INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 1, January 2026 | ISSN: 2456-4184 | IJNRD.ORG

Evaluating Different Machine Learning Models for Heart Disorders Prediction

[1]Tikeshwar Gajpal, [2]Dr. Hemlata Sinha,

[1]Assistant Professor, [2]Professor,
[1]Department of Electronics & Telecommunication Engg.
[1]Shri Shankaracharya Institute of Professional Management and Technology, Raipur, India

***Abstract*:** As heart disease remains one of the leading causes of death worldwide, it is crucial to develop effective techniques for its early detection and prevention. By examining intricate links in patient health data, machine learning has become a potent method for predicting cardiac disease. Using health indicators including age, blood pressure, cholesterol, and lifestyle factors, and this work emphases on using machine learning techniques for heart disease prediction. Data preprocessing, which includes handling missing values, feature scaling, and categorical data encoding, is the first step in the process. Numerous machine learning models, such as Random Forest, Logistic Regression, Support Vector Machines (SVM), Naive Bayes, and XG-Boost, are employed to assess whether a patient is at risk of developing heart disease. The assessment metrics like accuracy rate, precision rate, recall rate, and F1-score are used to measure model performance, ensuring both reliability and clarity in interpretation. Our findings show that cardiac disease can be accurately predicted by machine learning models, with sophisticated algorithms obtaining strong performance and high accuracy. This study demonstrates how incorporating machine learning into healthcare systems might give doctors data-driven insights for early diagnosis and individualized treatment plans.This study intends to improve patient outcomes through early and precise forecasts and lessen the worldwide burden of heart disease by utilizing machine learning models.

Keywords: Heart Disease Expectation, Logistic Regression, Support Vector Machine, Confusion Matrix, KNN, Decision Tree, Random Forest, Naive Bayes, XG-Boost.

Introduction

Heart disease, or cardiovascular disease, ranks among the top causes of death globally, taking millions of lives annually. Heart disease includes various conditions that impact the heart, such as coronary artery disease, irregular heartbeats (arrhythmias), and heart failure. As per the World Health Organization, early diagnosis and protective measures are essential to lessen the impact of heart disease. [9]. Precise heart disease prediction enables early medical intervention, which can greatly improve patient outcomes and reduces overall healthcare cost.

In recent years, advancements in technology and data science have enabled healthcare practitioners to leverage machine learning (ML) as a tool for predictive analytics [8]. By analyzing large datasets of patient health metrics, ML models can identify complex designs and interactions that are tough to discern through traditional statistical approaches. These models can show a vital role in forecasting heart disease threat and guiding personalized treatment plans.

The important purposes of this study are:

1. To examine how machine learning technique can be applied to guess the risk of heart disease based on patient documents.
2. To investigate the efficiency of different models, including Logistic Regression, Random Forest, and Support Vector Machines, by measuring key performance parameters such as accuracy rate, precision rate, recall rate, and F1-score.
3. To tackle issues like data imbalance, selecting relevant features, and ensuring model interpretability in order to enhance the correctness and consistency of predictions.
4. To provide a framework for integrating machine learning-based prediction models into real-world healthcare systems.

This paper focuses on utilizing publicly available Heart Disease Dataset, to improve and authenticate machine learning approaches for heart disease prediction. The research covers:

- Preprocessing of healthcare data
- Implementation and comparison of various machine learning models to identify the best effective method.
- Apply evaluation metrics like the accuracy rate, precision rate, confusion matrix, and recall rate to assess the performance of the model.

Literature Survey

[1] Srinivas Kanakala et al. focused on predicting heart failure using several classification algorithms, including Naive Bayes, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Random Forest. The efficiency of each method was assessed based on accuracy rate, F1 score, and the space under the ROC curve. After thorough evaluation, Decision Trees and Random Forest proved to be more effective than the others, with Random Forest attaining the maximum accuracy rate and the largest ROC area, highlighting its strong capability to differentiate between positive and negative heart failure cases.

[2] Bhavesh Dhande et al. presented an innovative approach for identifying key features using machine learning techniques, aimed at enhancing the prediction of multiple diseases. This method was implemented using the Google Colab platform along with essential Python libraries. The dataset was analyzed using various models, including Ensemble Classifiers, XGBoost, Decision Tree, Logistic Regression, KNN, Random Forest, AdaBoost, and multiple types of SVCs such as Sigmoid, Polynomial, RBF, and Linear.

[3] Raparthi Yaswanth et.al. in march 2020 proposed in which dataset, sourced from the UCI repository, includes 303 records, with 164 classified as negative and 139 as positive. Among the tested algorithms KNN, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine and Neural Networks delivered the best outcome. The numerous assessment metrics such as accuracy rate, precision rate, recall rate, F-score, and ROC were used to compare the different models.

[6] Heart disease prediction using Support Vector Machine method enhances online consultations by analyzing patient data for accurate risk assessment. SVM is employed to estimate the probability of heart illness by analyzing a range of associated risk factors.

[8] Sumit Sharma et al. in January 2020 suggested that Talos is a technique used for hyperparameter optimization. In this study, the Talos optimizer is utilized in combination with the Keras library. Keras is a deep learning neural network library. Keras supports the implementation of both Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including their hybrid forms. Talos achieved higher accuracy (90.76%) compared to other optimization methods (Logistic Regression, KNN, SVM, Naïve Bayes, Random foster).

[10] Kallepalli Reshma et al. collected data from Kaggle.com and performed preprocessing on it. They designed a multi-disease prediction system by machine learning model. For diabetes prediction, they applied the SVM model and gain an accuracy of 78%. The same algorithm was used to predict Parkinson's disease, resulting in an accuracy rate of 89%. In heart illness prediction, they utilized logistic regression and obtained an accuracy of 85%.

[11] The UCI repository is the source of the disease dataset. This dataset includes 303 rows and 14 columns, and each row represents a single record. They discovered that KNN and Logistic Regression are more effective in diagnosing heart conditions.

[12] Dr. Rajani P.K and colleagues utilized a heart disease dataset from Kaggle to apply various machine learning processes for calculation. The algorithms investigated include Random Forest, XG-Boost, K-Nearest Neighbors, Logistic Regression, and Support

Vector Machines. These models were developed using Python in the Google Colab environment. To measure performance, metrics such as accuracy rate, precision rate, recall rate, and F1-score were used. Training and testing were carried out with different data splits, including 60:40, 70:30, and 80:20 ratios. Among all the algorithms, XG-Boost demonstrated the maximum accuracy in heart disease prediction.

[13] D Venkatesh et al proposed hybrid machine learning model in which SVM and KNN models are used that gives better result compared to individual. The dataset contains 303 patients' records.

[14] Ramanathan G. worked on two different datasets the Cleveland Dataset and the Framingham Dataset—that were both downloaded from Kaggle were analyzed. They found that the Gradient Boosting method performed best on the Framingham dataset and the Decision Tree model performed best on the Cleveland dataset.

[15] Sanjana Chaudhari et al. evaluate the performance and found that Random Forest model has the peak correctness of 90.63%. The dataset is taken from IEEE Dataport.org that has 1189 patient data in which 628 are heart disease patient and 561 are the not heart disease patient.

Material & Method

A. Dataset:-Dataset is taken from kaggel.com. This dataset includes 303 rows and 14 columns in which last column is target that defined whether patient have disease (integer value =1) or not (integer value =0). Below figure showing dataset information with column name and its data type.

```
# getting some info about the data
heart_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Figure 1 Dataset Information

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 2 First Five Data

Chest pain type (cp) is categorized into four types and encoded numerically (0-3). The chest pain is a significant factor in predicting heart attacks.[7]

Cholesterol (chol) is integer value. It can block arteries, increasing risk of heart attack.

Fasting Blood Sugar (fbs) is blood glucose level after at least 8 hours of fasting.

Oldpeak is provide insight into how the heart reacts to stress or exercise.

Resting ECG (testecg) is ECG recorded while the patient is at rest.

Thal refer to result of thalium stress test,which evaluates blood flow to the heart muscle during exercise.

B. Data cleaning: - This process is an essential initial stage in any study involving data processing or modeling. It involves refining and adjusting the data to make it more accessible for analysis and interpretation.

C. Feature engineering: - Feature engineering is a key aspect of deep learning, focused on extracting meaningful attributes from a dataset. It involves transforming raw data into useful features, which enhances model performance and leads to improved accuracy. [8]

D. Learning algorithms: - We utilized various kinds of learning techniques are discussed here. **Logistic Regression: -** Logistic regression is a classification method used for predicting categorical outcomes. The sigmoid function plays a central role in logistic regression. [8]

KNN: It is a non-parametric and supervised machine learning model. It guesses the segment of a new data point by examining the labels of its adjacent neighbors. The KNN algorithm works by following a series of defined steps.

1. Select a suitable value for K where K is the count of nearest neighbors to identify.
2. Compute the distance between the new data point and each point in the dataset.
3. Recognize the K nearest neighbors to the unknown data point from the training data.
4. Forecast the output of unknown data point using the most common class. That class is the prediction.
5. Halt

In this method data is separated into training and test data sets. The model construction and training is done using the training dataset. K-value is chosen which is the square root of the number of observations. Then test data is predicated on the design model. [8]

Support Vector Machine (SVM):- It is supervised deep learning method which can be used both for classification and regression. It is employed to manage class imbalance, which occurs in machine learning model when the number of positive and negative occurrences differs significantly, often resulting in reduced classifier effectiveness. It aims to identify the most optimal decision boundary, known as a hyperplane that split up data points belonging to dissimilar groups. [8]

Decision Tree: - In machine learning, decision tree is a flowchart-style approach that represents the decision-making process. It is a supervised learning approach used in both classification and regression problems. The model starts with a root node, which splits into branches based on certain features, leading to decision nodes and finally to leaf nodes that indicate the predicted result. The outcomes generated by decision trees are often binary, as yes such or no. [5]

Random Forest Classifier: - it is a supervised machine learning method primarily used for classification, however it can also handle regression tasks. It generates multiple decision trees and merges their results to produce more correct and consistent calculations. [5]

Naive Bayes: - The Naive Bayes model is especially useful in medical diagnosis of heart patients due to its simplicity and the absence of complex iterative parameter estimation. Despite its straightforward nature, the Naive Bayes classifier is widely favored because it often delivers better results than more advanced classification methods. [4]

XG-Boost: - XG-Boost (Extreme Gradient Boosting) is a dominant and efficient supervised machine learning method that builds on decision tree algorithms. It performs well in both classification and regression tasks. It aims to improve the forecasts over time by correcting the errors in the previous model. The ability to handle complex relationships in data, feature importance analysis, and robustness against over fitting when hyper parameters are properly tuned are just a few advantages of gradient boosting machines. In order to achieve optimal performance, it necessitates meticulous adjustment and can be computationally costly.[4]

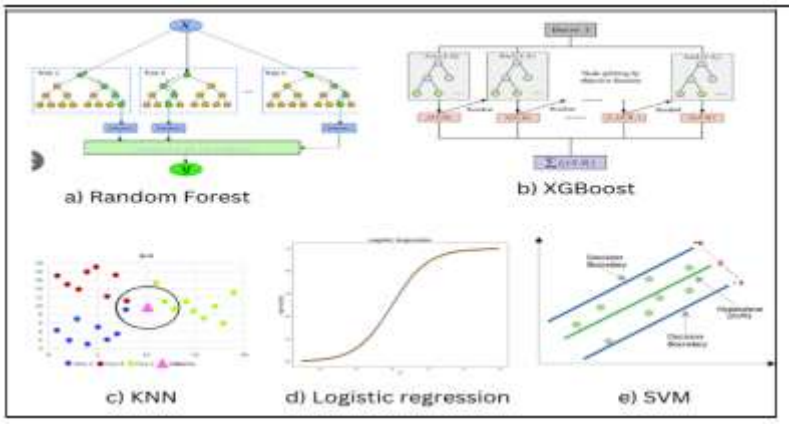

Figure 3 Graphical Representation of Machine Learning Model [12]

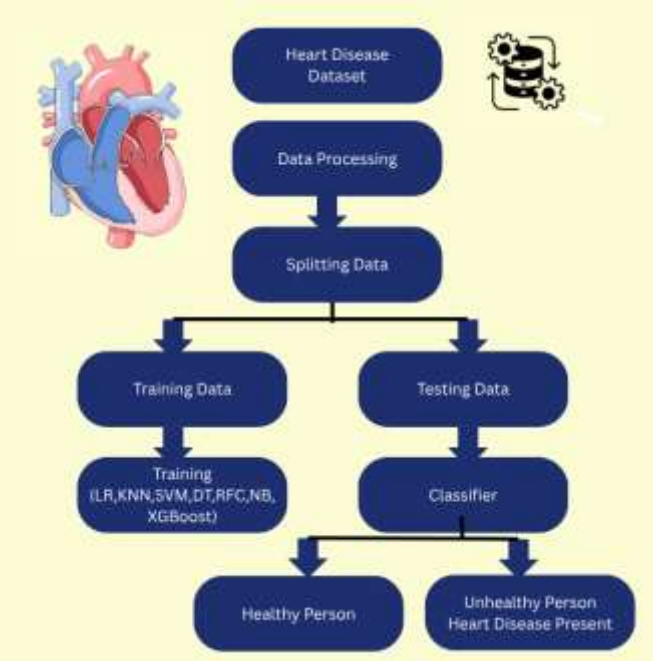

Figure 4 Flow chart

Experimental Result

The details of software used in analysis are as follows

Python, Pandas, Jupyter Notebook, Scikit-learn, Seaborn, Numpy, Matplotlib

The training and test data are the two categories into which the data is divided. In the following steps, the models are trained using these two. It is then used on several classification models, as illustrated below:

Table 1 Evaluation of numerous machine learning processes

S.N.	Model	Accuracy Score of Test data %	Recall Score of Test Data%	Precision Score of Test Data %	F1 Score of test Data %	Confusion matrix of Test Data	Cross_val_Score(cv=3) %
1	Logostic Regression	81.96	84.3	81.8	83	[[23 5] [6 27]]	83.5
2	KNN(5)	62.29	64.7	66.6	65.5	[[16 12] [11 22]]	63.03
3	KNN(3)	62.2	65.6	63.6	64.6	[[17 11] [12 21]]	60.72
4	SVM	81.9	80.5	87.8	84.05	[[21 7] [4 29]]	54.45
5	Decision Tree	78.6	79.4	81.8	80.5	[[21 7] [6 27]]	71.94
6	RandomForestClassifier	78.6	79.41	81.8	80.59	[[21 7] [6 27]]	78.87
7	Naïve Bayes Classification	81.96	89.28	75.757	81.96	[[25 3] [8 25]]	80.52
8	XGBoost	75.409	78.12	75.75	76.923	[[21 7] [8 25]]	81.18

Accuracy reflects how precise the predictions are by indicating the percentage of correctly classified instances. It shows the overall proportion of correct guesses made by a model or algorithm [9]. A higher accuracy signifies improved model performance. The table shows that Logistic Regression, SVM, and Naive Bayes achieved the highest precision of 81.96%, outperforming the other models. Recall score shows out of all positive cases, how many correctly predicted [9]. Recall score column shows Naïve Bayes has the highest

value 89.28. Precision Score shows out of all predicted positive, how many actually positive. Precision score column shows that SVM has the highest value 87.8% and KNN has the lowest value. F1 score shows the balance between Recall and Precision value especially when data is imbalanced. SVM has the highest value 84.05% and KNN has the lowest 64.3 %. The cross-validation score column reflects how the data is divided into multiple subsets to estimate the model's efficiency on unseen data. Among the models, Logistic Regression attained the maximum score of 83.5%, while SVM recorded the lowest at 54.45%.

Based on these result Logistic Regression, XG-Boost and Naive Bayes appear to be attractive model for use in real time world environment.

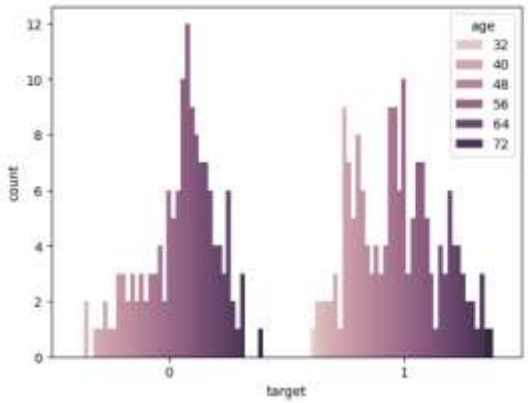

Figure 6 Heat Map Age Target

This figure showing target value with respect to age, patient have disease (integer value =1) or not (integer value =0)

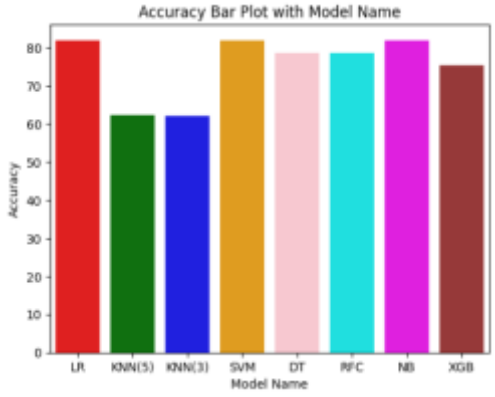

Figure 7 Accuracy Bar Plot

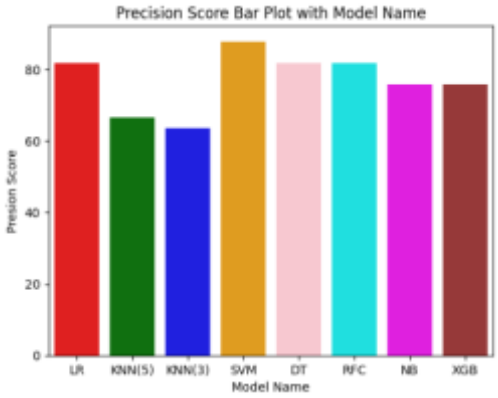

Figure 8 Precision Bar Plot

Figure 6 showing Accuracy bar Plot in which Logistic Regression, SVM and Naïve Bayes the highest value has 81.96%. Figure 7 showing precision bar plot in which SVM has the highest value 87.8%

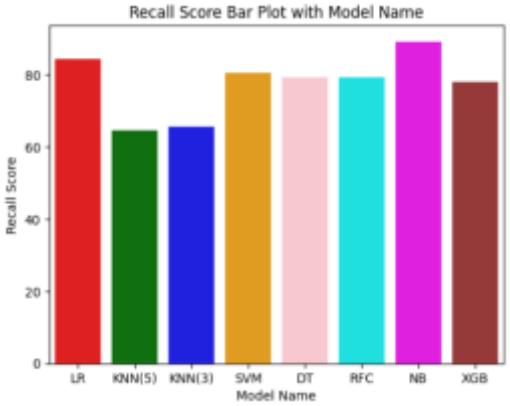

Figure 9 Recall Bar Plot

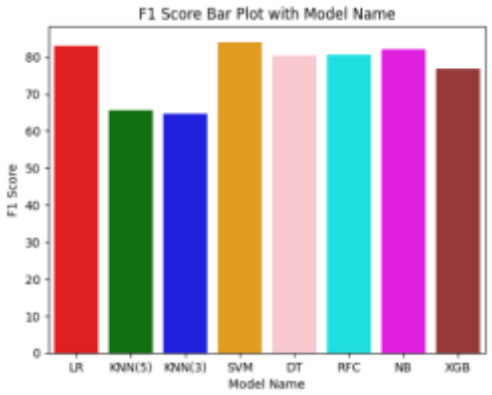

Figure 10 F1 Score Bar Plot

Figure 8 showing Recall bar plot in which Naïve Bayes has the highest value 89.28% and figure 9 showing F1 score in which SVM has the highest value 84.05%

Conclusion

To ensure the heart failure prediction model is reliable and broadly applicable, it is important to test its performance on bigger and more dissimilar datasets. Using to real clinical data can offer valuable perceptions into the models robustness and practical use. Collaborating with healthcare institutions and participating in data-sharing initiatives can help acquire comprehensive datasets, enabling the model to train and be evaluated on more realistic and varied patient data. Intensifying the dataset to include a wider range of patient characteristics enhances the systems predictive correctness and supports its effectiveness in real-world healthcare situations.

REFERENCE

[1] Srinivas Kanakala, Vempaty Prashanthi "Comparative analysis of heart failure prediction using machine learning models"International Journal of Informatics and Communication Technology (IJ-ICT) Vol. 13, No. 2, August 2024, pp. 297~305 ISSN: 2252-8776, DOI: 10.11591/ijict.v13i2.pp297-305

[2] Bhavesh Dhande, Kartik Bamble, Sahil Chavan, Tabassum Maktum "Diabetes & Heart Disease Prediction Using Machine Learning " on ITM Web of Conferences **44**, 03057 (2022) ICACC-2022

[3] Raparthi Yaswanth, Y. Md. Riyazuddin "Heart Disease Prediction using Machine Learning Techniques" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-5, March 2020

[4] Patricia Rufes, J Sorna Jenita, Mabel Rakshitha , Arthika Infanta , Divya M " Heart Disease Prediction Using Machine Learning" International Research Journal on Advanced Engineering Hub e ISSN: 2584-2137 Vol. 02 Issue: 03 March 2024

[5] Govardhan Logabiraman, D.Ganesh, M. Sunil Kumar, A. Vinay Kumar, Nitin Bhardwaj "Heart disease prediction using machine learning algorithms" MATEC Web of Conferences ICMED 2024

[6] Balakrishnan Duraisamy, Rakesh Sunku, Krithik Selvaraj, Vishnu Vardhan Reddy Pilla, Manoj Sanikalaa "Heart disease prediction using support vector machine."Multidisciplinary Science Journal,15th December 2024

[7] Romeo Jousef A Laxamana , Joan Marie Vale "Heart Attack Prediction using Machine Learning Algorithms " J. Electrical Systems 20-5s (2024): 1428-1436[ResearchGate]

[8] Sumit Sharma, Mahesh Parmar "Heart Diseases Prediction using Deep Learning Neural Network Model" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-3, January 2020

[9] Sashank Yadav, Aman Singh, Veena Jadhav, Dr. Rohini Jadhav " Heart Disease Prediction using Machine Learning" July 2022| IJIRT | Volume 9 Issue 2 | ISSN: 2349-6002 [ResearchGate]

[10] Kallepalli Reshma, Pasumarthi Niharika, Javvadi Haneesha, Kodithala Rajavardhan, Sana Swaroop"Multi Disease Prediction System Using Machine Learning" International Research Journal of Modernization in Engineering Technology and Science Volume:06/Issue:02/February-2024

[11] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath "Heart disese prediction using machine learning algorithm" ICCRDA 2020

[12] Dr. Rajani P.K, Kalyani Patil, Bhagyashree Marathe, Prerna Mhaisane, Atharva Tundalwar
"Heart Disease Prediction using Different Machine Learning Algorithms" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 11 Issue: 9s July 2023 [ResearchGate]

[13] D Venkatesh, T Saravanan, D Raghavaraju, M Vijaya Bhaskar, S Vasundra "Prediction of heart disease using machine learning and hybrid methods" International Conference ICOTL 2023

[14] Ramanathan G. Jagadeesha S. N. "Prediction of Coronary Artery Disease using Machine Learning – A Comparative study of Algorithms" International Journal of Health Sciences and Pharmacy (IJHSP), ISSN: 2581-6411, Vol. 7, No. 2, December 2023. [ReasearchGate]

[15] Sanjana Chaudhari, Mr. Chandra Shekhar Gautam, Dr. Akhilesh A. Waoo "Optimizing Heart Disease Prediction Accuracy using Machine Learning Models" IJARESM ISSN: 2455-6211, Volume 12, Issue 6, June-2024

[16] David Lapp Heart Disease Dataset link https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset