**Interim Progress Report on**

**An Application Development for Symptom-Based Heart Disease Detection using Machine Learning**

Table of Contents

[Introduction 3](#_Toc96832911)

[Project Aim 3](#_Toc96832912)

[Research Questions 3](#_Toc96832913)

[Project Objectives 4](#_Toc96832914)

[Literature Review 4](#_Toc96832915)

[Previous Researches 4](#_Toc96832916)

[Comparison of Previous Research 5](#_Toc96832917)

[Progress Summary 7](#_Toc96832918)

[Data Collection 7](#_Toc96832919)

[Data Background 8](#_Toc96832920)

[Attributes of Data 8](#_Toc96832921)

[Proposed Methodology 9](#_Toc96832922)

[Data Analysis 10](#_Toc96832923)

[Issues for Research 10](#_Toc96832924)

[Ethical and Legal Considerations 10](#_Toc96832925)

[Professional Issue 10](#_Toc96832926)

[Social Issue 11](#_Toc96832927)

[Research Planning 11](#_Toc96832928)

[Selection of Tool 11](#_Toc96832929)

[Selection of Technologies 11](#_Toc96832930)

[Selection of Algorithms 11](#_Toc96832931)

[Project Planning 12](#_Toc96832932)

[References 13](#_Toc96832933)

[Appendix 14](#_Toc96832934)

# Introduction

Heart Disease is one of the most seen diseases in humans that can happen for various health conditions (Biswas, et al., 2020). There are no particular symptoms by checking which it can be said that the patient is affected with heart disease, Rather, the heart disease can be detected with the combination of symptoms like the number of cholesterols in blood, amount of sugar in blood etc. So, the determination should be done to check which symptom is most influential to detect heart disease and whether there will be any relationship with age or not (Keya, et al., 2021). This kind of determination will be done in this project by applying statistical methods such as correlation and feature selection. Using the feature selection method, the highly influential symptoms will be considered for the detection of heart disease (Schulte, et al., 2018).

In this research and project, heart disease will be identified using the historical data of the patients and concerning the symptoms as features (Katarya & Srinivas, 2020). In this process, first, the important features will be identified using correlation and then by applying machine learning classifiers, the heart attack will be predicted. As there will be multiple algorithms selected, all those will be applied to detect heart disease and the performances will be recorded in terms of the matrices such as accuracy, precision, recall, f1-scores etc (Biswas, et al., 2020). By comparing the performances, the best model will be detected. Finally, with the application of the best model the detection of heart disease will be done using the Graphical User Interface where the patient will input the values of the symptoms. The backend model will take those symptoms and detect whether the patient has a heart attack or there will be a possibility of a heart attack in future. The detailed report will be presented when the patient will check the result.

## Project Aim

The project aims to design an application through which heart disease can be detected based on the symptoms of the patients. Heart Disease can be seen for different symptoms and health conditions and it varies by the person (Atallah & Al-Mousa, 2021). In this project, the application interface will be designed by taking machine learning in the backend to detect the heart disease of the patients based on their symptoms. When a patient will input their symptoms, the application will detect the symptoms and predict whether the patient has heart disease or not and the outcome will be shown as a detailed report (Abdulhadi & Al-Mousa, 2021).

## Research Questions

The research questions have been prepared to satisfy the aim of the project and to conduct the project successfully. The research questions are as follows:

1. Which symptoms are highly effective to detect the disease?
2. How the users o the application will be facilitated by getting the result of disease testing?
3. Which algorithm will be the best to detect disease?

## Project Objectives

The objectives have been prepared in the view of fulfilling the aim and answering the research questions successfully. The objectives of the project are discussed below in steps:

1. To study and obtain the idea about disease prediction approaches ad methodologies from the existing research models.
2. To determine the algorithm and the methods that the researchers have applied in their experiments.
3. To select the heart disease data and study the features and understand the symptoms and the effect on disease.
4. To select and apply the algorithms of machine learning and determine the best model to detect disease by comparing the classification matrices such as accuracy, precision etc. along with signifying the improvement of the research by comparing the performances of the best model with the existing models.
5. To design the Graphical User Interface to detect disease by applying the best model in the backend.
6. To Visualize the outcome of the disease prediction in the form of a detailed evaluation report.

# Literature Review

Different Scholars and researchers have asserted their ideas and approaches for the identification of heart disease. There are several ways to detect heart disease. However, symptoms play a key role in the detection process. So, many researchers have observed the symptoms and analyzed them for different people to identify heart disease. In this section, some important research papers will be reviewed and discussed regarding the method & approach along with the challenges that they have faced & proposed solutions in addition to the data that they have used.

## Previous Researches

Khan & Mondal (2021) have approached the model to identify heart disease at the early stage. They have selected the heart disease dataset from UCI and applied their selected classifiers of machine learning. However, they have found the data contains several missing values with special characters. They have faced the primary challenge to handle the noise of the data along with facing the less number of observations in the dataset. To overcome the challenge, they have preprocessed the data and cleaned it. Secondly, they have resampled the data to get more data instances. They have applied Support Vector Machine and Naïve Bayes on the data and got 72% accuracy using SVM for heart disease prediction.

Haq et al. (2021) have used the heart disease dataset from UCI for the detection of heart attacks. Primarily, they have faced a challenge in this research because of the data quality and the huge number of missing values. To overcome the challenge, they have first replaced the data with the missing values (as the data noise have been found as a special character) and finally replaced the missing values by average. After preprocessing of the data, they have applied six classifiers of machine learning and attained 86% accuracy using Support Vector Machine for the identification of heart disease.

Kavitha et al. (2021) have proposed a hybrid model using machine learning classifiers for the identification of heart disease. They have selected the heart disease dataset from UCI for the detection of heart attacks. They have initially faced the challenge as they have found that not cases, single classifiers will be enough to detect heart disease sophistically. They have understood this as they have applied the individual classifier where they have got lower accuracy. To overcome this challenge, they have combined the classifiers (random forest and decision tree) and applied the data to classify and detect a heart attack. In this experiment, they have got 88% accuracy in the detection of heart attack using a hybrid machine learning approach.

Mohan et al. (2021) have applied the classifiers of machine learning namely Decision Tree, Support Vector Machine and Naïve Bayes on the selected heart disease data from the UCI. They have faced the primary challenge to get higher accuracy because of the high amount of noise and missing values in the data. To overcome this challenge, they have first cleaned the data and normalized it. After that, they have applied the selected classifiers and detected heart disease. They have got 89% accuracy using a support vector machine.

Patil & Mathur (2021) have applied the bagging and boosting methods on the classifiers of machine learning to predict heart attacks. They have primarily selected the heart disease dataset from the UCI data vault and applied the selected classifiers. However, they have not attained significant results and accuracy in this application. To overcome the problem, they have applied bagging and boosting methods on the selected classifiers that are Random Forest, K-Nearest Neighbour, Decision Tree and Naïve Bayes. In this experiment, they have achieved 92% accuracy with the help of machine learning and the bagging & boosting methods.

## Comparison of Previous Research

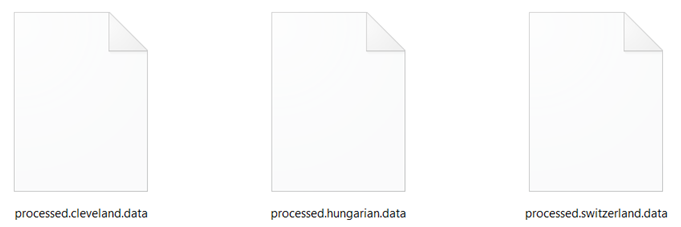
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Authors | *Method /Algorithm* | *Data* | *Problem Addressed* | *Proposed Solution* | *Accuracy* | *DOI* |
| Khan & Mondal (2021) | Support Vector Machine and Naïve Bayes | Heart Disease Data from UCI | They have faced the primary challenge to handle the noise of the data along with facing the less number of observations in the dataset | To overcome the challenge, they have preprocessed the data and cleaned it. Secondly, they have resampled the data to get more data instances | 72% | 10.1109/ICECE51571.2020.9393055 |
| Haq et al. (2021) | Six Classifiers | Heart Disease Data from UCI | They have faced a challenge in this research because of the data quality and the huge number of missing values | To overcome the challenge, they have first replaced the data with the missing values and finally replaced the missing values by average | 86% | 10.1109/ICCWAMTIP47768.2019.9067519 |
| Kavitha et al. (2021) | Hybrid Classifiers (Random forest and Decision tree) | Heart Disease Data from UCI | They have initially faced the challenge as they have found that not cases, single classifiers will be enough to detect heart disease sophistically | To overcome this challenge, they have combined the classifiers and applied the data to classify and detect a heart attack | 88% | 10.1109/ICICT50816.2021.9358597 |
| Mohan et al. (2021) | Decision Tree, Support Vector Machine and Naïve Bayes | Heart Disease Data from UCI | They have faced the primary challenge to get higher accuracy because of the high amount of noise and missing values in the data | To overcome this challenge, they have first cleaned the data and normalized it | 89% | 10.1109/ISCON52037.2021.9702314 |
| Patil & Mathur (2021) | Random Forest, K-Nearest Neighbour, Decision Tree and Naïve Bayes with bagging & boosting methods | Heart Disease Data from UCI | They have not attained significant results and accuracy in this application | To overcome the problem, they have applied bagging and boosting methods on the selected classifiers | 92% | https://doi.org/10.32628/IJSRSET218486 |

# Progress Summary

The progress of the research work wil be discussed in this section. Presently, the research has been initiated and the data has been collected. As the project has the aim to design the application with the machine learning backend for heart disease prediction, so, it has two significant segments. The first segment will contain the research issues from where the most effective model for heart disease prediction will be done with the highest accuracy. The second segment of the project will contain the development of an application where the users can input the symptom values to get the prediction result of whether the user has the possibility of a heart attack. Presently, the first segment of the project has been initiated and the primary data analysis has been done.

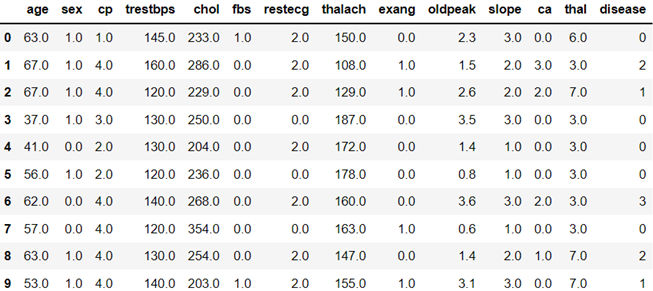
## Data Collection

Data is the main issue and the centre of interest in the research. In this project and research, the heart disease dataset has been selected from the UCI repository by following the brief literature review (UCI, 1988). It has been seen that many researchers have used the dataset and thus, it can be considered to be reliable data. So, this has been selected for the research, The data has been collected in the form of three DATA files and those are shown below:



**Fig-1 Collected Data Files**

These three data files have been combined and the final data has been formed. This data will be used for the purpose of analytics and prediction of heart disease. The data is shown below:



**Fig-2: Processed Data**

In this processed data, it has been seen that the target feature that is “disease” is with five labels. So, the labels of the target features have been transformed into binary labels. This has been done as per the aim of the research which is to detect whether a user has the possibility of a heart attack which should be in the binary labels. The final data is shown below:



**Fig-3: Final Data**

### Data Background

The selected data contains 720 instances for the patients with 13 symptoms and one target feature. So, in the data, there are a total of 14 features. The data has been collected for three different locations namely Cleveland, Hungary and Switzerland. Out of 720 instances, Cleveland contains 303 observations, Hungarian contains 294 observations and Switzerland contains 123 observations.

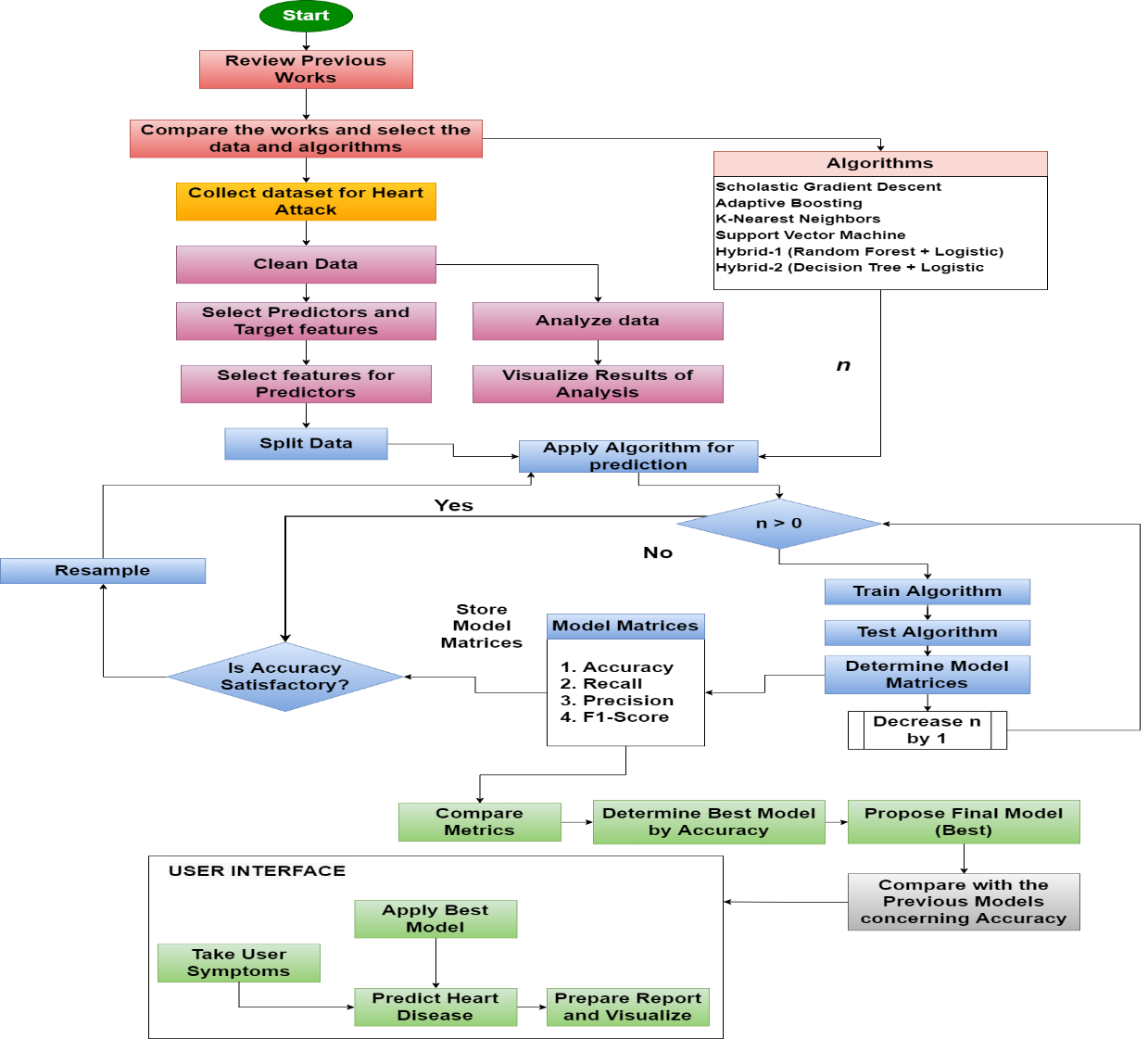
### Attributes of Data

The attributes of the data are discussed as follows:

* age: The attribute of heart disease data indicates the age of the patient who has been taken into observation to identify heart disease.
* sex: The attribute of heart disease data indicates the gender of the patient who has been taken into observation to identify heart disease.
* cp: The attribute of the heart disease dataset indicates the type of chest pain that the patients have faced and observed. This is an important attribute that accelerates the identification process of heart disease.
* trtbps: The attribute of the heart disease dataset that implied the blood pressure of the patients while they were taking rest.
* chol: The attribute of the heart disease dataset indicates the value of cholesterol in the blood of the patient who has been taken into observation to identify heart disease.
* fbs: The attribute of the heart disease dataset indicates the blood sugar level in the body of the patients who has been taken into observation to identify heart disease.
* restecg: The attribute of the heart disease dataset indicates the electrocardiographic results at the resting condition of the patients who has been taken into observation to identify heart disease.
* thalachh: The attribute of the heart disease dataset indicates the maximum heart rate that has been achieved from the patients while at the time of testing.
* exng: The attribute of the heart disease dataset indicates the value of agina while at the time exercise by the patients.
* oldpeak: The attribute of the heart disease dataset indicates the ST depression level that has been induced in the body of the patient at the time of exercise.
* slope: The attribute of the heart disease dataset that indicates the depression level of the patient who has been taken into observation to identify heart disease.
* ca: The attribute of the heart disease dataset indicates the count of the major vessels of the patients who were under the test for the determination of heart attack.
* thal: The attribute of the heart disease dataset indicates the type of defect that has been found in the heart of the patients who has been taken into observation to identify heart disease.
* disease: This is the dependent attribute of the data and will be used as the target data to predict whether the patient has heart disease. The attribute contain two labels that are labelled as 1 for heart disease and 0 for a healthy heart.

## Proposed Methodology

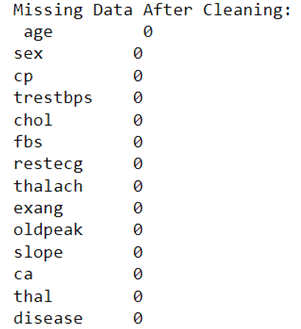
The methodology for the research and the design of the user interface (application) to predict heart disease is shown below:



**Fig-4:Proposed Methodology**

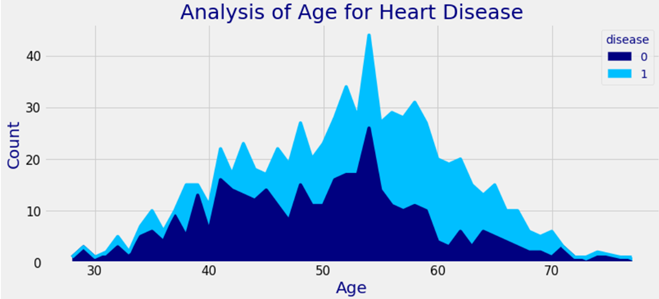
## Data Analysis

The data has been read and analysed initially. In this context, the missing values of the data have been first detected and removed as follows:



**Fig-5: Missing Values**

After that, the data analysis has been done for the age variable to check the relationship between the patient’s age and the possibility of a heart attack. The outcome is shown below:



**Fig-6: Analysis of Age and Heart attack**

# Issues for Research

## Ethical and Legal Considerations

The ethical and legal issues for the detection of heart disease are discussed below:

1. The research will be done based on the data that will be selected from an open-source data vault.
2. No permission will be required to access and use the data in the research.
3. No payment will be required to access and use the data in the research.
4. The research will be done by selecting the open-source programming tool and thus, no subscription fee will be required there.
5. The research will be conducted based on the features found in the data.

## Professional Issue

The professional issues for the detection of heart disease are discussed below:

1. The research and the design of the user interface will be done professionally.
2. The report of the project will be written by following the guidelines provided by the university.
3. The citation and referencing will be done as per the rule of the university.
4. If an idea will be taken from any external resources, the reference will be cited properly

## Social Issue

The social issues for the detection of heart disease are discussed below:

1. No manual labour will be incorporated in this research for the collection of healthcare data.
2. The resesrch will help to identify the heart disease based on the symptoms of the patients.

# Research Planning

The plan of the research will be discussed in this section regarding the selection of tools, technology, algorithms etc.

## Selection of Tool

Python 3 programming language will be chosen as the tool for programming and designing the user interface through which heart disease can be predicted. The important reasons for which this tool has been taken into choice are discussed below:

1. It is the cross-platform programming language
2. It is an interpreted programming language
3. It is the open-source programming language
4. It has a large library support

## Selection of Technologies

The technologies that have been selected in this project are as follows:

1. Data analysis and Visualization
2. Statistical analysis
3. Data Analytics
4. Machine Learning and Classification
5. Graphical User Interface

## Selection of Algorithms

The following algorithms in machine learning are planned to be applied in this research for the detection of heart disease:

1. Scholastic Gradient Descent
2. Adaptive Boosting
3. Support Vector Machine
4. Logistic Regression
5. Random Forest
6. Decision Tree

# Project Planning

The complete plan of the project has been mentioned below in the form of the Gantt Chart:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Day 1-10** | **Day 11-20** | **Day**  **21-35** | **Day**  **36-40** | **Day**  **41-45** | **Day**  **45-65** | **Day**  **66-70** | **Day 71-74** |
| **Introduction** |  |  |  |  |  |  |  |  |
| **Select algorithms and data ideas from previous researches** |  |  |  |  |  |  |  |  |
| **Collect the heart disease data and analyse** |  |  |  |  |  |  |  |  |
| **Select machine learning classifiers (algorithms)** |  |  |  |  |  |  |  |  |
| **Train and test the algorithms and predict disease in the heart and compared the performances of all classifiers to find the best model** |  |  |  |  |  |  |  |  |
| **Design the GUI and apply the best model to predict heart disease by symptom input** |  |  |  |  |  |  |  |  |
| **Present the detailed result of heart disease detection** |  |  |  |  |  |  |  |  |
| **Concluding the research** |  |  |  |  |  |  |  |  |

# References

Abdulhadi, N. & Al-Mousa, A., 2021. Diabetes Detection Using Machine Learning Classification Methods. *IEEE Access,* pp. 715-719.

Ahmed, M. R. et al., 2021. *A Cloud Based Four-Tier Architecture for Early Detection of Heart Disease with Machine Learning Algorithms.* [Online]   
Available at: https://ieeexplore.ieee.org/document/8781022

Atallah, R. & Al-Mousa, A., 2021. Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method. *IEEE Access,* 9 Nov.pp. 660-669.

Biswas, R. et al., 2020. Identification of Pathological Disease in Plants using Deep Neural Networks - Powered by Intel® Distribution of OpenVINO. *Computing Analytics and Networks (Indo-Taiwan ICAN) 2020 Indo – Taiwan 2nd International Conference,* pp. 45-48.

Chitnis, G., Bhanushali, V. & Ranade, A., 2020. A Review of Machine Learning Methodologies for Dental Disease Detection. *IEEE Access,* pp. 145-151.

Haq, A. U. et al., 2021. Identifying the Predictive Capability of Machine Learning Classifiers for Designing Heart Disease Detection System. *IEEE Access,* 9 Nov.pp. 1-5.

Katarya, R. & Srinivas, P., 2020. Predicting heart disease at early stages using machine learning: A survey. *In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE.,* pp. pp. 302-305.

Kavitha, M. et al., 2021. Heart Disease Prediction using Hybrid machine Learning Model. *6th International Conference on Inventive Computation Technologies (ICICT),* pp. 1-5.

Keya, M. S. et al., 2021. Measuring the Heart Attack Possibility using Different Types of Machine Learning Algorithms. *IEEE Access,* 9 Nov.pp. 852-863.

Khan, M. I. H. & Mondal, M. R. H., 2021. Effectiveness of Data Driven Diagnosis of Heart Disease. *IEEE Access,* 9 Nov.pp. 3998-4011.

Mohan, N., Jain, V. & Agrawal, G., 2021. Heart Disease Prediction Using Supervised Machine Learning Algorithms. *5th International Conference on Information Systems and Computer Networks (ISCON),* pp. 1-4.

Patil, M. D. & Mathur, D. H., 2021. Prediction of Cardiovascular Disease on Different Parameters Using Machine Learning. *International Journal of Scientific Research in Science, Engineering and Technology,* pp. 52-57.

Schulte, R., Krug, J. & Rose, G., 2018. Identification of a signal for an optimal heart beat detection in multimodal physiological datasets. *Computing in Cardiology,* pp. 1-6.

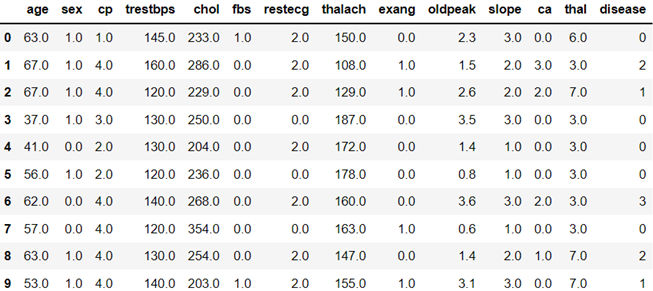
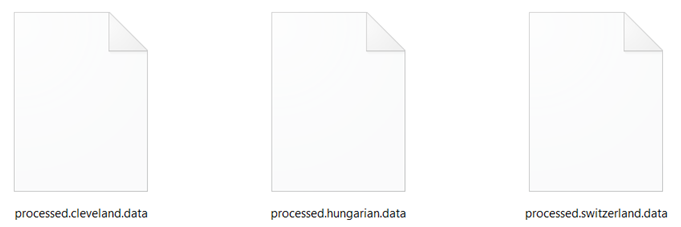
Tao, R. et al., 2018. Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods. *IEEE Transactions on Biomedical Engineering, 66(6).,* pp. 1658-1667.

UCI, 1988. *Heart Disease Data Set.* [Online]   
Available at: https://archive.ics.uci.edu/ml/datasets/heart+disease  
[Accessed 2022].

# Appendix

**Data**

**Data Link**: https://archive.ics.uci.edu/ml/datasets/heart+disease



**Proposed Method**

