# A detailed examination of machine learning techniques for predicting heart illness


| | |
|---|---|
| Journal: | *Transactions on Pattern Analysis and Machine Intelligence* |
| Manuscript ID | Draft |
| Manuscript Type: | Regular |
| Keywords: | Cardiovascular disorders, KNN |
| | |

# A detailed examination of machine learning techniques for predicting heart illness

Dr. ThiruKrishna JT[1], Associate Professor , Dayananda Sagar Academy of Tech & MGMT, Bengaluru, India
Prakruthi HR[2], Nayana Sagar[3] ,Moulya G[4] , Priyanshu Singh[5]
Students, Dayananda Sagar Academy of Technology & Management, Bengaluru, India
drthirukrishna@dsatm.edu.in[1],prakruthihr04@gmail.com[2], nayanas.1dt19is086@gmail.com[3], moulyag.1dt19is079@gmail.com[4], singhpriyanshu073@gmail.com[5]

**ABSTRACT**

The prenatal recognition of CVDs can help high-risk patients choose whether to change the way they live , which can lessen their severity. Using consistent techniques of machine learning, research has sought to identify the most significant risk variables for heart disease as well as effectively estimate the total risk. In order to produce an accurate predictive algorithm for heart disease, the latest research has looked at bringing together these methods using techniques like machine learning (ml) algorithms. These findings recommend a framework for assessing the precision of implementing particular outcomes from the use of decision trees, k-near neighbor, and logistic regression and SVM on the Cleveland Heart Disease Database.

**Keywords**

Cardiovascular disorders, decision trees, logistic regression (LR), machinelearning(ML) with support vector machines (SVM),KNN

## 1. INTRODUCTION

According to statistics from the WHO, coronary heart disease is the leading cause of fatalities globally, resulting in 17.9 million fatalities [1]. The biggest lifestyle risk factors for cardiovascular disease and stroke include poor eating habits, inactivity, cigarette smoking, and excessive drinking [1]. A cardiac event occurs when the heart's capacity to pump blood is compromised by arterial plaque formation. A stroke can happen when there is a thrombus in an artery that blocks the flow of blood to the brain [2]. As a result of the symptoms' resemblance to those of other conditions and potential confusion with aging symptoms, diagnosing patients can be challenging for health providers. Heart disease is difficult to pin down due to a number of risk factors that are connected to it, such as high cholesterol levels, diabetes, high blood pressure, irregular heartbeat, and numerous other factors.Relevant coronary artery disease prediction and early reconnaissance have become crucial to boosting patient rates of survival.

ML has now established itself as a key instrument in the healthcare sector for aiding in patient diagnosis. The bulk of the time, the existing methods for anticipating and diagnosing cardiac disease rely on practitioners' assessments of the medical history of a patient, symptoms, and results from health screenings. Clinical evaluations and other patient information are openly accessible and expanding daily in databases used by the healthcare sector today. The severity of the disease is assessed using a variety of methods, including the K-Nearest Neighbour Technique (KNN), Decision Trees (DT), random forest algorithm (RF), and naive Bayes (NB) algorithms []. Because heart disease has a complex nature, it needs careful supervision. Failure to do so might harm the heart or result in a premature death.

Numerous techniques have been attempted to acquire knowledge using well-known ML techniques for heart disease prediction. In order to establish a prediction model, numerous analyses have been done throughout this investigation using an assortment of methods as well as by linking multiple strategies.

## 2. LITERATURE SURVEY

| AUTHOR | PURPOSE | TECHNIQUE USED | ACCURACY |
|---|---|---|---|
| 1.Gudadhe et al. | Created a diagnosis system for HD | •SVM algorithm | 80.41% |

| | | | |
|---|---|---|---|
| | diagnosis utilizing recurrent neural networks with multiple levels and support vector machines (SVM). | •Neural networks | |
| 2. Palaniapp\an et al | A system for professional medical diagnosis was proposed for heart disease identification using ANN,NB and DT | •Artificial Neural Networks<br>•Decision Trees (DT)<br>• Navies Bays (NB). | •88.12%<br>•86.12%<br>•80.4% |
| 3. S. U J. K. A. KHAN and A. SABOOR | The authors suggested a feature correlation-based NN-based forecasting of coronary heart disease (CHD) research. The present research utilised of the KNHANES-VI sample produced by the Korean Institute for Disease Prevention and Control. | •Fast Conditional Mutual Information method for selecting features (FCMIM)<br>•FCMIM-SVM | 92.37% |
| 4. Waqar et al. | Suggested using deep learning based on SMOTE. Without feature selection, the author balanced the dataset using the SMOTE technique. A deep neural network was trained and tested to predict the absence and presence of a cardiac arrest using the balanced dataset, | Deepneural network | 96% |
| 5. Fitriyani et al | developed a methodology for HD prediction that combines hybrid synthetic minority over-sampling technique-edited nearest neighbor (SMOTE-ENN) and density-based spatial clustering of applications with noise (DBSCAN). | DBSCAN<br>SMOTE-ENN<br>XG BOOST<br>CLASSIFIER | 95.9% |
| 6. MOHAN et al. | Developed hybrid machine learning strategy for HD detection. He also put forth a novel methodology for choosing important characteristics from the information for machine learning classifiers to use in training and testing. | HYBRID ALGORITHM | 88.07% |

# 3. PROPOSED FRAMEWORK

## 3.1 Algo Description & Equations

### 3.1.1 Support Vect. Machine

A type of model known as SV machine is employed in classification and regression analysis to examine data and identify trends. When your data contains precisely two class, S.V.M is employed. By locating the ideal hyperplane that differentiates all of the information's data points in a particular category from those in another, . The mathematical model is accurate to a greater extent the more distance separating the two groups.The inner region of a margin cannot contain any points. The data points on the margin's edge are the support vectors.SVM is a

computational modeling strategy that represents tough, practical issues. It is based on mathematical functions. Support The training data is translated into kernel space using vector machines. There are several other kernel spaces that may be employed, including the linear (dot product) kernel, quadratic kernel, polynomial kernel, radial basis function kernel, multilayer perceptron kernel, etc. Moreover, there are other ways to put SVM into practice, including least squares, sequential minimal optimisation, and quadratic programming. The difficult part of SVM is choosing a kernel and a technique such that your model isn't overly optimistic or pessimistic..

It is debatable if the selected kernel is RBF or linear because the CHDD comprises a substantial number of instances and characteristics. Despite the nonlinear relationship between characteristics and class labels, RBF kernel performance may not be enhanced by the sheer amount of features. It is advised to test both kernels before choosing the one that is more effective.

Assume that the information for the trained samples is Data=yi, xi;i=1, 2,3,..., n, where xi Rn represents the ith vector and yi Rn defines the target element. The linear Support Vector Machine (SVM) is used to identify the ideal hyperplane with the shape f(x)=wTx+b, where w is a multidimensional parameter vector and b is an interval. By fixing the subsequent optimizations issue, this is achieved:

$$Min_{w,b,\xi_i} \frac{1}{2}w^2 + C\sum_{i=1}^{n}\xi_i$$
$$s.t.\ y_i\left(w^{\mathrm{T}}x_{\mathrm{i}} + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \quad \forall_i \in \{1, 2, \ldots, m\}$$

### 3.1.2 Decision Trees(DT)

A decision tree(DT) is an approach for categorizing or forecasting what will happen to data using a regression analysis or classifier.When there is continuous data, regression is utilised , while classification is used when the characteristics are clustered. One of the key techniques for data mining is the decision tree. A root node, branches, and leaf nodes make up a decision tree. Follow the path from the root node to a leaf node to assess the data.

A purity index, which will divide the nodes as stated in the training section, must be used to generate decision trees. Each of the 297 tuples is assessed for heart disease using the CHDD decision tree, which results in a positive or negative judgment for each.The reliability, precision, and degree of sensitivity of the predicted outcome are evaluated by comparing them to the initial selection parameter in the CHDD in order to check for erroneous positives or misleading negatives. The splitting parameter that has been utilised further demonstrates the value of each attribute. The trees are built utilising inputs with high entropy for the training instance of D. The top-down recursion division and conquering (DAC) method is used to swiftly and easily produce these trees. To get rid of the unnecessary samples, D is pruned.

$$Entropy = -\sum_{j=1}^{m} p_{ij}\log_2 p_{ij}$$

### 3.1.3 Logistic Regression

It is typical to refer to this kind of statistical framework as a log it model, and it is frequently utilised in reclassification and analytical forecasting. Based on a variety of factors that are independent, logistic regression estimates the possibility that a situation, such as voting or not voting, is going to occur.Men are more likely than women to get heart disease, according to the results of the logistic regression analysis. The risk factors for CHD are age, daily cigarette consumption, and systolic blood pressure. Yet, neither the total cholesterol tierl nor the blood glucose level have changed much.

### 3.1.4 Naive Bayes

For task classification like categorizing texts, the Naive Bayes classification model is a supervised artificial intelligence methodology. Furthermore, it belongs to the family of generative learning methods, which duplicates the input distribution within an identified group or categories.The NB algorithm could recognise the features associated with heart disease. It displays the potential for each of the 15 input attributes for the predetermined condition.

### 3.1.5 Random Forest

Leo Breiman and Adele Cutler created the widely used method for machine learning known as random forest modelling, which integrates the outcomes of numerous decision trees to get one final decision. Its versatility and effectiveness, which can handle issues with regression and classification, are what fuel its widespread usage.The random forest (RF) methodology is used in ROC curve. For both the true positive rate and the rate of false positives at various sensitivity settings, the area under the ROC curve is shown. The simulation properly determined whether an individual had coronary artery bypass graft or not, according to the ROC curve's AUC measurement of 93.3%.

### 3.1.6 K-nearest neighbors

The k-nearest neighbours technique, commonly referred to as KNN or k-NN, is a classifier developed using supervised learning that anticipates or groups how a single data point will be categorized.KNN is a simple classifier in which samples are categorized according to the class of their closest neighbor. High volume is a characteristic of medical databases. Classification may result in less accurate results if the data collection contains redundant and unnecessary properties.

## 3.2 Architecture

The general layout of the software architecture is shown in Fig 3.2.1. The proposed system's core modules are made

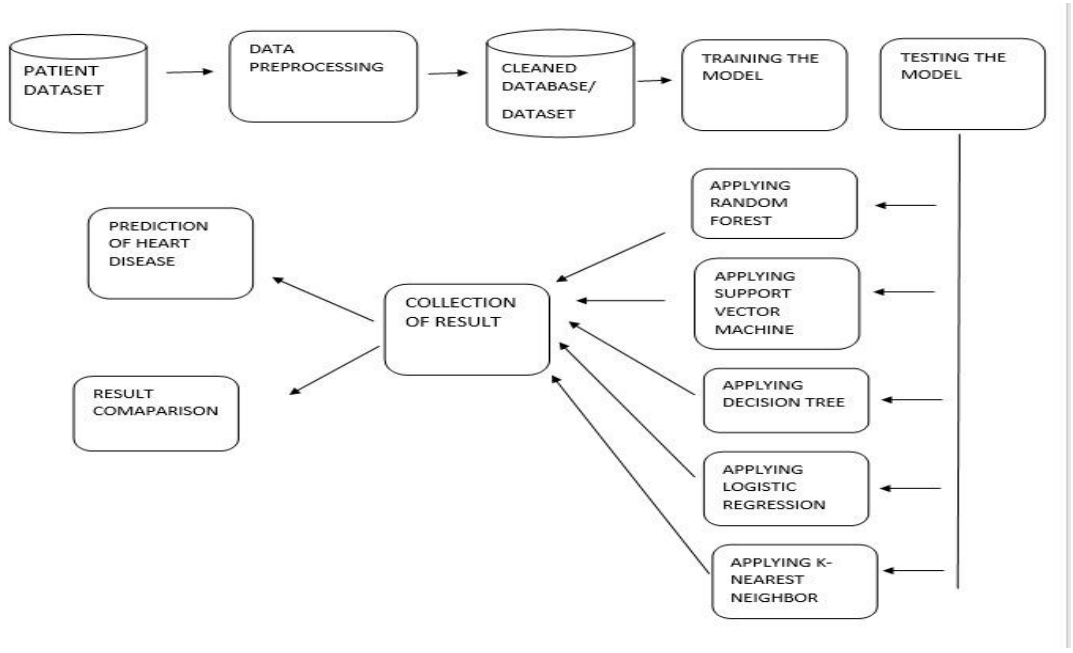


fig 3.2.1 Architecture of HD prediction system

from :

**Knowledge of Domain Dataset**

The system is given a dataset as input, which is covered in depth . It also has the option of accepting human entries

**Processing of Data**

Data preparation is the process of transforming data so that it may be used for future analysis.

**Module**

It discusses the algorithmic methodology used on the system to provide very precise findings. In machine learning techniques, we employ SVM ,KNN,Random forest ,Decision tree as algorithmic approaches.

**Assessment and Implementation**

Information about the outcome is included in the concluding analysis modules. Our method compares and draws conclusions based on quantifiable results like after getting a confusion matrix, the levels of sensitivity, specificity, accuracy, true positive rate, and the false-positive rate.

### 3.3 Flowchart

A flowchart is a diagrammatic representation of the steps of a process in a sequence . It's a tool that may be utilized for describing a variety of procedures, such as a production process, service procedure, or a strategy for a project.

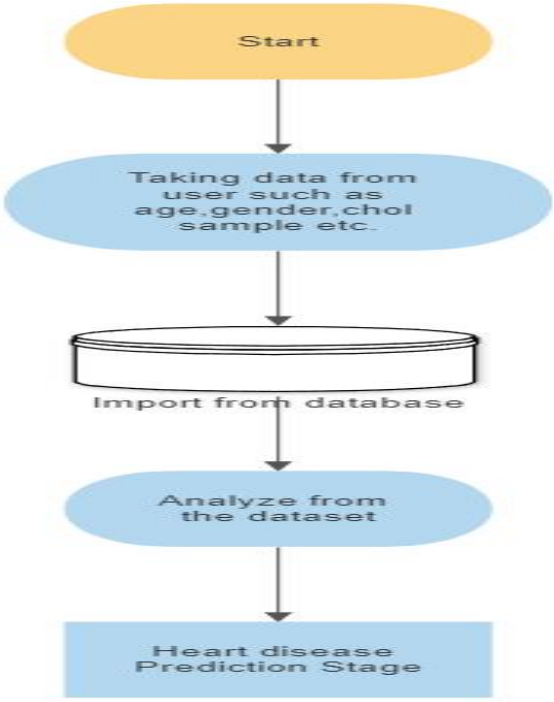


Fig.3.3.1 showing the flow of implementation

## 4. Results & Discussion

### 4.1 Simulation Para

The random variable inputs in simulation are typically not precisely understood, although the model is frequently. Inputs are precisely known in machine learning, but the model is unknown before training. The variations in production are relatively slight. Both provide an output, but there are several sources of uncertainty.The final results are detected by the medical experts .

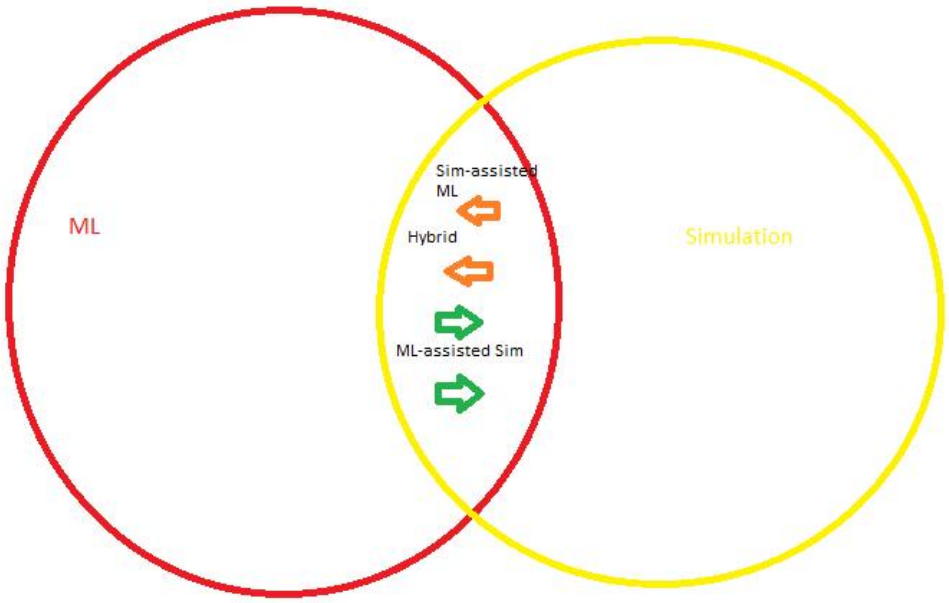


Fig 4.1.1 showing ML and simulation relation

## 4.2 Output Screenshot

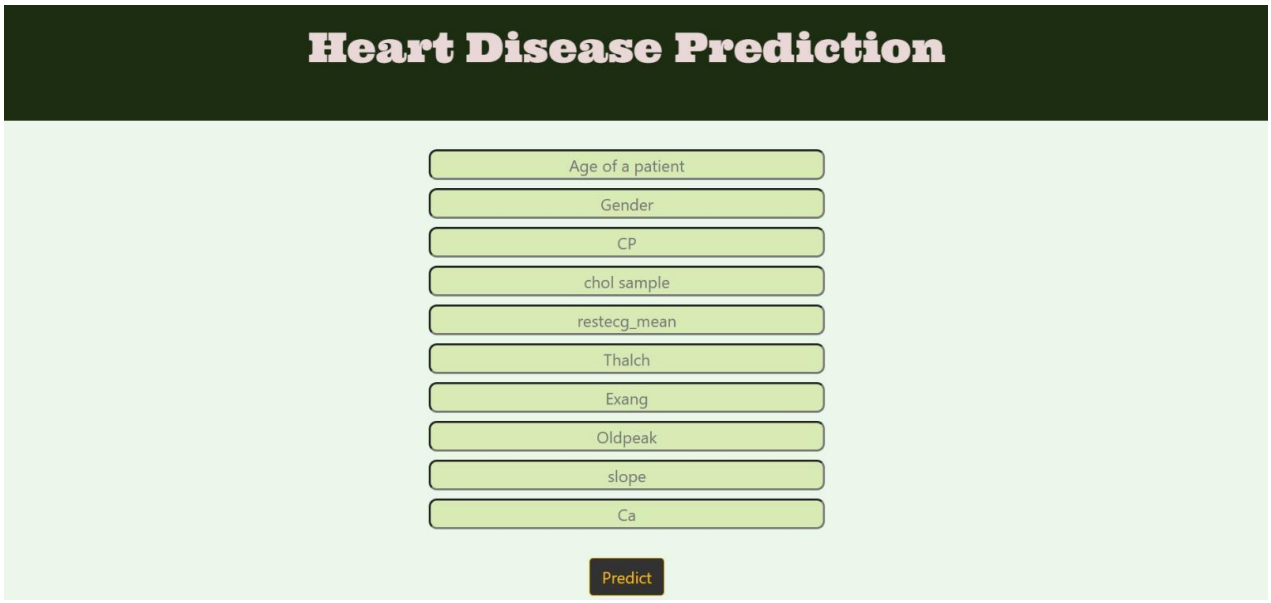


Fig4. 2.1 : Some standard parameters are present and by entering the values against each parameter , The algorithm predicts the type of heart disease.

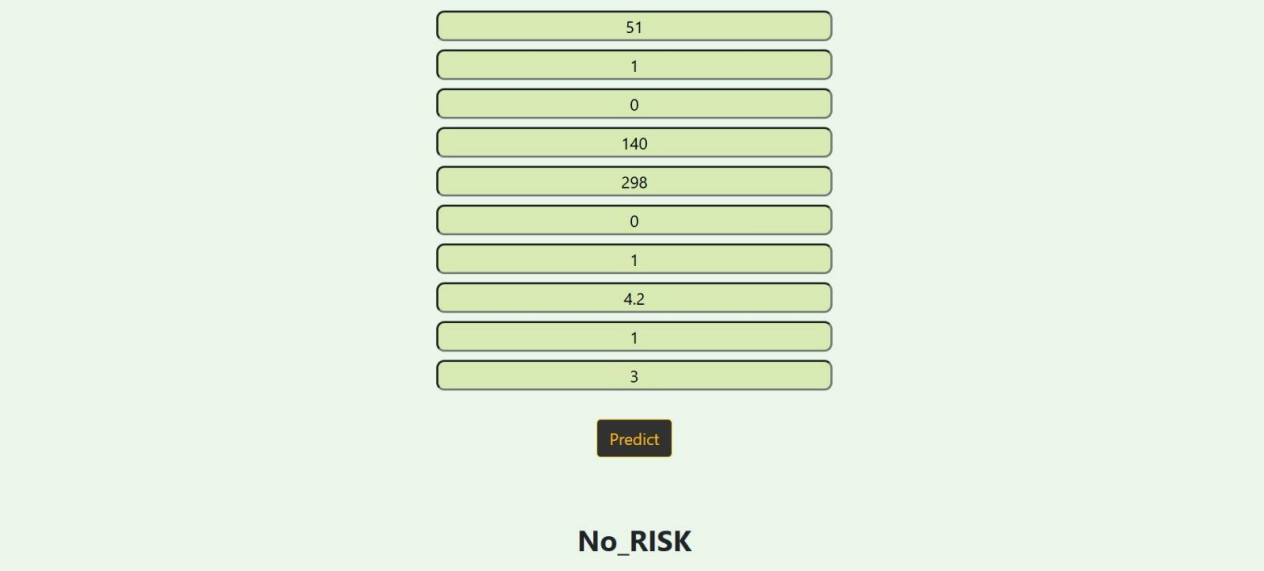


Fig 4.2.2: The figure contains values against each parameter and the algorithm has detected no risk in the patient's heart condition.

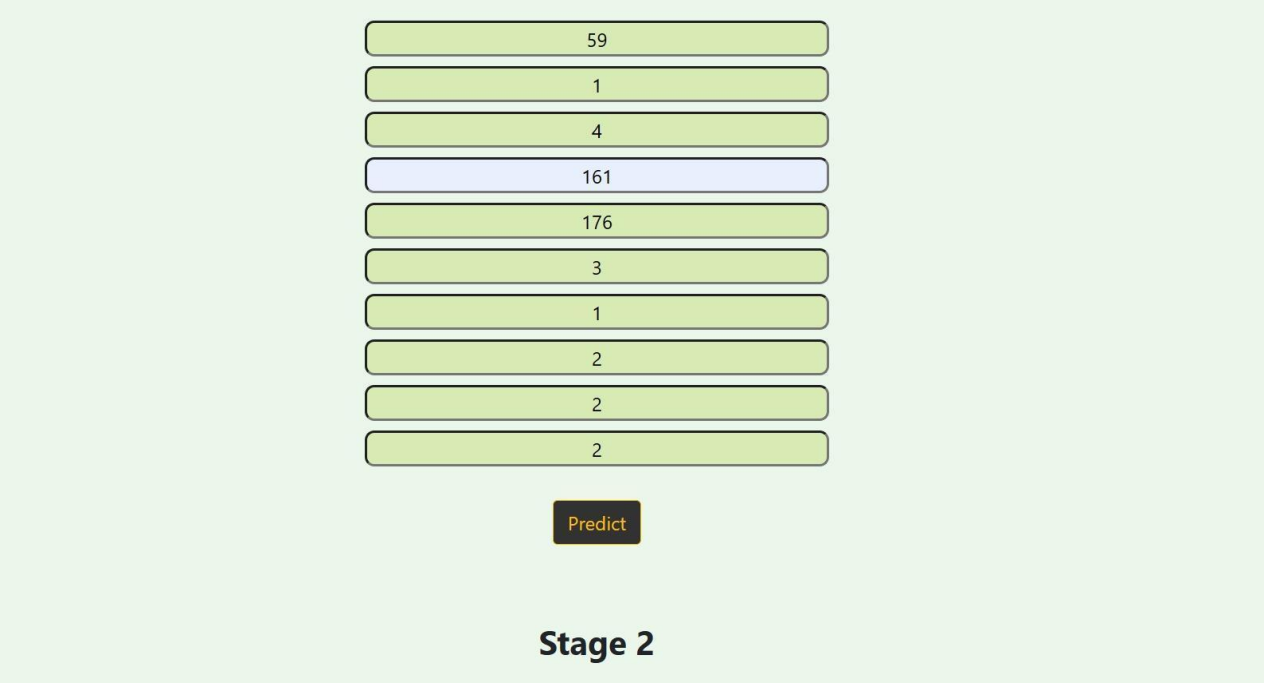


Fig 4.2.3:The figure contains values which have been detected as stage 2 heart condition.

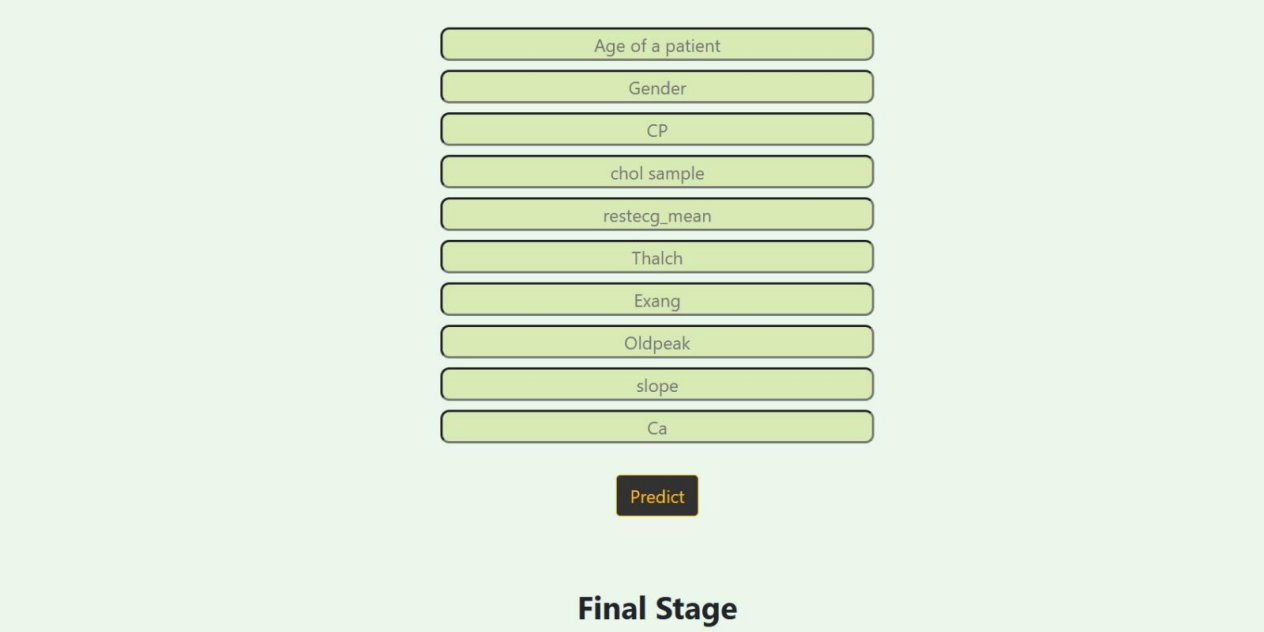


Fig 4.2.4: The figure shows the entries which have the highest risk of having heart disease.

## 5. APPLICATION

One of the most deadly and curable chronic diseases, heart disease is a leading cause of mortality in both economically developed and underdeveloped nations. If the patient is identified early on and receives the appropriate care, the harm can be significantly mitigated. Hence, early identification can help people make lifestyle adjustments and, if necessary, provide optimal medical care.Cardiovascular disease prediction supports practitioners in making more accurate health decisions for their patients By processing enormous volumes of complicated health data and exposing clinically meaningful information regarding CVDs, machine learning techniques not only help doctors make more efficient and precise clinical choices but also considerably advance clinical understanding.It is a realistic alternative for limiting and comprehending heart clinical symptoms is through detection using machine learning (ML) .

## 6. CONCLUSION AND FUTURE WORK

By using the review of literature we can draw a conclusion that combinational and more advanced algorithms are to be used to increase the efficiency and accuracy of the predicting system to detect the heart diseases in the earlier stages.

The purpose of the study was to discover if it would be possible to identify potential risks for heart disease using patient questionnaires that contained history subjective and examination-based objective health data. In order to make a precise prediction of cardiac illness, this research offers a framework that combines decision trees, random forests, logistic regression, and support vector machines. This paper offers recommendations for training and testing the system, resulting in the best effective model among the various rule-based combinations, using the Heart Disease database. This research also suggests comparing the various results, including sensitivity, specificity, and accuracy. The system will need to be developed using the above approaches, and this will require training and testing the system. It also includes development of a tool to estimate a potential patient's illness risk. Future study on this topic may combine various methods for machine learning in order to improve prediction tools in order to improve the accuracy of coronary artery disease prediction and get a deeper knowledge of the crucial factors, new feature-selection algorithms may also be developed.

## 7. REFERENCES

1.C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2022,

2.V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques,"2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 177-181.

3.A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL) Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu School of Computing Sciences and Engineering, VIT University Vellore – 632014, Tamil Nadu, India.

4.NHANES: https://www.cdc.gov/nchs/nhanes/index.htm

5.WHO, https://www.who.int/health-topics/cardiovascular- diseases#tab=tab_1

6.S. Singh, and R. Zeltser, "*Cardiac Risk Stratification,"* in:

*StatPearls*, Treasure Island (FL): StatPearls Publishing, 2020.

7.A. Pandya, et al., "A comparative assessment of non-laboratory- based versus commonly used laboratory-based cardiovascular disease risk scores in the NHANES III population," *PLoS One*, vol. 6, no. 5, pp. e20416, May 2011.

8.C.Y. Wang, et al., "Cardiorespiratory fitness levels among US adults 20-49 years of age: findings from the 1999-2004 National Health and Nutrition Examination Survey," *Am J Epidemiol*., vol. 171, no. 4, pp. 426-435, Feb. 2010.

9.P.L. Tsou, and C.J. Wu, "Sex-Dimorphic Association of Plasma Fatty Acids with Cardiovascular Fitness in Young and Middle-Aged General Adults: Subsamples from NHANES 2003-2004," *Nutrients*, vol. 10, no. 10, 1558, Oct. 2018.

10.S.S. Yoon, et al., "Trends in the Prevalence of Coronary Heart Disease in the U.S.: National Health and Nutrition Examination Survey, 2001-2012," *Am. J. Prev. Med*., vol. 51, no. 4, pp. 437-445, Oct. 2016.

11.R. Moonesinghe, et al., "Prevalence and Cardiovascular Health Impact of Family History of Premature Heart Disease in the United States: Analysis of the National Health and Nutrition Examination Survey, 2007-2014," *J. Am. Heart Assoc*., vol. 8, no. 14, e012364, July 2019.

12.K. Jindai, et al., "Multimorbidity and Functional Limitations Among Adults 65 or Older, NHANES 2005–2012," *Prev. Chronic Dis*., vol. 13, 160174, Nov. 2016.

13.S.Heyden, et al., "Angina Pectoris and the Rose Questionnaire,"*Arch. Intern. Med*., vol. 128, no. 6, pp. 961–964, 1971.

14.A. Koyanagi, et al., "Correlates of physical activity among community-dwelling adults aged 50 or over in six low- and middle-income countries," *PLoS ONE*, vol. 12, no. 10, e0186992, Oct. 2017.

15.W.-H. Weng, "Machine Learning for Clinical Predictive Analytics," in: *Leveraging Data Science for Global Health*. L. A. Celi et al. (eds.), 2020, ch. 12