IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SURVEY ON HEART DISEASE PREDICTION USING MACHINE LEARNING

[1] Yuvraj Nikhate, [2] M. V. Jonnalagedda,

[1]Student, [2]Associate Professor,

[1] Department of Information Technology,

[1] SGGS IE & T, Nanded, India

Abstract: Heart Disease is the one of major causes of death globally. Around 17.9 million people die each year. Cardiovascular diseases include disorders of the heart and blood vessels. Four out of five cardiovascular disease deaths are due to heart attacks. One-third of these deaths occur prematurely under the age of seventy. The major number of deaths have occurred in developing countries. India is one of them.

For heart disease diagnosis we need cardiologists, which are in limited number in developing countries. Also, the tests for cardiovascular diseases are quite expensive; sometimes out of the budget for common people. Early detection is important in case of heart disease with less expensive prediction techniques. As we know, now-a-days Machine Learning algorithms are used for predicting various diseases. They are also used for predicting Heart Disease. This paper deals with the survey of Machine Learning algorithms used for predicting heart disease, the importance of attributes to predict the disease and selection of important attributes for predication.

Index Terms **- Heart Disease, important attributes, Linear Regression, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, 10-fold cross-validation supervise machine learning algorithms, Feature selection.**

I. INTRODUCTION

Machine learning (ML) is the subdomain of artificial intelligence (AI). Today we are using ML in day to day life**.** ML based computer programs can access data and use it to learn themselves. It means past experience is used for prediction in ML. ML algorithms are of four types: Supervised Learning in which direct supervision is involved developer label the dataset restricts the boundaries of algorithm, Unsupervised Learning supervision is not required, semi supervised machine learning both type supervised and unsupervised used in combine format and Reinforcement Learning exploration of thing one by one fist event take as input for next event. In this paper the focus is on supervised machine learning algorithms.

In supervised machine learning algorithms, as the name indicate there is presence of a supervisor who train the machine for prediction. In other words we train the machine with the help of a labeled dataset. Labelled dataset is the one which is tagged with the correct answer or class which can be labeled after predicted by machine. In our prediction system, we are predicting the labelled data into two categories: having a heart disease or not having a heart disease. Heart disease dataset from UCI machine learning repository is used. This dataset contain 76 various attributes and 303 instances. Attribute selection is an important factor for the accuracy of result as more relevant data can predict results accurately. Selection of attributes from the dataset needs proper domain knowledge that can help select fewer attributes to predict results accurately. Out of the 76 attributes present in the Cleveland Heart disease dataset, 14 attributes are selected for prediction

In Cleveland Heart disease dataset for six instances some values are missing. The remaining 297 instances have complete values. Hence most of the researchers use the 297 instance values for prediction of heart disease. For machine learning size of the data set is an important point for the accuracy of the model. Bigger size dataset with data without noise and data with no missing values are usefully reliable. But we face problem if the dataset size is small. For better accuracy, more pattern in the dataset should be explored. The cross-validation is used for it**.**

For any supervise Machine Learning algorithm dataset is divided in two-parts: first one is training part and the second one is the testing part. Commonly training size is 70% and testing size is 30% but for special analysis size may vary. In general training using larger amount of data and testing against a small amount of data account for accuracy of a ML algorithm**.**

As shown in Fig. 2, the first two steps are importing the dataset and pre-processing the data. In pre-processing of the data we remove or replace the missing values in the dataset before applying it to a supervised machine learning algorithm then the cross-validation technique is applied.

The attributes selected from the dataset [10] are very simple and are regularly taken during the medical examination of a patient. This is done mainly to keep the cost of heart disease prediction low. Age and sex both are the personal information of the patient. Blood pressure and the chest pain can be examined by doctor without any additional requirement of the test. For attribute such as fasting blood sugar, blood test is required and for examining attributes such as slope and old peak ST, ECG is needed. So maximum attributes used in the system are obtained by normal testing of patient. No expensive tests are needed for predication, hence the system is cost-effective.

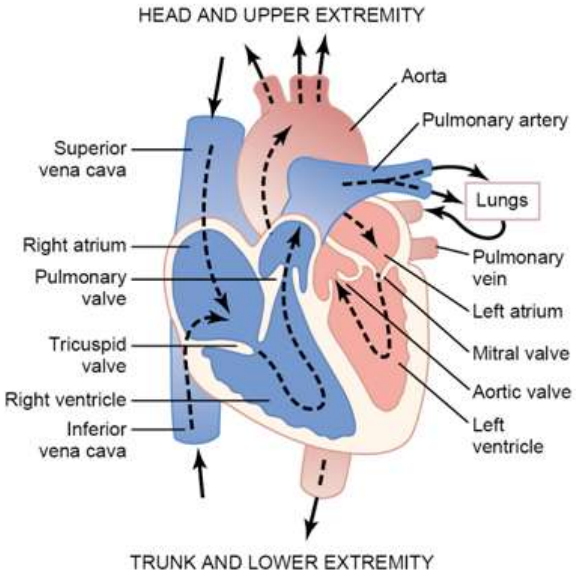

Fig. 1: Human heart

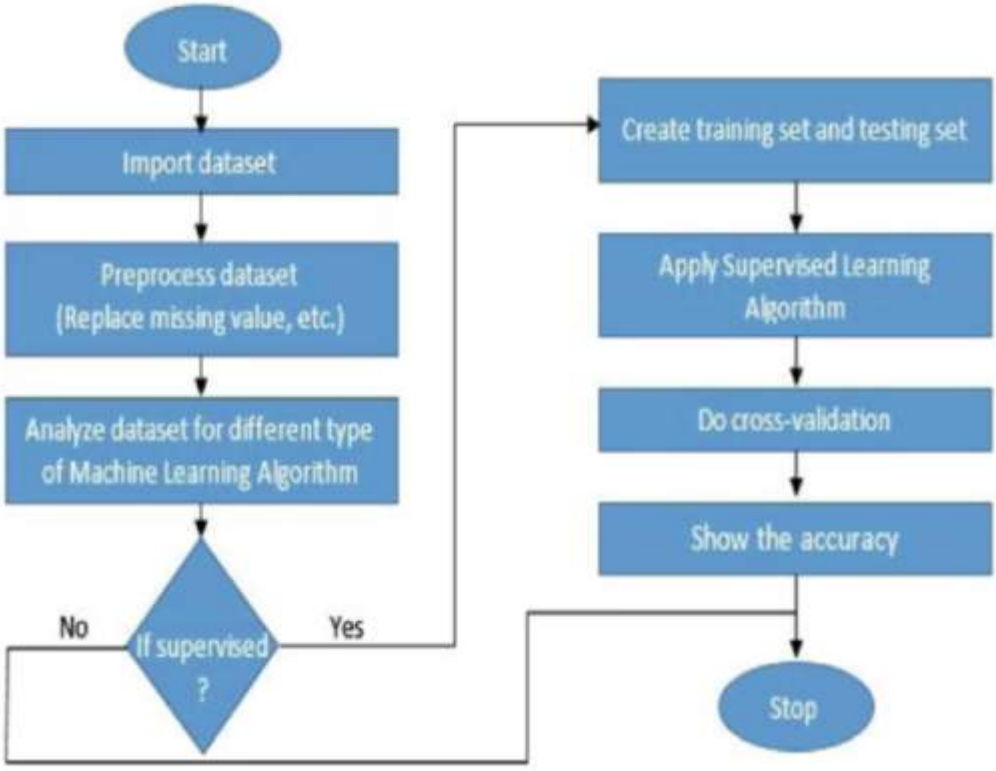

Fig. 2: Flow chart of ML system with cross validation

In Cleveland Heart disease dataset for six instances some values are missing. The remaining 297 instances have complete values. Hence most of the researchers use the 297 instance values for prediction of heart disease. For machine learning size of the data set is an important point for the accuracy of the model. Bigger size dataset with data without noise and data with no missing values are usefully reliable. But we face problem if the dataset size is small. For better accuracy, more pattern in the dataset should be explored. The cross-validation is used for it.

For any supervise Machine Learning algorithm dataset is divided in two-parts: first one is training part and the second one is the testing part. Commonly training size is 70% and testing size is 30% but for special analysis size may vary. In general training using larger amount of data and testing against a small amount of data account for accuracy of a ML algorithm.

As shown in Fig. 2, the first two steps are importing the dataset and pre-processing the data. In pre-processing of the data they remove or replace the missing values in the dataset before applying it to a supervised machine learning algorithm then the cross-validation technique is applied.

The attributes selected from the dataset [10] are very simple and are regularly taken during the medical examination of a patient. This is done mainly to keep the cost of heart disease prediction low. Age and sex both are the personal information of the patient. Blood pressure and the chest pain can be examined by doctor without any additional requirement of the test. For attribute such as fasting blood sugar,

blood test is required and for examining attributes such as slope and old peak ST, ECG is needed. So maximum attributes used in the system are obtained by normal testing of patient. No expensive tests are needed for predication, hence the system is cost-effective.

Table 1: Dataset description [10]

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain type 1 = typical angina, 2 = a typical angina, 3 = non-anginalpain, 4 = asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg / dl:1 – True, 0 = False
Exang Maximum	Continuous	heart rate achieved Discrete Exercise induced angina: 1 = Yes, 0 = No
Thalach	Continuous	Maximum heart rate achieved
Old peak ST	Continuous	Old peak ST Depression induced by exercise relative to rest
Ca	Continuous	Number of major vessels coloured by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal, 6 = fixed defect, 7= reversible defect
Slope	Discrete	The slope of the peak Exercise segment : 1 = up sloping, 2 = flat, 3 = down sloping
Restecg	Discrete	Resting electrocardiographic results Value 0: normal Value 1: having ST-T wave abnormality
Class (Target)	Discrete	classes: 0 = No Presence, 1=Presence of heart disease

Predication with supervised machine learning is done by dividing the dataset in two sets: independent and dependent attributes. First set of attributes, the independent attributes are the ones whose value does not depend on the value of other attributes. In the dataset, first 13 attributes are independent attributes. The last one, which is commonly called as a class or the target attribute is a dependable attribute. Its value depends on the other independent attributes. In our case target is the last attribute which talks about having a heart disease or not having a heart disease.

2. LITERATURE SURVEY

The literature survey consists of first four papers surveyed from medical background. They are studied to establish the importance of various features and their typical values. Being from medical background, no machine learning algorithms is applied for the heart disease detection and hence are not included in the survey table (Table 3). The literature survey after these four papers is from engineering background where different features are considered and various machine learning algorithms are applied for automatic detection of presence of heart disease. The numbers in the square bracket below corresponds to the numbers mentioned in the 'References'.

[1] In this paper the authors talk about Atherosclerotic Cardiovascular disease (CVD) which is age dependent. Atherosclerotic CVD starts at a very young age and progresses over time. Age is the traditional non-modifiable factor in heart disease. The risk of heart disease is high during the lifetime, but, at the age of 70 years it gets reduced as compared to the risk at the age of 50 years. It also depends on other risk factors that remain unchanged in the life such as smoking and drinking a lot of alcohol.

Age group 70 and above have a shorter period of time left to develop the disease due to lower burden on cardiovascular system and due to their genetic makeup. It indicates that heart disease is a function of age and the risk is lower at the age of 70 and above but higher at age group 32 – 62.

[2] In this, the authors deal with the impact of gender on heart disease basically comparison between male and female. Heart disease develops in the female 7 to 10 years late than the male which implies that male can get the heart disease earlier than female. But still the major reason for death in female above 65 years is heart disease. Past twenty years the heart disease midlife (35 to 54 years) for women has increased.

Why women are more protected to heart disease than men? The answer is 'Estrogen' hormone. It is a hormone secreted by ovaries. Estrogen regulates the metabolic activity of lipids, inflammatory markers and the coagulant system. Effect of the vasodilatory action is reduced risk of atherosclerosis. Early menopause increases the risk of heart disease and it is observed that the early menopause women have two years less life than the normal or late menopause women.

Smoking at an early age less than 50 years affects more in women than men and it also increases acute myocardial infarction in females than in men.

After menopause normal body weight increase in the first few years. Increased weight and increased changes of diabetes type 2 it is observed that female with diabetes are 50% at more risk to heart disease than men.

During menopause total cholesterol and low-density lipoprotein (LDL) levels rise 10 to 14% and HDL remain the same. Systolic blood pressure increases rapidly in menopaused women as compared to the same age of men.

[3] In this paper author's deal with two risk factors blood pressure and cholesterol. Let's first consider blood pressure. Normal and normal high consider as normal for a person but the normal high person has to take care of himself as the normal high may get converted to hypertension in near future. The Hypertension stages are from one to four and are given in table 2 below.

As we have already seen above, age act as a major factor in heart disease as with others such as high BP, high cholesterol, chest pain, high sugar level. Bad habits such as drinking a lot of alcohol and smoking, unhealthy lifestyle, lack of body activity results in more risk of heart disease. The habits such as smoking and drinking cause the thickness and stiffness of the vessel resulting into major problem in the circulation of blood and have adverse effects on heart.

Table 2: Blood pressure

SR.	Person	Systolic, mm Hg	Diastolic, mm Hg
1	Normal	130	85
2	High normal	130 -139	85- 89
3	Hypertension stage I	140 - 159	90 – 99
4	Hypertension stage II –IV	≥160	≥100

The second factor is cholesterol. Cholesterol is of two types: HDL and LDL, both together is called total Cholesterol. For normal person 200mg/dl is considered normal. For borderline it is in the range 200 to 239 mg/dl more than this value is considered as the high value of Cholesterol. High value means a high risk of heart disease. High-density lipoprotein (HDL) cholesterol is also known as good cholesterol because it helps to remove other forms of cholesterol from serum. A higher level of HDL means a low risk of heart disease. For men the value of HDL less than 40mg/dl and for women less than 50mg/dl indicates the risk of heart disease. Normal value for men and women is min 60mg/dl. More than 60mg/dl indicates less risk of heart disease obviously.

Low-density lipoprotein (LDL) is also called bad cholesterol because it gets deposited on the walls of blood vessels. Due to this, the blood vessels get narrow and blood pressure increases on the walls of the blood vessels. Indirectly LDL increase the risk of heart disease.

LDL normal level is considered between 100 to 129 mg/dl for a normal person. From 130 to 159 mg/dl is called borderline high. More than that called to higher 160 to 189 mg/dl and very high 190 and above. Both higher and very higher are considered to be a risk for heart disease.

So from [2] and [3] we can conclude that gender, cholesterol levels and blood pressure values play an important role in detection of heart disease and hence should be considered as features for the detection system.

[4] In this paper authors deal with ST-segment depression during treadmill electrocardiography (ECG). For this study, 150 number of subjects were selected with a low likelihood of coronary heart disease (out of hundred are normal and 50 with non-anginal chest pain) and another group 150 subjects was selected with a high likelihood of coronary heart disease.

The total 300 subjects are then divided into four groups and are made to do the treadmill exercise.

First group is of 100 normal subjects i.e. with low likelihood of coronary heart disease. There were 81 men and 19 women with an age between 47-60 years. It was observed that there was no chest pain during treadmill exercise.

Second group subjects are with non-anginal chest pain. They are 50 in number with 33 men and 17 female with age ranging between 47-60 years. It was observed that even there was no chest pain in these subjects during treadmill exercise.

Group 3 was with subjects having history of clinical angina. There were 50 subjects with 31 men and 19 female having age range 61-71 years. All the subjects in this group developed typical chest pain during treadmill exercise.

Fourth group subjects were with catheterization-proved coronary disease. They are 100 in number with 84 male and 16 female having age group between 58-67 years.

For all groups ECG was taken. The sensitivity of an ST segment/heart rate slope partition of 2.4 μV/beats/min was 95% and the sensitivity of a ΔST segment/heart rate index partition of 1.6μV/beats/min was 91%. Analysis of ΔST segment/heart rate index and ST

segment depression can markedly improve the clinical usefulness of the treadmill exercise and ECG. It can be concluded in general that if the ST slop < 2.4, then the person is not having a heart disease and if ST slope > 2.4 then the person can have a heart disease.

[5] In this paper authors used two algorithms: hill climbing and decision tree. Before applying the classification algorithms, the data is pre-processed. The data set used is Cleveland data set. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an open-source data mining tool which is used to fill the missing values in the data set. Hill climbing algorithm is then used to find the best subset of rules. The parameters and their values used are - Confidence: minimum confidence value is 0.25, MinItemsets: the minimum number of item-sets per leaf is two, Threshold: a value of 10 is used to find the best subset of rules for the hill-climbing algorithm by the authors.

A decision tree is constructed in the top-down approach for each level and a node is selected by a test for the actual node chosen using a hill-climbing algorithm. The decision tree knows how to generate the basic rules. First generated rules are referred to as original rules and from these rules Pruned Rules are generated. From these rules without duplicates and classified rules are obtained. And finally a small number of rules called Class wise Rule distribution is generated. The accuracy of the system is about 86.7%.

[6] In this paper author has implemented hybrid machine learning for heart disease prediction. The data set used is Cleveland data set.
The first step is data pre-processing step. In this the tuples are removed from the data set which have missing the values. Attributes age and sex from data set are also not used as the authors think that it's personal information and has no impact on predication. The remaining 11 attributes are considered important as they contain vital clinical records. They have proposed own Hybrid Random Forest Linear Method (HRFLM) which is combination of Random Forest (RF) and Linear method (LM).

In HRFLM algorithm the authors have used four algorithms. First algorithm deals with partitioning the input dataset. It is based on decision tree which is executed for each sample of the dataset. After identifying the feature space, the dataset is split into the leaf nodes. Output of first algorithm is Partition of data set. After that in second algorithm they apply rules to the data set and output here is the classification of data with those rules. In third algorithm features are extracted using Less Error Classifier. This algorithm deals with finding the minimum and maximum error rate from the classifier. Output of this algorithm is the features with classified attributes. In forth algorithm they apply Classifier which is hybrid method based on the error rate on the Extracted Features. Finally they have compared the results obtained after applying HRFLM with other classification algorithms such as decision tree and support vector machine.

In result as RF and LM are giving better results than other, both the algorithms are put together and new unique algorithm HRFLM is created. The accuracy of HRFLM initially increased with number splits and then has become constant at a particular level. The accuracy obtained is 88.7% which higher than the SVM and decision tree. The authors suggest further improvement in accuracy by using combination of various machine learning algorithms and also by concentrating on developing novel feature selection techniques which would help in extracting significant features.

[7] In this paper, the authors propose a system containing two models based on linear Support Vector Machine (SVM). The first one is called L1 regularized and the second one is called L2 regularized. First model is used for removing unnecessary features by making coefficient of those features zero. The second model is used for prediction. Predication of disease is done in this part. To optimize both models they proposed a hybrid grid search algorithm. This algorithm optimizes two models based on metrics: accuracy, sensitivity, septicity, the Matthews correlation coefficient, ROC chart and area under the curve.

They used Cleveland data set. Data splits into 70% training and 30% testing used holdout validation. There are two experiments carried out and each experiment is carried out for various values of C1, C2 and k where C1 is hyperparameter of L1 regularized model, C2 is hyperparameter of L2 regularized model and k is the size of selected subset of features. First experiment is L1-linear SVM model stacked with L2-linear SVM model which is giving maximum testing accuracy of 91.11% and training accuracy of 84.05% for value of C1 = 0.200, k = 11 and C2 = 0.500. The second experiment is L1-linear SVM model cascaded with L2-linear SVM model with RBF kernel. This is giving maximum testing accuracy of 92.22% and training accuracy of 85.02% for value of C1 = 0.060, k = 8 and C2 = 400.00 and G (Hyperparameter of the L2-SVM model with RBF kernel) = 0.015. They have obtained an improvement in accuracy over conventional SVM models by 3.3%.

[8] In this paper authors deal with various supervised machine learning algorithms such as Random Forest, Support Vector Machine, Logistic Regression, Linear Regression, Decision Tree with 3 fold, 5 fold and 10 fold cross-validation techniques.
They have used Cleveland data set having 303 tuples, with some tuples having missing attributes. In the preprocessing of data they just removed the missing value tuple from the data set which are six in number and then from the remaining 297 tuples, they divided the data as training 70% and testing 30%.

First algorithm applied is Linear Regression. In this, they have defined the dependency of one attribute over others which can be linearly separated from each other. Basically the classification takes place with the help of the group of attributes used for binary classification. They have obtained best results in 10 fold which is 83.82%. Logistic regression classification is done using a sigmoid function. This algorithm applied for heart disease prediction shows maximum accuracy with 3 and 5 fold cross-validation and it is 83.83%. Support Vector Machine is the classification algorithm in supervised machine learning. In this the classification is done by hyperplane. The maximum accuracy achieved by SVM in 3 fold cross-validation is 83.17%.

For Decision Tree in this paper, the authors have used different number splits and different number of leaf nodes to find the maximum accuracy. With 37 number splits and 6 leaf nodes maximum accuracy is achieved which is 79.12%. When used with cross-validation, accuracy achieved by the decision tree 79.54% with 5 fold. Random forest algorithm used on nonlinear data set gives better results as compared to the decision tree. Random forest is the group of decision tree created by the different root nodes. From this group of decision tree, voting can be done first and then classification can be done from the one getting maximum votes**.** Authors have used different number splits, different number of tree per observation and different number of folds for cross-validation. For random forest, 85.81% accuracy is achieved by 20 Number of splits, 75 Number of trees and 10 number of folds.

[9] In this paper authors deal with machine learning algorithms such as decision tree and Naive Bayes algorithm for prediction of heart disease. In first algorithm the decision tree is built using certain conditions which gives True or False decisions. Other algorithms like SVM, KNN are results based on vertical or horizontal split conditions depends on dependent variables. But decision tree for a tree like structure having root node, leaves and branches base on the decision made in each of tree Decision tree also help in the understating the importance of the attributes in the dataset.

They have also used Cleveland data set. Dataset splits in 70% training and 30% testing. This algorithm gives a 91% accuracy.

The second algorithm is Naive Bayes. It is used for classification. It can handle complicated, nonlinear, dependent data and hence is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy.

Table 3: Survey Table

Author	Method	Description	Accuracy
[5]Purushottam et al	Efficient Heart Disease Prediction System	Classification rules generated by Decision tree algorithm	86.7%
[6]Senthilkumar Mohan et al	Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	Combining the characteristics of Random Forest (RF) and Linear Method (LM).	88.7%
[7]Liaqat Ali et al	An Optimized Stacked SVM Based Expert system for the Effective Prediction of Heart Failure	In this method two models SVM model is linear and *L*1 regularized and SVM is L2 regularized	92.22%
[8]Singh, Yeshvendra K. et al	Heart Disease Prediction System Using Random Forest	Random Forest (RF) with cross-validation	85.81%
[9]Santhana Krishnan. J et al	Prediction of Heart Disease Using Machine Learning Algorithms	Two algorithms are used separately: Decision tree & Naive Bayes algorithm	Decision tree = 91% Naive Bayes =87%

3. CONCLUSION

In this paper, we surveyed on two things: the first part of the study is finding important factors affecting the heart disease, the important attributes and their minimum support values for no heart disease. Then use of machine learning algorithm for prediction of heart disease. From risk factors we have selected number of attributes and their minimum value for normal and diseased person. Values of the attributes more than the minimum value means you have a risk of heart disease. That means a person who have diabetes, smoking habit last 10 years, hypertension, chest pain, ST-T depression, older age more than 60 years, women with early menopause, over dirking, higher values of cholesterol more than 200mg/dl and person with blockage in heart vessels are more likely to be the heart disease person.

Second part of the studies of supervised machine algorithms for prediction of heart disease. In this various algorithms studied are Support vector machine, Decision tree, Random forest, Linear regression and Naive Bayes classifier. Lots of work is done in this area. The dataset is quite old and has no new attributes added in it. There is no cleaning and pruning of data. Uncleaned and missing values in the dataset has no use for classification and prediction. Moreover, no one has worked on the size of the dataset. The small size of the data set is a problem for machine learning algorithms. Large size of the dataset is needed for better prediction.

REFERENCES

[1] Dhingra, Ravi, and Ramachandran S. Vasan, "Biomarkers in cardiovascular disease: Statistical assessment and section on key novel heart failure biomarkers." *Trends in cardiovascular medicine* 27.2 (2017): 123-133.

[2] Maas, Angela HEM, and Yolande EA Appelman. "Gender differences in coronary heart disease." *Netherlands Heart Journal* 18.12 (2010): 598-603

[3] Wilson, Peter WF, et al, "Prediction of coronary heart disease using risk factor categories." *Circulation* 97.18 (1998): 1837-1847.

[4] Kligfield, Paul, Olivier Ameisen, and P. M. Okin. "Heart rate adjustment of ST segment depression for improved detection of coronary artery disease." *Circulation* 79.2 (1989): 245-255.

[5] Purushottam, Kanak Saxena and Richa Sharma, "Efficient heart disease prediction system." *Procedia Computer Science* 85 (2016): 962-969.

[6] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques." *IEEE Access* 7 (2019): 81542-81554.

[7] Ali, Liaqat, et al, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure." *IEEE Access* 7 (2019): 54007-54014.

[8] Singh, Yeshvendra K., Nikhil Sinha, and Sanjay K. Singh, "Heart Disease Prediction System Using Random Forest", *International Conference on Advances in Computing and Data Sciences*. Springer, Singapore, 2016.

[9] Santhana Krishnan. J, Geetha S., "Prediction of Heart Disease Using Machine Learning Algorithms", ***2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)***

[10] http:archive.ics.uci.edu/ml/datasets/Heart+Disease