International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7208
***Article Received**: 26 April 2023 **Revised:** 14 June 2023 **Accepted:** 30 June 2023*

Heart Disease Prediction Using Ensemble Model and Hyperparameter Optimization

Jonathan Ivan[1], Simeon Yuda Prasetyo[2]
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
jonathan.ivan@binus.ac.id
[2]Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
simeon.prasetyo@binus.ac.id

Abstract— Heart disease is a major global health concern that responsible for significant mortality rates, killing 17.9 million people each year on average. To overcome this problem, machine learning can assist in forecasting the occurrence of heart disease, aiding in its prevention and treatment. This paper explores several classification models to forecast heart disease. This paper also utilizes the hyperparameter tuning method via grid search cv to enhance the accuracy of the models. Finally, the experiment concludes with an ensemble vote on all hyperparameter-tuned classification models. The x-gradient boost and random forest classifier deliver the best outcomes, with an accuracy of 88.04% and 89.13% before hyperparameter optimization, and 92.39% after hyperparameter optimization. These results show that machine learning models are capable of forecasting the risk of heart disease. These models may assist healthcare professionals in identifying individuals at risk of heart disease, enabling preventative measures to be taken. It is essential to note that this study focuses solely on classification models and may not represent the entire population. Further research is required to determine the predictability of heart disease in diverse populations.

Keywords- Heart Disease; Hyperparameter Optimization; Ensemble Learning; Machine Learning.

I. INTRODUCTION

The major cause of death worldwide in recent years has been cardiovascular. According to WHO [1] every year, 17.9 million individuals die from cardiovascular disease. Heart disease is a common condition that affects people in their middle or old years and frequently results in fatal complications, as a result, heart disease is responsible for one-third of all fatalities globally [2]. When modern technology and medical professionals are unavailable, the diagnosis and treatment of heart disease are very challenging, even though that an accurate diagnosis and appropriate treatment can save the lives of many individuals [3]. To accurately predict heart disease, machine learning is required.

Diagnosis of heart disease is difficult due to various contributing risk factors [4]. Therefore the main goal of this paper is to make the best possible heart disease predictions using machine learning that will use hyperparameter optimization and ensemble voting to try to improve the prediction results.

In this paper [5], conducted similar studies using several classification algorithms and using several ensemble algorithms which include boosting, bagging, stacking, and majority vote. After conducting research, the results of the best models that use the ensemble algorithm with an accuracy rate of 86.32%.

Gupta and Seth [6], conducted similar research but there is a slight difference where they use the feature selection algorithm and also added the K-Nearest Neighbor classification algorithm. After training and testing the model they made, an accuracy of 98.38% was obtained using the majority vote algorithm.

In this paper [7], conducted similar research but this time using a tuning hyperparameter. After training and testing are obtained by the best model accuracy that has been tuned, namely random forest with a gain accuracy of 80.95%.

In this paper [8], conducted similar research, but conducted the research in more detail such as conducting feature selection, irregular oversampling, synthetic minority oversampling, and adaptive synthetic sampling approach. After training and testing, it was found that the best result from the stacking classifier method with an accuracy of 99.00%.

Previous study has shown that employing a ML algorithm to forecast the risk of heart disease yields good results. This paper was made to forecast heart disease, but with a different dataset to broaden the sample population. The dataset utilized

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7208
***Article Received**: 26 April 2023 **Revised:** 14 June 2023 **Accepted:** 30 June 2023*

in this research paper exhibits a higher level of complexity compared to the UCI dataset utilized by previous researchers. This assertion can be substantiated by examining the presence of duplicate data within each dataset and the overall sample size contained within them.

II. METHODOLOGY

A. Dataset

The dataset used for this paper is the dataset heart failure prediction taken from Kaggle [9]. The Proportion for the training set and testing set is 80 : 20. This dataset has 11 features and 1 output with a total of 918 patients. Table I describe its attributes and description.

TABLE I. TABLE TYPE STYLES

Attribute	Description
Age	The patient's age
Sex	The patient's sex
Chest pain type	TA stands for Typical Angina, ATA stands for Atypical Angina, NAP stands for Non-Anginal Pain, ASY stands for Asymptomatic
Resting BP	Resting Blood Pressure (mm HG)
Cholesterol	Serum cholesterol(mm / dl)
Fasting BS	1 stands for a patient who has FBS > 120 mg / dl, 0 stands for otherwise
Resting ECG	Resting Electrocardiogram (Normal stands for normal, ST stands for ST-T abnormality, LVH stands for probable of confirmed left ventricular hypertrophy according to Estes's criteria)
Max HR	Max HR stands for maximum heart rate that achieved between 60 – 202
Exercise Angina	Y stands for yes, N stands for no
Oldpeak	Oldpeak stands for numerical value according to depression
ST Slope	The peak exercise's slope ST section. Up means upsloping, Flat means flat, and Down means downsloping.
Heart Disease	1 stands for Heart Disease, 0 stands for Normal

B. Data Preprocessing

Based on the dataset for this paper, there are some variables that have categorical datatype. Therefore, encoding is needed to solve this problem. Categorical variables can be translated into numerical values and easily fitted to a machine learning model using the encoding technique.

Label encoder is a feature provided by sklearn, to make it easier for us to transform text into numeric form so that it can be processed by machine learning algorithms [10]. After using label encoder, values will range from 0 to n-1 in each variable, where n stands for total distinct values in each variable. Fig. 1 shows visualization of label encoder.

Figure 1. Label Encoder visualization.

C. Experimental Process Design

The major contribution of this paper is to predict cardiac disease using machine learning classification algorithms. Several ML algorithms were trained in this paper. To get better performance, hyperparameter optimization will be used namely gridsearchCV with 5 fold cross validation for the validator. The performance of each model, is evaluated using the heart disease Kaggle dataset using several performance metrics such as accuracy, recall, precision, F1-score. The XGBoost and RF with gridsearchCV HPO are the best models with gain accuracy of 92.39% for those 2 models. Fig. 2 shows this paper's workflow.

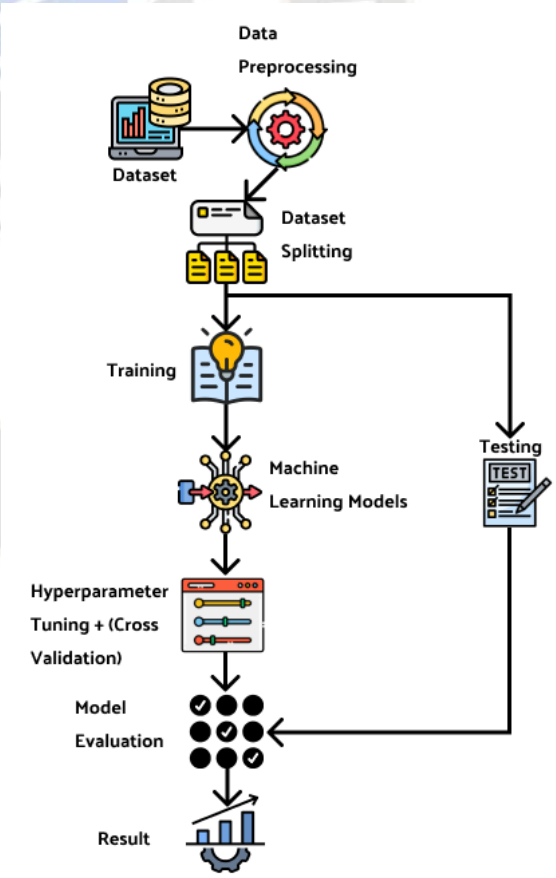

Figure 2. Heart disease prediction workflow.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7208
***Article Received**: 26 April 2023 **Revised:** 14 June 2023 **Accepted:** 30 June 2023*

D. Machine Learning Algorithm

1) Decision Tree (DT)

One of the oldest and most extensively used ML techniques is the decision tree (DT). A decision tree is a decision logic design that evaluates and compares then classification results of data items using a tree structure. A decision tree has multiple tiers of nodes. The root or parent node is the highest level, while the others are known as child nodes [11].

2) Random Forest (RF)

The RF algorithm is a ML technique that may address many problems. Because the forecast is based on the mixture of all decision trees, random forest can be defined as an ensemble decision tree. Random forest is useful for large datasets or high-dimensionality data since it can generate accurate predictions in those cases [12].

3) Logistic Regression (LR)

Logistic Regression, as it can be seen from the name, is a machine learning model that is often used for regression tasks. However, this model can also be used for binary classification tasks, where there are only two classes for the result. The way this model works is that it calculates the probability of a data being one of the two classes available by utilizing logistic/sigmoid function. The function will optimize any values found in-between the two classes available as the classification labels, which would result in the model being able to predict the probability of the target data being one of the two classes [13].

4) Gradient Boosting (GB)

Gradient boosting is widely recognized as one of the most effective supervised ML algorithm due to its impressive performance in solving complex classification and regression problems. This versatile algorithm works by iteratively training a collection of weak predictive model, likely decision tree to create a single strong predictor [14].

5) Extreme Gradient Boosting (XGB)

XGB is one of implementation of the ensemble learning. XGB use a set of weak algorithm that have been combined to improve the accuracy. XGB is a gradient boosting and decision tree enhancement that can be used for many problems. XGB trains a tree by adding a tree and separating the features in each iteration [15].

Since XGB is a decision tree based technique, sub sample and maximum depth are used to avoid overfitting and enhance model performance [16]. Learning rate control how much weight is given to trees and is employed to slow the network's rate of acclimatization to training set. The regularization concept in XGB objective function helps with prediction function selection and model complexity control. Equation (1) shows objective function of XGB.

$$Obj= \sum_{i=1}^{n} L(\hat{y}_i, y_i)+ \sum_{i=1}^{k} R(f_i) \quad (1)$$

When $\hat{y}_i$ stands for predicted label and y_i stands for actual label. L stands for the loss function, which assesses how well the model performs on train data. R(f) stands for training tree's function complexity. Function of tree f(x) must first be definedin order to determine the complexity.

$$f(x)= w_{q(x)},\ w \in R^T,\ q: R^M \rightarrow \{1,2\ldots, T\} \quad (2)$$

Here, number of leaves is denoted by T, and w represents leaf score vector, q is a function instances mapping to related leaf. Equation (3) shows Model complexity's formula.

$$R(f)=its+\alpha(\|w\|)+\frac{1}{2}\lambda(\|w\|^2) \quad (3)$$

Here, γ stands for value of every leaf , T stands for total amount of tree leaves, λ and γ are the hyperparameters and constant coefficients. $\|w\|^2$ stands for L2 leaf weight norm governed by λ, and $\|w\|$ stands for L1 leaf weight norm governed by α.

XGBoost initiates by generating an initial prediction to establish the residual value. The residual value is calculated by subtracting the actual data from the projected data. Subsequently, boost constructs a weak learner, which takes the form of a decision tree, using the previously obtained residual value, [16]. By utilizing the residuals to guide the decision tree's construction, boost can effectively analyze the patterns contained within these residuals, thereby improving the overall accuracy of the model. This iterative process continues until the residuals converge. In the end, the predictions from the decision tree are combined to yield the final prediction produced by boost. Fig. 3 shows the XGB algorithm architecture.

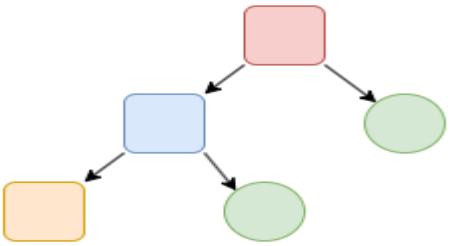

Figure 3. The Extreme Gradient Boosting algorithm architecture.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7208
***Article Received**: 26 April 2023 **Revised:** 14 June 2023 **Accepted:** 30 June 2023*

E. Performance Metrics

After training and testing the model will display the output of the results of testing the model that has been done. The performance metrics used in this paper are accuracy, precision, recall, F1-score. Mathematical representation of those metrics are [17].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{F1-Score} = \frac{2TP}{2TP+FN+FP} \quad (7)$$

III. RESULT AND DISCUSSION

All the calculations that have been done in this paper, were done on a laptop with the following detail : AMD Ryzen 7 5800H with Radeon Graphics, 3.8GHz and 8 GB RAM, using Google Colab with the python programming language. Table II shows comparation between common machine learning models before and after HPO. Table III shows comparation between ensemble machine learning models before and after HPO. Table IV shows parameters before and after hyperparameter optimization for each model.

TABLE II. COMPARATION BETWEEN COMMON MACHINE LEARNING MODELS AND AFTER HYPERPARAMETER OPTIMIZATION.

Method	Metrics	Default	Tuning
Decision Tree	Accuracy (%)	85.87	86.96
	Precision (%)	87.76	89.80
	Sensitivity (%)	86.00	86.27
	F1 - Score (%)	86.87	88.00
Logistic Regression	Accuracy (%)	85.87	86.41
	Precision (%)	86.73	86.73
	Sensitivity (%)	86.73	87.63
	F1 – Score (%)	86.73	87.18

According to the results in table II, the hyperparameter optimization method in the decision tree and logistic regression models has been shown to boost the accuracy of each model. Although not statistically significant, this approach performs well.

TABLE III. COMPARATION BETWEEN ENSEMBLE MACHINE LEARNING MODELS AND AFTER HYPERPARAMETER OPTIMIZATION.

Method	Metrics	Default	Tuning
Random Forest	Accuracy (%)	88.04	92.39
	Precision (%)	93.88	94.90
	Sensitivity (%)	85.19	91.18
	F1 - Score (%)	89.32	93.00
Gradient Boosting	Accuracy (%)	89.67	91.30
	Precision (%)	88.78	91.84
	Sensitivity (%)	91.58	91.84
	F1 - Score (%)	90.16	91.84
Extreme Gradient Boosting	Accuracy (%)	89.13	92.39
	Precision (%)	88.78	92.86
	Sensitivity (%)	90.62	92.86
	F1 - Score (%)	89.69	92.86
Ensemble Voting	Accuracy (%)	89.13	89.67
	Precision (%)	88.78	89.80
	Sensitivity (%)	90.62	90.72
	F1 - Score (%)	89.69	90.26

According to the results in table III, the hyperparameter optimization method in ensemble machine learning models has been shown to boost the accuracy of each model. This method performs better on these four models than on the preceding two. Because these four models are ensemble models built from some simple models, they are more powerful than the preceding two models.

For ensemble voting, using several machine learning models that have also been used in this paper, such as DT, RF, LR, GB, XGBoost, either using default parameters or using optimized parameters and also with hard voting type.

TABLE IV. PARAMETERS BEFORE AND AFTER HYPERPARAMETER OPTIMIZATION FOR EACH MODEL

Method	Parameter	Default	Tuning
Decision Tree	Criterion	Gini	Gini
	Max depth	None	5
	Min samples split	2	6
	Min samples leaf	1	1
	Max features	None	Log2
Logistic Regression	Penalty	L2	L2
	C	1.0	0.03359 8182862 83781
	Solver	Lbfgs	Newton-cg
Random Forest	Criterion	Gini	Entropy
	Max depth	None	None
	Min samples split	2	10
	N estimators	10	100
Gradient Boosting	Learning rate	0.1	0.1
	Max depth	3	3
	Max features	Auto	Sqrt
	Min samples leaf	1	4
	Min samples split	2	5
	N estimators	100	100
Extreme Gradient Boosting	Learning rate	0.3	0.01
	Max depth	6	3
	N estimators	100	1000

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7208
***Article Received**: 26 April 2023 **Revised:** 14 June 2023 **Accepted:** 30 June 2023*

Table IV shows the default parameters as well as the parameters after optimization for each model. Even though it only optimizes a few parameters, the parameters optimized in this work are crucial parameters for each model. Table III shows that the best models after hyperparameter optimization are random forest and extreme gradient boosting with gain accuracy of 92.39% respectively.

Fig. 3 and 4 shows ROC Curve and AUC score for each model before hyperparameter optimization and after hyperparameter optimization respectively. ROC curve is a graph to seeks the connection between TPR and FPR. Otherwise, AUC calculates the area beneath the full ROC curve in two dimension.

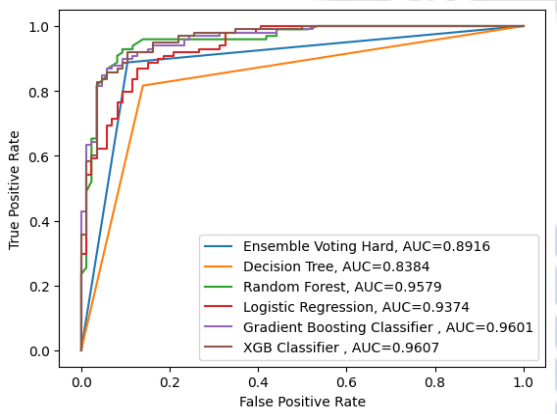

Figure 4. ROC AUC score before hyperparameter optimization.

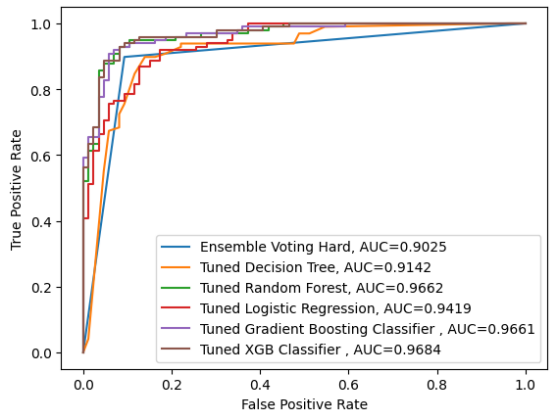

Figure 5. ROC AUC score after hyperparameter optimization.

Based on Fig. 4, XGB and random forest get an AUC score of 0.9684 and 0.9662 respectively, indicating the models perform well. Fig. 5 shows confusion metrics for XGB after hyperparameter optimization. Fig. 6 shows confusion metrics for random forest classifier after hyperparameter optimization.

Figure 5 depicts True Negative number 79, False Positive number 7, False Negative number 7, and True Positive number 91. On the other side, figure 6 depicts True Negative number 77, False Positive number 9, False Negative number 5, and True Positive number 93. Where there are zeros and ones on the X and Y axes, zero denotes a class with a healthy heart and one, a class with heart failure.

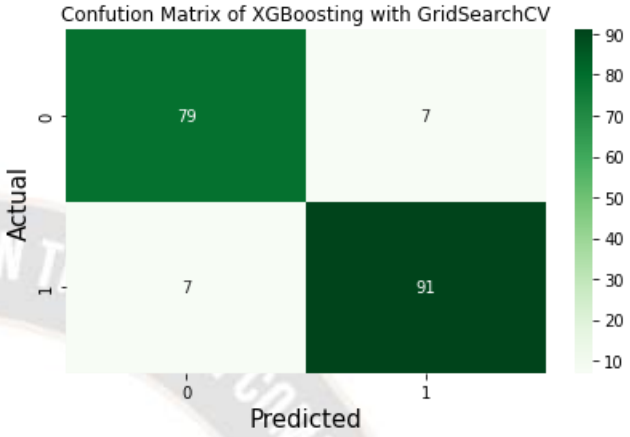

Figure 6. Confusion metrics of tuned Extreme Gradient Boost.

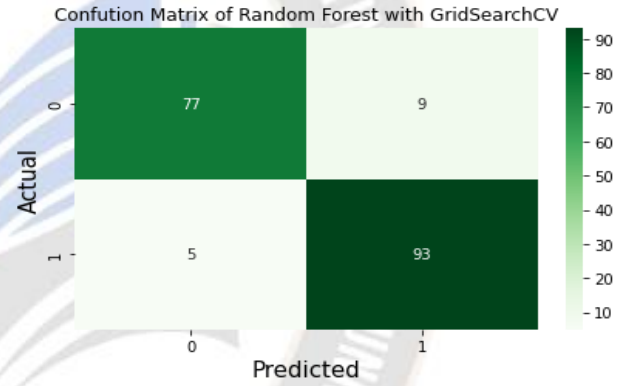

Figure 7. Confusion metrics of tuned Random Forest.

Additional corroborating evidence suggesting the favorable progress of this research is the comparation with other scholar work that used same dataset. Table V shows comparation with other scholar work.

TABLE V. COMPARATION WITH OTHER SCHOLAR WORK.

Paper	Model	Metrics	Score
[18]	Artificial Neural Network (MLP Classifier)	Accuracy (%)	92.03
		Precision (%)	91.70
		Sensitivity (%)	91.91
		F1 - Score (%)	91.80
Proposed Method	Tuned Extreme Gradient Boosting	Accuracy (%)	92.39
		Precision (%)	92.86
		Sensitivity (%)	92.86
		F1 - Score (%)	92.86
	Tuned Random Forest	Accuracy (%)	92.39
		Precision (%)	94.90
		Sensitivity (%)	91.18
		F1 - Score (%)	93.00

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7208
***Article Received**: 26 April 2023 **Revised:** 14 June 2023 **Accepted:** 30 June 2023*

Based on table V which contains a comparison of the results with previous studies, it can be proven that this experiment went well. Proposed method get better result, whereas in the previous experiment, the method used was Neural Network and in this experiment the method used was basic machine learning classification

IV. CONCLUSION

In this paper, several classification machine learning algorithm have been used to predict heart disease. There are many ways that can be used to improve the accuracy of each model. Hyperparameter tuning using gridsearchCV is a method that used in this paper. Based on the experiment that has been done, tuned XGB and tuned random forest are the best models with gain accuracy of 92.39% respectively on the heart failure prediction dataset.

Future study should consider more thorough hyperparameter optimization in addition to using more powerful machine learning models like neural network and others to gain better performance.

.

REFERENCES

[1] WHO, "Cardiovascular diseases." Accessed: Apr. 12, 2023. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Inform Med Unlocked, vol. 16, Jan. 2019, doi: 10.1016/j.imu.2019.100203.

[3] 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.

[4] 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.

[5] P. Puvar, N. Patel, A. Shah, R. Solanki, and D. Rana, "Heart Disease Detection using Ensemble Learning Approach," International Research Journal of Engineering and Technology, 2021, [Online]. Available: www.irjet.net

[6] P. Gupta and D. D. Seth, "Improving the Prediction of Heart Disease Using Ensemble Learning and Feature Selection," International Journal of Advances in Soft Computing and its Applications, vol. 14, no. 2, pp. 36–48, 2022, doi: 10.15849/IJASCA.220720.03.

[7] P. Premananthan, S. Prasanth, and K. Mauran, "HYPER PARAMETER TUNED ENSEMBLE APPROACH FOR HEART DISEASE PREDICTION."

[8] K. Rohit Chowdary, P. Bhargav, N. Nikhil, K. Varun, and D. Jayanthi, "Early heart disease prediction using ensemble learning techniques," in Journal of Physics: Conference Series, Institute of Physics, 2022. doi: 10.1088/1742-6596/2325/1/012051.

[9] David W. Aha, "Heart Failure Prediction Dataset," Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction (accessed Apr. 12, 2023).

[10] N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 229–241, Jun. 2021, doi: 10.1016/j.ijcce.2021.12.001.

[11] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Comput Biol Med, vol. 136, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

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

[13] P. Anbuselvan, "Heart Disease Prediction using Machine Learning Techniques." [Online]. Available: www.ijert.org

[14] R. Kannan and V. Vasanthi, "Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease," in SpringerBriefs in Applied Sciences and Technology, Springer Verlag, 2019, pp. 63–72. doi: 10.1007/978-981-13-0059-2_8.

[15] R. Valarmathi and T. Sheela, "Heart disease prediction using hyper parameter optimization (HPO) tuning," Biomed Signal Process Control, vol. 70, Sep. 2021, doi: 10.1016/j.bspc.2021.103033.

[16] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 7, pp. 4514–4523, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.013.

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

[18] S. Y. Prasetyo, "Prediksi Gagal Jantung Menggunakan Artificial Neural Network," Jurnal SAINTEKOM, vol. 13, no. 1, pp. 79–88, Mar. 2023, doi: 10.33020/saintekom.v13i1.379.