# Cardiovascular Disease Prediction Using Machine Learning Models

Atharv Nikam
*School of Computer Science and Engineering*
*Dr. Vishwanath Karad MIT World Peace University*
Pune, India
https://orcid.org/0000-0002-4988-3479

Sanket Bhandari
*Department of Computer Engineering*
*Pimpri Chinchwad College of Engineering*
Nigdi Pune, India
https://orcid.org/0000-0003-0179-7771

Aditya Mhaske
*School of Computer Science and Engineering*
*Dr. Vishwanath Karad MIT World Peace University*
Pune, India
https://orcid.org/0000-0001-7735-1579

Shamla Mantri
*School of Computer Science and Engineering*
*Dr. Vishwanath Karad MIT World Peace University*
Pune, India
https://orcid.org/0000-0001-9155-2860

2020 IEEE Pune Section International Conference (PuneCon) | 978-1-7281-9600-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/PuneCon50868.2020.9362367

*Abstract*— **Cardiovascular diseases are one of the most vital causes of fatality. Cardiovascular disease prediction is a critical challenge in the area of clinical data analysis. Machine learning and Neural Networks are more promising in assisting decide and predict from the massive data produced by healthcare. We have noted different features had used in recent developments of the machine learning model. In this paper, we proposed machine learning techniques to predict cardiovascular disease using features. BMI is one of the highlighting features we used for prediction. BMI is important in predicting cardiovascular disease. The main focus of the article is the effect of BMI on the prediction of cardiovascular disease. The model has proposed with different features as well as regression and classification techniques. We conclude that BMI is a significant factor while predicting cardiovascular disease.**

***Keywords—Machine Learning (ML), Cardiovascular Diseases Prediction (CVD), Body Mass Index (BMI), Decision tree classifier, Neural Network (NN)***

## I. Introduction

According to WHO, every year, twelve million losses of life happen globally due to Heart diseases. More than half of the mortality in the United States and other nations happened due to cardiovascular diseases. It is one of the main reasons for death in various countries. It has considered the primary reason for death in adults. Coronary heart disease and Cardiovascular disease are classes of heart diseases. The term cardiovascular disease includes several contingencies that affect the heart, blood vessels, and circulating and pumping blood throughout the body. Cardiovascular diseases result in several illnesses, disabilities, and death. The diagnosis of diseases is an essential and complex responsibility in medicine.

In developed countries, cardiovascular diseases are one of the leading reasons for death worldwide. A high risk of CVD is also because of factors like high blood pressure, obesity, stress, diabetes, alcohol, cholesterol, and smoking that can prevent also treat with salutary practices changes. However, other risk factors can be ungovernable such as age, gender, and history of the family. Initial phase detection of cardiovascular diseases can prevent the mortality rate; people are not conscious of the causes of the cardiovascular disease earlier due to the absence of awareness. Health care organizations are trying to diagnose the disease at the initial phase. Most of the time, the disease is noticed at the last phase or after death. We aim to diagnose the disease at an initial phase [1-2]

Using machine learning algorithms, we can recognize the disease at an initial stage and help to cure disease with a conventional diagnosis. To develop the model, we have used different machine learning algorithms to resolve the problem; we also tried to find the significant factor while predicting cardiovascular disease. Various issues, such as absent value, outlier, and recognizing important dimensions, are well handled by decision trees.

## II. Literature Review

Various studies are present which focuses on heart disease prediction where the diagnosis is done by using different techniques of data mining.[3] As per the research done by the research group in [4], the decision tree classifier shows quite good performance as compared to all other models where the performance of a model is evaluated in terms of classification accuracy. In [5], the focus is on developing a system to help medical professionals to evaluate the risk of heart disease of a patient based on the patient's clinical data. In [12], Heart disease prediction is done using machine learning where the parameters used are Age, Sex, Blood Pressure, Heart Rate, Diabetes, Hyper cholesterol, Body Mass Index (obesity). In [13], ANN, KNN, k-means, and K-medoids algorithms are trained on the Cleveland dataset for heart disease.

Body Mass Index (BMI) is one of the most significant factors responsible for Heart Disease. Higher BMI in childhood increases the risk of Coronary Heart Disease in adulthood.[6].[7], shows a study based on the observations of BMI in 2.3 Million Adolescents and their Cardiovascular Death in Adulthood. Here the data is measured from 1967 through 2010 which concluded that Increased BMI in adolescence increases the risk of cardiovascular mortality in adulthood. The study conducted in [9], focuses on the effect of the higher fat mass index and BMI on the risk of various cardiovascular conditions, which provides evidence that higher BMI increases the risk of aortic valve stenosis.

For heart failure patients, the relation between BMI and mortality is U-shaped [10] having a BMI of 32 to 33 kg/ m2 as the lowest one. Also, the increase in BMI through puberty contributes to increased risk of adult stroke, adult IS and, ICH in men.[8][15] which signifies the effect of BMI on the risk of heart disease. [16] provides evidence about an association between the excess gain in BMI between childhood to adult ages and increased risk of CVD (Cardio Vascular Disease).

## III. Proposed Method

According to Figure 1, the system is implemented using five major steps in prediction cardiovascular disease.

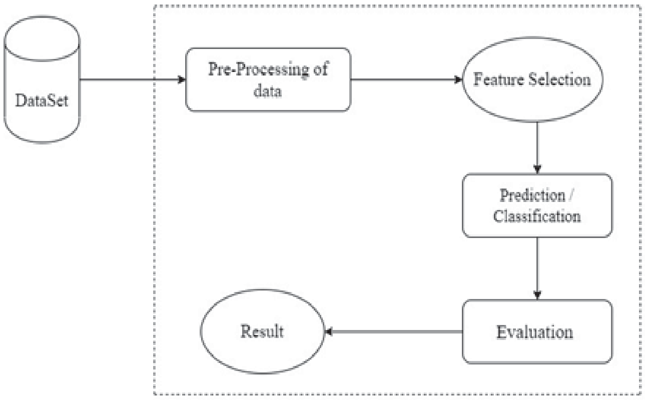

Fig. 1. Experiment Workflow with dataset

## IV. Method Explanation

### A. Data Description

There are three important features of input classified as Objective, Examination, and Subjective.

1. Objective features for factual information from patient;
2. Examination features for results collected after medical examination;
3. Subjective features for information collected from the patient.

TABLE I. Feature Description

| **Feature** | **Value Type** | **Variable Type (Feature classification)** | **Variable** |
|---|---|---|---|
| Age | int (days) | Objective | age |
| Glucose | 1: normal,<br>2: above normal,<br>3: well above normal | Examination | gluc |
| Height | int (cm) | Objective | height |
| Cardiovascular disease (Presence) | binary | Target Variable | cardio |
| Weight | float (kg) | Objective | weight |
| Physical activity | binary | Subjective | active |
| Gender | categorical code | Objective | gender |
| Alcohol intake | binary | Subjective | alco |
| Systolic blood pressure | int | Examination | ap_hi |
| Smoking | binary | Subjective | smoke |
| Diastolic blood pressure | int | Examination | ap_lo |
| Cholesterol | 1: normal,<br>2: above normal,<br>3: well above normal | Examination | cholesterol |

Features used in the implementation are:

Age, Glucose, Height, Cardiovascular disease (Presence), Weight, Physical activity, Gender, Alcohol intake, Systolic blood pressure, Smoking, Diastolic blood pressure, Cholesterol are the features of dataset used in model for description refer (table no.1) and BMI is added using calculation

### B. Data Pre-processing

The dataset is consisting of 12 rows and 70000 columns (patient records). After removing similar records, the remaining 68975 patient records used. For the features of the provided dataset, the binary classification and the multiclass parameter is proposed. To examine the presence or absence of the heart disease, the multiclass parameter is used. The value 1 indicates that the patient has heart disease. State 0 specifies that the heart disease is absent in the patient. The medical records are transformed into the detection values during the pre-processing. The data pre-processing of the patient records designated that the 34135 cases designate the value of 1 indicating that the heart disease is present. The value of 0 shown by the rest 34840 records signifies the absence of the heart disease.

### C. Feature Selection and Adding a new feature

A new feature BMI is appended to the dataset. The BMI is computed by using the height data and weight data for each patient. It is a weight divided by the square of height. In this dataset, height measured in centimeter and weight measured in kilograms. To calculated BMI first, we have to convert height from centimeters to meter and then proceed for further calculations. Selection of feature BMI used to build a model with more accuracy [11]

Formula:

$$[\mathrm{bmi}] = [\mathrm{weight(kg)}] / [\mathrm{height(m)}]^2 \qquad (a)$$

While developing model, the formula used is:

["bmi"] = ["weight"]/ (["height"]/100)**2 (b)

### D. Experimental setup for evaluation

We have split the dataset into an 80:20 ratio. That is, the training set size is 80%, and the testing set size is 20% of the entire dataset. The training set used to develop a model, while the testing set utilized to assess the predictive model's performance.

We developed our first model (model A) with the given dataset and did not make any changes while training Second model (model B) we add column BMI. All train and test accuracies as well as train and test scores are in percentages.

- Model A (with given dataset) divided into train_A and test_A and their accuracies are denoted by Score_train_A , Score_test_A respectively

- Model B (with addition of column BMI) divide into train_B and test_B and their accuracies are denoted by Score_train_B , Score_test_B respectively

### *E. Classification*

Classification means classifying the data in different groups based on the similarities present in different data points. Here classification is used in the prediction of heart disease. Various machine learning models are available, but in the proposed method, any one of the following algorithms or models can be used [12] The given problem is a classification and regression problem. We are using various algorithms to find (or predict) the relation between the target variable (i.e. survived or not) with other variables (BMI, Gender, Age, and more). The following algorithms are used

(1) Logistic Regression: It is a predictive analysis technique which is used when the target variable is dichotomous (binary). Logistic Regression model explains the relationship between one dependent binary variable and one or more independent variables. It predicts the probability of the target value.

(2) k-Nearest Neighbors algorithm: k-Nearest Neighbors algorithm is a classification algorithm.[13] The class of a particular data point is determined based on the class which is most common among its k nearest neighbors where k is a small positive integer.

(3) Naive Bayes: It is a set of supervised learning algorithms based on applying Bayes' theorem. The naïve assumption being the conditional independence between every pair of features.[14]

(4) Neural Networks: A neural network is a series of algorithms which recognizes underlying relations in a training data set through a process that vaguely mimics the way of working of human brain. Basically, neural networks are a system of neurons which are either artificial or organic in nature. Neural network adapts to the changing input which allows it to generate best output.

(5) Decision Tree Classifier: This classifier model applies a Decision Tree as a predictive model. It organizes the characteristics (tree branches) to inferences about the target value (tree leaves). The classification trees are the tree models in which the target parameter can acquire a finite set of values. In these tree frameworks, the class labels are signified by the leaves, and the branches describe the concurrences of features that guide to those class labels. The regression trees are the decision trees in which the target parameter can take the continuous values (generally real numbers).

(6) XGB Classifier with HyperOpt: The XGBoost classifier is an ensemble tree approach. The principle of boosting weak learners (CARTs generally) is implemented by this approach. The gradient descent architecture is used by this principle. Gradient Boosting Machines (GBM) structure enhances the XGBoost through the systems advancements and algorithmic improvements. Here we are tuning the hyperparameters of the XGB Classifier model using the HyperOpt and 10-fold cross-validation.

(7) LGBM Classifier with HyperOpt: The Light GBM classifier is based on the decision tree algorithms. It is a disseminated, high-speed, and highly efficient gradient boosting structure. The other boosting procedures split the tree level-wise or depth-wise, but the LGBM method divides the tree leaf-wise with the best fit. This leaf-wise method can decrease more loss than level-wise procedures when growing on the same leaf. Thus, the Light GBM algorithm results in greater precision, which can seldomly be accomplished by any of the present boosting methods. Here we are tuning the hyperparameters of the LGBM Classifier model using the HyperOpt and 10-fold cross-validation.

### *F. Evaluation*

The classification models were developed using 13 features. The train and test accuracies were calculated for each model. The evaluation of the model was based on seven different classifiers. After Classifying both models, A and B, BMI is a significant factor to increase the accuracy of the model.

## V. Result

To evaluate the efficiency of the classification, the predicted type was compared with the original type. In the propounded methodology, for classification objective, the dataset was examined by seven different classifiers mentioned above (1 -7). The decision tree classifier generated the highest accuracy of 73.12% as shown in Table II. Table II represents train and test accuracies. Here because of BMI, accuracy of model B is Increased in almost every model. Hence BMI plays an essential role in prognosticating cardiovascular disease. The decision tree classifier tested with the highest accuracy. To find out which technique predict cardiovascular disease with more accuracy we used different algorithms such as, Neural Networks, K-Nearest Neighbors, Naive Bayes, Logistic Regression, Decision Tree Classifier, XGB classifier, and LGBM classifier and their result are shown in Table II

Moreover, as shown in Figure-4, 'BMI' is amongst the most prominent characteristics to prognosticate cardiovascular disease. There is a bit more difference in testing and training accuracies of LGBM Classifier and XGB Classifier because of overfitting of models. And by the result of the XGB Classifier, we assigned the importance of the features (Figure-4). The Figure-2 shows the performance comparison and the error between Model A and Model B for each classification method. Feature Importance (Figure-4) is defined according to XGB Classifier.

TABLE II. ACCURACY OF MODEL WITHOUT CONSIDERING BMI

| No. | Model | Train Accuracy (%) | Test Accuracy (%) | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| 1. | Decision Tree | 72.76 | 72.89 | 0.73 | 0.73 | 0.73 |
| 2. | LGBM | 72.85 | 72.60 | 0.74 | 0.73 | 0.72 |
| 3. | XGB | 77.35 | 72.81 | 0.73 | 0.73 | 0.73 |
| 4. | Neural Network | 69.15 | 68.99 | 0.71 | 0.69 | 0.69 |
| 5. | Logistic Regression | 64.66 | 64.40 | 0.64 | 0.64 | 0.64 |
| 6. | K-Nearest Neighbors | 72.35 | 71.13 | 0.72 | 0.71 | 0.71 |
| 7. | Naive Bayes | 70.82 | 70.50 | 0.72 | 0.71 | 0.70 |

TABLE III. ACCURACY OF MODEL WITH CONSIDERING BMI

| No. | Model | Train Accuracy (%) | Test Accuracy (%) | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| 1. | Decision Tree | 72.68 | 73.13 | 0.73 | 0.73 | 0.73 |
| 2. | LGBM | 80.12 | 72.95 | 0.73 | 0.73 | 0.73 |
| 3. | XGB | 83.52 | 72.82 | 0.73 | 0.73 | 0.73 |
| 4. | Neural Network | 71.85 | 72.22 | 0.72 | 0.72 | 0.72 |
| 5. | Logistic Regression | 71.65 | 71.94 | 0.72 | 0.71 | 0.71 |
| 6. | K-Nearest Neighbors | 72.13 | 70.60 | 0.72 | 0.71 | 0.71 |
| 7. | Naive Bayes | 70.27 | 70.26 | 0.72 | 0.71 | 0.70 |

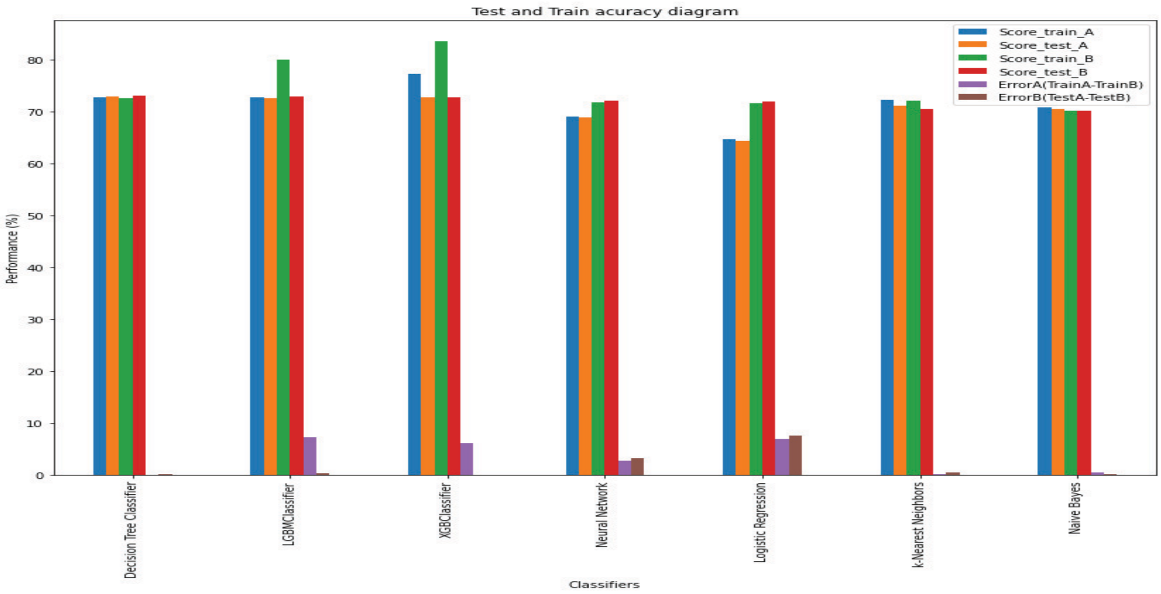

Fig. 2. Error between Model A and Model B *(Prediction and Detection)*

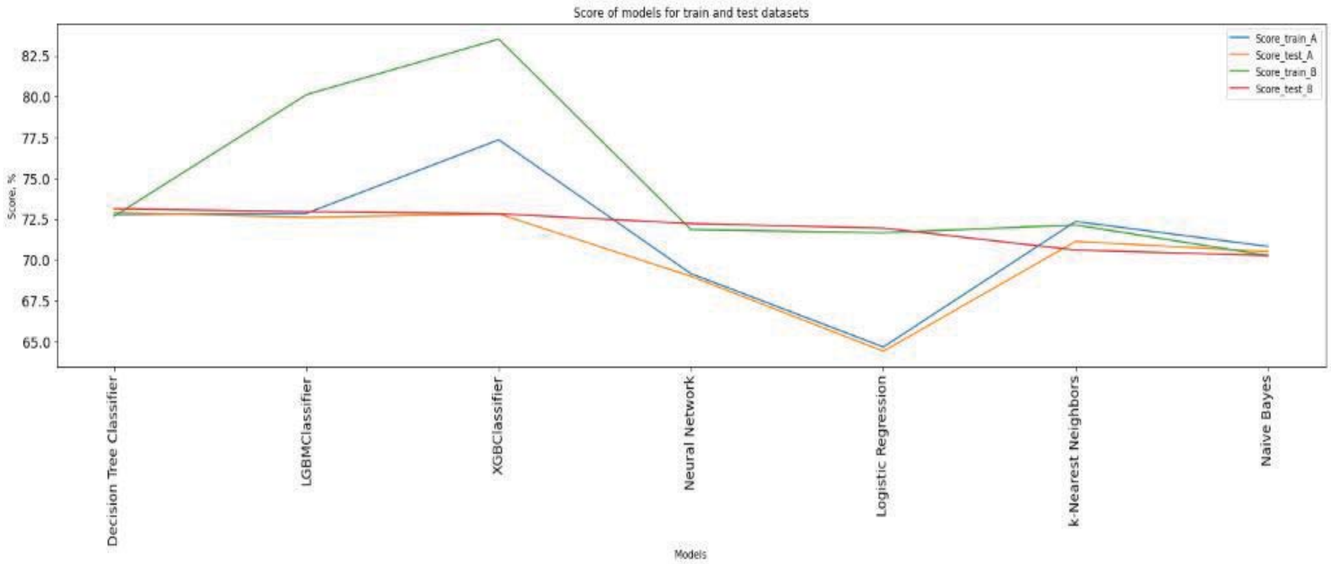

Fig. 3. Comparison after assessing test and train sets of both models

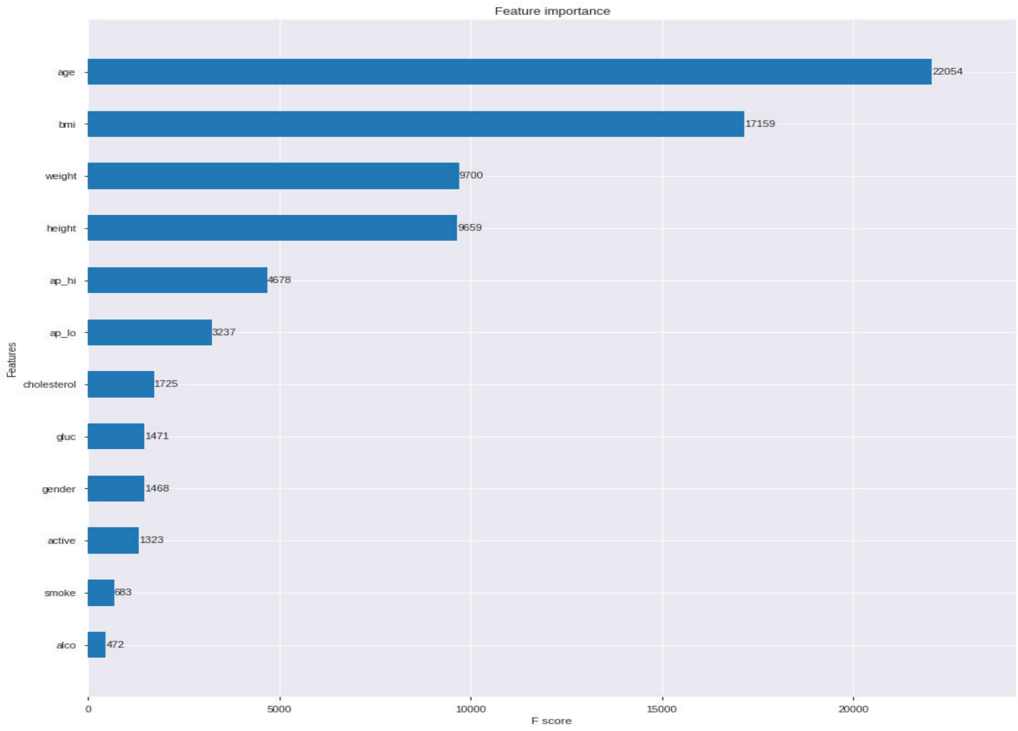

Fig. 4. Feature Importance

## VI. Conclusion

Heart disease prediction is essential as well as challenging work in the medical Field. Nevertheless, the mortality rate can be reduced if the disease is recognized at the initial stages, and precautions and proper treatment are possible. This paper illustrates various automated computerized Cardiovascular Disease Prediction methodologies, which can be performed by Supervised Learning plus Classification and Regression methods. The algorithms are tested using various features.

Accurate prognostication of the diseases is the goal of the proposed method. The decision classifier approach proved to be very efficacious to predict the diseased using features like age, BMI, cholesterol, and more. Adding feature BMI improved the accuracy of prediction. Thus, by assessing the results, the suggested approach generates a more precise prediction of cardiovascular diseases. The Decision tree algorithm found out more efficient and tested with the highest accuracy. XGB classifier used to identify the importance of each feature in the prognostication of heart disease

## VII. Discussion

In the proposed models, for some algorithms, the testing accuracy is slightly greater than training accuracy. Generally, the testing accuracy should be less than that of the training accuracy. As test data is the data unseen by the model, and train data is the data that the model uses to train itself. Also note that the difference is minimal. For the Decision Tree model, without considering BMI, this difference is 0.13. While considering BMI, this difference for Decision Tree model is 0.45, for Neural Network, it is 0.37, and for Logistic Regression model, it is 0.29. So, a couple of hard train samples could create this bias.

## REFERENCES

[1] N. Townsend, M. Nichols, P. Scarborough and M. Rayner, "Cardiovascular disease in Europe-epidemiological update 2015", *European heart journal*, vol. 36, no. 40, pp. 2696-2705, 2015.

[2] Y. Roche, Risques médicaux au cabinet dentaire en pratique quotidienne: Identification des patients évaluation des risques prise encharge: prévention précautions, 2010.

[3] S. Chaitrali, Dangare and S. Apte Sulabha, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", *International Journal of Computer Applications (0975-888)*, vol. 47, no. 10, pp. 44-48, June 2012.

[4] Purushottam, K. Saxena and R. Sharma, "Efficient heart disease prediction system using decision tree," International Conference on Computing, Communication & Automation, Noida, 2015, pp. 72-77, doi: 10.1109/CCAA.2015.7148346.

[5] A. H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng and E. J. Lin, "HDPS: Heart disease prediction system," 2011 Computing in Cardiology, Hangzhou, 2011, pp. 557-560.

[6] Baker JL, Olsen LW, Sorensen TI. Childhood body-mass index and the risk of coronary heart disease in adulthood. N Engl J Med. 2007;357(23):2329–37.

[7] Twig G, Yaniv G, Levine H, Leiba A, Goldberger N, Derazne E, Ben-Ami Shor D, Tzur D, Afek A, Shamiss A, et al. Body-mass index in 2.3 million adolescents and cardiovascular death in adulthood. N Engl J Med. 2016;374(25):2430–40.

[8] C. Ohlsson, M. Bygdell, A. Sonden, C. Jern, A. Rosengren, J.M. Kindblom BMI increase through puberty and adolescence is associated with risk of adult stroke Neurology, 89 (2017), pp. 363-369

[9] Susanna C Larsson, Magnus Bäck, Jessica M B Rees, Amy M Mason, Stephen Burgess, "Body mass index and body composition in relation to 14 cardiovascular conditions" in UK Biobank: a Mendelian randomization study, European Heart Journal, Volume 41, Issue 2, 7 January 2020, Pages 221–226,

[10] Zhang, J., Begley, A., Jackson, R. *et al.* Body mass index and all-cause mortality in heart failure patients with normal and reduced ventricular ejection fraction: a dose–response meta-analysis. *Clin Res Cardiol* 108, 119–132 (2019).

[11] A. M.A. and P. A. Thomas, "Comparative Review of Feature Selection and Classification modeling," 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, 2019, pp. 1-9, doi: 10.1109/ICAC347590.2019.9036816.

[12] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.

[13] O. Terrada, B. Cherradi, A. Raihani and O. Bouattane, "Classification and Prediction of atherosclerosis diseases using machine learning algorithms," 2019 5th International Conference on Optimization and Applications (ICOA), Kenitra, Morocco, 2019, pp. 1-5, doi: 10.1109/ICOA.2019.8727688.

[14] F. Mendonca, R. Manihar, A. Pal and S. U. Prabhu, "Intelligent Cardiovascular Disease Risk Estimation Prediction System," 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, 2019, pp. 1-6, doi: 10.1109/ICAC347590.2019.9036738.

[15] B. S. Larsen, S. Winther, M. Böttcher, L. Nissen, J. Struijk and S. E. Schmidt, "Correlations of first and second heart sounds with age, sex, and body mass index," 2017 Computing in Cardiology (CinC), Rennes, 2017, pp. 1-4, doi: 10.22489/CinC.2017.141-408.

[16] Bjerregaard, L., Adelborg, K. and Baker, J., 2020. Change in body mass index from childhood onwards and risk of adult cardiovascular disease. *Trends in Cardiovascular Medicine*, 30(1), pp.39-45.

[17] M. Abdar, U. R. Acharya, N. Sarrafzadegan and V. Makarenkov, "NE-nu-SVC: A New Nested Ensemble Clinical Decision Support System for Effective Diagnosis of Coronary Artery Disease," in IEEE Access, vol. 7, pp. 167605-167620, 2019, doi: 10.1109/ACCESS.2019.2953920.