IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Special Issue 2, December 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

# Heart Disease Prediction Using Effective Machine Learning Techniques

**Dr.M.Deepamalar[1], Mrs.R.Padma[2], C. Deepika[3]**

Assistant Professor, Paravathy's Arts and Science College, Dindigul, TamilNadu, India[1]

Associate Professor, Paravathy's Arts and Science College, Dindigul, TamilNadu, India[2]

PG Scholar, Paravathy's Arts and Science College, Dindigul, TamilNadu, India[3]

**ABSTRACT:** In today's era heart disease is the main reason for deaths. WHO-World Health Organization has foreshadow that 12 million people die every year because of heart diseases. Some heart diseases are cardiovascular, heart attack, coronary and knock. Knock is sort of heart disease that occurs due to strengthening, blocking or lessening of blood vessels in which drive through the brain or it can be initiated by high blood pressure(HP). This considering both male and female category and the ratio may vary according to the region, this ratio is considered for the people of age group 25-69. This does not designate that the people with other age group will not be affected by heart diseases. This problem may also start in early age group also and predict the cause and disease is a major challenge now a days. Here, I have discussed various algorithms and tools used for prediction of heart diseases. K Neighbors Classifier(KNN), Support Vector Classifier (SVM), Decision Tree Classifier(DTC), Random Forest Classifier(RF) algorithm.

**KEYWORDS:** Classification, Heart Disease, Machine Learning.

## I. INTRODUCTION

The heart doesn't function properly, that will distress the other parts of the human body such as brain, kidney etc. Heart disease is a kind of disease which effects the functioning of the heart. Now a days heart disease is the primary reason for deaths. WHO-World Health Organization has anticipated that 12 million people death every year because of heart diseases. Some heart diseases are cardiovascular, heart attack, coronary and knock. Knock is a sort of heart disease that occurs due to strengthening, blocking or lessening of blood vessels, which drive through the brain or it can also be initiated by high blood pressure. The major challenge that the Healthcare industry faces is superiority of facility. Diagnosing the disease correctly and providing effective treatment to patients will define the quality of service. Poor diagnosis causes disastrous consequences that are not accepted. Records or data of medical history is very large, but these are from many dissimilar foundations. The interpretations that are done by physicians are essential components of these data. The data in real world might be noisy, incomplete and inconsistent, so data preprocessing will be required in directive to fill the omitted values in the database. Even if cardiovascular diseases is found as the important source of death in world in ancient years, these have been announced as the most avoidable and manageable diseases. The whole and accurate management of a disease rest on the well-timed judgment of that disease.

**Data Set**

| | |
|---|---|
| **Age** | age in years |
| **Sex** | (1 = male; 0 = female) |
| **Cp** | chest pain type |

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

| | |
|---|---|
| **Trestbps** | resting blood pressure (in mm Hg on admission to the hospital) |
| **Chol** | serum cholestoral in mg/dl |
| **Fbs** | (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) |
| **Restecg** | resting electrocardiographic results |
| **Thalach** | maximum heart rate achieved |
| **Exang** | exercise induced angina (1 = yes; 0 = no) |
| **Old peak** | ST depression induced by exercise relativeto rest |
| **Slope** | the slope of the peak exercise ST segment |
| **Ca** | number of major vessels (0-3) colored by flourosopy |
| **Thal** | 3 = normal; 6 = fixed defect; 7 = reversable defect |
| **Target** | 1 or 0 |

**Data Preprocessing:**

The dataset obtained is not completely accurate and error free. Hence, I will carry out the following operations on it.

**Data Cleaning:**

NA values in the dataset are the major setback for us as it will reduce the accuracy of the prediction profoundly so,I will remove the fields which do not have values. I will substitute it with the mean value of the column. This way, I will remove all the values in the dataset.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

**Feature Scaling:**

The range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without feature scaling. The range of all features should be scaled so that each feature contributes approximately proportionately to the final distance. I will scale the various fields in order to get them closer in terms of values. Ex: Age has just two values i.e. 0,1 and cholesterol has high values like 100. In order to get them closer to each other I will need to scale them.

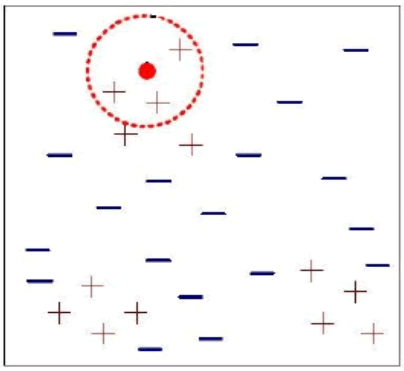
**Factorization**:

This process assigned a meaning to the values so that the algorithm doesn't confuse between them. Ex: Assigning meaning to 0 and 1 in the age section, the algorithm doesn't consider 1 as greater than 0 in that section.

**K – Nearest Neighbors**

In 1951, Hodges et al. introduced a nonparametric technique for pattern classification which are popularly known the K-Nearest Neighbour rule. K-Nearest Neighbour technique is one of the most elementary and very effective classification techniques. K-Nearest neighbor (KNN) is a simple, lazy and nonparametric classifier and preferred when all the features are continuous. KNN makes no assumptions about data and is generally be used for classification tasks when there is very less or no prior knowledge about the data distribution.

KNN algorithm involves finding the k nearest data points in the training set to the data point for which a target value is unavailable and assigning the average value of the found data points to it. KNN gives an accuracy of 83.16% when the value of k is equal to 9 while using 10-cross validation technique. In KNN with Ant Colony Optimization performs better than other techniques with an accuracy of 70.26% and the error rates is 0.526. Ridhi Saini et al. has obtained a efficiency of 87.5% , which is very good. KNN is also called as case-based reasoning and had been used in many applications like pattern recognition, statistical estimation. Classification is obtained by identifying the nearest neighbor to determine the class of an unknown sample.



**KNN Classification**

1) Find the k number of instances in the dataset that is closest to instances
2) These k number of instances then vote to determine the class of instances

The Accuracy of KNN depends on distance metric and K value

**Support Vector Machine**

Support Vector Machine is an extremely popular supervised machine learning technique (having a pre-defined target variable) which can be used as classifier as well as a predictor. For classification, it finds a hyper-plane in the feature space that differentiates between the classes. SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide possible. The test data points are then mapped into that same space and are classified based on which side of the margin they fall. Shan Xu et al. have used SVM to achieve an accuracy of 98.9% in People's Hospital dataset. SVM performs the best with 85.7665% of correctly classified instance and in SVM are used with boosting technique to give an accuracy of 84.81%. HoudaMezriguiet al.has used SVM to attain measure value of 93.5517. In SVM classifies the pixel variation with an accuracy of 92.1% helping to identify the affected regionaccurately.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

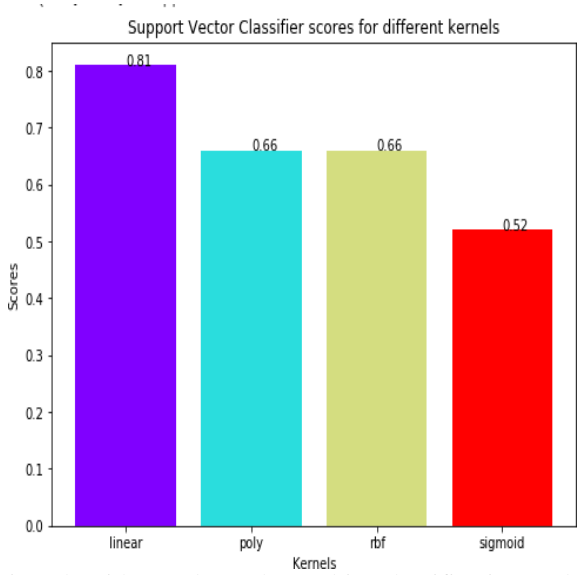
Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Support vector machine (SVM) is supervised learning method that analyze data used for classification and regression analysis. It is given a set of training data, marked as belonging to either one of two categories, an SVM training algorithm then builds model that assigns new examples to one category or the other, making it a no probabilistic binary linear classifier.

SVM model representation of the examples for points in space, mapped so that the examples separate categories are divided by a clear gap thatis as wide as possible. The points are separated based on hyper plane is separated them. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.

**Decision Tree**

**Decision Tree**




Decision tree is a of supervised learning algorithm and mostly used in classification problems. It performs effortlessly with continuous and categorical attributes. This algorithm divides the population into two or more similar sets based on the most significant predictors.

Decision Tree algorithm, first calculates the entropy of each and every attribute. Then dataset is break with the help of the variables or predictors with maximum information gain or minimum entropy. These two steps are performed recursively with the remaining attributes. In decision tree has the worst performance with an accuracy of 77.56% but when decision tree is used with boosting technique it performs better with an accuracy of 82.18%. In decision tree performs very poorly with a correctly classified percentage of 42.8964% in also uses the same dataset but usedthe J48 algorithm for implementing Decision Trees and the accuracy thus obtained is 67.7% is less but still an improvement on the former. Renu Chauhan et al.have obtained an accuracy of 71.42% . M.A. Jabbar et al. have used alternating decision trees with principle component analysis to obtain an accuracy 92.2%. Kamran Farooq et al. have achieved the best results on using decision tree-based classifier combined with forward selection achieves weighted accuracy of 78.4603%. The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree and applying the tree to the dataset. There are many popular decision tree algorithms. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility and accuracy. The information gain is based on the decrease in entropy after a dataset is split attribute. A decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one.

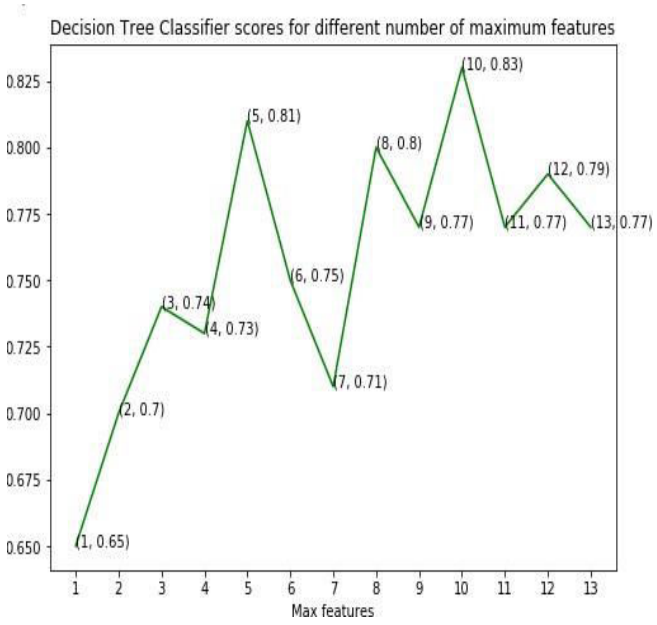
|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India




**Random Forest**

Random Forest is popularly supervised machine learning algorithm. RF can be used for both regression and classification tasks but generally performs better in classification tasks. Random Forest technique considers multiple decision trees before giving an output, and basically an ensemble of decision trees. This technique is based on the belief that more number of trees would converge to the right decision.It uses a voting system and decides the class whereas in regression it takes the mean of all the outputs of each of the decision trees. It works well with large datasets with high dimensionality and Random forest performs exceptionally well. In Cleveland dataset, random forest had a significantly higher accuracy of 91.6% than all the other methods. In People's Hospital dataset, it achieves an accuracy of 97%. In Random forest has achieved measure of0.86.

Random forest is used to predict coronary heart disease and it obtains an accuracy of 97.6%. Random forest is an ensemble classifier which combines bagging and random selection of features and it can handle data without preprocessing.

It has been used in prediction and probability estimation. This consists of many decision trees and outputs the class, which is the mode of individual trees class. It is one of the most accurate classifier. It produces a highly accurate classification for many data sets. Especially for heart disease dataset.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

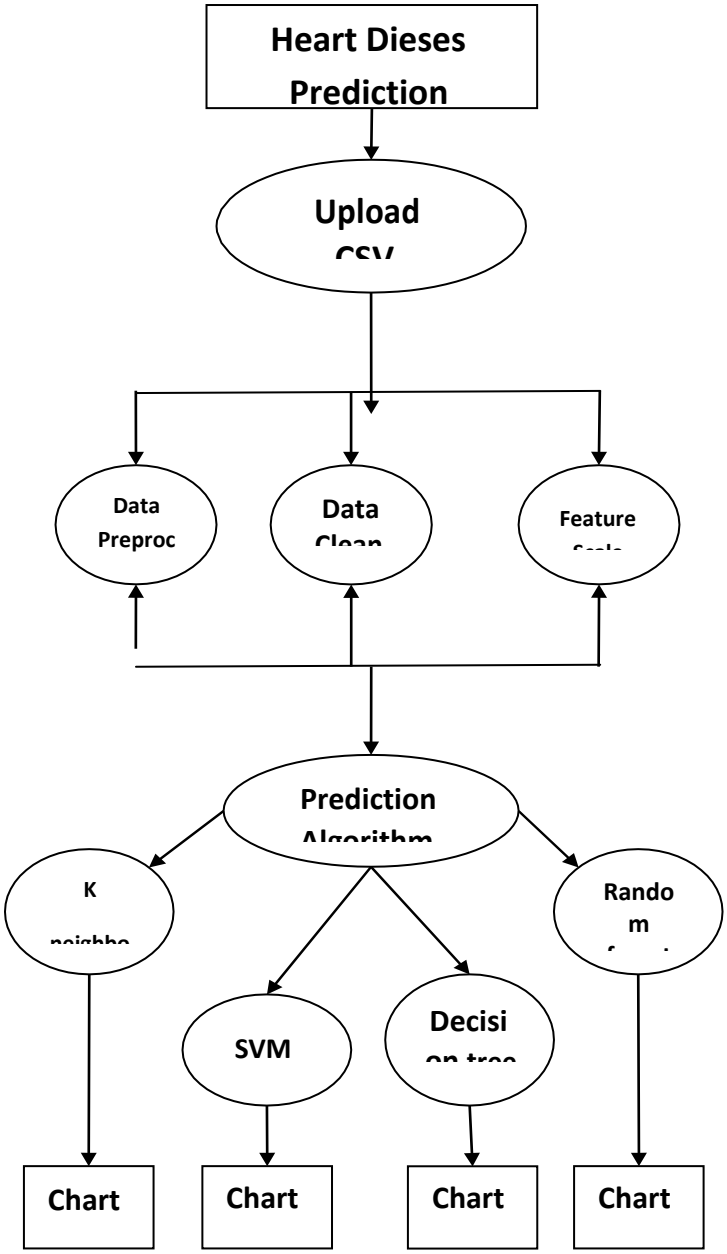
4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

**Implementation Diagram**




## II. 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. Ma-chine learning techniques used in this work to process raw data and provide a new and novel discernment towards heartdisease.

Heart disease prediction is challenging and very important in the medical field. The mortality rate can drastically controlled if the disease is detected at the early stages and preventative measures is adopted. Further extension of study is highly desirable to direct the investigations in real-world datasets instead of just theoretical approaches. I have studied various classification algorithms that can be used for classification of heart disease databases. In future, structure can be further upgraded by creating various combinations.

The models are trained and validated against a test dataset. Lift Chart and Classification Matrix methods is used to

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

evaluate the effectiveness of the models. Models are able to extract patterns in response to the predictable state. The most effective model to predict patients with heart disease appears to be Random Forest Classifier, with its own strength with respect to case of model interpretation, access to detailed information and accuracy.

## REFERENCES

[1]. V. Krishnaiah, G. Narasimha, N. Subhash Chandra, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review" IJCA 2016.
[2]. ]K.Sudhakar, Dr. M. Manimekalai "Study of Heart Disease Prediction using Data Mining", IJARCSSE 2016.
[3]. NagannaChetty, Kunwar Singh Vaisla, NagammaPatil, "An Improved Method for Disease Prediction using Fuzzy Approach", ACCE 2015.
[4]. VikasChaurasia, Saurabh Pal, "Early Prediction of Heart disease using Data mining Techniques", Caribbean journal of Science and Technology,2013
[5]. ShusakuTsumoto," Problems with Mining Medical Data", 0-7695- 0792-1 I00@ 2000IEEE.
[6]. Y. AlpAslandoganet. al.," Evidence Combination in Medical Data Mining", Proceedings of the international conference on Information Technology: Coding and Computing (ITCC'04) 0-7695-2108- 8/04©2004IEEE.
[7]. Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.
[8]. Franck Le Duff, CristianMunteanu, Marc Cuggiaa, Philippe Mabob,"Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, page no. 1256-1259,2004.
[9]. Boleslaw Szymanski, Long Han, Mark Embrechts, Alexander Ross, KarstenSternickel,Lijuan Zhu, "Using Efficient Supanova Kernel For Heart Disease Diagnosis", Proc. ANNIE 06, intelligent engineering systems through artificial neural networks, vol. 16,page no. 305-310,2006.
[10]. Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer 2006,Vol:345, page no. 721- 727.
[11] Ramadoss and Shah B et al."A. Responding to the threat of chronic diseases in India". Lancet. 2005; 366:1744–1749. doi:10.1016/S0140-6736(05)67343-6.
[12] Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011
[13] homse Kanchan B and MahaleKishor M. et al. "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis", 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication.
[14] R.Kavitha and E.Kannan et al."An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Techniquein Data Mining ", 2016.
[15] Shan Xu ,Tiangang Zhu, Zhen Zang, DaoxianWang,JunfengHuandXiaohui Duan et al. "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework", 2017 IEEE 2nd International Conference on Big Data Analysis.
[16] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis and Vijay K. Mago et al. " Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map", 978-1- 5090-0626-7/16/$31.00 c 2016IEEE.
[17] Kanika Pahwa and Ravinder Kumar et al. "Prediction of Heart Disease Using Hybrid Technique For Selecting Features", 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics(UPCON).
[18] SeyedaminPouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. " A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth2017
[19] Hanen Bouali and JalelAkaichi et al. "Comparative study of Different classification techniques, heart Diseases use Case.", 2014 13th International Conference on Machine Learning andApplications
[20] SeyedaminPouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. " A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth2017.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com



Scan to save the contact details