B.M.S COLLEGE OF ENGINEERING BENGALURU
Autonomous Institute, Affiliated to VTU

A Technical Seminar Report based on review of Research Publication / Patent

HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

Submitted in partial fulfillment for the award of degree of

Bachelor of Technology
in
Computer Science and Engineering

Submitted by:
MADHUKESH BHAT M
1BM18CS051

Work carried out at

Internal Guide

Sheetal V A
Assistant Professor
B.M.S College of Engineering

Department of Computer Science and Engineering
B.M.S College of Engineering
Bull Temple Road, Basavanagudi, Bangalore 560 019

2019-2020

B.M.S COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

I, Madhukesh Bhat M (1BM18CS051) student of 7th Semester, B.E, Department of Computer Science and Engineering, B.M.S College of Engineering, Bangalore, hereby declare that, this technical seminar entitled "**HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES**" has been carried out under the guidance of **SHEETAL V A,** Assistant Professor, Department of CSE, BMS College of Engineering, Bangalore during the academic semester Jan-May 2020. I also declare that to the best of our knowledge and belief, the technical seminar report is not from part of any other report by any other students.

Signature of the Candidate

Madhukesh Bhat M (1BM18CS051)

B.M.S COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Technical Seminar titled "**HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES"** has been carried out by Madhukesh Bhat M (1BM18CS051) during the academic year 2021-2022.

Signature of the guide
SHEETAL V A
Assistant Professor
Department of Computer Science and Engineering
BMS College of Engineering, Bangalore

ABSTRACT

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyse huge complex medical data, helping healthcare professionals to predict heart disease. This research paper presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This model aims to envision the probability of developing heart disease in the patients. It can also be helpful to the medical practitioners at their clinic as decision support system. The results portray that the highest accuracy score is achieved with K-nearest neighbor.

CHAPTER 1: INTRODUCTION

1.1 Overview

Healthcare is one of the primary focus for humanity. According to WHO guidelines, good health is the fundamental right for individuals. It is considered that appropriate health care services should be available for regular checkup of one's health. Almost 31% of all deaths are due to heart related disease in all over the world. Early detection and treatment of several heart diseases is very complex, especially in developing countries, because of the lack of diagnostic centers and qualified doctors and other resources that affect the accurate prognosis of heart disease. With this concern, in recent times computer technology and machine learning techniques are being used to make medical aid software as a support system for early diagnosis of heart disease. Identification of any heart related illness at primary stage can reduce the death risk. Various ML techniques are used in medical data to understand the pattern of data and making prediction from them. Healthcare data are generally massive in volumes and complex in structure. ML algorithms are capable to handle the big data and mine them to find the meaningful information. Machine Learning algorithms learn from past data and do prediction on real time data. This sort of ML framework for coronary illness expectation can encourage cardiologists in taking quicker actions so more patients can get medicines within a shorter timeframe, thus saving large number of lives.

Machine Learning is a branch of AI research and has become a very popular aspect of data science. The Machine Learning algorithms are designed to perform a large number of tasks such as prediction, classification, decision making etc. To learn the ML algorithms, training data is required. After the learning phase, a model is produced which is considered as an output of ML

algorithm. This model is then tested and validated on a set of unseen real time test dataset. The final accuracy of the model is then compared with the actual value, which justify the overall correctness of predicted result.

Lots of efforts has already been done to predict the heart disease using the ML algorithms by many researchers, but this is an additional effort to do the experiment on benchmarking UCI heart disease prediction dataset while comparing the five popular ML technique to check the most accurate ML technique.

Various data mining techniques such as regression, clustering, association rule and classification techniques like Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine and K-nearest neighbor are used to classify various heart disease attributes in predicting heart disease. In this research, a dataset from the UCI repository have been taken. The classification model is developed using classification algorithms for prediction of heart disease. A discussion of algorithms used for heart disease prediction and comparison among the existing systems is made here. Further research and advancement possibilities is also mentioned in here.

1.2 Motivation

Over the last decade, heart disease or cardiovascular remains the primary basis of death worldwide. An estimate by the World Health Organization, that over 17.9 million deaths occur every year worldwide because of cardiovascular disease, and of these deaths, 80% are because of coronary artery disease and cerebral stroke. The vast number of deaths is common amongst low and middle-income countries. Many predisposing factors such as personal and professional habits and genetic predisposition accounts for heart disease. Various habitual

risk factors such as smoking, overuse of alcohol and caffeine, stress, and physical inactivity along with other physiological factors like obesity, hypertension, high blood cholesterol, and pre-existing heart conditions are predisposing factors for heart disease. The efficient and accurate and early medical diagnosis of heart disease plays a crucial role in taking preventive measures to prevent death. This leads to the proposed system which determines the best Machine Learning algorithm to predict the occurrence of heart disease in patients beforehand.

1.3 Objective

Heart disease affects millions of people, and it remains the chief cause of death in the world. Medical diagnosis should be proficient, reliable, and aided with computer techniques to reduce the effective cost for diagnostic tests. Data mining is a software technology that helps computers to build and classify various attributes. The objective of this research is to use different classification techniques to predict heart disease and determine the best of them.

CHAPTER 2: LITERATURE SURVEY

The work of Devansh Shah, et. al. [1] presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes and K-nearest neighbor. Of these algorithms, K-Nearest Neighbor classifier turned out to be the best one with an accuracy of 78.94% on testing set with K = 2 neighbors.

In the research [2], Rohit Bharthi et al. proposed three methods in which comparative analysis was done and promising results were achieved. The conclusion which was found is that machine learning algorithms performed better in this analysis. For the 13 features which were in the dataset, K-Neighbors classifier performed better in the ML approach when data preprocessing is applied.

In the research of Vijeta Sharma et. al., [3] Machine Learning algorithms such as Random Forest, Support Vector Machine (SVM), Naive Bayes and Decision tree have been used for the development of model. In the research they have also tried to find the correlations between the different attributes available in the dataset with the help of standard Machine Learning methods and then using them efficiently in the prediction of chances of Heart disease. Result showed that compared to other ML techniques, Random Forest gives more accuracy in less time for the prediction.

In this paper [4], the traditional and proposed system was implemented by Emrana Kabir Hashi et. al., to predict Cleveland heart disease dataset. Without hyperparameters tuning, the LR, KNN, SVM, DT, and RF classifiers provide an accuracy rate of 88.52%, 90.16%, 88.52%, 81.97%, and 85.25% respectively. However, with the hyperparameters tuning approach, the LR, KNN, SVM, DT, and RF classifiers in the refined set takes the accuracy rate 90.16%, 91.80%, 90.16%, 86.89%, and 85.25% respectively. Hence, it is concluded that the proposed model is more efficient.

CHAPTER 3: METHODOLOGY/TECHNIQUES OR ALGORITHM USED

This research aims to foresee the odds of having heart disease as probable cause of computerized prediction of heart disease that is helpful in the medical field for clinicians and patients. To accomplish the aim, the use of various machine learning algorithms on the data set and dataset analysis have been discussed in this research. This result additionally depicts which attributes contribute more than the others to anticipation of higher precision. This may spare the expense of different trials of a patient, as all the attributes may not contribute such a substantial amount to expect the outcome.

Dataset Used:

For this study, a dataset from UCI Machine learning repository have been used. It comprises a real dataset of 300 examples of data with 14 various attributes (13 predictors; 1 class) like blood pressure, type of chest pain, electrocardiogram result, etc. Four algorithms to get reasons for heart disease and create a model with the maximum possible accuracy have been used in this research.

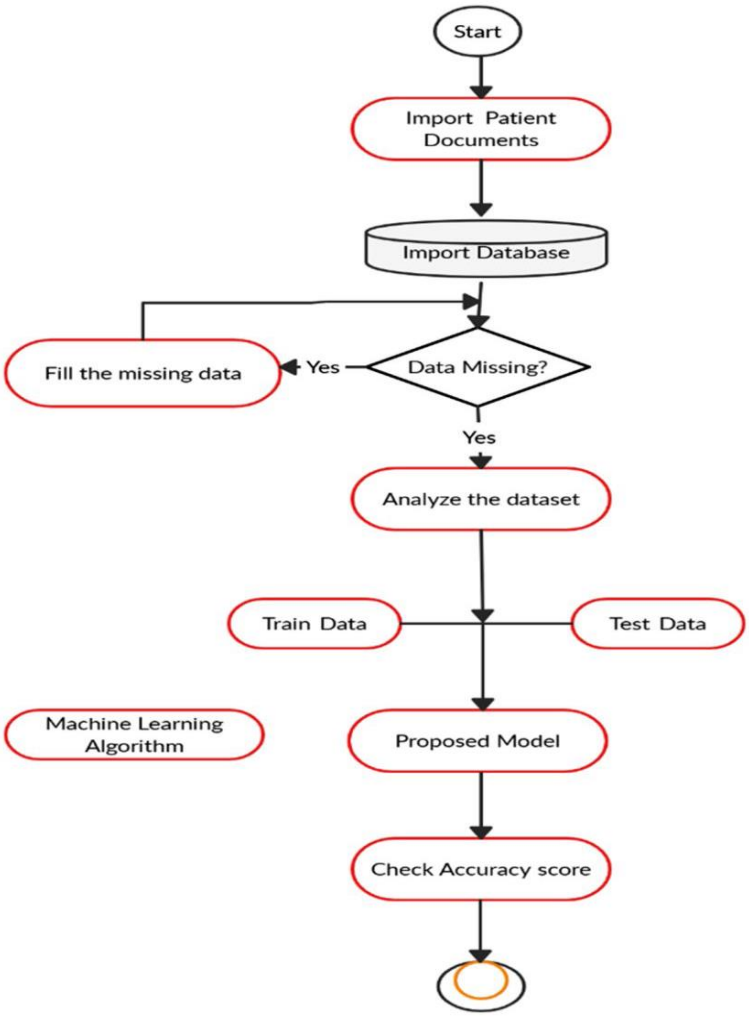
Data preprocessing:

The real-life information contains large numbers with missing and noisy data. These data are pre-processed to overcome such issues and make predictions vigorously. Figure 1 shown below explains the sequential chart of our proposed model.

- *Cleaning:* The collected data usually has noise and missing values. To get an accurate and effective result, these data need to be cleaned in terms of noise and missing values are to be filled up.

- *Transformation:* It changes the format of the data from one form to another to make it more comprehensible. It involves smoothing, normalization, and aggregation tasks.
- *Integration:* The data may not be acquired from a single source but varied sources, and it has to be integrated before processing.
- *Reduction:* The data gained are complex and require to be formatted to achieve effective results.

The data are then classified and split into training data set and test data set which is run on various algorithms to achieve accuracy score results.



Sequential chart of proposed model

Algorithms used:

1. Naïve Bayes' Classifier:

   Naïve Bayes classifier is a supervised algorithm. It is a simple classification technique using Bayes theorem. It assumes strong (Naive) independence among attributes. Bayes theorem is a mathematical concept to get the probability. The predictors are neither related to each other nor have correlation to one another. All the attributes independently contribute to the probability to maximize it. It is able to work with Naïve Bayes model and does not use Bayesian methods. Many complex real-world situations use Naive Bayes classifiers.

   $$P(X/Y) = \frac{P(Y/X) \times P(X)}{P(Y)}$$

   where, $P(X/Y)$ is the posterior probability, $P(X)$ is the class prior probability, $P(Y)$ is the predictor prior probability, $P(Y/X)$ is the likelihood, probability of predictor.

   Naïve Bayes is a simple, easy to implement, and efficient classification algorithm that handles non-linear, complicated data. However, there is a loss of accuracy as it is based on assumption and class conditional independence.

   An accuracy of 83.49% has been achieved in Naïve using all 13 attributes of Cleveland dataset.

2. Decision Tree:

   Decision tree is a classification algorithm that works on categorical as well as numerical data. Decision tree is used for creating tree-like structures. Decision tree is simple and widely used to handle medical dataset. It is easy to implement and analyse the data in tree-shaped graph. The decision tree model makes analysis based on three nodes.

   - Root node: main node, based on this all other nodes functions.

- Interior node: handles various attributes.
- Leaf node: represent the result of each test.

This algorithm splits the data into two or more analogous sets based on the most important indicators. The entropy of each attribute is calculated and then the data are divided, with predictors having maximum information gain or minimum entropy:

$$Entropy(S) = \sum_{i=1}^{c} -P_i log_2 P_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

The results obtained are easier to read and interpret. This algorithm has higher accuracy in comparison to other algorithms as it analyzes the dataset in the tree-like graph. However, the data may be over classified and only one attribute is tested at a time for decision-making.

An accuracy of 71.43% has been achieved by the decision tree by Chauhan et al., whereas the accuracy obtained was very poor about 42.8954% in Buoli H et al.

3. K – Nearest Neighbor (K-NN)

The K-nearest neighbors algorithm is a supervised classification algorithm method. It classifies objects dependant on nearest neighbor. It is a type of instance-based learning. The calculation of distance of an attribute from its neighbors is measured using Euclidean distance. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them, and is possible to fill the missing values of data using K-NN. Once the missing values are filled, various prediction techniques apply to the data set. It is possible to gain better accuracy by utilizing various combinations of these algorithms.

K-NN algorithm is simple to carry out without creating a model or making other assumptions. This algorithm is versatile and is used for classification, regression, and search. Even though K-NN is the simplest algorithm, noisy and irrelevant features affect its accuracy. In a study by Pouriyeh et al., 83.16% accuracy was achieved with value K = 9.

4. Random Forest Algorithm:

   Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the more number of trees give higher accuracy. The three common methodologies are:

   - Forest RI (random input choice)
   - Forest RC (random blend)
   - Combination of forest RI and forest RC

   It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets and more trees, results are unaccountable.

   Random forest algorithm has obtained an accuracy of 91.6% with Cleveland dataset in Xu S et al.. Using People's dataset, an accuracy of 97% was achieved.

5. Support Vector Machine (SVM):

   Support Vector Machine is a classification technique of Machine learning to, which is used to analyze data and discover patters in classification and regression analysis. SVM is typically mull over when data is characterized as two class problem. In this strategy, data is characterized by finding the best hyper plane that isolates all data points of one class to the other class. The

higher separation or edge between the two classes is, the better is the model, considered. The data points lying on limit of the margin are called as support vectors. The actual basis of SVM is mathematical methods used to design complex real-world problems. We have chosen SVM for this experiment because our dataset - Cleveland Heart Disease Dataset CHDD has multi class to predict based on various parameters. In SVM, the mapping of training data is to be done with a function called kernel (Kernels of SVM), these are - linear kernel, quadratic kernel, polynomial kernel, Radial Basis Function kernel, Multilayer Perceptron kernel, etc. Apart from the kernel's functionalities in SVM, few more methods are available such as quadratic programming, sequential minimal optimization, and least squares.

While building up the model with SVM, most challenging thing is kernel selection and method selection to evade the issue of overfitting and underfitting. Since our dataset is having enormous number of parameters and instances too. So, we had choice of selecting the RBF or linear kernel. Thus, final model developed by SVM requires tested and validated against actual data.

CHAPTER 4: DESCRIPTION OF TOOL SELECTED

1. Scikit-learn:

   Scikit-learn is the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

2. Numpy:

   NumPy is the package for scientific computing in Python. NumPy arrays eases advanced mathematical and other types of operations on large numbers of data.

3. Pandas:

   Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with relational or labeled data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.

4. Matplotlib:

   Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.

5. Jupyter Notebook:

   The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document.

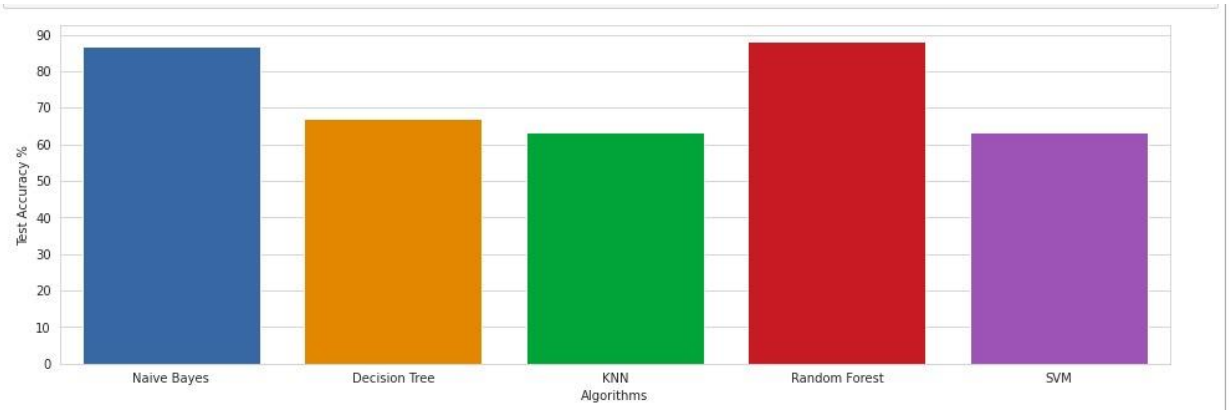
CHAPTER 5: DETAILED DESCRIPTION OF MODULES IMPLEMENTED

Aim of this research is to predict whether or not a patient will develop heart disease. This research was done on supervised machine learning classification techniques using Naïve Bayes, Decision Tree, Random Forest, K-nearest Neighbor and Support Vector Machines on UCI repository.

Dataset was classified and split into a training set (75%) and a test set (25%). Pre-processing of the data is done and supervised classification techniques mentioned above are applied to get accuracy score. The accuracy score results of different classification techniques were noted using for training and test data sets. Percentage accuracy scores are depicted in the below figure for different algorithms.

	model	train_accuracy	test_accuracy
0	Naive Bayes	81.06	86.84
1	Decision Tree	100.00	67.11
2	KNN	77.97	63.16
3	Random Forest	100.00	88.16
4	SVM	66.08	63.16

A comparison of accuracy of different algorithms used is shown below.



After this, Cross Validation techniques and Hyper Parameter Tuning was performed on the dataset to improve the testing accuracies of the implemented algorithms.

- K-fold Cross Validation:

  Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

  The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

  Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

  The general procedure is as follows:

  1. Shuffle the dataset randomly.
  2. Split the dataset into k groups
  3. For each unique group:
     a. Take the group as a hold out or test data set
     b. Take the remaining groups as a training data set
     c. Fit a model on the training set and evaluate it on the test set
     d. Retain the evaluation score and discard the model
  4. Summarize the skill of the model using the sample of model evaluation scores

  It is a popular method because it generally results in a less optimistic estimate of the model skill than other methods, such as a simple train/test split.

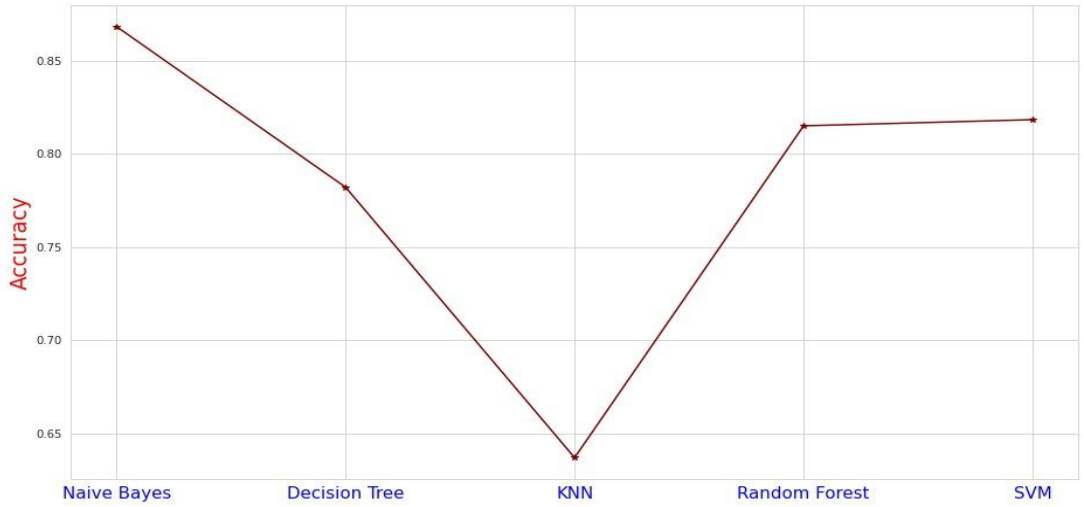
- Hyper Parameter Tuning:

  Hyperparameter tuning works by running multiple *trials* in a single training job. Each trial is a complete execution of your training application with values for our chosen hyperparameters, set within limits you specify. We can keep track of the results of each trial and makes adjustments for subsequent trials. When the job is finished, we can get a summary of all the trials along with the most effective configuration of values according to the criteria we specified.

Percentage accuracy scores after performing these two techniques are depicted in the below figure for different algorithms.

	model	best_score	best_parameter
0	Naive Bayes	0.868421	None
1	Decision Tree	0.782178	{'criterion': 'entropy', 'max_features': 'sqrt'}
2	KNN	0.636964	{'algorithm': 'auto', 'n_neighbors': 5}
3	Random Forest	0.815182	{'n_estimators': 50, 'random_state': 100}
4	SVM	0.818482	{'C': 50, 'kernel': 'linear'}

A comparison of accuracy of different algorithms, after hyperparameter tuning used is shown below. As it is clearly seen, the accuracies of different algorithms have been improved after performing cross validation and hyperparameter tuning techniques on the dataset.



CHAPTER 6: NEW LEARNINGS FROM THE TOPIC

During this research, I learnt about

- The importance of machine learning domain in the medical field.
- The benefits of predicting heart disease before hand.
- Few data preprocessing methods used on the dataset.
- Five machine learning models implemented to predict the heart disease.
- Cross validation and Hyperparameter tuning techniques used to improve the accuracy of algorithms implemented.

REFERENCES AND ANNEXURES

[1] Devansh Shah, Samir Patel, and Santosh Kumar Bharti, "Heart Disease Prediction using Machine Learning Techniques", Springer Nature Singapore Pte Ltd 2020, SN Computer Science (2020) 1:345

[2] Rohit Bharthi, Adithya Kamparia et. al., "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", Hindawi, Computational Intelligence and Neuroscience, Volume 2021

[3] Vijeta Sharma, Shrinkala Yadav, and Manjari Guptha, "Heart Disease Prediction using Machine Learning Techniques", 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)

[4] Emrana Kabir Hashi and Md. Shahid Uz Zaman, "Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction", Journal of Applied Science & Process Engineering, Vol. 7, No. 2, 2020