Predicting Heart Diseases through Feature Selection and Ensemble Classifiers

Predicting Heart Diseases through Feature Selection and Ensemble Classifiers

Shivangi Diwan[1], Gajendra Singh Thakur[2], Sunil K. Sahu[2], Mridu Sahu[1] and N. K. Swamy[2]

[1]National Institute of Technology, Raipur, India.
[2]School of Science, ISBM University Nawapara (Kosmi), Block & Tehsil- Chhura, Gariyaband, Chhattisgarh- 493996, India.

Email- shivangi.diwan10@gmail.com

Abstract. Heart diseases or Cardiovascular Diseases are the leading cause of death globally. Amid the Covid-19 pandemic, the toll has further increased and is prevalent among all age groups. The reasons are associated with various side effects of lockdown or socio-economic affairs. It becomes extremely important to strengthen our research on diagnosis systems to timely and accurately identify the disease. This paper is an attempt to predict a healthy or heart patient using ensemble machine learning methods depending on selected features. The proposed model shows that after performing feature selection the ensemble models give optimum accuracy with significantly lesser features.

1. Introduction

The burden of cardiovascular disease (CVD) in India is one of the highest in the world. The number of CVD fatalities in India is expected to increase from 2.26 million in 1990 to 4.77 million in 2020 [1]. If diagnosed early these CVDs can be treated and hence lives can be saved. Although the interpretation of medical data is a challenging task because of its complexity, the diagnosis can be accurately done using Artificial Intelligence (AI). Machine learning techniques like ensemble technique provide a great scope of enhancement in the approach to perform any classification task. The proposed model uses ensemble machine learning techniques like Classification And Regression Trees(CART), Gradient Boosting Machines(GBM), Adaboost, k-Nearest Neighbors, Multilayer perceptron, Stochastic Gradient Descent, Support Vector Classifiers, Naïve Bayes for prediction and also minimized the feature and reevaluated the model [2]. A comprehensive dataset for heart disease prediction is used which comprises essential features like exercise-induced angina, chest pain type, fasting blood sugar, resting ecg , st slope etc. [3]. For heart disease prediction, Ghosh P et al used the LASSO and Relief techniques for selection of feature and used the ensemble method of machine learning to create hybrid classifiers like K-Nearest Neighbors

Bagging Method (KNNBM), Random Forest Bagging Method (RFBM) and Gradient Boosting Boosting Method (GBBM) and more for training and testing [4]. For the prediction of cardiac illnesses, Bharti R et al used Logistic Regression, KNeighbors Classifier, and Random Forest Classifier [5]. Li J et al proposed Relief, MRMR, LASSO, and LLBFS as the basic feature selection techniques and Fast Conditional Mutual Information (FCMIM) a unique feature selection approach for solving feature selection problems. The system uses the LOSO cross-validation approach to determine the optimum hyperparameters. The Cleveland heart disease dataset was used in this study. The results showed that ANN with Relief is the most accurate detecting system. The most appropriate characteristics include chest discomfort and exercise-induced angina [6]. Using a hybrid random forest with a linear model, Mohan S et al introduced a novel way for selecting the most appropriate features and classification using machine algorithms (HRFLM), which boosted prediction accuracy to 88.7% [7]. Mienye I et al proposed the multiple CART models which are combined into a homogenous ensemble using an accuracy-based weighted aging classifier ensemble, which is a variation of the weighted aging classifier ensemble (WAE). The strategy guarantees that the best possible results are attained. On the Cleveland and Framingham datasets, the experimental results achieved classification accuracies of 93 percent and 91 percent, respectively[8]. The proposed work aims to create a high-performing ensemble learning model for predicting heart disease risk using the optimal feature selection method.

2. Methodology

The dataset is first preprocessed and the machine learning models predicting the appropriate target with concise features is the target of the work. The flow of the work can be shown in Figure 1 shown below.

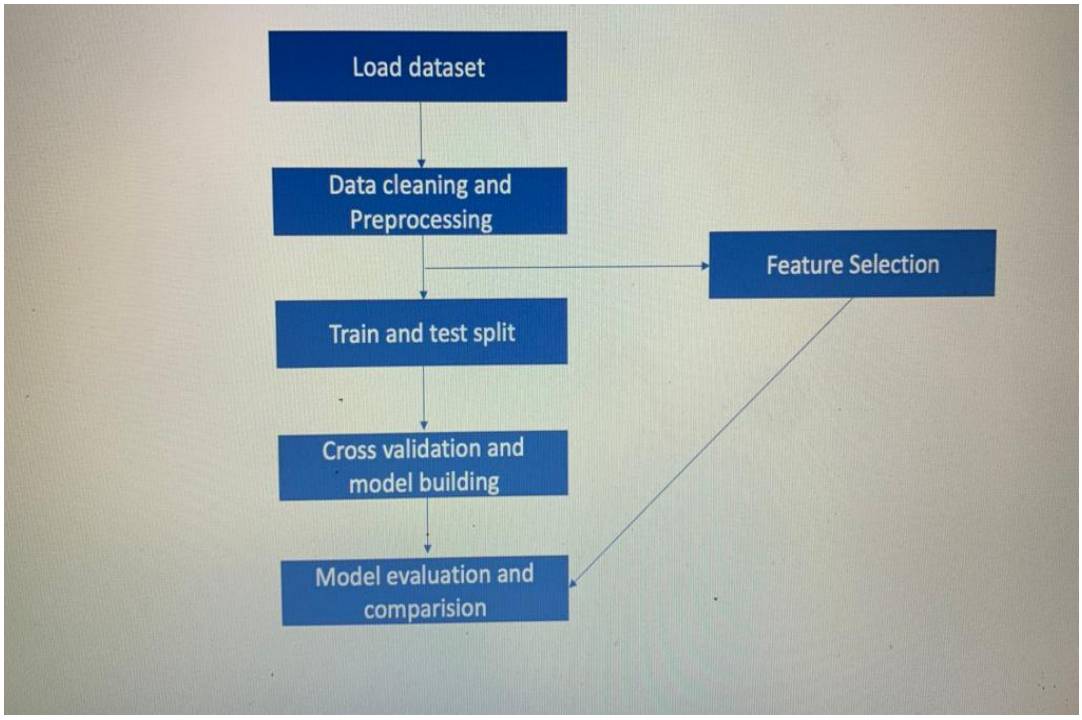

Figure 1: Flow chart of the proposed work.

2.1. Dataset

The dataset is comprehensive and is obtained from the combination of five very popular datasets which are the Long Beach VA dataset, Switzerland dataset, Hungarian dataset, Statlog dataset, and Cleveland dataset, It consists of 1190 records and 11 features. This dataset has been presented by Alizadehsani et al and is briefly described below in Table 1[3]. The 11 features with description and the data type are shown in the table.

Table 1: Comprehensive dataset

Feature	**Description**	**Datatype**
Age	Patients age in year	Numeric
Sex	Gender,Male-1,Female-0	Nominal
Chest Pain Type	1-typical, 2-typical angina,3-non- anginal pain, 4-asymptomatic	Nominal
resting bp s	at resting mode, blood pressure (mm/HG)	Numeric
Cholesterol	Serum cholestrol in mg/dl	Numeric
Fasting blood sugar	If greater than 120 mg/dl 1, otherwise 0	Nominal
Resting ECG	Normal-0, ST-T wave abnormality -1, Left ventricular hypertrophy-2	Nominal
Max heart rate	Maximum heart rate	Numeric
Exercise angina	0 represents No whereas 1 represents Yes	Nominal
Oldpeak	Exercise-induced ST-depression in comparison with the state of rest	Numeric
ST slope	0: Normal 1: Upsloping 2: Flat 3: Downsloping	Nominal

2.2. Preprocessing

The dataset is first cleaned and preprocessed as the name of the columns are changed, features are encoded into categorical variables, and the null values are dropped. Exploratory data analysis is done and Figure 2-5 shows the distribution of numerical features. Figure 2 is a scatter plot of the maximum heart rate achieved and age.

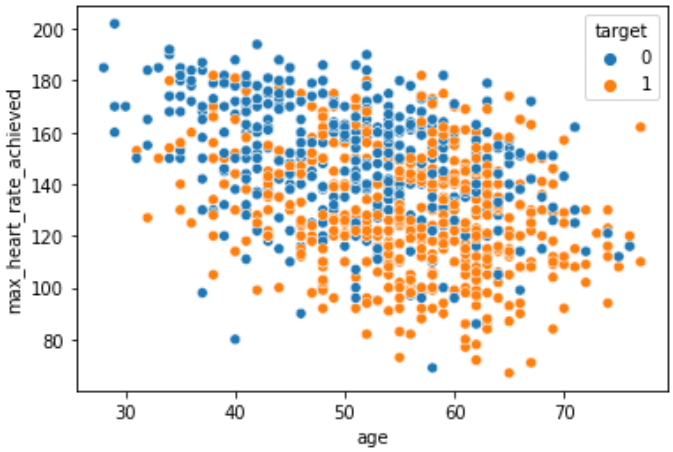

Figure 2: Scatterplot of maximum heart rate achieved and the age

From the above figure, outliers can be easily spotted.

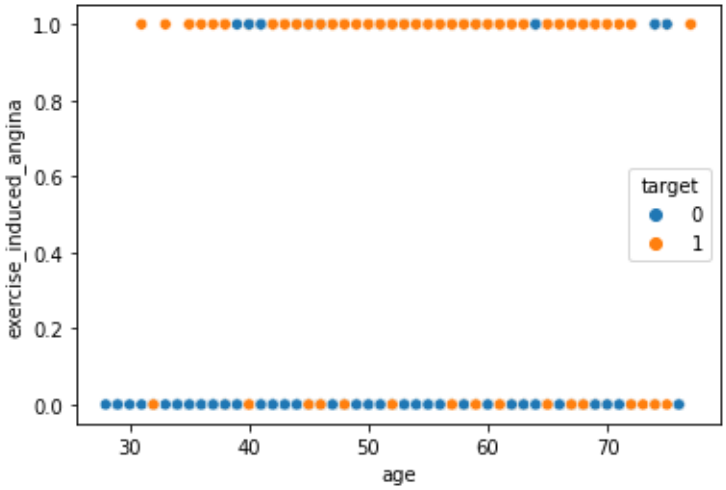

Figure 3: Scatter plot of exercise-induced angina and age.

Here target value 0 depicts No angina is present whereas, 1 depicts the presence of angina.

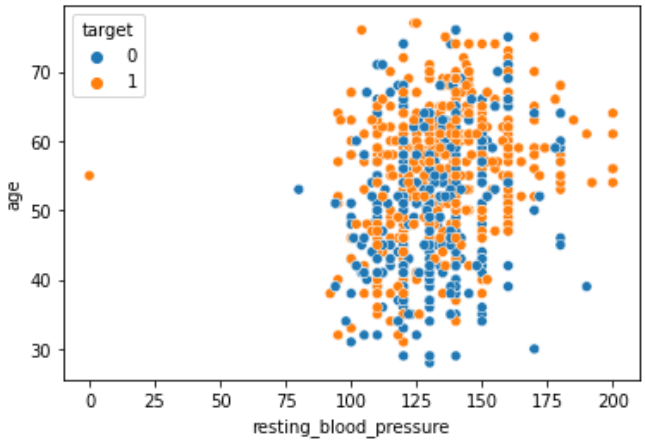

Figure 4: Scatter plot of age vs resting blood pressure

Patient with age between 50-60 is found to be heart patient with resting blood pressure less than 25, so the outliers are present here.

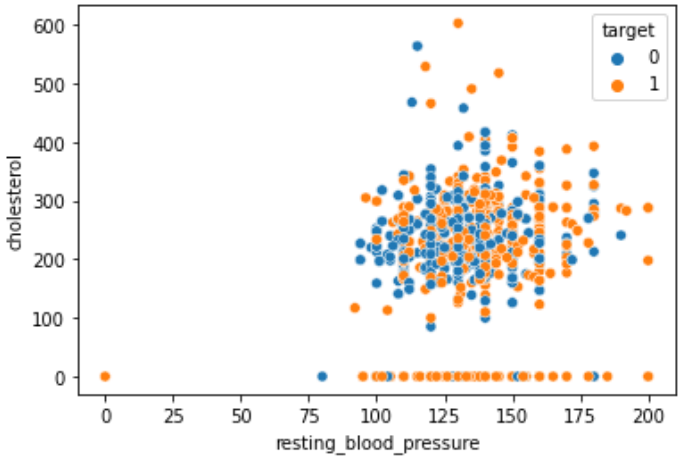

Figure 5: Scatter plot of Cholesterol and resting blood pressure

A heart Patient with no cholesterol and no resting blood pressure is an outlier. The outliers are removed systematically using the z-score of numeric columns in the dataset and defining the threshold for filtering. We use 80 percent train data and 20% test data to train and test the data. MinMax Scalar is used to normalize the features [2]. We developed various baseline models, performed 10-fold cross-validation, then utilized the ensemble approach to choose and determine the top-performing baseline models. The models deployed in the proposed work are Adaboost, Classification And Regression Trees(CART), Gradient Boosting Machine(GBM), kNN, Multilayer Perceptron(MLP), Stochastic Gradient Descent(SGD), Support Vector Classifier(SVC), and Naïve Bayes(NB). The models are evaluated based on performance metrics like accuracy, sensitivity, specificity, F1 score, Log loss, Mathew correlation coefficient.

2.3. Feature Selection

Different features are selected using feature selection algorithms like Pearson's Correlation coefficient method, Chi-selector, Recursive Feature Elimination(rfe) selector, Embedded lr selector as the embedded algorithm is used for selecting feature, logistic regression L2 penalty for feature selection, Light GBM is used as a selector. The majority voting feature selection is used to select the features with majority voting. A total of 7 features are selected, after which models are again evaluated and compared based on performance metrics. Table 2 shows the majority voting of the features.

Table 2: Majority voting feature selection (T=true and F=false)

S.No	Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	Light GBM	Total
1.	st_slope_flat	T	T	T	T	T	T	6
2.	st_depression	T	T	T	T	T	T	6
3.	Max_heart_rate_achieved	T	T	T	F	T	T	5
4.	Exercise_induced_angina	T	T	T	F	T	T	5
5.	Cholesterol	T	F	T	T	T	T	5
6.	Age	T	T	T	F	T	T	5
7.	st_slope_upsloping	T	T	T	F	T	F	4
8.	Sex	T	T	T	T	F	F	4
9.	Chest_pain_type_non_anginal_ pain	T	T	T	T	F	F	4
10.	Chest_pain_type_typical_ angina	T	T	T	T	F	F	4
11.	resting_blood_pressure	F	F	F	F	T	T	2

2.4. Performance metrics

The relevant assessment criteria for the proposed work are

2.4.1. Accuracy= $\frac{TP+TN}{TP+FN+TN+FP}$

It is the measure of correct classifications [9], [10].

2.4.2. Specificity = $\frac{TN}{TN+FP}$

2.4.3. Sensitivity = $\frac{TP}{TP+FN}$ *=Recall*

The ratio of the number of accurate positive forecasts to the total number of actual positive cases is known as sensitivity[11].

2.4.4. Precision=$\frac{TP}{TP+FP}$

The ratio of correct positive predictions to the number of positive results expected is known as precision.

2.4.5. Confusion Matrix

It's an NxN matrix that helps evaluate a machine learning model's performance in a classification problem[12].

2.4.6. AUC

To discriminate among the classes AUC is a summary of ROC and evaluates a classifier's capacity.

2.4.7. ROC

TPR is shown against FPR on the Receiver Operator Characteristic (ROC) curve, which is used to evaluate binary classification problems[13].

2.4.8. F1 Score=$\frac{2(Recall \times Precision)}{Recall+Precision}$

The harmonic mean of precision and sensitivity is the F-score[8].

2.4.9. Log Loss

Using logarithmic loss, the performance of a classification model with a probability value between 0 and 1 as a prediction input is evaluated. The purpose of our machine learning system is to reduce the value to the smallest amount achievable. The log loss would be 0 in a perfect model[6].

2.4.10. Mathew Correlation Coefficient (MCC) = $\frac{(TP \times TN - FP \times FN)}{\sqrt{((TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN))}}$

MCC has a top value of +1 and the worst value of -1. The statistical rate results in a high score only if the forecast produces better results in all areas (TP, FN, TN, FP)[14] [15].

3. Results and Discussion

The findings show the performance of several models as well as metrics. The performance metrics of different baseline models individually before and after minimizing the feature using feature selection are as shown in Tables 3 and 4 respectively. The best performing model is Classification And Regression Trees(CART) with 87.65% accuracy and 0.75 MCC value.

Table 3: Performance metrics of models before feature selection

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log loss	MCC
MLP	0.863829	0.86400	0.878048	0.848214	0.870967	0.863131	4.70321	0.726949
SVC	0.855319	0.85600	0.869918	0.839285	0.862903	0.854602	4.99717	0.709874
SGD	0.855319	0.83969	0.894308	0.812500	0.866141	0.853404	4.99710	0.710740
Ada Boost	0.834042	0.83333	0.853658	0.812500	0.843373	0.833079	5.73203	0.667176
CART	0.876595	0.88524	0.878048	0.875000	0.881632	0.876523	4.26227	0.752775
GBM	0.876595	0.87903	0.886178	0.866071	0.882590	0.876125	4.26228	0.752578
kNN	0.863829	0.86991	0.869918	0.857142	0.869918	0.863530	4.70320	0.727061
NB	0.838297	0.84000	0.853658	0.821428	0.846774	0.837543	5.58506	0.675725

Table 4: Performance metrics of models after feature selection

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log loss	MCC
MLP	0.804255	0.78947	0.853658	0.75000	0.820312	0.801829	6.76087	0.608312
SVC	0.825531	0.825396	0.845528	0.803571	0.835341	0.824549	6.02598	0.650091
SGD	0.719148	0.890410	0.52824	0.928571	0.663265	0.728513	9.700879	0.493269
Ada Boost	0.795744	0.800000	0.813008	0.776785	0.806451	0.794896	7.054813	0.590351
CART	0.868085	0.883333	0.8617886	0.875000	0.872427	0.868394	4.556226	0.736147
GBM	0.846808	0.848000	0.8617886	0.830357	0.854838	0.846072	5.291111	0.692799
kNN	0.519148	0.536585	0.5409836	0.495575	0.538775	0.518279	16.60820	0.036572
NB	0.510632	0.528000	0.5409836	0.477876	0.534412	0.509429	16.90215	0.018884

The confusion matrices of the best performing models CART and GBM are shown in Figures 6 and 7 respectively.

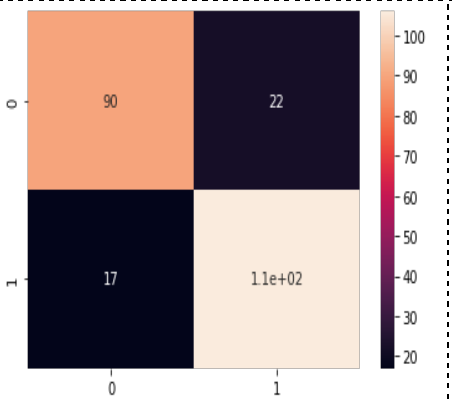

Figure 6. Confusion matrix of CART

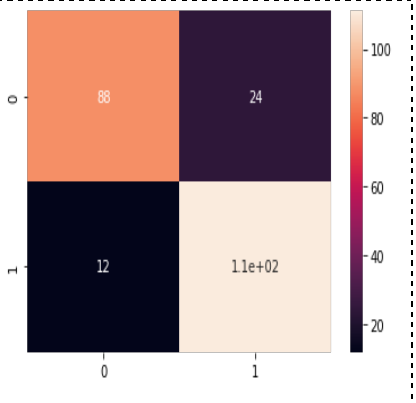

Figure 7. Confusion matrix of GBM

The Precision-Recall curve is shown in Figure 8 and the ROC curve is shown in Figure 9.

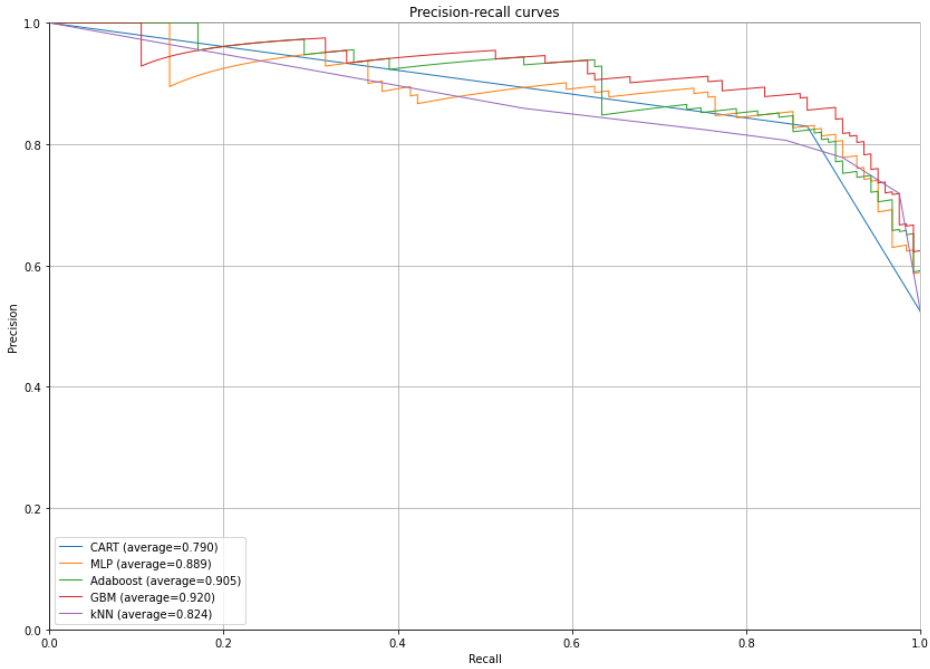

Figure 8: Precision-Recall curves for CART, MLP, Adaboost, GBM, and kNN before feature selection

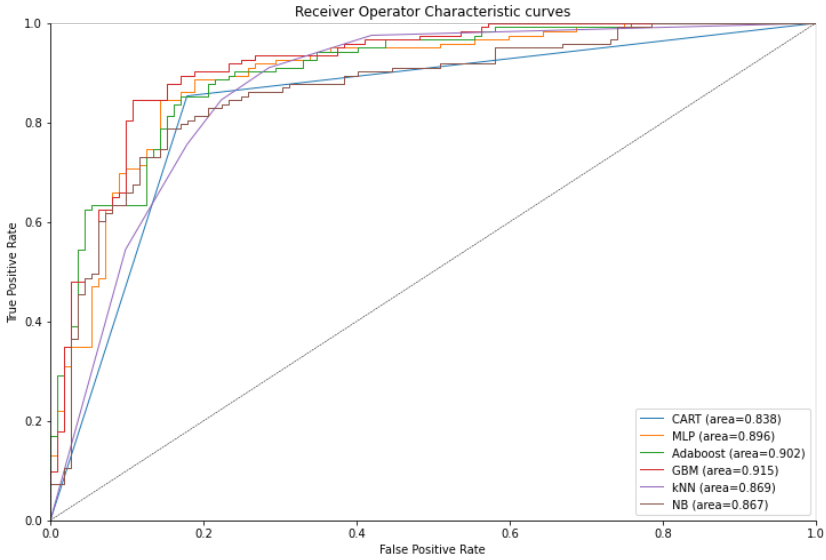

Figure 9: ROC curves for CART, MLP, Adaboost, GBM and kNN, and Naïve Bayes before feature selection

The comparison of the proposed work with the other research/study is shown in Table 5.

Table 5: Comparative analysis

Research/Study	**Algorithm used**	**Accuracy(%)**
J P Li et al	Logistic Regression	84
	KNN	59
	ANN	60
	SVM	57
	Naïve Bayes	75
	Decision Trees	70
Katarya et al	Logistic Regression	93.4
	Naïve Bayes	90.1
	SVM	92.3
	KNN	71.42
	Decision Trees	81.31
	Random Forest	95.6
	ANN	92.3
	DNN	76.9
	MLP	75.42
Proposed work	MLP	80.4
	SVC	82.5

SGD	71.2
Ada Boost	79.5
CART	**87**
GBM	84.7
kNN	51.9
NB	51

4. Conclusion and Future work

This research investigates ensemble machine learning algorithms for predicting cardiac disease. The best performing models are found to be CART and GBM. The most relevant feature voted as the majority are st_slope_flat and st_depression with the highest number of votes whereas, maximum heart rate achieved exercise induced angina and cholesterol as the second-highest vote. In the future, we will implement deep learning algorithm on the dataset to enhance the performance in disease prediction and classification.

References

[1] M. D. Huffman *et al*., "Incidence of cardiovascular risk factors in an Indian urban cohort: Results from the New Delhi Birth Cohort," *Journal of the American College of Cardiology*, vol. 57, no. 17. Elsevier USA, pp. 1765–1774, Apr. 26, 2011. doi: 10.1016/j.jacc.2010.09.083.

[2] Y. M. C Zhang, *Ensemble machine learning: methods and applications*. 2012.

[3] R. Alizadehsani *et al*., "A database for using machine learning and data mining techniques for coronary artery disease diagnosis," *Scientific Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1038/s41597-019-0206-3.

[4] P. Ghosh *et al*., "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[5] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2021, 2021, doi: 10.1155/2021/8387680.

[6] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[8] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, Jan. 2020, doi: 10.1016/j.imu.2020.100402.

[9] R. Layton, *Learning data mining with python*. 2015.

[10] R. Katarya and S. K. Meena, "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis," *Health and Technology*, vol. 11, no. 1, pp. 87–97, Jan. 2021, doi: 10.1007/s12553-020-00505-7.

[11] S. N. Khan *et al*., "Comparative analysis for heart disease prediction," *International Journal on Informatics Visualization*, vol. 1, no. 4–2, pp. 227–231, 2017, doi: 10.30630/joiv.1.4-2.66.

[12] J. Emakhu, S. Shrestha, and S. Arslanturk, "Prediction System for Heart Disease Based on Ensemble Classifiers."

[13] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics in Medicine Unlocked*, vol. 26, Jan. 2021, doi: 10.1016/j.imu.2021.100655.

[14] P. N. , S. M. , K. v an, *Introduction to dats mining*. Pearson Education India, 2016.

[15] G. Bonaccorso, *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd, 2018.