An Efficient Heart Disease Prediction System based on Supervised Machine Learning Methods

M.Kalaivani[1],Dr.S.Anitha[2]
[12]Assistant professor, Post Graduate Department of Computer Science,
Dwaraka Doss Goverdhan Doss Vaishnav College,
Arumbakkam, [1]email - kalaivani@dgvaishnavcollege.edu.in
[2]email anitasenthil –@gmail.com

Abstract: Heart illness is the most frequent disease that can lead to grave situations in a human life. Every year more number of people gets affected by heart disease around the world. Prediction and finding of heart disease has always been a critical and difficult task for the medical field. To overcome the issues, Machine Learning (ML) methods played a vital role to predict and detect heart disease very accurately and quickly. In this paper, an innovative approach is proposed to predict heart disease using different ML methods. The proposed method consists of three stages. First to preprocess the heart disease Dataset, which consists of checking the missing values, data scaling and calculate the correlation between the predictors. In the second stage apply the supervised classification methods such as Logistic Regression (LR), KNN, Random Forest (RF) and AdaBoost to train the model. Classification metrics are used to evaluate the model and choose the best classifier for heart disease prediction. Also comparison can be done using the metrics of each classifier. As a result of the experiment, it is concluded that the RF classifier has attained more accuracy of 95.08% , ROC-AUC score of 0.95.

Keywords: Heart disease, preprocess, supervised classification algorithms, Classification metrics.

I. INTRODUCTION

The primary role of the heart is to circulate blood to all parts of the body. In the event that the heart doesn't work properly, it will affect different organs of the human body. The main reason is due to blockage or tightening down of the coronary arteries. Heart disease symptoms include dizziness, shortness of breath, anxiety, physical body weakness, vomiting, indigestion etc. There are various types of heart disease, each of which affects the heart in a unique way. Due to a scarcity of testing tools and a physician shortage, proper diagnosis and treatment of heart patients is delayed[1].Hence it has become important for diagnosis and prediction of heart disease quickly and efficiently. The efficient and correct diagnosis of heart disease plays an important role in taking measures to prevent death.

In this paper, the research work carried out the performance of various algorithms such as LR, KNN, RF, and Adaboost classifiers by combining all the significant features of the dataset to predict the disease in the early stage.

The research paper is framed as follows. In Section 2, the reviews of literature are narrated. In section 3, Materials and methods are discussed with the various techniques for analyzing heart diseases. In Section 4, Empirical results are described. Section 5 concludes the research and exhibits the further enhancement of the work.

II. LITERATURE REVIEW

F.Z.Abdeljaouad et al,[2] have proposed a new hybrid methodology that combines Principal Component Analysis(PCA) as feature extraction with ML techniques and performance evaluations are calculated. Ch.Satyanarayana et al, [3] have proposed the OMLR algorithm, which is used for detecting the severity of heart disease. The proposed method

International Journal of Computing Algorithm
Volume: 10 Issue: 01 June 2021
ISSN: 2278-2397

produced 92% accuracy for detecting heart disease. A NFG approach has been proposed to analyze diseases using GA to optimize the result that will increase the classification accuracy [4].

Cameron R.Olsen et al,[5] have presented a method using both Supervised Learning such as Regression, Classification and Unsupervised Learning such as clustering methods to find the diseases. Syed Nawaz Pasha et al,[6] have presented a method to predict cardiovascular disease using Multilayer perceptron and it is evaluated using SVM, Decision tree KNN. R.Indrakumari et al. [7] have discussed Exploratory Data Analysis(EDA) for predicting heart disease by k-means clustering method. In this research the two diagnostic methods are classified as Invasive diagnostic (ID) and Non-invasive diagnostic (NID). Firstly, ID helps to cut the skin, connective tissues and mucus membrane. Similarly, NID are used to diagnose diseases without opening the skin as second. Hlaudi Daniel Masethe et al,[8] have carried out research to predict heart disease by J48, Naïve Bayes, CART, Bayes Net and REPTREE with accuracy of 99%. Senthil Kumar et al,[9] have developed an efficient method NMF-HC for non-negative matrix factorization with hierarchical clustering algorithms.

Pronob Ghosh et al [10] have proposed Relief univariate filter method and LASSO techniques to retrieve significant features and build a model using Decision Tree Bagging Method to predict cardiovascular disease. AditiGavhane et al[11] developed an application using Multilayer perceptron to predict heart disease and produced more accurate results.

III. MATERIALS AND METHODS

This section elaborates the various Methods used in this research work. The details of the dataset and an architectural view of the research method have been demonstrated.

3.1 DATASET

Heart disease data set has been taken for this analysis. It has 303 samples with 14 attributes. Table - 1 shows the details of the dataset. Sex and Age are significant features because Male has higher risk factors for coronary heart disease than females. The remaining features indicates the level of disease.

Table -1: Details of Dataset

Sl.No	Feature Code	Description	Domain of Value
1	Age	Age of the person in years	29-77
2	Sex	Sex of the Patient	1-Male, 0-Female
3	CP	1. typical angina, 2. atypical angina, 3. non-anginal pain, 4. asymptomatic	1,2,3,4
4	Trestbps	mm HG	94-200
5	Chol	Mg/dl	126-564
6	Fbs	>120Mg/dl	1-true 0-false
7	restecg	1.Normal, 2.Having ST-T wave abnormality 3.Left ventricular hyperthrophy	0,1,2
8	thalach		71-202
9	exang		1-yes, 0-no
10	oldpeak		0-6.2
11	Slope	1.Upsloping, 2.Flat, 3.Downsloping	0,1,2
12	Ca		0-4
13	Thal	Normal ,Fixed defect ,Reversible defect	3,6,7
14	Class		0-no heart disease,1-heart disease

3.2. PROPOSED METHOD

The proposed method is structured into three steps such as follows.

1. Preprocessing the Model
2. Classification Model Construction
3. Model Evaluation.

The General architecture of the proposed system is rendered in Figure-1

International Journal of Computing Algorithm
Volume: 10 Issue: 01 June 2021
ISSN: 2278-2397

3.2.1 DATA PREPROCESSING

Data preprocessing steps helps to increase the purity of the data. As the first phase of the preprocessing work replaces the missing values by the mean value. In the second phase, Z-Score normalization is used to normalize the data. This technique is used to perform normalization of the feature. For Attribute subset selection, Pearson correlation coefficient method is utilized for removing irrelevant and redundant features in the heart disease dataset[12]. The linear relationship between two features, which can vary between 1 and -1are found.

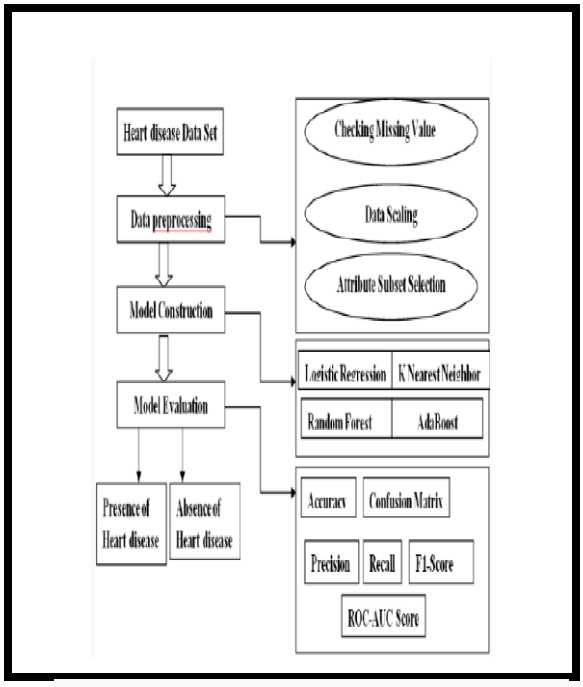


Figure-1: Framework of the proposed system

3.2.2 MODEL CONSTRUCTION

Machine learning can be an effective tool to apply in the medical field to predict and detect the most important clinical features [13]. ML algorithms are effectively used to deal with large number of variables for medical diagnosis and predicting diseases [14]. It uses various algorithms to optimize the operations by analyzing input data for producing the results within a limited range [15]. In this proposed method, to perform model construction using ML algorithms such as LR, KNN, RF, and AdaBoost. Dataset has been divided into 80% and 20% with respect to train and test data to validate the ability of the research method.

LOGISTIC REGRESSION(LR)

A supervised machine learning model called logistic regression is used to predict the likelihood of a target variable. The goal of the LR algorithm is to obtain the best fit for representing the relation between the response variable and predictors that are diagnostically acceptable. The target variable is binary, which means it only contains data defined as 1 or 0, indicating whether a patient has heart disease or not.

K NEAREST NEIGHBORS(KNN)

It is a non-parametric, supervised learning algorithm that uses an instance-based method[16]. It categories data based on their proximity to their nearest neighbors. It predicts the target class based on how similar that particular data is to the model's training data. Initialize and define the K value for each data point. Using the Euclidean distance, calculate the distance between test and training data. Based on the K^{th} minimal distance, sort the distance and identify the closest neighbors. The data is classified based on the prediction value of the nearest neighbors.

RANDOM FOREST(RF)

International Journal of Computing Algorithm
Volume: 10 Issue: 01 June 2021
ISSN: 2278-2397

RF is a supervised classification that builds decision trees on data points, extracts predictions from each one, and then votes on the best option. The goal of voting to merge the decision trees is to select the highest projected tree, which can improve the model's accuracy [17]. It essentially employs a technique known as bagging, which generates a large number of decision trees and then combines them to get a finer output[18].

ADABOOST

AdaBoost is an ensemble boosting classifier that combines many classifiers to improve the model's accuracy. Initial task is to set the weights of each instance and train the data sample such that it can accurately predict uncommon observations. The weighted average of the weak classifiers is calculated and the total of the weighted predictions is used to make predictions. The resultant dataset from training phase is fed into the classifier for transforming weak learners into strong learners.[19]

3.2.3 MODEL EVALUATION

To evaluate the model using classifier metrics such as Confusion matrix, Accuracy, Sensitivity, Specificity and ROC-AUC Curve are used. These evaluation metrics are used for checking the robustness and efficiency of the algorithm. Finally to validate the accuracy and select the best model to detect the disease effectively.

IV. EMPIRICAL RESULTS

This section contains the Empirical analysis and findings of predicting heart disease. A core i3 processor with 4 GB capacity of RAM and python libraries such as pandas, Scipy and Matplotlib were used in the Jupyter Notebook Web application environment. In the first phase, preprocessing the heart disease Dataset then it is fed into the classification methods such as LR, KNN, RF and AdaBoost. In this dataset the target variable 0 is considered as the absence of heart disease (+ve) whereas 1 is considered as the presence of heart disease(-ve). Confusion Matrix is an important measure to examine the effectiveness of a classification model with four measures such as T_Pos,T_Neg,F_Pos and F_Neg. The accuracy of a classifier is a vital parameter which gives the percentage of the overall correct predictions divided by the total number of sample values. Accuracy can be calculated by using the formula

$$\text{Accuracy} = (T\_Pos + T\_Neg)/(T\_Pos + T\_Neg + F\_Pos + F\_Neg) \quad (1)$$

True Positive Rate(TPR-Sensitivity) indicates the proportion of (+ve) sample values that are correctly classified as (+ve). Sensitivity can be calculated using the formula

$$\textbf{Sensitivity(Recall)} = T\_Pos/(T\_Pos + F\_Neg)$$

True Negative Rate or Specificity shows the proportion of (-ve) samples that are correctly classified as (-ve). TNR can be calculated as

$$\textbf{Specificity} = T\_Neg/(T\_Neg + F\_Pos)$$

Precision indicates the proposition of the correctly predicted values actually turned out to be a (+ve) value of the model. Precision can be calculated as

$$\textbf{Precision} = T\_Pos/(T\_Pos + F\_Pos)$$

F1 measure is the combination of two metrics such as Precision and Recall. It can be defined as

$$\textbf{F1-Score} = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall})$$

International Journal of Computing Algorithm
Volume: 10 Issue: 01 June 2021
ISSN: 2278-2397

The ROC and AUC graphs represent the performance of a classifier. The x-axis depicts the FPR, while the y-axis depicts the TPR.

4.1 EVALUATION OF RESULTS

The performance of four classifiers namely LR, KNN, RF, and AdaBoost was analyzed and compared the results. Random Forest classifier had the highest classification accuracy of 95.08 percent, while the AdaBoost model had the second highest classification accuracy of 90.16 percent. The accuracy score of four classifiers is shown in Figure 2.Confusion matrix prediction results displayed in Figures-3,4,5 and 6 for LR, KNN, RF and Adaboost respectively. All the four models' True Negative value is high which indicates that the presence of heart disease is correctly predicted. Figure-7 shows the ROC-AUC score of four models. Random Forest models outperformed high AUC value of 0.950 which is closer to 1. Figure-8 specifies the classification metrics of four classifiers such as Precision, Recall, F1-Score.Random Forest classifier and Adaboost classifier methods have shown maximum precision of 99% and 94% respectively. Random Forest has the highest F1-score of 95% compared with all other techniques.

Figure -2 Accuracy Score of Four Classifiers

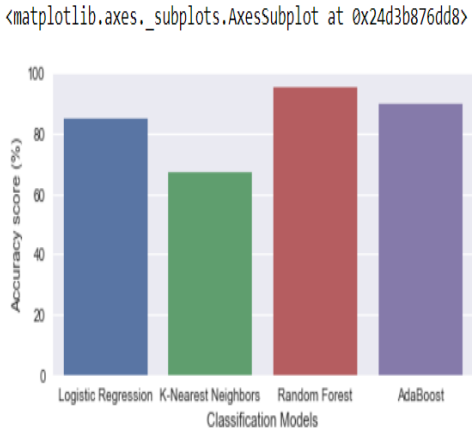
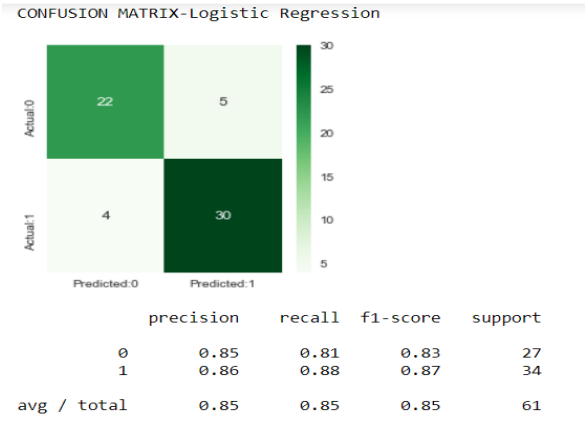


Figure-3 Confusion Matrix of Logistic Regression

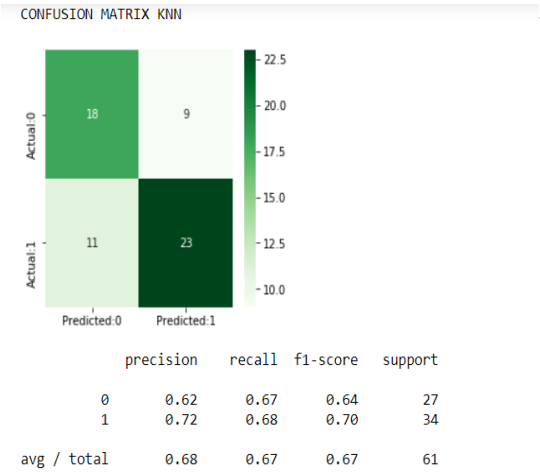


	precision	recall	f1-score	support
0	0.85	0.81	0.83	27
1	0.86	0.88	0.87	34
avg / total	0.85	0.85	0.85	61

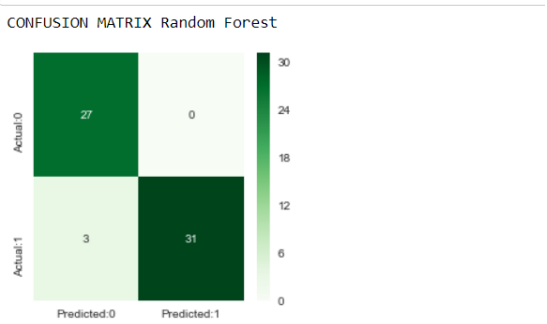
Figure-4 Confusion Matrix of KNN

Figure-5 Confusion Matrix of Random Forest

International Journal of Computing Algorithm
Volume: 10 Issue: 01 June 2021
ISSN: 2278-2397



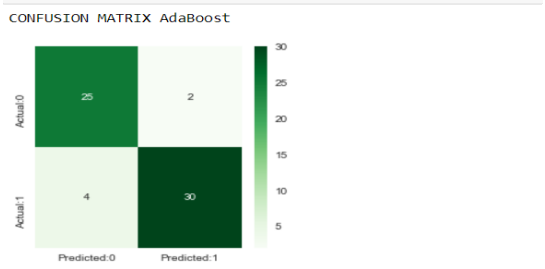
	precision	recall	f1-score	support
0	0.62	0.67	0.64	27
1	0.72	0.68	0.70	34
avg / total	0.68	0.67	0.67	61



	precision	recall	f1-score	support
0	0.90	1.00	0.95	27
1	1.00	0.91	0.95	34
avg / total	0.96	0.95	0.95	61

Figure-6 Confusion Matrix of Adaboost Score of four Classifier model

Figure-7 Comparison of ROC-AUC



	precision	recall	f1-score	support
0	0.86	0.93	0.89	27
1	0.94	0.88	0.91	34
avg / total	0.90	0.90	0.90	61

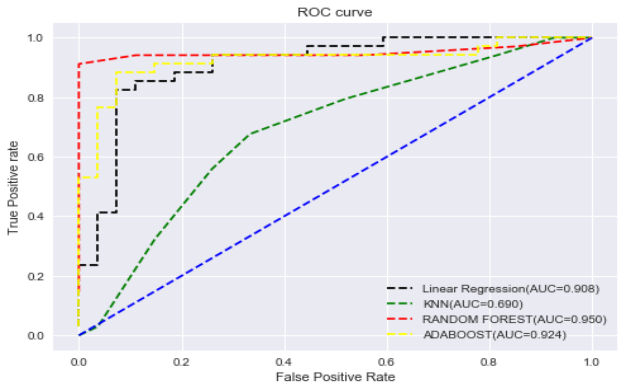
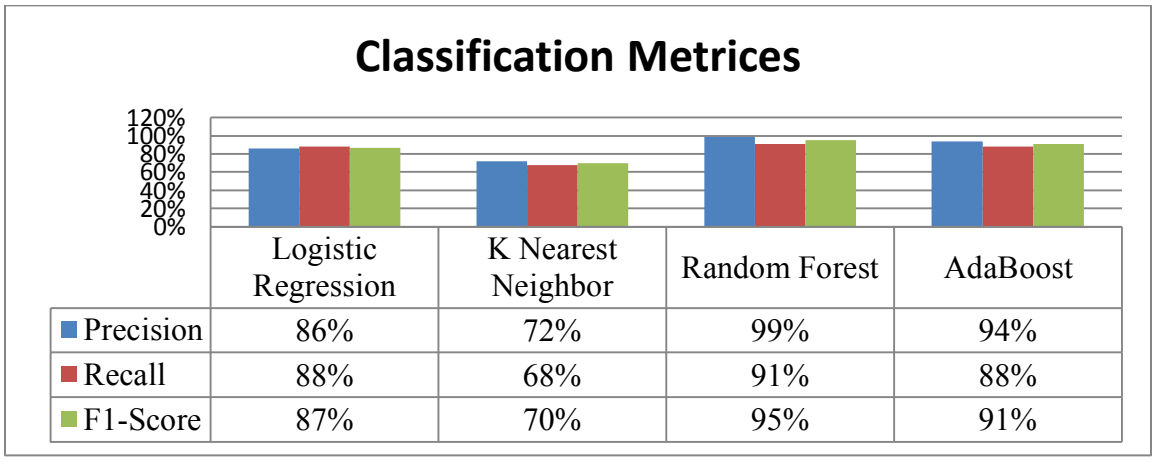


Figure -8: Classification metrics of Four Classifiers



	Logistic Regression	K Nearest Neighbor	Random Forest	AdaBoost
Precision	86%	72%	99%	94%
Recall	88%	68%	91%	88%
F1-Score	87%	70%	95%	91%

International Journal of Computing Algorithm
Volume: 10 Issue: 01 June 2021
ISSN: 2278-2397

V. CONCLUSION

The major objective of the study is to find the best reliable method of predicting heart disease. In this research paper, supervised machine learning methods were used in the prediction of heart disease. The presented method uses preprocessing in the first step and model construction in the second step using four classifiers namely LR,KNN,RF and AdaBoost. To evaluate the model using classification metrics. Random Forest supervised learning classifier has achieved a greater accuracy of 95.08% respectively. The proposed method is extremely valuable in assisting doctors for predicting heart disease successfully. Different deep learning approaches are used in the future research to improve prediction.

VI. REFERENCES

[1] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative Artificial Neural Networks-Based Decision Support System for Heart Diseases Diagnosis," *J. Intell. Learn. Syst. Appl.*, vol. 05, no. 03, pp. 176–183, 2013.

[2] F. Z. Abdeldjouad, M. Brahami, and N. Matta, *A* Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques, vol. 12157 *LNCS. Springer International Publishing*, 2020.

[3] N. Satyanandam and C. Satyanarayana, "Heart Disease Detection Using Predictive Optimization Techniques," *Int. J. Image, Graph. Signal Process.*, vol. 11, no. 9, pp. 18–24, 2019.

[4] S. Nikam, P. Shukla, and M. Shah, "Cardiovascular Disease Prediction Using Genetic Algorithm and Neuro-Fuzzy System," *Int. J. Latest Trends Eng. Technol.*, vol. 8, no. 2, pp. 104–110, 2017.

[5] C. R. Olsen, R. J. Mentz, K. J. Anstrom, D. Page, and P. A. Patel, "Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure: Machine learning in heart failure," *Am. Heart J.*, vol. 229, pp. 1–17, 2020.

[6] S. N. Pasha, D. Ramesh, S. Mohmmad, A. Harshavardhan, and Shabana, "Cardiovascular disease prediction using deep learning techniques," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 981, no. 2, 2020.

[7] R. Indrakumari, T. Poongodi, and S. R. Jena, "Heart Disease Prediction using Exploratory Data Analysis," *Procedia Comput. Sci.*, vol. 173, no. 2019, pp. 130–139, 2020.

[8] H. D. Masethe and M. A. Masethe, "Prediction of heart disease using classification algorithms," *Lect. Notes Eng. Comput. Sci.*, vol. 2, pp. 809–812, 2014.

[9] A. V. S. Kumar, "Heart Disease Prediction Using Data Mining preprocessing and Hierarchical Clustering," *Int. J. Adv. Trends Comput. Scienec Eng.*, vol. 4, no. 6, pp. 7–18, 2015.

[10] 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.

[11] M. Balakrishnan, A. B. Arockia Christopher, P. Ramprakash, and A. Logeswari, "Prediction of Cardiovascular Disease using Machine Learning," *J. Phys. Conf. Ser.*, vol. 1767, no. 1, pp. 1–7, 2021.

[12] A. Wosiak and D. Zakrzewska, "Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis," *Complexity*, vol. 2018, 2018.

[13] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–16, 2020.

[14] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cogn. Comput. Eng.*, vol. 2, no. November 2020, pp. 40–46, 2021.

[15] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease

prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–16, 2019.

[16] V. Sharma, S. Yadav, and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," *Proc. - IEEE 2020 2nd Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2020*, vol. 1, no. 6, pp. 177–181, 2020.

[17] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 261–268, 2019.

[18] I. Javid, A. K. Z. Alsaedi, and R. Ghazali, "Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 540–551, 2020.

[19] I. Yekkala, S. Dixit, and M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization," *Proc. 2017 Int. Conf. Smart Technol. Smart Nation, SmartTechCon 2017*, pp. 691–698, 2018.