International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Research on Comparative Analysis of Classification Algorithms for Heart Disease

Prof. Poonam Pathekar (Professor), Yash Surana(Student), Prajit Thube(Student),

Shraddha Solapure(Student), Ruchika Jadhav(Student)

Poonamjaikar@gmail.com, yash.surana.zeal@gmail.com, prajitthube@gmail.com, shraddhasolapure@gmail.com, ruchikajadhav2432002@gmail.com

Computer Department Zeal Collage of Engineering and Research, Pune, Maharashtra, India

Abstract : Heart disease, also referred to as cardiovascular disease, is the leading cause of death globally over the past few decades. It includes a variety of disorders that have an impact on the heart. Numerous risk factors for heart disease are linked to the requirement for timely access to accurate, trustworthy, and practical methods for early diagnosis and disease management. Data mining is a popular method for processing vast amounts of data in the healthcare industry. In order to forecast cardiac disease, researchers evaluate vast amounts of complex medical data using a variety of data mining and machine learning approaches. This study proposes a model based on supervised learning methods such as decision trees, and numerous heart disease-related variables. K is for closest neighbour, The random forest algorithm and the support vector machine. It makes use of the current dataset from the UCI heart disease patient repository's Cleveland database. There are 303 instances and 76 attributes in the collection. Only 14 of these 76 attributes—which are crucial to demonstrating the effectiveness of various algorithms—are tested. The purpose of this study work is to estimate the likelihood that patients will develop heart disease.

Keywords: Heart disease,* SVM; Decision Tree; Random Forest;, *Data Mining, KNN

I. INTRODUCTION

Over the past decade, heart disease, or cardiovascular disease, remains the leading cause of death worldwide. According to the World Health Organization, more than 17.9 million people worldwide die from cardiovascular diseases every year, and 80% of these deaths are caused by coronary artery disease and stroke [1]. High mortality rates are common in low- and middle-income countries [2]. Many predisposing factors, such as personal and occupational habits and genetic predisposition, contribute to heart disease. Several common risk factors contribute to heart disease, such as smoking, excessive alcohol and caffeine consumption, stress and physical inactivity, as well as other physiological factors such as obesity, hypertension, high blood cholesterol, and existing heart disease. Effective, accurate and early medical diagnosis of heart diseases plays a crucial role in the implementation of measures to prevent death. Data mining examines vast data sets to extract important decision-making information hidden from past archival collections for future analysis. The medical field contains a huge amount of patient data. This data must be mined by various machine learning algorithms. Healthcare professionals analyze this information to arrive at an effective diagnostic decision for healthcare professionals. Medical data mining with classification algorithms provides clinical assistance through analysis. It tests classification algorithms to predict heart disease in patients. Data mining is the process of extracting valuable information from huge databases. Various data mining techniques such as regression, clustering, association rule, and classification techniques such as decision tree, random forest, and K-nearest neighbors, SVM are used in heart disease prediction to classify various heart disease attributes. Comparative analysis is used for classification techniques. In this research I took dataset from kaggle. A classification model was developed to predict heart disease using classification algorithms. In this article, the algorithms used to predict heart diseases are discussed, and the existing systems are compared.

Heart disease is one of the biggest problems today due to health and lifestyle choices. Our main goal in the project is to see the possibilities of heart diseases and the most important factors affecting it. The heart pumps blood to all parts of our body. If it does not work properly, the brain and many other organs stop working and eventually the person dies. According to the World Health Organization, heart disease kills 17.7 million people every year, which is 31 percent of all deaths worldwide. Heart disease was also the leading cause of death in India. Heart disease claimed 1.7 million lives in India in 2016, according to the Global Burden of Disease Study released on September 15, 2017. According to the WHO, India lost $237 billion due to heart-related or cardiovascular diseases between 2005 and 2015. Therefore, accurate and feasible prognosis of heart disease is essential. Medical organizations around the world collect information on various heart-related problems. However, these datasets are too large for the human mind and can be easily explored using various machine learning techniques. Several machine learning algorithms have recently proven to be very useful in reliably predicting the presence or absence of heart disease. Early prediction of heart

disease would save lives. To achieve this goal, several risk factors related to heart disease must be considered and anticipated. To study these variables.

II.LITERATUE REVIEW

1. "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, and Parneet Singh, Volume 2021, Article ID 8387680, July 2021. To compare the findings and analysis of the UCI Machine Learning Heart Disease dataset, various machine learning methods including deep learning are employed.
2. Santhana Krishnan J. et al. suggested utilising decision trees and the Naive Bayes algorithm in their study "Prediction of Heart Disease Using Machine Learning Algorithms" to predict heart disease. The decision tree algorithm builds the tree based on specific circumstances that result in true or false choices. The outcomes of algorithms like SVM and KNN are based on split conditions that can be vertical or horizontal, depending on the dependent variablesUsing some techniques, the data set is divided into 70% training and 30% testing. The accuracy of this method is 91%. Naive Bayes, the second algorithm, is used for categorization. Since it can handle complex, nonlinear, and dependent data, the heart disease dataset—which is similarly complex, dependent, and nonlinear in nature—is seen as a good fit. 87% accuracy is provided by this method.
3. "A Survey on Prediction Techniques of Heart Disease Using Machine Learning" by Mangesh Limbitote and Pushkar Patil June 2020, Volume 9, Issue 6, ISSN 2278-0181 thorough examination of pertinent medical methods to forecast heart disease. accuracy was 91.3% with random forest and 82.30% with SVM.
4. Rishabh Khera, Sartak Agrawal, and Harshit Jindal June 2020, "Heart disease prediction using machine learning algorithms." Using the patient's medical history, a system has been developed to predict whether or not the patient will be diagnosed with a cardiac condition.
5. "Using machine learning for heart disease prediction" by Dhai Eddine Salhi, Abdelkamel Tari, and M-Tahar Kechadi, published in February 2021. used data analytics to find and anticipate the sufferers of diseases. Three data analytics techniques were used on data sets of various sizes after a pretreatment stage in which the most pertinent attributes were chosen.

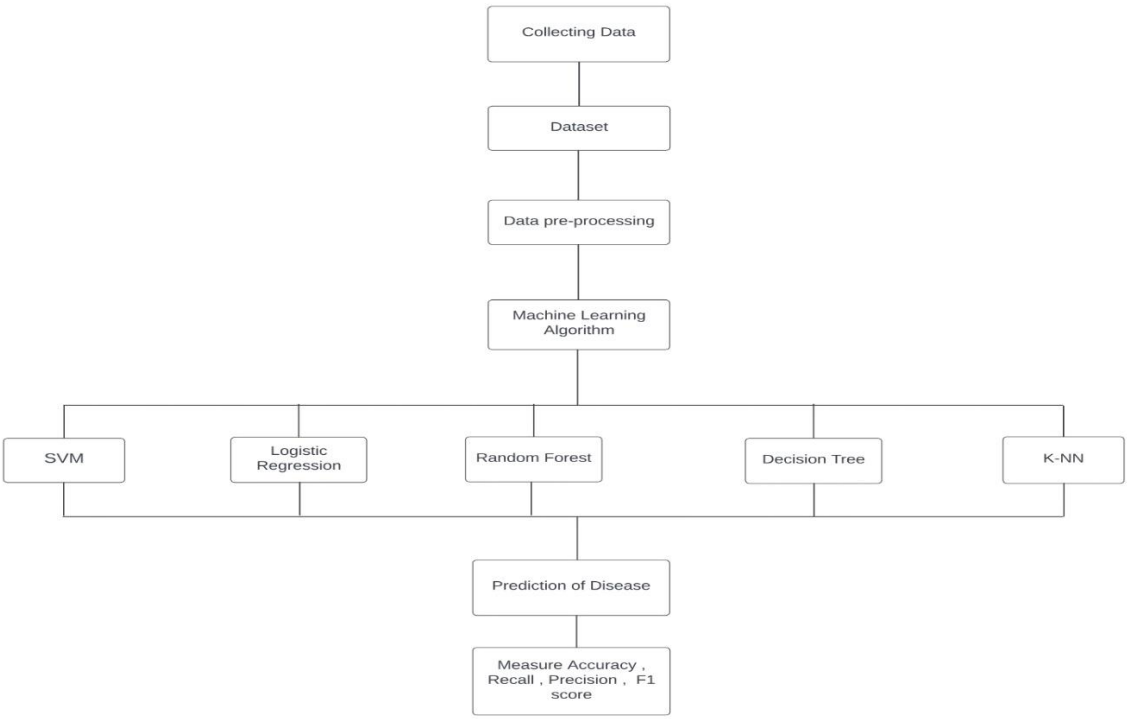
III.PROPOSED SYSTEM

A. Problem Description

The goal is to create a machine learning system for heart disease prediction. The system should be able to extract facts and trends that can aid in making objective health decisions. The user should be able to enter specific patient-related data so that an accurate prediction can be created. Any non-technical individual should be able to use the user interface if it is simple and straight forward.

B. Methodology

The first step in the processed approach is to obtain the data for this download from Kaggle, which has been thoroughly vetted by academics. There are numerous steps in this process, as depicted in the block diagram.



C. Description of the Dataset:
Information about the data set

The Public Health Dataset, which was used for this study, was created in 1988 and comprises four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It has 76 properties, including the one that was anticipated; however, all published experiments only mention using a portion of 14 of these. The "target" field alludes to the patient's having a heart illness. 0 means there is no disease, while 1 means there is a disease. Table 1 displays the first four rows and all dataset features without any preprocessing. Now the characteristics that are used in this research are discussed, along with their uses or similarities:

(i) Age (patient's age in years; sex (male: 1; female: 0).
(ii) Type C chest discomfort
(iii) Trestbps, or resting blood pressure, measured at the time of hospital admission (in mm Hg). If your blood pressure is within the typical range of 120/80, everything is great; however, if it is slightly higher than it should be, you should endeavor to bring it down. Make positive lifestyle adjustments.
(iv) Chol—serum cholesterol reveals the concentration of triglycerides. Another lipid that can be detected in the blood is triglycerides. It should be less than 170 mg/dL (however, this can vary amongst labs).
(v) Fasting blood sugar (FBS) greater than 120 mg/dl (v) (1 true) Less than 100 mg/dL (5.6 mmol/L) and between 100 and 125 mg/dL (5.6 and 6.9 mmol/L) are considered normal.
(vi) Restecg—results from a resting electrocardiogram
(vii) Thalach reached his highest heart rate. Your age less 220 determines your maximum heart rate. Exercise-induced angina

(viii) 1 yes for this. Reduced blood supply to the heart is the cause of angina, a particular type of chest discomfort. A sign of coronary artery disease is angina.
(ix) Exercise-induced Oldpeak-ST depression as compared to rest
(x) Target (T)—angiographic disease state; no illness = 0, disease = 1, etc.
(xi) Slope: the steepness of the exercise's ST segment peak.
(xii) Ca: fluoroscopically colored main vascular count (0–3)
(xiii) unknown but very likely thalassemia (three normal cells, six fixed defects, and seven reversible defects).

Table 1 Attributes and details of dataset of heart disease

Sr. no.	Attribute	Representative icon	Details
1	Age	Age	Patients age, in years
2	Sex	Sex	0 = female; 1 = male
3	Chest pain	Cp	4 types of chest pain (1—typical angina; 2—atypical angina; 3—non-anginal pain; 4—asymptomatic)
4	Rest blood pressure	Trestbps	Resting systolic blood pressure (in mm Hg on admission to the hospital)
5	Serum cholesterol	Chol	Serum cholesterol in mg/dl
6	Fasting blood sugar	Fbs	Fasting blood sugar > 120 mg/dl (0—false; 1—true)
7	Rest electrocardiograph	Restecg	0—normal; 1—having ST-T wave abnormality; 2—left ventricular hypertrophy
8	MaxHeart rate	Thalch	Maximum heart rate achieved
9	Exercise-induced angina	Exang	Exercise-induced angina (0—no; 1—yes)
10	ST depression	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope	slope of the peak exercise ST segment (1—upsloping; 2—flat; 3—down sloping)
12	No. of vessels	Ca	No. of major vessels (0–3) colored by fluoroscopy
13	Thalassemia	Thal	Defect types; 3—normal; 6—fixed defect; 7—reversible defect
14	Num(class attribute)	Class	diagnosis of heart disease status (0—nil risk; 1—low risk; 2—potential risk; 3—high risk; 4—very high risk)

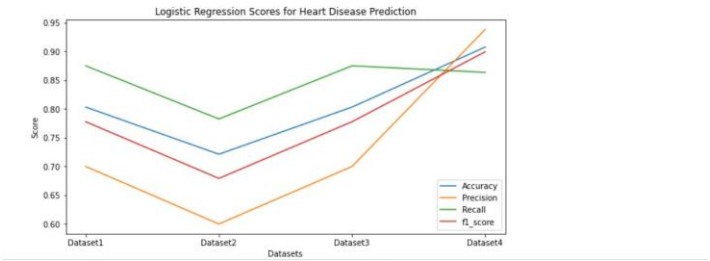
D. Data Pre-processing:
The data from real-world situations contains a lot of numbers that are missing or have noisy data. To get around these problems and create strong predictions, these data have already been processed. The sequential chart of our suggested model is shown in Figure 1. Data cleaning typically includes noise and missing values. These data need to be cleansed of noise and the missing values filled in in order to produce an accurate and useful result. Transformation is the process of converting data from one format to another so that it is easier to understand. It entails responsibilities for aggregation, normalization, and smoothing. Integration The data must be integrated before processing because it may come from multiple sources rather than just one. Reduction The data acquired is intricate and needs to be structured to produce useful results. Following classification, the data are divided into training and test sets, which are subjected to a variety of algorithms to produce accuracy score results.

IV.MODLES

A. *Logistic Regression,*

Assigning observations to a discrete set of classes is done using the classification process known as logistic regression. Logistic regression changes its output using the logistic sigmoid function to generate a probability value that may then be translated to two or more discrete classes, in contrast to linear regression, which produces continuous number values.

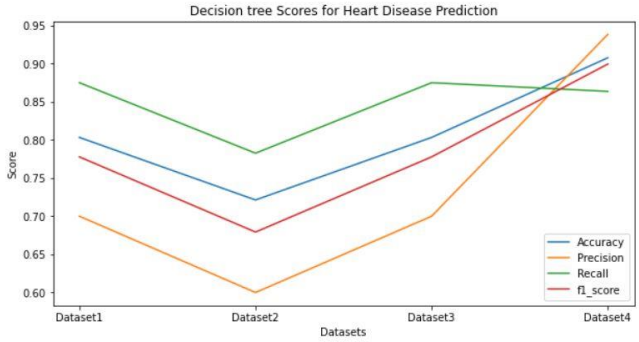
```
datasets = ['Dataset1', 'Dataset2', 'Dataset3', 'Dataset4']
accuracies = [accuracy1, accuracy2, accuracy3 , accuracy4]
precisions = [precision1, precision2, precision3, precision4]
recalls = [recall1, recall2, recall3, recall4]
f1_score =[f1_score1,f1_score2,f1_score3, f1_score4]

# Plot the scores
plt.figure(figsize=(10, 5))
plt.plot(datasets, accuracies, label='Accuracy')
plt.plot(datasets, precisions, label='Precision')
plt.plot(datasets, recalls, label='Recall')
plt.plot(datasets,f1_score,label='f1_score')
plt.xlabel('Datasets')
plt.ylabel('Score')
plt.title('Logistic Regression Scores for Heart Disease Prediction')
plt.legend()
plt.show()
```



B. Decision Tree[DT]

DT is an algorithm that, in spite of arithmetic data, categories parameters. DT produces a structure that resembles a tree. Due to DT's simplicity, it has been used to analyze a number of huge medical data sets. The analysis is based on tree nodes. The leaf node Denote the completion of each test. Node inside: handle a variety of elements Root Node: The main node Based on the primary node, other nodes operate. Applying this algorithm will divide the data into two or more parallel sets. Then, each parameter's entropy is determined. Next, divide the data by a predict or with a high information gain or one with a low entropy.

Gain (S, A) = Entropy (S1.vs.yValues (Alues(A) |Sv| |S| Entropy Entropy (Sropy(S) = c i=1 Pi log2 Pi.

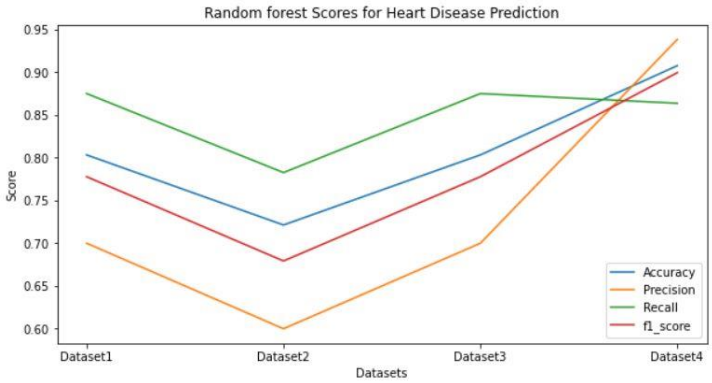


C. Random forest [RF]

The RF method relies heavily on supervised learning. In many different domains, it serves as a classifier. Using more trees in this way creates a forest. The accuracy increases as the number of trees increases. Additionally, regression tasks employ it. However, it does so effectively when the task is classified. and might outweigh misguided values. Three RF approaches exist: Forest RC (random blend), Forest RI (random input), and RC and RI together.

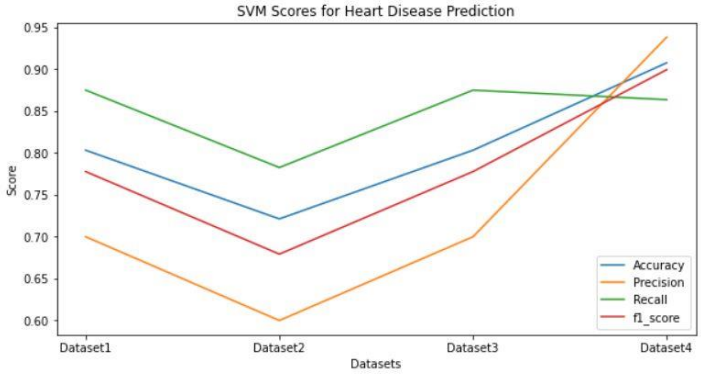
Regression using logs (LR):

The supervised ML learning approach is called LR. As can be seen in Fig. 5, the relationship between the dependent and independent variables is formed. This method is known as linear regression since the relationship between variables "a" and "b" is indicated by an equation of a line, which is linear in nature.



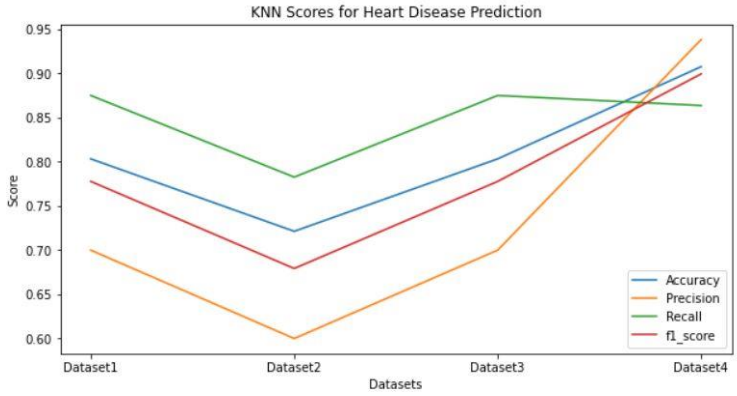
D. Support Vector Machine

SVM is one kind of ML technique that focuses on the idea of a hyperplane. This data point can be used to classify a hyperplane, especially in n-dimensional space [13]. (Xa, Ya) is a training sample for a data set with the variables a = 1, 2, 3, etc., with Ya serving as the target vector and Xa serving as the it vector. When choosing a support vector for a hyperplane, there are many different options available. For instance, if a line is used, the method is known as a linear support vector.



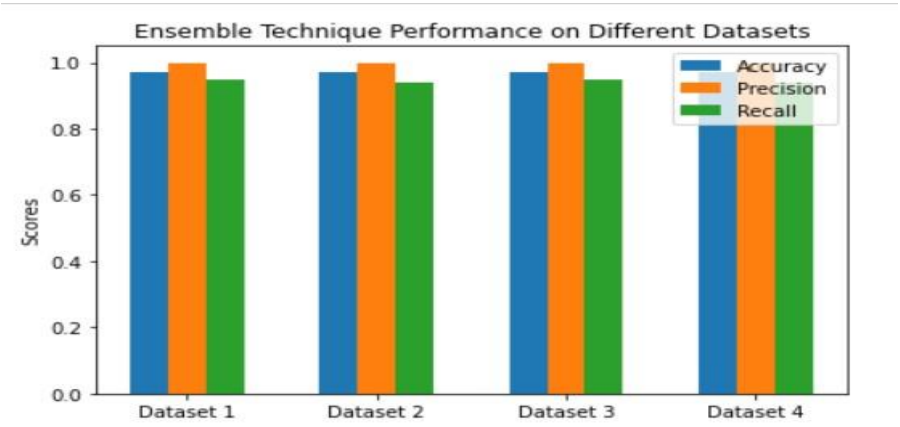
E. K-Nearest Neighbors

K-Nearest Neighbours fared poorly in this method because no training was done prior to testing; instead, KNN classifies test data directly from the dataset. Tree of Decisions (ID3): It provided a range and turned the continuous value's data into categorical values during the training phase. The performance of the classifier was impacted and, as a result, it forecast the incorrect class label when the test data pattern comprised values outside of this specified range.



B. RESULT AND ANANLYSIS

This paper compares the performances of the classification algorithms in the prediction of heart disease. It tries to find out the best classifier for this task. In the experimental dataset, 13 attributes were used. But all the attributes are not equally emphasized for detecting heart disease. For this reason, a feature selection method was presented that removes the irrelevant attributes which are not highly correlated with the other features used for classification. Each classification algorithm gives a noticeable performance while using the selected 13 attributes in the prediction of heart disease. Among the studied classifiers, Logistic Regression performs better than other classification algorithms. Binary class problem is solved to identify whether the patient has heart disease or not. It is recommended to solve the multiclass problem for detecting heart disease by dividing heart disease patients into various classes.



C. CONCLUSION

This study offers a thorough understanding of machine learning methods for categorizing cardiac disorders. In order to forecast the treatment that can be given to patients, classifiers play a key role in the healthcare business. In order to identify effective and precise methods, the existing methodologies are examined and contrasted. Machine learning approaches dramatically increase the accuracy of cardiovascular risk prediction, allowing for the early diagnosis of patients who can then receive preventative care.

Conclusion: Machine learning algorithms have enormous potential for predicting heart-related or cardiovascular disorders. The purpose of this study is to forecast a patient's risk of developing heart disease. This study used K-nearest neighbor, decision trees, random forests, and support vector machines to classify data from the UCI repository using supervised machine learning techniques. Through the WEKA tool, numerous experiments employing various classification algorithms were carried out. Research was conducted using an 8th-generation Intel Corei7 with 16 GB of RAM and an 8750H processor that can run at 4.1 GHz. A training set and a test set were created from the categorized data set. The data are pre-processed, and to obtain an accuracy score, supervised classification methods such as Naive Bayes, decision trees, K-nearest neighbors, and random forests are used.

Using Python programming, the accuracy score results of several classification algorithms were reported for the training and test data sets.

D. REFERENCES

1. "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, and Parneet Singh,
2. Volume 2021, Article ID 8387680, July 2021.
3. "Heart Disease Prediction Using Machine Learning Techniques", Devansh Shah, Samir Patel, and Santosh Kumar Bharti, October 16, 2020
4. "A Survey on Prediction Techniques of Heart Disease Using Machine Learning," by Mangesh Limbitote and Pushkar Patil, Vol. 9, Issue 6, June 2020, ISSN 2278-0181.
5. Rishabh Khera, Sartak Agrawal, and Harshit Jindal June 2020, "Heart disease prediction using machine learning algorithms."
6. "Using machine learning for heart disease prediction" by Hai Eddine Salhi, Abdelkamel Tari, and M-Tahar Kechadi was published in February 2021.
7. Banana.U. Rindhe, Manisha Darade, Nikita Ahire, and Rupali Patil, Volume 5, Issue 1, May 20: "Heart Disease Prediction Using Machine Learning."
8. Survey on Prediction and Analysis of the Occurrence of Heart Disease Using Data Mining Techniques, P. K. Chala Beyene, 2018. 165–174 are included in the International Journal of Pure and Applied Mathematics, 118. Gupta Y, Ahmed R. K. A., and Kautish S. K. (2017), "Application of data mining and knowledge management with special reference to medical informatics," in International Journal of Medical Laboratory Research.
9. Jyoti Arora and Amandeep Kaur (2019) The study is titled "Heart Disease Prediction Using Data Mining Techniques." IJARCS is an acronym for International Journal of Advanced Research in Computer Science.