evaluate performance of different machine learning algorithm using Naive Bayes, KNN, Decision Tree, ANN. Among them ANN gave the highest accuracy of 85.3%. While Naïve Bayes and KNN gave almost 78% and Decision Tree gave 80%.Paper [4] use WEKA tool for measuring performance of different machine learning algorithm. ANN with PCA was used to speed the performance. It shows accuracy of 94.5% before applying of PCA but after applying of PCA it gives accuracy of 97.7%. So, a big difference is noticed. [4]. Cardio Vascular Disease was predicted using machine learning algorithms such as Random Forest, Decision tree SVM(support vector machine) and KNN while highest accuracy of 85% was achieved by implementing Random forest machine learning algorithm.[5].

According to study, artificial neural network showed the best accuracy of 84.25 % in contrast to other models and it was found that in spite of other models showed higher accuracy than ANN while this model with lower accuracy was chosen as a final model to make sure the balance between precision and transparency of the model used for predicting the heart disease. [6].Hidden naïve Bayes algorithm can be used to predict heart disease and it achieved 100% with respect to accuracy and dominated naïve Bayes. [7]

Methodology:

The main purpose of the proposed method is to predict the occurrence of heart disease for early detection of the disease in a short time. In our approach, we are using different data mining techniques and machine learning algorithms, Naïve Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest to predict the heart disease based on some health parameters.

Data is analyzed using Anaconda Navigator's jupyter Notebook. It is an open source software where we can implement multiple machine learning algorithms by importing libraries. We can also download the needed libraries by anaconda prompt. It allows us to create live code, perform visualizations, process data and plot graphs.

Dataset for implementation

We have used built in dataset from UCI Machine learning repository for predicting heart disease.

This database contains 14 attributes listed below:

Age

Sex

Cerebral palsey/chest pain (CP)

Blood Pressure(bps) in mm HG

Cholesterol in mg

Fasting blood sugar test (fbs)

resting electrocardiographic results

thalach (Maximum heart rate achieved)

exang (Exercise induced Angina)

oldpeak (ST depression induced by exercise relative to rest)

slope(the slope of the peak exercise ST segment)

ca(Number of major vessels)

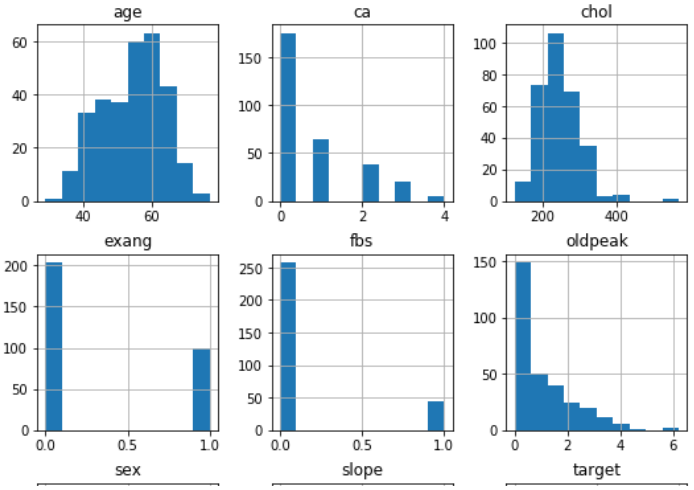
thal(Reversible defect)

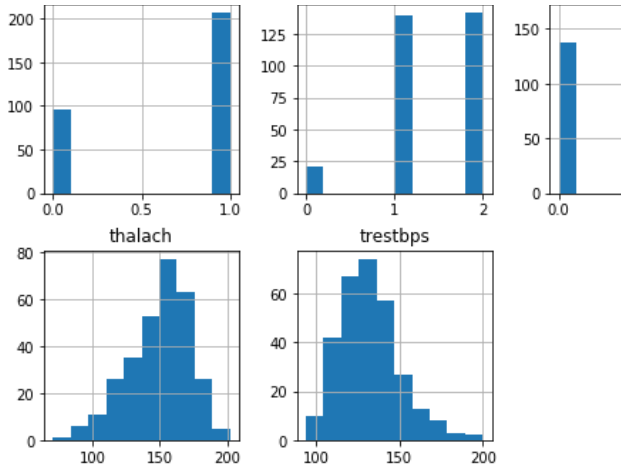
Target(0,1)

Data splitting:

Data is splitted into training and testing data. 25 % data is used for testing purpose while 75 % data is used for training purpose. We performed data normalization for removing Nan values.

Visualization of data:





The performance and the accuracy of each experiment is evaluated by standard metrics such as TP rate, TN rate, precision, recall and F-measure which are calculated by Confusion Matrix which is known as predictive classification table. All these measures will be used to compare the performance of these selected and implemented algorithms.

ALGORITHMS USED FOR EXPERIMENTS

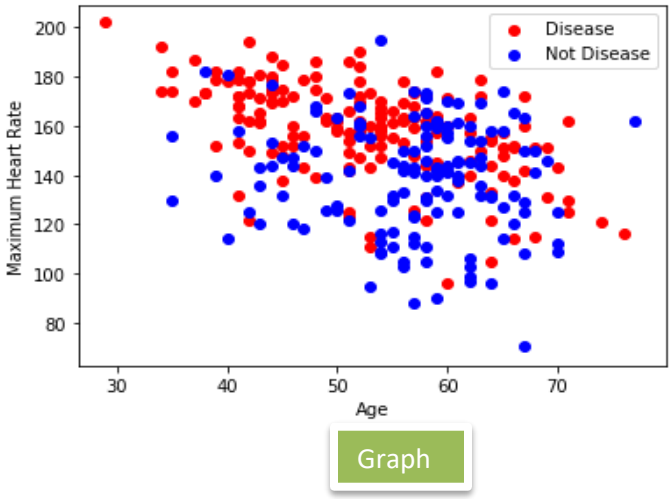
- k Nearest Neighbor (KNN):

KNN is the machine learning algorithm and is most commonly used algorithm .It is preferred when parameters are continuous. In KNN, classification is done by predicting the nearest neighbor .It is preferred over other classification algorithm due to its simplicity and high speed .It can be used to solve both classification and regression problem. The algorithm takes the heart disease data set and classifies whether a person has heart disease or not. KNN captures the idea of by calculating the distance between points on a graph. We used KNN to classify and predict people with heart disease based on parameters such as age, sex etc. It does not need training data for model generation because the training data is used in testing stage. It stores all the cases and then classifies new data according to the nearest neighbor.

KNN has two stages:

1. Find the k number of instances in the dataset
2. Use the k instances to find the nearest neighbor.

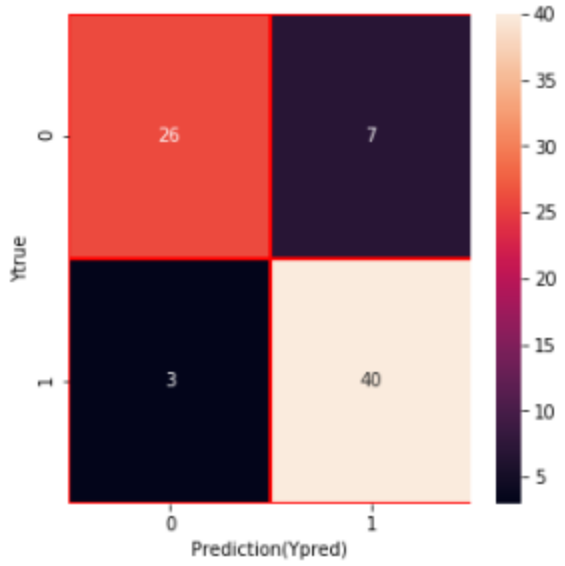
Visualization:



Graph visualizes the patients who have heart disease with red dots while the blue dots represent the patient who are not suffering from heart disease.

Confusion Metrics:

The precision of KNN depends on the distance metric and the K value. It is measured by confusion metrics.



It shows that it has 26 true negative rate, 7 false positive, 3 false negative while true positive cases are 40.

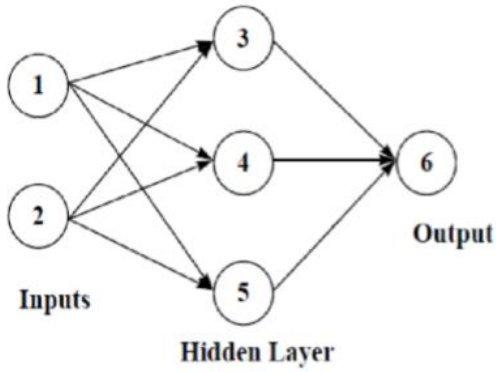
KNN Model Result

	precision	recall	f1-score	support
0	0.90	0.79	0.84	33
1	0.85	0.93	0.89	43
accuracy			0.87	76
macro avg	0.87	0.86	0.86	76
weighted avg	0.87	0.87	0.87	76

We have achieved accuracy of 87 % by implementing this model.

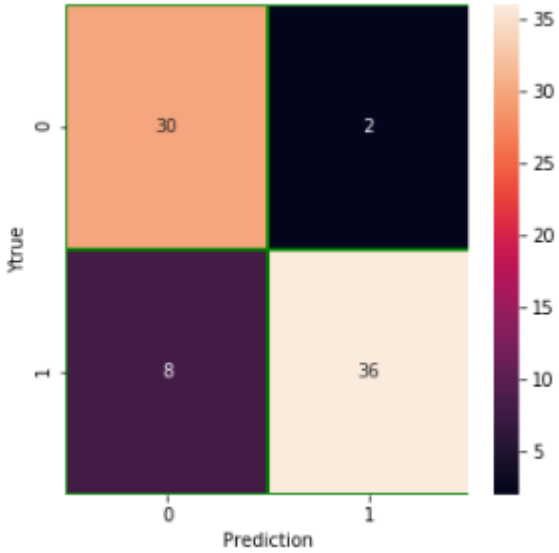
- **Artificial Neural Network**

Artificial Neural Networks are the human neurons type network structure which consists of number of nodes that are connected through directional links where each node represents a processing unit and the links between them designate the casual relation between them [4]. Artificial Neural Networks utilized in clinical decision making and helps the doctors to analyze and make decision efficiently and accurately. A Neural Network start with an input layer where each node of input layer is connected to the nodes of hidden layers and the nodes of hidden layer may connects to an output layer. This classification technique is becoming potent tool in data mining and may be utilized for different purposes in descriptive and predictive data mining [4]. The sample artificial neural network is shown in Figure below.



Artificial Neural Networks (ANN) are components of a computing system designed to simulate, analyses and processes information the way the human brain does. It has self-learning capabilities that enable them to produce better results as more data become available.

Confusion Matrix:



It shows that it has 30 true negative rate, 2 false positive, 8 false negative while true positive cases are 36.

ANN Model Result

	precision	recall	f1-score
0	0.79	0.94	0.86
1	0.95	0.82	0.88
accuracy			0.87
macro avg	0.87	0.88	0.87
weighted avg	0.88	0.87	0.87

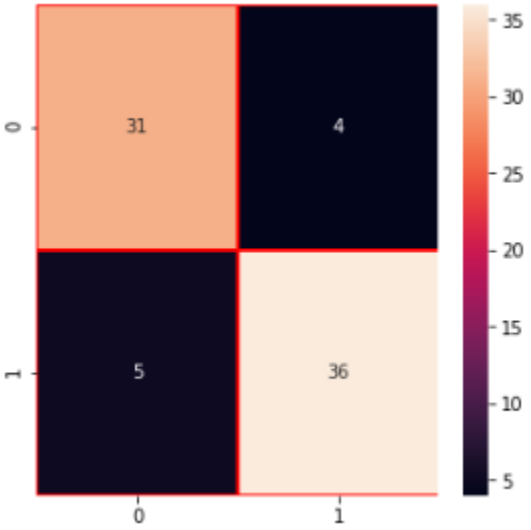
This model gives accuracy of 87 %.

- **Naive Bayes:**

Naive Bayes is used for a classification based on Bayes' theorem. Occurrences of Particular characteristics of a class are independent of the presence or absence of other characteristics according to the naive Bayesian classifier theorem. It is a robust classifier for predicting heart disease. Naive Bayes is used to compute posterior probability of each class based on conditional probability of classifying data sets [4]. The equation is given as follows. $(C|X)$ =P X|C ∗ P(C) P(X)

Where X is the instance to be predicted, and C is the class value for instance the above-given formula or equation helps to determine the class in which feature expected to categorize.

Confusion Matrix:



It shows that it has 31 true negative rate, 4 false positive, 5 false negative while true positive cases are 36.

Naive Bayes Model Result

	precision	recall	f1-score	support
0	0.86	0.89	0.87	35
1	0.90	0.88	0.89	41
accuracy			0.88	76
macro avg	0.88	0.88	0.88	76
weighted avg	0.88	0.88	0.88	76

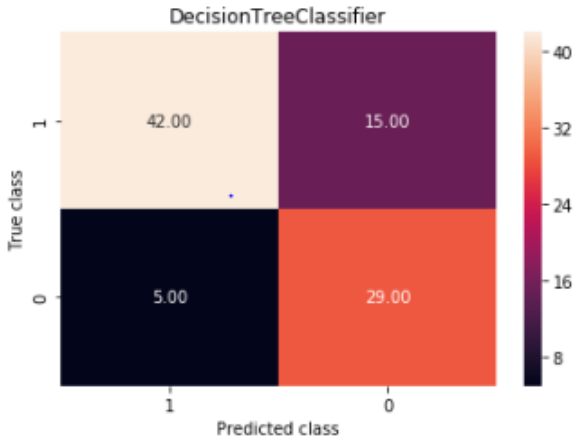
This model gives accuracy of 88 %.

- **Decision Tree**

A decision tree is a type of a supervised learning algorithm classifier, which is easy to understand. They deal with numerical and categorical data. The decision tree resembles the tree structure consisting of internal nodes, branches, and leaf nodes in which each branch represents the values of a given data set, internal nodes Tests on a given attribute and the Leaf nodes show the class to predict or indicate the results of the result. The classification rule starts from the root node to the leaf nodes, depending on the predictive attribute and the given rules.

Commonly used decision tree algorithms are CART, ID3, C4.5, J48 and CHAID are very important in the prediction of diseases [3].

Confusion Matrix:



It shows that it has 42 true negative rate, 15 false positive, 5 false negative while true positive cases are 29.

Decision Tree Result

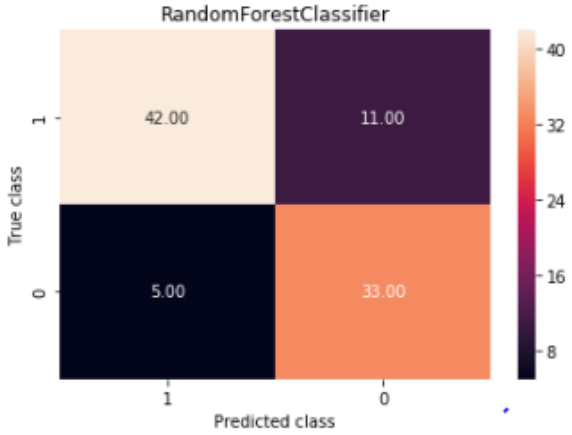
	precision	recall	f1-score	support
0	0.85	0.66	0.74	44
1	0.74	0.89	0.81	47
accuracy			0.78	91
macro avg	0.79	0.78	0.78	91
weighted avg	0.79	0.78	0.78	91

This model gives accuracy of 78 %.

- **Random forest**

Random forest is also a type of supervised learning. It can be used both for classification and regression. It is also the most flexible and user friendly algorithm. A forest is made up of trees. It is said that the more trees it has, the more robust a forest is. Random forests create decision trees on randomly selected data samples, obtain predictions from each tree, and select the best solution by voting.

Confusion Matrix:



This matrix shows that it has 42 true negative rate, 11 false positive, 5 false negative while true positive cases are 33.

Random Forest Result:

	precision	recall	f1-score	support
0	0.87	0.75	0.80	44
1	0.79	0.89	0.84	47
accuracy			0.82	91
macro avg	0.83	0.82	0.82	91
weighted avg	0.83	0.82	0.82	91

We have achieved accuracy of 82 % as shown in classification report.

DISCUSSION:

The focus of our study was on using data mining techniques in healthcare for heart disease.
We performed some experiments on our data set of heart disease by applying five data mining algorithms. Through implementation of different classification algorithms we try to find out that which algorithm is best in predicting heart disease .And which one gives the best accuracy. There are five experiments we performed and these experiments are designed for the same purpose, the purpose is to compare the

results of KNN, Neural Networks, Decision Tree, and Naive Bayes and Random Forest.

COMPARISON OF IMPLEMENTED ALGORITHMS:

The purpose of our study was to use machine learning algorithms for heart disease in healthcare. So for this we performed experiment by using different algorithms on heart disease patients. Through implementation we can know which classification algorithm is best for predicting heart disease.

After the implementation of different algorithms the second step is the comparison between different machine learning algorithms used in these experiments and choose the best one which gives most accuracy. In order to do comparison of these experiments different performance measures are used for example, Accuracy True Positive, False Positive, False Negative, True Negative and ROC Curve is used.. Summary of algorithms is shown in the following table.

Table 1.1:

Algorithms	Accuracy	TN	FP	FN	TP
KNN	0.87	26	7	3	40
ANN	0.87	30	2	8	36
Naïve Bayes	0.88	31	4	5	36
Decision Tree	0.78	42	15	5	29
Random Forest	0.82	42	11	5	33

The above table 1.1show that the best accuracy on the given dataset is of 88% and lowest accuracy of 78%. The Naïve Bayes has the highest accuracy while the Decision Tree has the lowest accuracy.

By observing other performance measures that are used for result too. TP rate of KNN is 40, ANN is 36, Naïve Bayes is 36, Decision tree is 29, and Random forest is 33. This shows that the KNN has the highest TP rate and Decision Tree has the lowest TP rate. Similarly Decision tree has the highest FP rate of 15 and the ANN has the lowest FP rate of 2.

Based on the above comparison it can be seen that Naïve Bayes, KNN and ANN are good as they have nearly same accuracy and they have the best TP rate and KNN has the FP rate of 7 followed by naïve Bayes with FP rate of 4 and ANN with FP rate of 2.

As we all know that Heart Disease is a sensitive and critical disease which causes millions of death. Due to this we need to keep TP rate high and FP rate less. If the diagnosis of disease is correct and also done earlier then it is best to cure patients of suffering from that disease. So it is expected from algorithms to perform well. Accuracy is also matters to identify heart patients.

Conclusion:

Our study mainly focused on the use of data mining techniques in healthcare especially in detection of heart disease. Heart disease is a fatal disease which may cause death. Data mining techniques were implemented using the following algorithm, KNN, Neural Networks, Decision Tree, and Naive Bayes and Random Forest. We measured performance on the basis of Accuracy, TN, FP, FN and TP rate and in some algorithm.

We conducted five experiments with the same data set to predict heart disease. The result of all the implemented algorithm are

shown in tabular form for better understanding and comparisons .The experiment shows that Naive Bayes gives the highest accuracy which is 88% followed by ANN and KNN with accuracy of 87% . Our findings indicate that data mining can be used and applied in the healthcare industry to predict and diagnose the disease at early stages,

Future work:

Further research should be done to increase classification accuracy through the use of advanced algorithms such as Bagging, Vector Machine Support or table decision etc. Determine the performance of the predictions per algorithm and apply the proposed system to the area of interest. We can add more features needed to improve accuracy implementation of algorithms. Stakeholders should use it as dedicated tool to make better decisions. We did not change parameters in our implementation. In future, it can be improved and adjust by changing the parameters for the experiment.

In the future, more work can be done by using more data related to heart disease and by using different data reduction techniques. For better results and predictions of heart disease, high quality oriented datasets can be used which are free from inconsistencies.

References:

[1] Ujma Ansari, Jyoti Soni, Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", 258493784_Predictive_Data_Mining_for_Medical_Diagnosis_An_Overview_of_Heart_Disease_Prediction. March 2011Data Mining in Healthcare for Heart Diseases

[2] C. Beyene, P. Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", https://www.researchgate.net/publication/323277772_Survey_on_prediction_and_analysis_the_occurrence_of_heart_disease_using_data_mining_techniques, 118(8):165-173 · January 2018

[3] Muhammad Usama Riaz, SHAHID MEHMOOD AWAN, ABDUL GHAFFAR KHAN, "PREDICTION OF HEART DISEASE USING ARTIFICIAL NEURAL NETWORK", https://www.researchgate.net/publication/328630348_PREDICTION_OF_HEART_DISEASE_USING_ARTIFICIAL_NEURAL_NETWORK. October 2018

[4] Umair Shafique, Irfan Ul Mustafa, Haseeb Qaiser, Fiaz Majeed, "Data Mining in Healthcare for Heart Diseases",https://www.researchgate.net/publication/274718934_Data_Mining_in_Healthcare_for_Heart_Diseases. March 2015.

[5] Komal Kumar Napa, G.Sarika Sindhu, D.Krishna Prashanthi, A.Shaeen Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers", https://www.researchgate.net/publication/340885231_Analysis_and_Prediction_of_Cardio_Vascular_Disease_using_Machine_Learning_Classifiers, April 2020.

[6] Hossam Meshref, “Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach”,https://www.researchgate.net/publication/338428682_Cardiovascular_Disease_Diagnosis_A_Machine_Learning_Interpretation_Approach, January 2019.

[7] Jabbar Akhil, Shirina Samreen, “Heart disease prediction system based on hidden naïve Bayes classifier”, https://www.researchgate.net/publication/309735105_Heart_disease_prediction_system_based_on_hidden_naive_Bayes_classifier, October 2016.