2022 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, AUG. 22–25, 2022, XI'AN, CHINA

MACHINE LEARNING-BASED HEART DISEASE PREDICTION: A STUDY FOR HOME PERSONALIZED CARE

Goutam Kumar Sahoo[*,1]*,Keerthana Kanike*[*,2]*, Santos Kumar Das*[*,3] *and Poonam Singh*[*,4]

[*]Department of Electronics and Communication Engineering,
National Institute of Technology Rourkela, India
Email: [1]goutamkrsahoo@gmail.com, [2]keerthana.kanike112@gmail.com,
[3]dassk@nitrkl.ac.in, [4]psingh@nitrkl.ac.in

ABSTRACT

This study develops a framework for personalized care to tackle heart disease risk using an at-home system. The machine learning models used to predict heart disease are Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest and XG Boost. Timely and efficient detection of heart disease plays an important role in health care. It is essential to detect cardiovascular disease (CVD) at the earliest, consult a specialist doctor before the severity of the disease and start medication. The performance of the proposed model was assessed using the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. Compared to all machine learning algorithms, the Random Forest algorithm shows a better performance accuracy score of 90.16%. The best model may evaluate patient fitness rather than routine hospital visits. The proposed work will reduce the burden on hospitals and help hospitals reach only critical patients.

Index Terms— Cardio Vascular Disease (CVD), Data Mining, Healthcare, Machine Learning, Heart Disease Prediction.

1. INTRODUCTION

Heart disease is a type of disease that affects the heart. According to the World Health Organization (WHO), every year more than 17.5 million people die of heart disease [1]. In this state, the heart is usually unable to pump the required amount of blood to various parts of the human body to carry out the routine functions of the body, resulting in irreversible heart failure. Symptoms and signs of heart disease include chest pain, chest discomfort (angina), shortness of breath, and increase in the blood pressure. High cholesterol, high blood pressure or diabetes can also increase the risk of heart disease. There are many prevention methods to combat this disease, such as natural ways of stopping smoking, maintaining a healthy weight, adopting a healthy diet and practicing sports regularly. Prediction of heart disease before occurrence is as important as dealing with the disease. This can be done using different machine learning (ML) prediction techniques. The best model is decided based on the accuracy obtained on the test data. The most important thing to be considered in heart disease prediction is to reduce failure to identify patients with heart disease, which means reducing false negatives [2].

Exploratory Data Analysis (EDA) is the key step to identify important and relevant features to be used in the further process of modeling [3]. EDA includes univariate and bivariate analysis of features in graphical and tabular representations. Univariate analysis like histogram, box-plot, count plot gives a clear idea of each feature. Whereas bivariate analysis such as scatter plot, bar plot, two-way table, and correlation compare two features and mainly compare the features with the target to get a better relationship. Performance metrics like recall, accuracy, etc. play an important role in deciding the correct prediction model. The work by Bhatt *et al.* [4] used the data mining tool Weka to predict heart disease using two classification techniques with two different datasets. The J48 technique was applied to the Hungarian dataset, and Naive Bayes (NB) was applied to the echocardiogram database [5]. A classification accuracy of 82.3% was achieved using the Hungarian dataset with all features, with selected features outperforming an accuracy of 65.64%. The performance of the model depends on the deviation and bias of the dataset [6].

As per research on the machine learning for prediction of heart diseases [6] NB perform well with low variance and high biasness compared to K-Nearest Neighbour (KNN) with high variance and low biasness. The reason for less model performance in case of KNN is that with high variance and low biasness KNN suffers from over fitting. Arpaia *et al.* [7] reported an e-healthcare technology for home care that measures the risk of a cardiac patient using data acquired from the patients body. These data are experimented with a random forest (RF) classifier to classify cardiac risk, and the classification accuracy reported was 80%. A work by M. A. Khan [8], proposed an IoT framework based on modified deep convolutional neural network (MDCNN) to

978-1-6654-8547-0/22/$31.00

evaluate heart disease more accurately. The blood pressure and electrocardiogram (ECG) were acquired using wearable devices like smartwatch and heart monitor device. The acquired data were then transferred to long range (LoRa) cloud using LoRa gateway. This work predicts heart disease using machine learning techniques intended to be used in an embedded system for at-home personalized care as an early diagnosis of CVD patient.

The next part of the report is organized as follows: Section 2 presents the problem formulation and methodology. Section 3 explains the experimental results and analysis. The conclusion of the work is mentioned in Section 4.

2. PROBLEM FORMULATION AND METHODOLOGY

This study develops a framework for personalized care aimed at tackling the risk of CVD using an at-home system. The best machine learning models will be used to evaluate the patient's fitness instead of making regular visits to the hospital. This will reduce the burden on hospitals and help hospitals reach only critical patients. A system for early heart disease screening is shown in Fig. 1. Heart patient can predict heart disease at home using ML-based prediction system. If the system predicts the patient with heart disease, a computer-based alert will be generated suggesting the doctor to be consulted for diagnosis and will also store the estimated abnormal data related to CVD for future requirement.

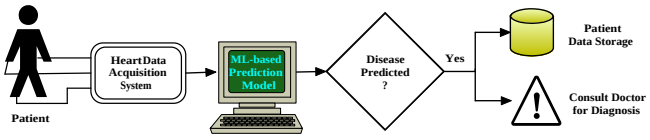


Fig. 1: Preliminary Heart Disease Investigation Framework

The main task is the methodology for the detection and prediction of heart disease. According to literature studies, this task can be classified as a classification problem to classify whether a person has heart disease or not. The task requires heart patient data to predict heart disease. For experimental work, a publicly available Cleveland heart disease dataset from the UCI Machine Learning Repository [9] is used. The dataset has 303 rows and 75 attributes, however all published experiments refer to using a subset of 14 of them. The patients are both male and female in the database. The patient's age ranges between 29-77 years. Most patients have chest pain typical of the angina type and the heart rate is between 71-262 bpm. This dataset is divided as two splits namely train and test split using python library scikit-learn. The step-wise methodology adopted for developing the heart disease prediction model is presented in Fig. 2 and discussed in detail.

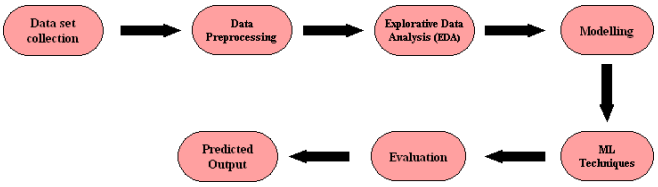


Fig. 2: Heart Disease Prediction Methodology

2.1. Dataset Collection

Data is an essential requirement for the prediction of heart disease. The performance of various ML algorithms will be evaluated using a publicly available dataset, the Cleveland Heart Disease Dataset from the UCI Machine Learning Repository. A brief description of each of the features is given in Table 1 for a better understanding. There are 14 features, of which 13 are considered independent variables/features, and the one, namely the Target feature, is known as the dependent variable/feature.

Table 1: Dataset details for heart disease prediction [9].

Sl. No.	Features	Description
1	Age	Age in years
2	Sex	male = 1 ; female = 0
3	Cp	The type of Chest pain categorized into 4 values
4	Trestbps	Level of blood pressure at rest (in mm Hg)
5	Chol	Serum cholesterol im mg/dl
6	Fbs	Blood sugar levels on fasting>120 mg/dl (1 = true; 0 = false).
7	Restecg	Results of an electrocardiogram while at rest
8	Thalach	The accomplishment of the maximum heart rate
9	Exang	Angina induced by exercise (1 = yes; 0 = no)
10	OldPeak	Exercise-induced ST depression in comparison to rest
11	Slope	ST segment measured in terms of the slope
12	Ca	Fluoroscopy coloured major vessels numbered from 0 to 3
13	Thal	Status of the heart (Normal = 3; fixed defect = 6 ; reversible defect = 7)
14	Target	Heart disease diagnosis Healthy = 0; Diseased = 1

2.2. Data Pre-Processing

In general, medical records that are not always complete may contain missing and unwanted data. Data pre-processing is used to remove the number of discrepancies associated with the data, remove duplicate records, normalize values, account for missing data, etc. The primary step in this data pre-processing is to check for null values and treat them by filling in or dropping them. After importing a dataset using the Python library pandas, common data pre-processing methods such as data cleaning, data transformation, efficient processing, and classification are performed. No unique method of data processing is used in this work.

2.3. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) provides the detailed features of the dataset through different graphs and tables. For example, Fig. 3 shows the relation of a categorical variable with four different categories namely "asymptotic", "non-anginal pain", "atypical angina", "typical angina". This plot mainly represents the count of people with "cp" divisions according

to target values. We can observe that people suffering from heart disease are mostly from the “cp” type of “non-anginal pain” and people without a heart disease are from “typical angina”.

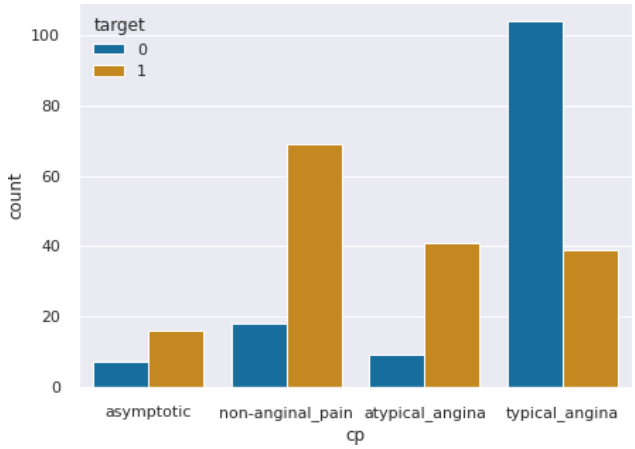


Fig. 3: Distribution of people according to ’cp’ Categories.

2.4. Modeling

In prediction problem, the system modeling is the process where the actual training happens using data mining and machine learning algorithms. A study using a publicly available dataset is made to explore the feasibility of predictive models for the early prediction of heart diseases. Different machine learning algorithms used for prediction process are: logistic regression (LR), K-Nearest Neighbour (KNN), support vector machine (SVM), decision tree (DT), random forest (RF), Naive Bayes (NB) and eXtreme Gradient Boosting (XG Boost). The step-wise approach of heart disease prediction modeling is presented in Algorithm 1.

Algorithm 1 : Algorithm for Heart Disease Prediction

1: Load the Cleveland heart disease dataset.
2: Apply the data pre-processing techniques.
3: Perform exploratory data analysis (EDA).
4: Divide the total data as feature variable and target variable.
5: Split the dataset into training and testing samples.
6: Build the model using machine learning algorithm.
7: Validate the model for prediction of healthy and diseased class.
8: Calculate the performance measurement parameters.
9: **If** (CVD detected), **Then** alert the patient to consult doctor and store the patient data locally.

2.5. Evaluation Metrics

After the modeling process, the main task is to evaluate the model on the test data to check whether a particular machine learning algorithm predicts the persons correctly with and without heart disease. This can be obtained using evaluation metrics of python library called scikit-learn. Using the confusion matrix, we can visualize the performance of computational intelligence techniques. In the confusion matrix, there are four classification performance indices, i.e., TP = True Positive (correctly identified), TN = True Negative (incorrectly identified), FP = False positive (correctly rejected), FN = False negative (falsely rejected). The expressions used to evaluate the various performance parameters are given in Equations (1) to (4).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. Experimental Setup

The experiment was implemented in Python 3.8 using a single computer (Acer Aspire A515-54G , Intel(R) Core(TM) i5-8265U CPU @1.60GHz, RAM 8 GB) with Windows 10.

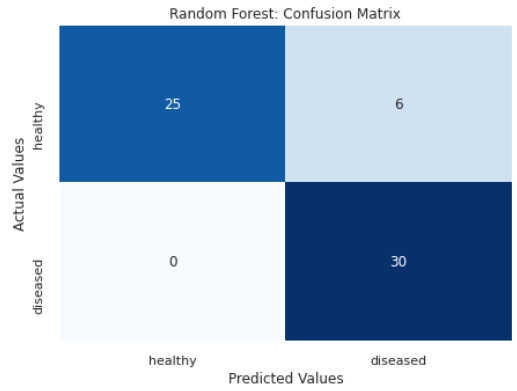
3.2. Result Analysis

The methodology adopted in training of the dataset using various ML algorithms provides predictive performance of heart disease in terms of evaluation metrics. The best three performances with high accuracy are analyzed in detail with the help of model performance learning curve, classification report and confusion matrix. The confusion matrix is generated by considering 20% of the real data (ie, test split) and made to run on a model trained with 80% real data (ie, train split). A confusion matrix is a form of table that represents the actual true and false values as well as the estimated true and false values. The training accuracy is generated using the train data, the test data, and the epochs, which tells us the accuracy for each epoch we run. A model learning curve is used to analyze the training process, which gives an idea of the model performance. Similarly, the training loss and testing loss curves tell about the error predicted from the actual values while training using the trained models.

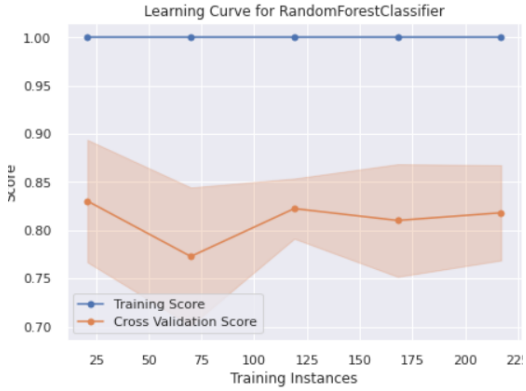
3.2.1. Random Forest Classifier Classification Report

The confusion matrix and training accuracy performance can be seen in Fig. 4. From the confusion matrix in Fig. 4a, the algorithm predicts 25 healthy individuals, whereas originally,

it was 31. The true positive rate (or recall) is 0.81, and the precision is 1.00. Similarly, the diseased class algorithm predicts 36 healthy individuals whereas originally, it was 30. The true positive rate (or recall) is 1.00, and the precision is 0.83. The overall prediction accuracy has been found to be 0.9016 or 90.16%. Fig. 4b shows that the training score for the random forest classifier remains constant throughout training, while the cross-validation score is finally below the starting score. Similarly, Fig. 4c represents the training loss performance curves.



(a) Confusion Matrix



(b) Training Accuracy

(c) Training Loss

Fig. 4: Random forest Model Performance.

3.2.2. Logistic Regression Classification Report

The confusion matrix and the training accuracy performances can be seen in Fig. 5. From the confusion matrix in Fig. 5a, the algorithm predicts 26 healthy persons, whereas initially, it was 31. The actual positive rate is 0.81, and the precision is 0.96. Similarly, the diseased class algorithm predicts 35 healthy persons, whereas initially, it was 30. The true positive rate is 0.97, and the precision is 0.83. The overall prediction accuracy is 88.52%. Fig. 5b contains two individual graphs, one representing the accuracy score on training data and the other on validation data for continuous training instances. One can see that the training score for Logistic Regression gradually decreases, whereas the cross validation score gradually increases. Similarly, Fig. 5c represents the training loss performance curves.

3.2.3. XG Boost Classification Report

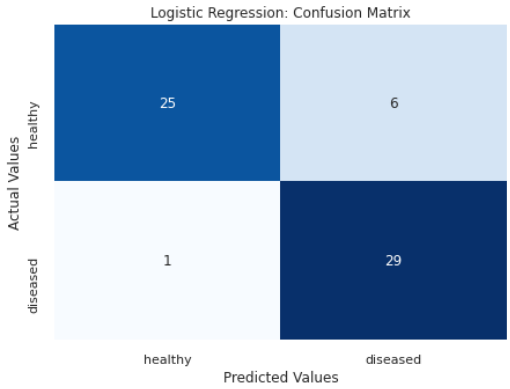
The confusion matrix and training accuracy performance can be seen in Fig. 6. From the confusion matrix in Fig. 6a, the algorithm predicts 28 healthy individuals, whereas originally, it was 31. The true positive rate is 0.77 and the accuracy is 0.86. Similarly, the diseased class algorithm predicts 33 healthy individuals whereas originally, it was 30. The true positive rate is 0.87 and the accuracy is 0.79. The overall prediction accuracy has been found to be 88.52%. The learning curve in Fig. 6b of the training data split using the XG Boost classifier shows that the training score for XG Boost gradually increases, while the cross-validation score does not increase gradually, but the overall initial increase in comparison. Similarly, Fig. 6c represents the training loss performance curves.

3.3. Performance Evaluation and Result Comparison

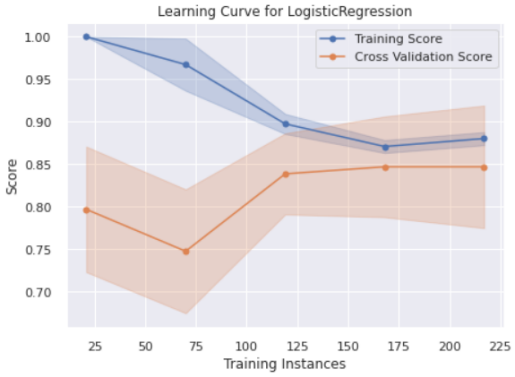
The evaluation metrics are calculated, and their representation can be visualized from Table 2. The metrics considered in this are Test accuracy, Precision, Recall, and F1-Score, which are the most important metrics to decide the better model. From the table, we can conclude that Random Forest has a maximum test accuracy of 90.16%.

Table 2: Performance Evaluation of Different ML Models.

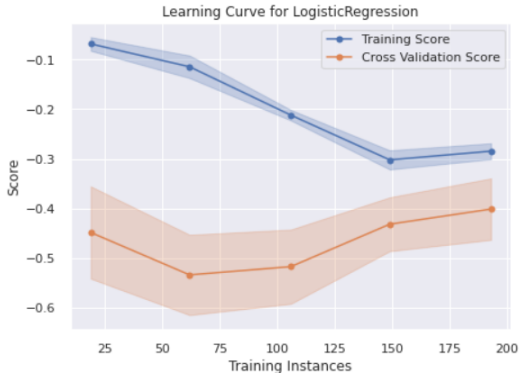
Algorithms	Test Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR	**88.52**	82.85	96.66	89.23
KNN	83.6	81.25	86.66	83.87
SVM	83.6	76.31	96.66	85.29
NB	75.4	67.44	96.66	79.45
DT	78.68	77.41	80	78.68
RF	**90.16**	83.33	100	90.9
XG Boost	**88.52**	84.84	93.33	88.88



(a) Confusion Matrix



(b) Training Accuracy



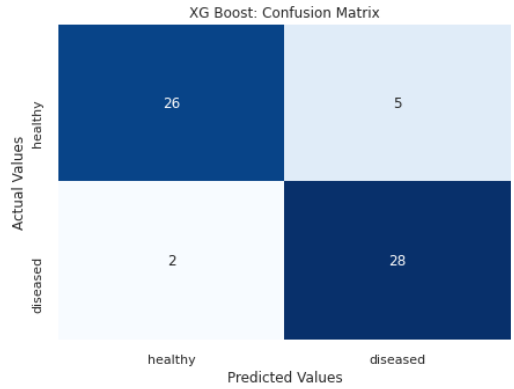
(c) Training Loss

Fig. 5: Logistic-Regression Performance Measures.

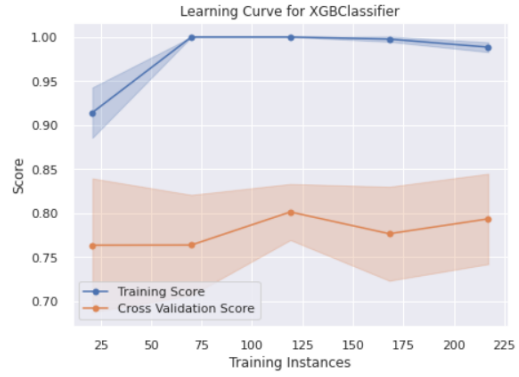
Table 3: 10-fold cross-validation to report std.

ML algorithms	Accuracy (%)	std
LR	82.183	0.067
KNN	80.15	0.058
SVM	83.03	0.07
DT	78.05	0.08
XG Boost	80.09	0.081
RF	77.61	0.085

Cross-validation (CV) is a technique for evaluating ML models. In K-fold CV the data is split into k-equal or nearly equal folds. A 10-fold cross validation is performed



(a) Confusion Matrix



(b) Training Accuracy

(c) Training Loss

Fig. 6: XG Boost Model Performance.

on each models created using different machine learning algorithms. The accuracy and standard deviation (std) of error are obtained, which can be seen from the Table 3. A low standard deviation means that the data is very closely related to the average, thus the model is reliable.

3.4. Comparative Study

Table 4 shows the accuracy result of the state-of-art techniques compared with that of our experimental study based on Cleveland Data-set. The performance of the LR technique in [8, 10, 13, 14] achieved accuracy performances of 85%,

Table 4: Comparison of our study with the existing methods.

Reference	Algorithms	Accuracy (%)
A. K. Dwivedi [10]	LR	85
Shah *et al.* [11]	RF	86.84
M. A. Khan [8]	LR	87.8
Ayon *et al.* [12]	RF	89.41
	LR	92.41
Tougui *et al.* [13]	LR	83.5
Bharti *et al.* [14]	LR	83.3
	RF	80.3
Divya *et al.* [15]	RF	89.97
	LR	92.01
Our Study	**LR**	**88.52**
	RF	**90.16**
	XG Boost	**88.52**

87.8%, 83.5%, and 83.3%, respectively, which are somehow similar and lower than the performance of our study. The study presented in [11, 12, 14, 15] used random forest and achieved accuracy performance of 86.84%, 89.41%, 80.30% and 89.97% respectively. However, our study on RF provides a better accuracy performance of 90.16%. The work done by Divya *et al.* and Ayon *et al.* [12, 15] shows improvement in LR techniques and poor performance on RF techniques. The present work does not show superiority over all the results of prior studies. However, this work achieved superior performance only on the RF algorithm, i.e., the maximum accuracy of 90.16%, and showed comparable performance to the LR and XG Boost algorithms. The XG Boost algorithm performance obtained is 88.52% and is the best on this dataset to our knowledge.

4. CONCLUSIONS AND FUTURE WORKS

This work presented a novel framework for heart disease prediction by applying ML techniques. The machine learning models used to predict heart disease are LR, KNN, SVM, NB, DT, RF, and XG Boost. Compared with all the ML algorithms, the RF algorithm shows superior performance accuracy of 90.16%. The best machine learning models can be deployed to evaluate the patient's fitness instead of making regular visits to the hospital. This will reduce the burden on hospitals and help hospitals reach only critical patients. The above predictions are essential to notify the doctor before the seriousness of the disease and to start the medication. This study develops a framework for personalized care aimed at tackling the risk of CVD using an at-home system. In the future, pre-trained deep learning model can be experimented to improve the prediction accuracy for low-cost embedded system applications.

5. REFERENCES

[1] Mariano Sanz et al., "Periodontitis and cardiovascular diseases: Consensus report," *J. Clinical Periodontology*, vol. 47, no. 3, pp. 268–288, 2020.

[2] Ramzi A Haraty et al., "An enhanced k-means clustering algorithm for pattern discovery in healthcare data," *Int. Journal Dist. Sensor Networks*, vol. 11, no. 6, pp. 615740, 2015.

[3] Charu C Aggarwal, *Data mining: the textbook*, Springer, 2015.

[4] Anurag Bhatt et al., "Data mining approach to predict and analyze the cardiovascular disease," in *Proc. 5th Int. Conf. Frontiers in Intelli. Comput. Theory Appl.* Springer, 2017, pp. 117–126.

[5] Dheeru Dua and C Graff, "UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2019)," 2019.

[6] Himanshu Sharma and MA Rizvi, "Prediction of heart disease using machine learning algorithms: A survey," *Int. Journal Recent Innov. Trends Comput. Commun.*, vol. 5, no. 8, pp. 99–104, 2017.

[7] Pasquale Arpaia et al., "Conceptual design of a machine learning-based wearable soft sensor for non-invasive cardiovascular risk assessment," *Measurement*, vol. 169, pp. 108551, 2021.

[8] Mohammad Ayoub Khan, "An iot framework for heart disease prediction based on mdcnn classifier," *IEEE Access*, vol. 8, pp. 34717–34727, 2020.

[9] UCI Machine learning repository., ," *Heart Disease Dataset.* [Online]. `Available:~https://archive.ics.uci.edu/ml/datasets/Heart+Disease`.

[10] Ashok Kumar Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing Appl.*, vol. 29, no. 10, pp. 685–693, 2018.

[11] Devansh Shah et al., "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, no. 6, pp. 1–6, 2020.

[12] Safial Islam Ayon et al., "Coronary artery heart disease prediction: a comparative study of computational intelligence techniques," *IETE J. Res.*, pp. 1–20, 2020.

[13] Ilias Tougui et al., "Heart disease classification using data mining tools and machine learning techniques," *Health Technol.*, vol. 10, pp. 1137–1144, 2020.

[14] Rohit Bharti et al., "Prediction of heart disease using a combination of machine learning and deep learning," *Comput. Intell. Neuroscience*, vol. 2021, 2021.

[15] K Divya et al., "An iomt assisted heart disease diagnostic system using machine learning techniques," in *Cogn. Internet Med. Things Smart Healthcare*, pp. 145–161. Springer, 2021.