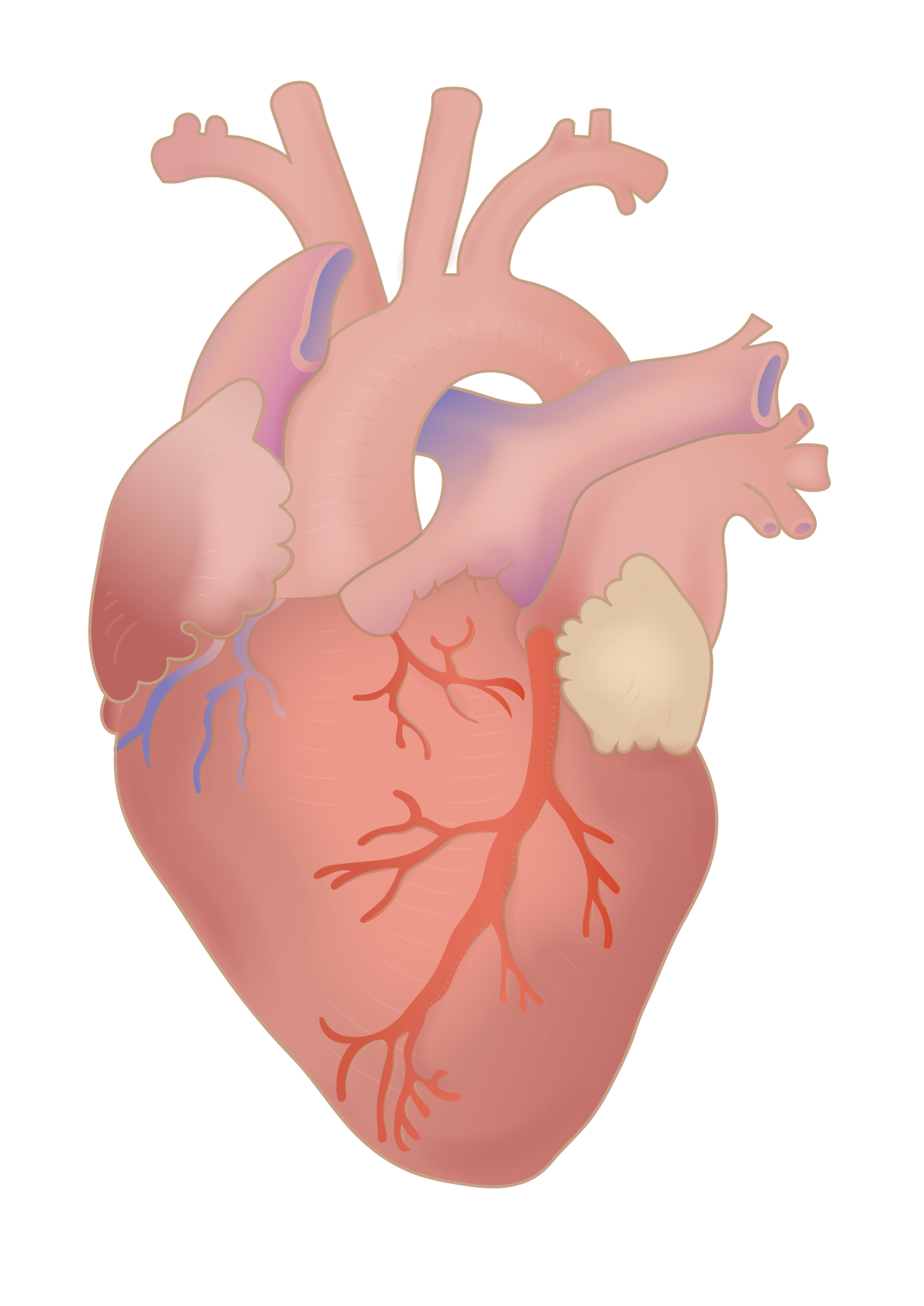
MINOR PROJECT REPORT

**Heart Disease Diagnosis**

**-under the supervision of Dr. Afzal Hussain Shahid**



ANUKUL KUMAR 1806093

ANKIT GOND 1806016

RITESH SINGH 1806092  
*6th Semester*

*CSE I*

# INTRODUCTION

The WHO states that cardiovascular diseases are one of the most prevalent causes of death globally and a large number of people die due to this disease every year. Cardiovascular diseases include deep vein thrombosis, peripheral arterial disease, coronary heart disease, congenital heart disease and rheumatic heart disease. 31% of the death of people is due to Heart disease around the globe every year.

Cardiovascular Disease is of two types.

* Heart Attack-It occurs when the heart blood vessels are suddenly blocked.
* Heart Failure-It results from coronary heart disease, hypertension, cardiomyopathy. Heart failure is basically when the heart is unable to maintain a strong blood flow and this results in chronic tiredness, resistance to physical activities and shortness of breath. Heart failure can be divided into three types.

1. right side heart failure
2. Left side heart failure
3. Congestive heart failure.

The aim of our project is to develop a model to predict heart diseases as detecting an anomaly in the heart can prove to be very beneficial and can provide some precious time to treat the patient .

# DATASET

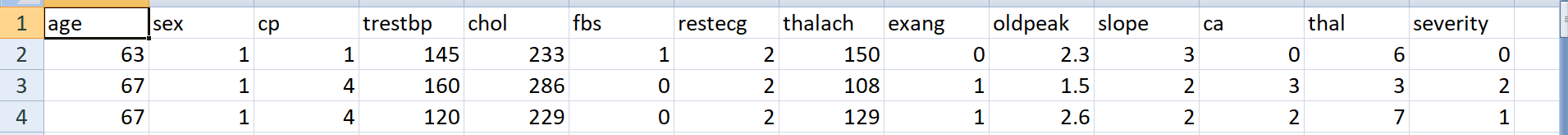
## Cleveland database

The dataset used is collected from the uci machine learning repository.

The dataset contains 303 rows.

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them.

The ‘severity’ field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).



**An instance of the dataset showing three rows and associated attributes**

## Attribute Information:

Only 14 attributes we use

-- 1. #3 (age) age in years

-- 2. #4 (gender) sex (1 = male; 0 = female)

-- 3. #9 (cp) chest pain::1 typical angina 2 atypical angina 3 non-anginal pain 4 asymptomatic

-- 4. #10 (trestbps) resting blood pressure (in mm Hg on admission to the hospital)

-- 5. #12 (chol) serum cholesterol in mg/dl

-- 6. #16 (fbs) (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

-- 7. #19 (restecg)

-- 8. #32 (thalach) maximum heart rate achieved

-- 9. #38 (exang) exercise induced angina (1 = yes; 0 = no)

-- 10. #40 (oldpeak) ST depression induced by exercise relative to rest

-- 11. #41 (slope)

-- 12. #44 (ca) number of major vessels (0-3) colored by fluoroscopy

-- 13. #51 (thal) 3 = normal; 6 = fixed defect; 7 = reversible defect

-- 14. #58 (num) \*\*this is the predicted attribute

# METHODOLOGY AND ANALYSIS

## A. Methodology

i. Data Preprocessing

Filling the missing data.

Selecting appropriate features for model creation.

Segregation of target data and feature data as training and test data.

Scaling the values in the data to be values between 0 and 1 in which and scale all the values before training the Machine Learning models.

ii. Applying Algorithms

Comparing 5-machine learning algorithms such as KNN,ANN,SVM, Naive Bayes and Random forest classifier to get the better accuracy to which highest parameter may cause disease.

For each algorithm, there is a pseudo code helpful to develop any kind of programming language. In python, there is a simple way to establish any kind of algorithm in which simple and short code easier to predict accuracy.

## B. Machine Learning Algorithm:

The algorithms used in this project are highly helpful to predict the accurate result to detect heart disease in which factors that cause a disease can be detected. The following algorithms have been built in the model.

**i. K-Nearest Neighbor algorithm (KNN):**

KNN is a supervised classifier that carry-outs a observations from within a test set to predict classification labels. KNN is one of the classification techniques used whenever there is a classification. It has a few assumptions including that the dataset has little noise, labeled and it should contain relevant features. By applying KNN in large datasets takes a long time to process. The accuracy gained with this algorithm is 68.3 %.

**ii. Artificial Neural Network (ANN):**

Artificial Neural Network (ANN) is a deep learning algorithm that emerged and evolved from the idea of Biological Neural Networks of human brains. A neural network having only one layer is called a perceptron. A multilayer perceptron is called Artificial Neural Network. An ANN can consist of any number of layers. Each layer can have a number of neurons. Each neuron is connected with each and every other neuron. The accuracy gained by this algorithm is 70.8.

**iii)Random Forest Classifier**

Random forest classifier is a powerful tool in the machine learning library. With this classifier, we will be able to get higher accuracy and training time should be less. Initially, we have to build a model and by splitting variables into training and test set. After splitting the data, train the dependent variables and predict the response. By using the random forest classifier, the accuracy of the predicted result is 62.4%.

**iv. Support Vector Machine (SVM):**

Support Vector Machine (SVM) is a Supervised Machine Learning Algorithm used for classification and/or regression.

“The goal of the SVM algorithm is, to create the best line or decision boundary that can segregate n-dimensional space into classes**”**, so that we can easily put the new data point in the correct category in the future. This decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.The accuracy gained with this algorithm is 67.21 %.

**v) Naive Bayes Classification:**

Naive Bayes classifiers are linear classifiers based on Bayes’ theorem. The model generated is probabilistic. It estimates conditional probability which is the probability that something will happen, given that something else has already occurred.Naive Bayes is a classification algorithm based on Bayes’ Theorem.The accuracy gained with this algorithm is 61.53 %.

# DATA

# 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | KNN | ANN | SVM | NAIVE BAYES | RANDOM FOREST |
| Accuracy | 68.3 | 70.8 | 67.21 | 61.53 | 62.4 |

# 

# RESULTS

All the models predicted heart disease and severity with a pretty decent accuracy while the ANN network stood out as the better one and predicted with an accuracy of 70.8% .

When a binary classification is performed that will only detect if there is a disease or not, using the ANN the model gains an improved accuracy of 91% and using SVM 89% and using Random Forest 88%.

# CONCLUSION

The model we developed can be used to detect the presence and severity of heart disease in patients.

#### Getting better:

The model can be made better at prediction by providing even large datasets and training the model.

The model can be also customized to perform only binary classification which will detect only the presence of disease rather than its severity which can be helpful and with an intervention of a professional after initial screening may protect some lives.