International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7430
***Article Received**: 12 May 2023 **Revised:** 25 June 2023 **Accepted:** 30 July 2023*

Heart Disease Prediction using Different Machine Learning Algorithms

Dr. Rajani P.K[1], Kalyani Patil[2], Bhagyashree Marathe[3], Prerna Mhaisane[4], Atharva Tundalwar[5]
[1]Department of Electronics and Telecommunication
Pimpri Chinchwad College of Engineering, Pune, India
e-mail: rajani.pk@pccoepune.org
[2]Department of Electronics and Telecommunication
Pimpri Chinchwad College of Engineering, Pune, India
e-mail: kalyani.patil21@pccoepune.org.
[3]Department of Electronics and Telecommunication
Pimpri Chinchwad College of Engineering, Pune, India
e-mail: bhagyashree.marathe20@pccoepune.org,
[4]Department of Electronics and Telecommunication
Pimpri Chinchwad College of Engineering, Pune, India
e-mail:prerna.mhaisane20@pccoepune.org
[5]Department of Electronics and Telecommunication
Pimpri Chinchwad College of Engineering, Pune, India
e-mail: atharva.tundalwar20@pccoepune.org

Abstract—Identifying a person's potential for developing heart disease is one of the most challenging tasks medical professionals faces today. With nearly one death from heart disease every minute, it is the leading cause of death in the modern era [4]. The database is taken from Kaggle. Various machine learning algorithms are used for prediction of heart disease detection here are Random Forest, XG-Boost, K- Nearest Neighbors (KNN), Logistic Regression, Support Vector Machines (SVM). All these algorithms are implemented using Python programming with Google collab. The performance evaluation parameters used here are Accuracy, precision, recall and Fi-score. Training and testing are implemented for different ratios such as 60:40, 70:30 and 80:20. From the analysis and comparisons of evaluation parameters of all the above algorithms, XG-Boost is having the highest accuracy and recall value. KNN having worst accuracy and recall amongst all. XG-Boost is having a training accuracy of 98.86, 98.74 and 97.68 for training and testing ratio of 60:40, 70:30 and 80:20 respectively. XG-Boost is having a testing accuracy of 95.85, 95.45 and 96.09 for training and testing ratio of 60:40, 70:30 and 80:20 respectively. So, XG-Boost algorithm can be used for obtaining the best prediction for heart disease. This type of heart disease prediction can be used as a secondary diagnostic tool for doctors, for best and fast prediction. This can help the early prediction of heart disease thus increasing the chances of the saving the life heart patient.

Keywords-Classification; K-Nearest Neighbor; XG Boost; Support Vector Machine; Random Forest; Logistic Regression.

I. INTRODUCTION

Heart disease is a serious problem that affects the entire world and accounts for many fatalities each year. Despite their best efforts, doctors still struggle to accurately forecast the risk of developing heart disease since it requires considerable skill and in-depth knowledge [8]. This paper intends to resolve this problem by creating an automated method for heart disease diagnostics that will improve medical effectiveness and lower expenses [21]. Heart disease, also known as cardiovascular disease, is brought on by a number of things, including bad lifestyle choices, smoking, drinking alcohol, and eating a lot of fat. The functioning of the heart and other crucial organs like the brain and kidneys can be severely impacted by these factors, which can also cause illnesses including hypertension, high blood pressure, diabetes, and strokes [9].

To accurately diagnose heart disease, researchers have employed various techniques, including K-Nearest Neighbor (KNN), Logistic Regression, XG-Boost, Random Forest, and Support Vector Machine (SVM) [16]. This paper will make use of these algorithms to efficiently predict the risk level of patients based on the parameters of their health. The prediction of heart disease requires a large amount of data, which is too complex and massive to process and analyze using conventional techniques [24]. The objective of this paper is to determine the most suitable machine learning technique for the prediction of heart disease through a comparison of various algorithms, with a focus on accuracy and computational efficiency. The following section provides further details of the techniques used in our study.

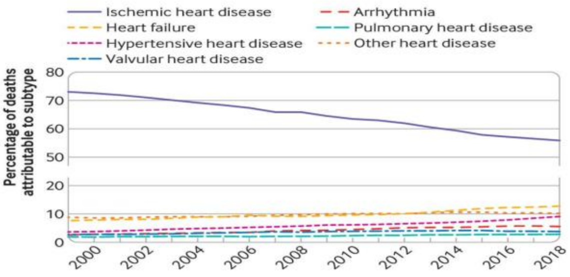

Figure 1. Graphical representation of deaths caused due to heart disease

According to the World Health Organization (WHO), cardiovascular diseases (CVDs) are the leading cause of death globally. In 2019, an estimated 17.9 million people died from CVDs, accounting for 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke [11]. CVDs affect people of all ages and backgrounds, but the majority of deaths (78%) occur in low- and middle-income countries. The main modifiable risk factors for CVDs include unhealthy diet, physical inactivity, tobacco use, and harmful use of alcohol. The prevalence of these risk factors is increasing globally, particularly in low- and middle-income countries where economic and social transitions have led to changes in lifestyle and dietary habits. The WHO has set a global target to reduce premature deaths from noncommunicable diseases (including CVDs) by one-third by 2030 through a range of interventions, including promoting healthy diets and physical activity, reducing tobacco use, and providing access to affordable and essential medicines and technologies [23]. These statistics highlight the significant impact of heart disease worldwide and the urgent need for preventive and treatment measures to address this public health issue [17].

II. METHOD

1. Software Used: Google Colab

Google Colab is a free cloud-based platform for machine learning research and development. It provides Jupyter notebook environment, which allows users to write and execute code, analyze data, and collaborate on projects. Colab supports multiple programming languages including Python, R, and others, and provides access to GPUs and TPUs for high-performance computation. It also includes many popular libraries and tools for data science, such as TensorFlow, PyTorch, and others, making it an accessible and convenient platform for beginners and experts alike. With Colab, users can easily import datasets, save their work, and share it with others, making it a useful tool for collaboration and knowledge sharing.

Age, sex, chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate reached, exercise-induced angina, oldpeak, slope of the peak exercise ST segment, and the number of major vessels highlighted by fluoroscopy are all examined in the dataset that was referred from Kaggle [25].

2. Machine Learning Algorithms

i. Logistic Regression (LR)

The algorithm for Logistic Regression model:

1. Start by choosing a hypothesis function that models the relationship between the input features and the output (binary) variable. A common choice is the logistic function (sigmoid), which maps the input features to a probability value between 0 and 1.
2. Choose a cost function that measures the difference between the predicted probability and the actual label (0 or 1) for each data point in the training set. A common choice is the cross-entropy loss function.
3. Use an optimization algorithm (e.g., gradient descent) to minimize the cost function by adjusting the model's parameters (weights and biases). The goal is to find the parameters that best fit the training data and minimize the error on the validation set.
4. To make a prediction for a new data point, pass the input features through the trained model and use the logistic function to convert the output to a probability value. If the probability is greater than a chosen threshold (e.g., 0.5), classify the new data point as belonging to the positive class (1); otherwise, classify it as belonging to the negative class as zero [5][10].

ii. Support Vector Machine (SVM)

The algorithm for SVM for a regression problem:

1. Start by selecting a kernel function. Popular choices include linear, polynomial, and radial basis function (RBF) kernels [20].
2. Choose a value for the regularization parameter C, which controls the trade-off between maximizing the margin and minimizing the regression error on the training set.
3. Use the training data to optimize the SVM's Parameters. This involves solving a quadratic programming problem that minimizes the regression error subject to the constraint that the data points are within a certain margin of the predicted values.
4. To make a prediction for a new data point, calculate its predicted value using the kernel function and the trained SVM parameters [1][5].

DOI: https://doi.org/10.17762/ijritcc.v11i9s.7430
***Article Received**: 12 May 2023 **Revised:** 25 June 2023 **Accepted:** 30 July 2023*

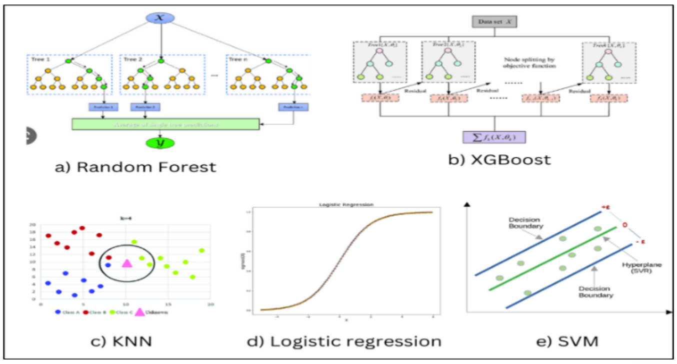

Figure 2. Graphical Illustration of Machine learning Algorithm

iii. K-Nearest Neighbour (KNN)

The algorithm for building a KNN model for a classification problem:

1. Start by choosing a value for K, which is the number of nearest neighbours to consider when making a prediction for a new data point.
2. Calculate the distance between the new data point and all the data points in the training set. The distance metric used can be Euclidean distance, Manhattan distance, or other distance measures.
3. Identify the K data points in the training set that are closest to the new data point based on the calculated distances [19].
4. To make a prediction for the new data point, take the majority vote of the class labels of the K nearest neighbours. If the K value is even and there is a tie, you can break the tie using some rules such as selecting the class label of the closest neighbour.
5. Repeat steps 2 to 4 for all new data points to classify [1][16].

iv. Random Forest (RF)

The algorithm for Random Forest:

1. Start by selecting a random sample of the data set.
2. Choose a random subset of the features from the data set.
3. Build a decision tree on the selected subset of data and features.
4. Repeat steps 1 to 3 several times to create a set of decision trees.
5. To make a prediction, pass a new data point through each decision tree in the forest and take the majority vote of the predictions made by the individual trees.
6. To improve the performance of the Random Forest, you can tune hyperparameters such as the number of trees, the maximum depth of each tree, and the size of the random subsets of features used for each tree [1][6][9].

V. XG-Boost

The algorithm for building an XGBoost model:

1. Start by initializing the model with some default hyperparameters (e.g., the learning rate, the number of trees, the maximum depth of each tree).
2. Use the training data to fit the model by iteratively adding decision trees to the ensemble. At each iteration, the algorithm calculates the gradient and hessian of the loss function (e.g., binary cross-entropy for classification, mean squared error for regression) with respect to the predicted values of the previous trees in the ensemble [14].
3. Fit a new decision tree to the residuals of the previous trees in the ensemble, using some method to determine the optimal split points and leaf node values. Popular methods include exact and approximate greedy algorithms, as well as regularization techniques to prevent overfitting.
4. Add the new decision tree to the ensemble, weighted by the learning rate, and update the predicted values for each data point in the training set.
5. Repeat steps 2 to 4 until the specified number of trees is reached, or until some stopping criterion (e.g., validation set error does not improve) is met.
6. To make a prediction for a new data point, pass the input features through the trained ensemble of decision trees, and sum the predicted values from each tree to obtain the final prediction [7][22].

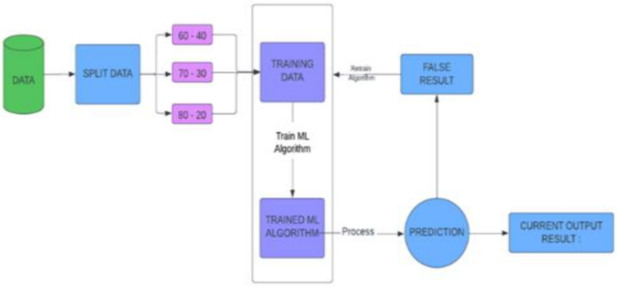

Figure 3. Block Diagram of Machine Learning Algorithm Implementation

III. RESULTS AND DISCUSSION

This section discusses, the evaluation the performance of five popular machine learning algorithms such as random forest, KNN, logistic regression, SVM, and XGBoost. Various evaluation metrics such as accuracy, precision, recall, and F1 score to determine the efficiency of each algorithm [18]. Precision measures the proportion of positive predictions that are actually correct, while recall measures the number of correct positive predictions out of all actual positive instances. The F1 score is a balance between precision and recall and is a commonly used metric for evaluating the performance of a classifier. This paper implemented using python programming with different ratios of training and testing data such as 60:40, 70:30 and 80:20 respectively. This will help in analysing the above-mentioned machine learning algorithms.

The terminologies used for the mathematical model given below are, TP (true positive), TN (true negative), FP (false positive), FN (false negative) [1][12][15].

Evaluation Parameters:

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN'} \quad \text{.... (1)}$$

$$\text{Recall} = \frac{TP}{TP+FN'} \quad \text{.... (2)}$$

$$\text{Precision} = \frac{TP}{TP+FP'} \quad \text{.... (3)}$$

$$\text{F1 Score}=\frac{2*Precision*Recall}{Precision+Recall} \quad \text{.... (4)}$$

The mathematical formula is used for the comparison of different evaluation parameters for various machine learning algorithms.

1. Accuracy Parameters of Algorithms for 60-Training 40-Testing

This can be observed in the Table 1 and Figure 4 as given below.

Table 1. 60-Training 40-Testing

Sr No.	Algorithm	Accuracy		Precision	Recall	F1 score
		Training	Testing			
1	LR	86.17	81.7	79.22	87.14	82.99
2	SVM	86.17	83.17	80.25	89.04	84.42
3	KNN	86.17	68.78	69.15	70.47	69.81
4	XGBOOST	98.86	95.85	96.61	95.23	95.92
5	RF	97.19	89.75	86.05	93.22	89.49

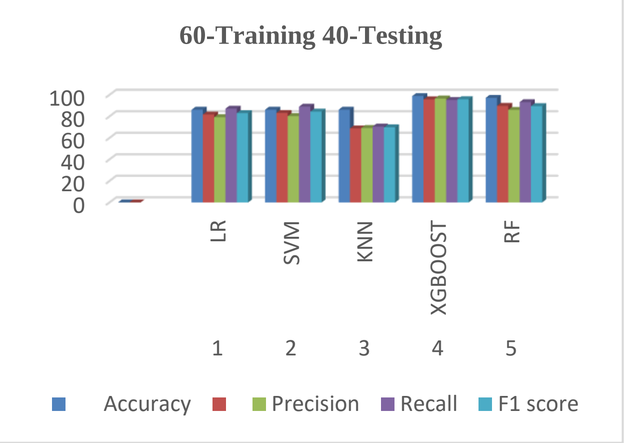

Figure 4. Comparison of Algorithms for 60:40 Ratio

With an accuracy of 98.86% in training, 95.85 in testing, and a recall value of 95.23%, the XG-Boost offers the most accuracy and recall value for the 60:40 ratio, while the KNN has the lowest accuracy and recall value, 86.17% in training, 68.78% in testing, and a recall value of 70.47%.

2. Accuracy Parameters of Algorithms for 70-Training 30-Testing

This can be observed in the Table 2 and Figure 5 as below.

Table 2. 70-Training 30-Testing

Sr No.	Algorithm	Accuracy		Precision	Recall	F1 score
		Training	Testing			
1	LR	86.47	79.22	76.4	86.07	80.95
2	SVM	86.47	80.19	76.79	87.97	82
3	KNN	86.47	67.85	69.03	67.72	68.37
4	XGBOOST	98.74	95.45	96.75	94.3	95.51
5	RF	98.14	93.18	89.67	96.52	92.97

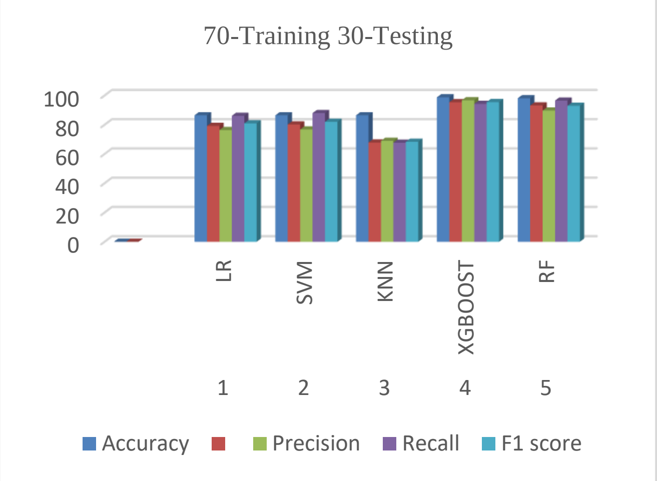

Figure 5. Comparison of Algorithms for 70:30 Ratio

According to the analysis above, the KNN has a low accuracy rate with 86.77% for training and 67.88% for testing for different machine learning algorithms at a ratio of 70:30 among all the methods. With an accuracy of 98.74% in training, 95.45% in testing, and a recall value of 95.23% for random forest, respectively, XG-Boost offers the highest level of accuracy.

3 Accuracy Parameters of Algorithms for 80-Training 20-Testing

This can be observed in the Table 3 and Figure 6 as below.

Table 3. 80-Training 20-Testing

Sr No.	Algorithm	Accuracy		Precision	Recall	F1 score
		Training	Testing			
1	LR	85.24	80.48	76.85	88.57	82.3
2	SVM	86.95	82.43	77.96	92.38	84.34
3	KNN	75.48	68.78	69.9	68.57	69.23
4	XGBOOST	97.68	96.09	96.19	96.19	96.19
5	RF	97.95	92.68	88.28	98	92.89

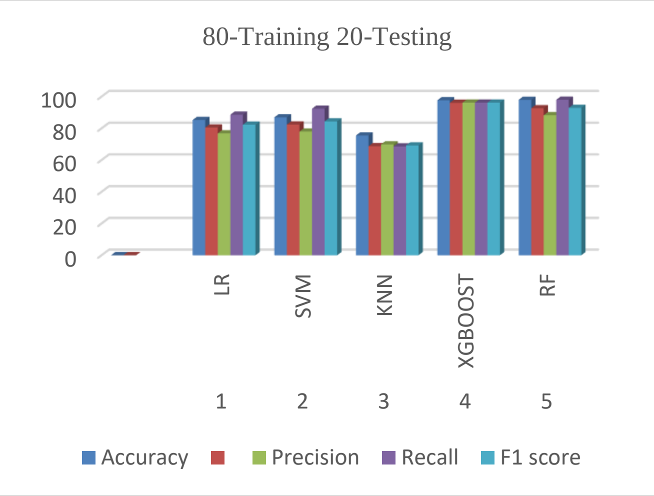

Figure 6. Comparison of Algorithms for 80:20 Ratio

According to the data above, the accuracy and recall for all machine learning algorithms for the ratio of 80:20 differ slightly. For XG-Boost, it is 97.95% for random forest during training and 96.09% during testing. In a similar manner, the recall value for random forest is maximal in each instance. In this instance, the accuracy and recall value for a KNN are both low, with training accuracy of 75.48%, testing accuracy of 68.78%, and recall value of 68.57%.

From the algorithms compared with different training and testing ratios, the XG Boost algorithm has maximum accuracy compared to other parameters in most of the cases.

IV. CONCLUSION

This paper compares the Heart disease prediction using the different classification algorithms available in machine learnings. Various machine learning algorithms are used for prediction of heart disease detection here are Random Forest, XG-Boost, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machines (SVM). All these algorithms are implemented using Python programming with Google collab. The generic heart disease databases used here is taken from Kaggle with 1025 samples [23]. Training and testing are implemented for different ratios such as 60:40, 70:30 and 80:20. The inputs given for 60:40 ratios were 613 samples for training and 410 samples for testing purpose. Similarly, for 70:30 ration, 716 samples for training and 307 samples for testing purpose and for 80:20 ration, 818 samples for training and 205 samples for testing purpose respectively. The comparison was carried out based on various evaluation parameters like accuracy, precision, recall, and F1-Score metrics. XG-Boost is having a training accuracy of 98.86, 98.74 and 97.68 for training and testing ratio of 60:40, 70:30 and 80:20 respectively. XG-Boost is having a testing accuracy of 95.85, 95.45 and 96.09 for training and testing ratio of 60:40, 70:30 and 80:20 respectively. The results of this comparison shows that XG boost is best compared to all other algorithms. The results also indicates that the KNN algorithm has the worst accuracy. XGBoost is designed to optimize the training objective function, which allows it to minimize both bias and variance simultaneously. This is achieved through several advanced techniques such as regularization, early stopping, and parallel processing, which make the algorithm both fast and accurate. It is also worth mentioning that unsupervised learning algorithms can be used to further analyze the results and gain deeper insights into the underlying patterns and relationships in the data. This type of heart disease prediction can be used as a secondary diagnostic tool for doctors, for best and fast heart disease prediction. Thus, it can increase the chances of saving the heart patient’s life.

REFERENCES

[1] Nagaraj M Lutimath, Chethan C, Basavaraj S Pol.,’ Prediction of Heart Disease using Machine Learning’, International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.

[2] Kavitha, B. S., & Siddappa, M., A Survey on Machine Learning Techniques to Predict Heart Disease, International Journal of Computer Science & Communication (ISSN: 0973-7391) Special Issue, Page 48-53, December 2020

[3] Dr. Poonam Ghuli, Heart Disease Prediction using Machine Learning, International Journal of Engineering Research & Technology ISSN: 2278-0181IJERTV9IS040614 (IJERT), Vol. 9 Issue 04, April-2020.

Rishabh Magar, Rohan Memane, Suraj Raut, Heart disease prediction using machine learning,2020 JETIR, Volume 7, Issue 6, June 2020

[4] A. H. Fauzi, A. F. Malik, and M. S. Rizal, "Medical Insurance Cost Prediction using Random Forest and Decision Tree," 2020 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta,

[5] Nair, K., Motagi, N., Narayankar, R., Rajani, P.K. (2023). Error Detection and Error Concealment of Medical Images Using Frequency Selective Extrapolation (FSE) Algorithm. In: Kaiser, M.S., Xie, J., Rathore, V.S. (eds) ICTCS2021, Lecture Notes in Networks and Systems, vol 401, pp 495–504, Springer, Singapore.

[6] Al-Fahoum, A. S., Al-Hazaimeh, H., & Al-Madi, N. (2020). An intelligent system for heart disease prediction based on machine learning. IEEE Access, 8, 27726-27738.

[7] Fajr Ibrahem Alarsan., and Mamoon Younes ‘Analysis and classification of heart diseases using heartbeat features and machine learning algorithms’,Journal Of Big Data,2019;6:81.

[8] Jayesh Kolhe, Guruprasad Deshpande, Gargi Patel, Rajani P.K, “Crop Decision using Various Machine Learning Classification Algorithms”, published in conference proceedings by Springer SIST. ISSN Number - 2190-3018,*Smart Innovation, Systems and Technologies*, 312, pp. 495–502, 2023 https://link.springer.com/chapter/10.1007/978-981-19-3575-6_49

[9] R. Gupta and S. Yusuf, “Challenges in management and prevention of ischemic heart disease in low socioeconomic status people in LLMICs,” BMC Medicine, vol. 17, no. 1, p. 209, 2019.

[10] T.Nagamani, S.Logeswari, B.Gomathy,Heart Disease Prediction using Data Mining with Mapreduce Algorithm,International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-8 Issue-3, January 2019

[11] Karim, M. A., & Majumdar, S. (2019). Diagnosis of heart disease using machine learning algorithms: A survey. Journal of Ambient Intelligence and Humanized Computing, 10(5), 1845-1867.

[12] Avinash Golande, Pavan Kumar T,International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019

[13] P. Umasankar and V. ,iagarasu, “Decision support system for heart disease diagnosis using interval vague set and fuzzy association rule mining,” in Proceedings of the International Conference on Devices, Circuits and Systems (ICDCS),pp. 223–227, Coimbatore, India, March 2018.

[14] Mr. Chala Beyene, Prof. Pooja Kamat, “Survey on Prediction and Analysis of the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure And Applied Mathematics, (2018).

[15] Rajesh N, T Maneesha, Shaik Hafeez and Hari Krishna, "Prediction of Heart Disease Using Machine Learning Algorithms," International Journal of Engineering & Technology, vol. 7, pp. 364-366, 2018.

[16] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In 2018 Second International Conference on Electronics,Communication and Aerospace Technology (ICECA) (pp. 1275-1278). IEEE.

[17] .N. Khateeb and M. Usman, “Efficient heart disease prediction system using k-nearest neighbor classification technique,” in Proceedings of the International Conference on Big Data and Internet of Fing (BDIOT), pp. 21–26, ACM, London UK, December 2017.

[18] Sharma, H., & Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends In Computing and Communication, 5(8),99-104.

[19] V. Krishnaiah, G. Narasimha, N. Subhash Chandra, “Heart Disease Prediction System using
Data Mining Techniques and Intelligent Fuzzy Approach: A Review” IJCA 2016.

[20] Thomas, J., & Princy, R. T. (2016, March). Human heart disease prediction system using data mining techniques. In 2016 international conference on circuit, power and computing technologies (ICCPCT).

[21] K.Sudhakar, Dr. M. Manimekalai “Study of Heart Disease Prediction using Data Mining”, IJARCSSE
2016.

[22] Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent F. Massin, “A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining”, (IJAER), 2012.

[23] Linkfor
dataset:https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset