Heart Disease Prediction Using Machine Learning Techniques

S. Rasswanth[1] and N. Muthurasu[2]

Department of Computer Science and Engineering[1,2]

SRM Institute of Science and Technology, Vadapalani, Chennai, India

Abstract*:* *Machine Learning has a wide variety of applications. The medical field is not an exception. Predicting the presence or absence of illnesses like heart disease, Parkinson's disease, and others may be greatly aided by machine learning. If this data can be accurately anticipated in advance, it might provide clinicians valuable insights about how to tailor their diagnosis and treatment for individual patients. We use Machine Learning algorithms to try to foretell the occurrence of heart disease in humans. We analyse existing classifiers and previously proposed classifiers such as Ada- boost and XG-boost to determine which can provide the higher precision, and we propose an ensemble classifier that performs hybrid classification by taking strong and weak classifiers due to its ability to have many training and validation samples*

Keywords: Decision tree, Random forest, logistic regression

I. INTRODUCTION

A recent study by the World Health Organization found that around 1.4 billion individuals worldwide die of cardiovascular disease each year. One of the leading causes of death and disability worldwide is heart disease. The ability to predict cardiovascular disease is recognised as a key topic in the field of data analysis. Globally, the prevalence of cardiovascular disease has continuously risen in recent years.

Many studies have been carried out to identify the most significant cardiovascular disease risk factors and to provide precise risk estimations. Of all the causes of mortality that don't have any overt signs before taking a life, cardiovascular disease has the greatest fatality rate.

High-risk individuals are significantly helped in making well-informed choices regarding lifestyle changes by an early diagnosis of heart disease. Machine learning has shown to be an effective tool for generating informed judgments and forecasts because of the enormous quantity of data generated in the healthcare business. By analysing patient data showing whether they already have heart disease, this study seeks to forecast the advent of heart disease.

Whether or not an artificial intelligence-based algorithm is implemented. Machine learning methods might be quite helpful in this area. Heart disease may show up in many ways, but everyone who is at risk for getting it has a similar set of underlying risk factors. By gathering data from various sources, categorising them into the relevant categories, and then analysing the data to extract the essential information, this method may be used to predict cardiac illness.

II. LITERATURE SURVEY

In a publication titled "Effective Heart Disease Prediction System," Purushottam et al.

[1] used the Cleveland dataset and preprocessed the data before using classification methods. Knowledge extraction is accomplished using Evolutionary Learning (KEEL), an open-source data mining technique that completes missing data values. In a top-down hierarchy, a selection tree, a node is picked at each level via a test for each legitimate node chosen using the hill-climbing method.

In a study titled "Prediction of Heart Disease Using Machine Learning Algorithms," Santhana Krishnan et al.

[2] used the decision tree and the Naive Bayes technique to predict heart disease. The decision tree algorithm creates a tree based on a set of criteria that may either be True or False. Algorithms like SVM and KNN provide results based on vertical or horizontal split conditions depending on the dependent variables. The structure of a decision tree, on the other hand, is formed by the decisions taken at each tree node. It has a root node, leaves, and branches. It is also feasible

to evaluate the importance of the characteristics in the dataset using a decision tree. Furthermore, the Cleveland data set was used.

In a work titled "Prediction of Heart Disease Using Machine Learning Algorithms" published by Sonam Nikhar et al.[3], the authors go into great length on the Nave Bayes and decision tree classifier algorithms that are specifically employed for the prediction of heart disease. The idea of using a predictive data mining technique on the same dataset resulted from research, and it was shown that Decision Tree is more accurate than Bayesian classifier.

In a work titled "Prediction of Heart Disease Using Machine Learning" published by Aditi Gavhane et al.[4], the multi-layer perceptron is utilised for dataset training and testing. There will be one input layer, one output layer, and one or more hidden layers between them in this approach. Each input node is connected to the output layer via hidden layers. Weights were allocated at random to this connection. Based on the criterion, a weight is applied to the remaining data, known as bias. Between nodes, there may be feedforward or feedback connections.

Avinash Golande et alarticle .'s "Heart Disease Prediction Using Effective Machine Learning Approaches"[5] uses a small number of data mining approaches to help physicians distinguish between various types of heart disease. Many times, methods like Naive Bayes, Decision trees, and K- Nearest Neighbor are used. Packing computation, Part thickness, consecutive insignificant streamlining and neural systems, straight Kernel selfarranging directing, and SVM are further novel characterization-based procedures (Bolster Vector Machine).

"Machine Learning Techniques for Heart Disease Prediction" was developed by Lakshmi Rao et al. [6] with more heart disease risk factors. As a result, cardiac illness is difficult to diagnose. The severity of cardiac disease in individuals is evaluated using a variety of neural network and data mining approaches.

A heart attack prediction system employing Deep learning methods and the Recurrent Neural System is utilised in "Heart Attack Prediction Using Deep Learning" by Abhay Kishore et al. [7], which aims to forecast the likely components of heart-related disorders in the patient. This model makes use of deep learning and data mining to create the most accurate and error-free model feasible. A number of heart attack prediction algorithms benefit from the solid basis this study offers.

Senthil Kumar Mohan et al"EffectiveHeartDisease .'s .'s Prediction Using Hybrid Machine Learning Techniques"

[8] aims to increase accuracy in the diagnosis of cardiovascular diseases. The algorithms utilised to provide a more accurate heart disease prediction model based on a hybrid random forest with linear model include KNN, LR, SVM, and NN (HRFLM).

Ajan N. Repaka and others Our suggested strategy for predicting the percentage of risk is more accurate than previous models, according to the testing data.

There are fewer manual processes when data is directly retrieved from electronic records. The quantity of services given is decreased by demonstrating that a large number of rules contribute to the most accurate prediction of heart disease and by reducing the number of services offered. Frequent pattern growth association mining is used on the patient's dataset to create a strong relationship.

III. SCOPE

The hardest part of cardiac disease is identification. Although there are technologies that may forecast heart disease, they are either costly or ineffective in determining the likelihood of heart disease in an individual. The mortality and overall effects of heart diseases may be reduced with early identification. However, because doing so calls for more tolerance, time, and experience, it is unrealistic to precisely monitor patients every day, and doctors cannot consult with patients around- the-clock. We can now employ a variety of machine learning techniques to uncover previously hidden patterns in the data since we have access to a large quantity of data. Medical data may include hidden patterns that may be utilised to diagnose illnesses.

IV. PROPOSED SYSTEM

The gathering of data and the identification of key properties are the first steps in the system's operation. The necessary format is then created by pre- processing the necessary information.

Next, training data and testing data are separated from the data. putting the processes into practise and training the model using the training data. Testing the system using the test data allows for the evaluation of its accuracy.

The following modules make up the system's implementation.

1. Data Set Collection
2. Choice of attributes
3. Pre-processing of Data
4. Making Sense of the Data

1. Data Set Collection

We start by gathering information from the famous Billroth Hospital in Tamil Nadu for our heart disease prediction system. The dataset is divided into training data and testing data after data collection. In contrast to the testing dataset, which is used to assess prediction models, the training dataset is used to train prediction models. 30% of the testing data and 70% of the training data are used in this project. The dataset for heart disease at UCI is used in this study. The dataset has 76 properties, 14 of which the system makes use of. accuracy of the model may be improved by data pre-processing.

2. Making Sense of the Data

Machine learning methods are used for classification while using EDA (exploratory data analysis), Xg boost, K nearest neighbours, Random forest, Logistic Regression, Decision tree, and Visualization. The algorithm with the best level of accuracy is chosen after a comparison of the algorithms to predict heart disease.

V. ALGORITHMS USED

A. Decision tree algorithm

It is a supervised technique that is frequently used to address classification problems, though it can also be applied to regression. It's a classifier set up like a tree, with internal nodes representing dataset features, branches representing the rules applied to those decisions, and leaf nodes representing the outcomes.

This tree has two different types of nodes called as the Decision Node and the Leaf Node. Leaf nodes reflect the outcomes of prior choices and do not have any more branches. The features of the supplied dataset serve as the basis for every decision or test that is made. It is a graphic tool for exploring all potential consequences of a decision or problem within predetermined bounds. The "Decision Tree" gets its name from the fact that it has a central node from which branches out in many ways as a tree does.

Our tree is built using the Classification and Regression Tree (CART) method. A Decision Tree just requires one question to start the process.

It posed yes-or-no questions, and then the tree was split into branches in accordance with the responses.

The Decision Tree Method is categorised as a supervised learning algorithm in the field of machine learning. Both a classification and a regression problem may be handled utilising it. With the leaf node representing a class label and the inner node conveying attributes, this algorithm's tree representation is able to meet the difficulty of developing a model that predicts the value of a target variable.

B. Choice of attributes

Selecting the most relevant traits for the prediction system is known as attribute or feature selection. This is used to raise the system's effectiveness. A variety of patient characteristics are chosen for prediction, including gender, the nature of the patient's chest discomfort, resting blood pressure, serum cholesterol, and blood pressure. In this approach, attributes are chosen using the correlation matrix.

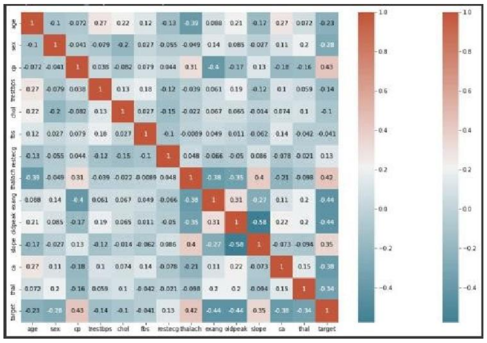


Fig 1: Heatmap to denote correlation between different attributes

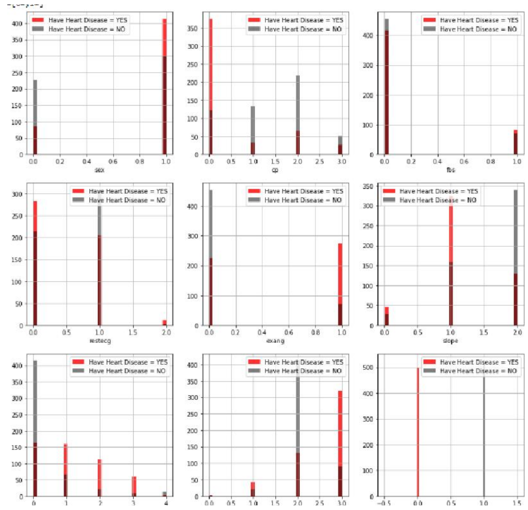


Fig2: Prediction against various attributes

C. Pre-processing of Data

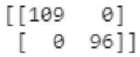
A crucial stage in the building of a machine learning model is data preparation. Data may not be accurate or in the format needed by the model at first, which might lead to incorrect conclusions. The data is changed into the necessary format during data preparation. It handles the missing values, redundancy, and noise in the dataset. Examples of data pre-processing step

It is essential to choose the best machine learning approach for the problem and dataset while creating a model. The decision tree's two key advantages are as follows: decision trees are simple because they mimic how people think when presented with a decision. The decision tree's shape resembles a tree, which makes its logic obvious at a look. The most challenging aspect of utilising a Decision Tree is determining what trait the root of each node should have. This is attribute selection in action. There are two typical approaches to selecting attributes:

Successfully Acquiring New Knowledge: When we use a Decision Tree node to split up the training samples into subgroups, the entropy changes. The gain in information may be used to measure an entropy rise or reduction. The entropy of a random variable, which also represents the impureness of any given collection of samples, measures how unreliable it is. Greater information is indicated by higher entropy.

Using a CART (Classification and Regression Tree) to analyse data is a dynamic learning technique that may produce either a regression or classification tree depending on the dependent variable.

Due to the fact that their algorithms start at the root of the tree, decision trees are helpful for generating predictions about the kind of data they include. This technique compares the values of the record (actual dataset) attribute with those of the root property to decide whether or not to go along a certain branch.



```
[[109   0]
 [  0  96]]
```

Fig 3: Confusion matrix – Decision tree algorithm

Random forest algorithm

In supervised settings, machine learning employs Random Forest. Decision Tree performance is enhanced by bagging, a machine learning classifier upgrade. It combines tree predictors, each of which depends on a separate randomly generated vector. There is a consistent distribution throughout all trees.

Random Forests selects the top predictors inside each node, rather than separating them according on the variables.

The learning complexity of Random Forests is O(M(dnlogn)). It can be applied to both classification and regression because of its adaptability. This approach is not only the most adaptable, but it is also the simplest to use. Simply put, a forest is made up of trees. It is a common misconception that a forest is more stable the more trees there are in it.

It serves as a suitable indicator of the feature's importance. Applications including recommendation systems, picture classification, and feature selection have all found success using Random Forests. It may be used for fraud detection, sickness prediction, and loan applicant categorization. It is the cornerstone of the Boruta algorithm, which pulls relevant information from a dataset.

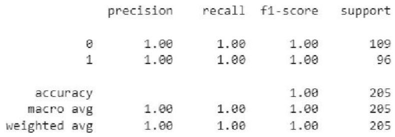
Random Forest is a common supervised learning method in the area of machine learning. It is useful for tackling classification and regression issues in machine learning. The approach is based on ensemble learning, which employs a number of classifiers to enhance model performance and address more challenging issues. It is a classifier that makes use of several decision trees that have been trained on various portions of a dataset before averaging the outcomes to increase the dataset's predictive accuracy. Instead of using a single decision tree to anticipate the output of a random forest, numerous separate trees' predictions are combined.

More trees in the forest equal more trustworthy outcomes and a lower likelihood of overfitting. Assumptions 198 Since the random forest employs a variety of decision trees to classify the dataset, it is likely that some of them will yield accurate results while others will not.

Logistic regression

A kind of supervised learning is logical regression, one of the most used machine learning algorithms. This method enables the prediction of the categorical dependent variable from a collection of independent factors. The goal of logistic regression is to forecast a categorical outcome variable.

The output must thus be a discrete or categorical number. It is possible that there are probability values. more than only these two extremes, such as True or False, Yes or No, 0 or 1, true or false, etc. The main difference between logistic and linear regression is in the application. While logistic regression may address classification problems, linear regression can address regression-related issues.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       109
           1       1.00      1.00      1.00        96

    accuracy                           1.00       205
   macro avg       1.00      1.00      1.00       205
weighted avg       1.00      1.00      1.00       205
```

Fig5: Metric evaluation –Random forest

The fitted "S"-shaped logistic function in logistic regression predicts two maximum values as opposed to a straight line (0 or 1). The logistic function curve shows the chance that an animal is fat given its weight, that malignant cells exist, etc.

Among the numerous machine learning algorithms, logistic regression stands out for its capacity to categorise fresh data and offer probabilities using both continuous and discrete datasets.

A few benefits of logistic regression are its simplicity, convenience of use, and, under some circumstances, its high training efficacy. These criteria also provide an explanation for how this model-training approach makes the best use of available computer resources. It is possible to deduce information about the importance of each characteristic from the learnt weights that were used to anticipate the parameters.

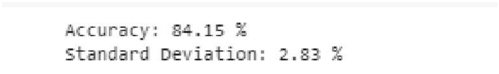
It also specifies if the inference is good or negative. We might do logistic regression to determine the link between the features. Unlike Decision Tree and Support Vector Machine, this approach enables models to quickly incorporate fresh data. It is possible to update using stochastic gradient descent. In addition to classification outcomes, Logistic Regression offers calibrated probabilities.

This is preferable to models that only provide a categorization at the end. We can determine which training examples are more accurate for the given problem if one training example has a 95% probability for a class and another has a 55% probability for the same class.

A statistical analysis model called logistic regression has limitations because it relies on independent characteristics to accurately predict uncertain events. This could lead to a model that struggles to predict outcomes when used with high-dimensional input.

The model has been "over-fit" to the training set, making it unable to reliably predict the results of the test set. When the model is trained using sparse training data but a large number of features, this often occurs. Regularization techniques should be investigated to prevent overfitting on high- dimensional datasets (but this makes the model complex). Under-fitting to the training data is conceivable with very high regularisation factors. The linear decision surface of the logistic regression restricts its use in nonlinear situations.

As a result, it is necessary to modify nonlinear features, which may be done by adding more characteristics to the dataset in order to make it linearly separable in higher dimensions. Non- Linearly Separatable Data: Logistic regression has difficulty expressing complicated connections. This method may be quickly outperformed by neural networks and other more sophisticated and powerful ones.



```
Accuracy: 84.15 %
Standard Deviation: 2.83 %
```

Fig7: Accuracy score – Logistic Regression

VI. VISUALIZATION

A comparison table between accuracies of all the four algorithms are tabulated below, hence K- nearest neighbours has the highest accuracy of 86.34%

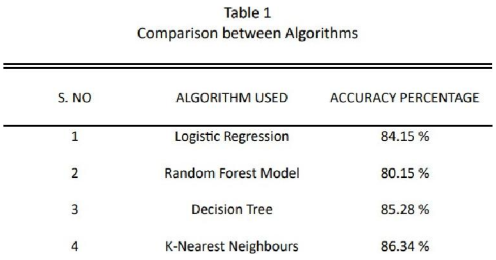


Table 1
Comparison between Algorithms

S. NO	ALGORITHM USED	ACCURACY PERCENTAGE
1	Logistic Regression	84.15 %
2	Random Forest Model	80.15 %
3	Decision Tree	85.28 %
4	K-Nearest Neighbours	86.34 %

Fig8: Comparison between Algorithms.

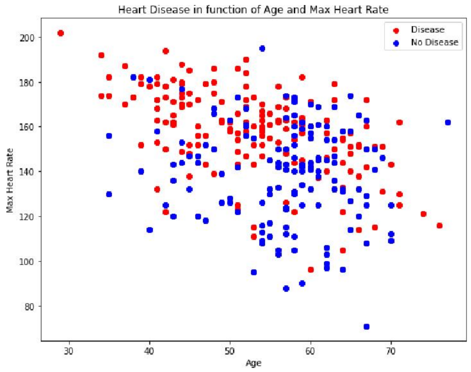


Fig9: Scatter plot - Heart disease in function of age and max heart rate

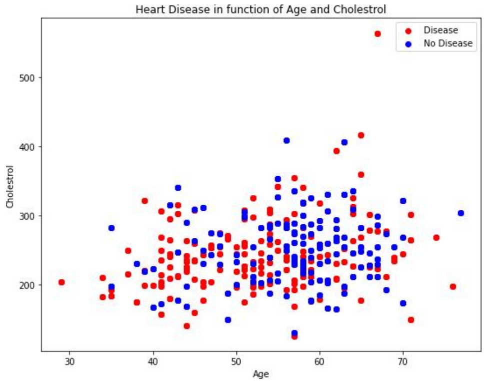


Fig10: Scatter plot - Heart disease in function of age and cholesterol.

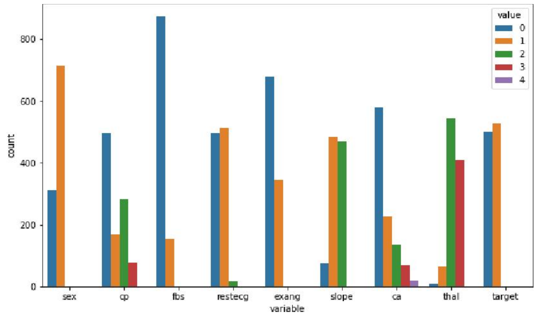


Fig11: Scatter plot - Heart disease in function of age and blood pressure

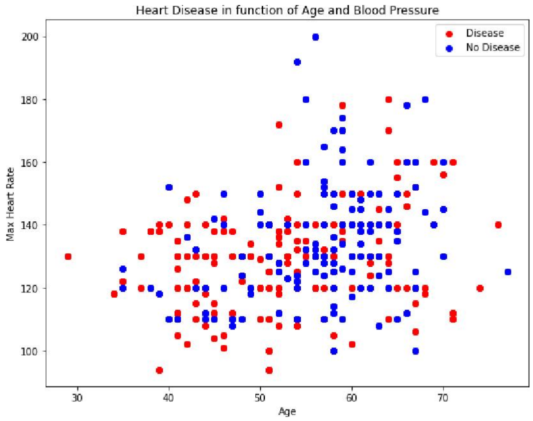


Fig12: Scatter plot - Heart disease in function of age and physical condition

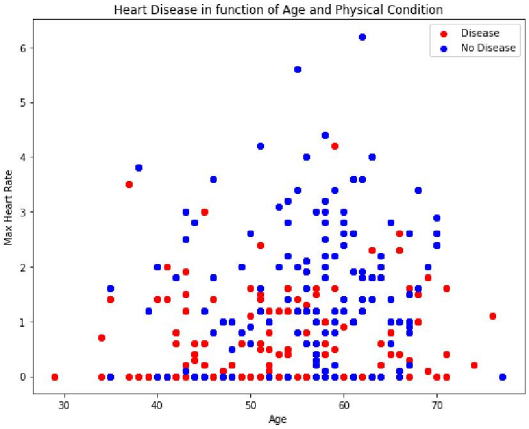


Fig13: Bar graph for categorical attributes

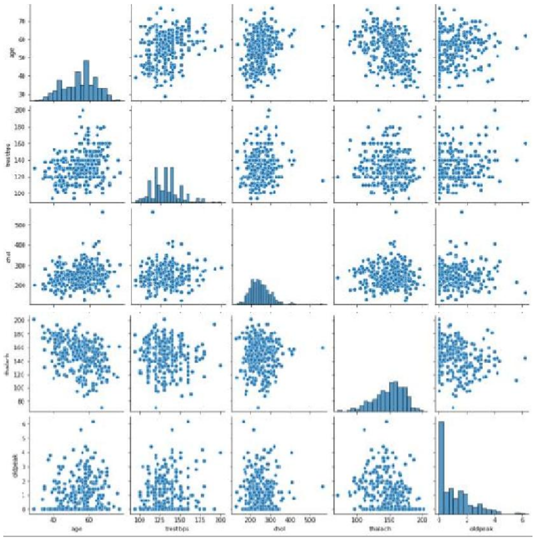


Fig14: Pair plot for various attributes

VII. CONCLUSION

The adoption of cutting-edge technology like machine learning to detect heart issues early will have a profound impact on society since heart illnesses are a significant cause of mortality in India and the rest of the world. Early detection of heart illness may aid high-risk individuals in making decisions regarding lifestyle adjustments, lessening repercussions and representing a major medical advancement. A rising number of people are given cardiovascular disease diagnoses every year.

This need quick identification and care. The medical community and its patients may benefit much from using the proper technology assistance in this field. Four of the seven machine learning algorithms used to the dataset and utilised to assess performance in this study are Decision Tree, Random Forest, Logistic Regression, and Extreme Gradient Boosting.

The 76 features in the dataset include the attributes that are thought to make patients more likely to develop heart disease, and 14 of the most important variables that are pertinent to assessing the system are selected from them. If all characteristics are present, the author's method will be less potent. The aim of attribute selection is to increase productivity.

In this situation, selecting n characteristics to assess the model will increase accuracy. Some dataset features are omitted since they have virtually the same association. Efficiency is greatly decreased if every aspect of a dataset is considered.

One prediction model is created based on a comparison of the seven machine learning techniques accuracy. In light of this, the goal is to use a number of evaluation measures that properly predict the illness, such as the confusion matrix, accuracy, precision, recall, and f1-score. The extreme gradient boosting classifier achieved the highest accuracy in a comparison of seven classifiers, scoring 86.34 percent.

REFERENCES

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications,47(10), 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using kmeans clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259- 64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757- 899X/1022/1/012072 9.

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International MutliConference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.