Heart disease prediction model with k-nearest neighbor algorithm

Tsehay Admassu Assegie
Department of Computer Science, College of Natural and Computational Science, Injibara University, Injibara, Ethiopia

Article Info

Article history:

Received Jan 12, 2021
RevisedAug 20, 2021
Accepted Sep 2, 2021

Keywords:

Automated diagnosis
Disease prediction
Heart disease prediction
KNN classifier
KNN model

ABSTRACT

In this study, the author proposed k-nearest neighbor (KNN) based heart disease prediction model. The author conducted an experiment to evaluate the performance of the proposed model. Moreover, the result of the experimental evaluation of the predictive performance of the proposed model is analyzed. To conduct the study, the author obtained heart disease data from Kaggle machine learning data repository. The dataset consists of 1025 observations of which 499 or 48.68% is heart disease negative and 526 or 51.32% is heart disease positive. Finally, the performance of KNN algorithm is analyzed on the test set. The result of performance analysis on the experimental results on the Kaggle heart disease data repository shows that the accuracy of the KNN is 91.99%.

Corresponding Author:

Tsehay Admassu Assegie
Department of Computer Science, College of Natural and Computational Science, Injibara University
Injibara, P.O.B: 40, Ethiopia
Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

Heart disease is a condition in which a waxy substance is formed in the coronary arteries. This accumulation of plague waxy substance in the arteries makes the blood pumping process to slow down and eventually causes death if not [1]. Heart disease is one of the causes of disease and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important research areas in clinical data analysis. Now a day, the amount of data in the healthcare centers is large. Machine learning algorithms are widely used in object recognition and disease diagnosis [2]. In disease diagnosis machine learning algorithm turns a large collection of healthcare dataset into information that can assist to make better decisions and predictions. Prediction of disease and developing machine-based diagnostics systems is one of the goals of machine learning research that gained importance in the medical research field in support of the health experts' herby improving the precision and accuracy in decision making process during the identification and diagnosis of a disease [3]-15]

One the major problem in heart disease diagnosis is the error during diagnosis process. These errors occur due to lack of experienced specialists in the medical field to accurately and precisely identify the heart disease. Literature survey [1]-[25], shows that the heart disease is still a serious issue which needs further research works in order to address the mortality rate caused by the disease. In this research, we proposed heart disease prediction model by employing k-nearest neighbor (KNN) algorithm to and this research is aimed to answer the following questions: i) What is the right distance measure that produces the optimal accuracy for the KNN on heart disease prediction? ii) What is the performance of KNN algorithm on prediction of heart disease? iii) What is the effect of the value of neighbors on the predictive accuracy of KNN on heart disease prediction?

2. RELATED WORK

Numerous research works have been conducted which has focus on heart disease identification by employing machine-learning algorithms. The research works applied different machine learning algorithms to develop a prediction model for classification of the heart disease. Some of the previous research works on heart disease prediction are discussed in this section. Gavhane *et al.* [4], Naïve Bayes, decision tree and random forest algorithms are applied to Cleveland heart disease dataset. The predictive performance of the algorithms is evaluated on the test dataset and random forest algorithm outperformed than the decision tree and Naïve Bayes algorithm.

Hasan *et al*. [5], Gaussian Naïve Bayes algorithm is applied to an online University of California, Irvine (UCI) heart disease data repository. The algorithm is evaluated against the predictive accuracy and the experimental analysis of result shows that the highest accuracy achieved by the Gaussian Naïve Bayes on prediction of the heart disease is 84.05%.

Ambekar and Phalnikar [6], a comparative analysis on the predictive performance of machine learning algorithms, such as Gaussian Naïve Bayes, Logistic regression, random forest and KNN is conducted heart disease dataset. The comparison result shows that logistic regression outperformed the other algorithms with better accuracy on prediction.

Pawlovsky [7], heart disease prediction model is proposed by employing convolutional neural network (CNN). The accuracy of the proposed heart disease prediction model is evaluated on test dataset and the analysis of the result shows that the CNN algorithm achieved a prediction accuracy of 65%. Zunaidi *et al.* [8], KNN is applied to heart disease observations collected from Wisconsin. The authors compared the performance of linear and non-linear support vector machine. The result of performance analysis shows that the KNN has predictive accuracy of 84.8% on the heart disease classification problem.

Jothi *et al.* [9], a comparative study on machine learning algorithms namely, decision tree, random forest and multi-layer perception is conducted on the Wisconsin heart disease data repository. The algorithms are evaluated against their accuracy on heart disease prediction and the result shows that multi-layer perception, neural network is better on prediction of the heart disease. Jabbar *et al*. [10], support vector machine is applied to the heart disease data repository to develop a heart disease prediction model. The authors applied feature selection to improve the prediction performance of the proposed model and result shows that the model has accuracy of 56.16%.

Assegie *et al.* [11], Naïve Bayes is employed to Wisconsin heart disease data repository to predict a heart disease. The maximum prediction accuracy achieved by using this model is 87%. A prediction accuracy of 87% is acceptable in machine learning and prediction system and hence, Naïve Bayes model is better in performance and prediction of heart disease.

3. RESEARCH METHOD

In this research, the researcher collected heart disease data from Kaggle data repository for training and testing the proposed KNN model. For implementation and experimental testing, the researcher employed Python 3.7 programming language. A statistical method that is Pearson's correlation analysis and data visualization as well as feature relationship measures are employed for identification and interpretation of heart disease data repository to find out the relationship between the class and the features in observations. To develop heart disease prediction, model the researcher employed KNN algorithm. Figure 1 demosntrates heart disease distribution in the dataset.

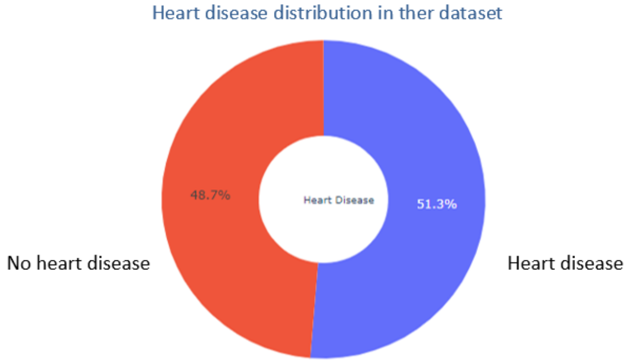


Figure 1. Heart disease patient and non-patient class distribution

3.1. Dataset description

In this study, Kaggle breast cancer data repository used in this study consists of 1025 observations and 13 features. Among the 1025 observations, 499 or 48.68% are heart disease negative and 526 or 51.32% are heart disease positive. The dataset has no missing feature values. Table 1 summarizes the details of the features of heart disease dataset. The dataset observations used in training is 80% and in testing 20% of the dataset is used.

Table 1 demonstrates heart disease dataset features employed for training and testing the KNN model. Figure 2 shows the distribution of heart disease patients against non patient class for different slopping conditions such as sup sloping, down slopping and flat. As demonstrated in the Figure 2, the number of patients is higher when the patient ST-T wave is up sloping.

Table 1. The Kaggle heart disease data repository features description

No.	Feature	Description
1	Age	The age of a person
2	CP	Chest paint (1= typical angina, 2= atypical angina, 3= non-angina pain, 4= asymptotic)
3	trestbps	Resting blood pressure
4	Chol	Serum cholesterol in mg/dl
5	fbs	Fasting blood sugar
6	restecg	Resting electrocardiographic results (values 0, 1, 2)
7	thalach	Maximum heart rate achieved
8	exang	*Exercise induced angina (1= yes, 0= no)*
9	Oldpeak	ST depression induced by exercise relative to rest
10	slope	The slope of the peak exercise ST segment
11	ca	Number of major vessels (0-3) colored by fluoroscopy
12	Thal	Thalassemia (3= normal, 6= fixed defect, 7= reversible defect)
13	target	Class label (1= patient, 0= not patient)

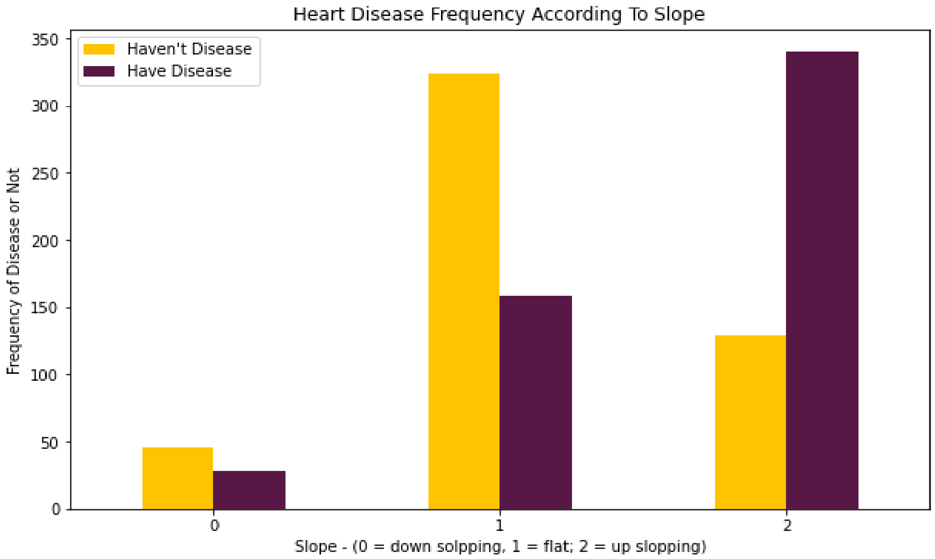
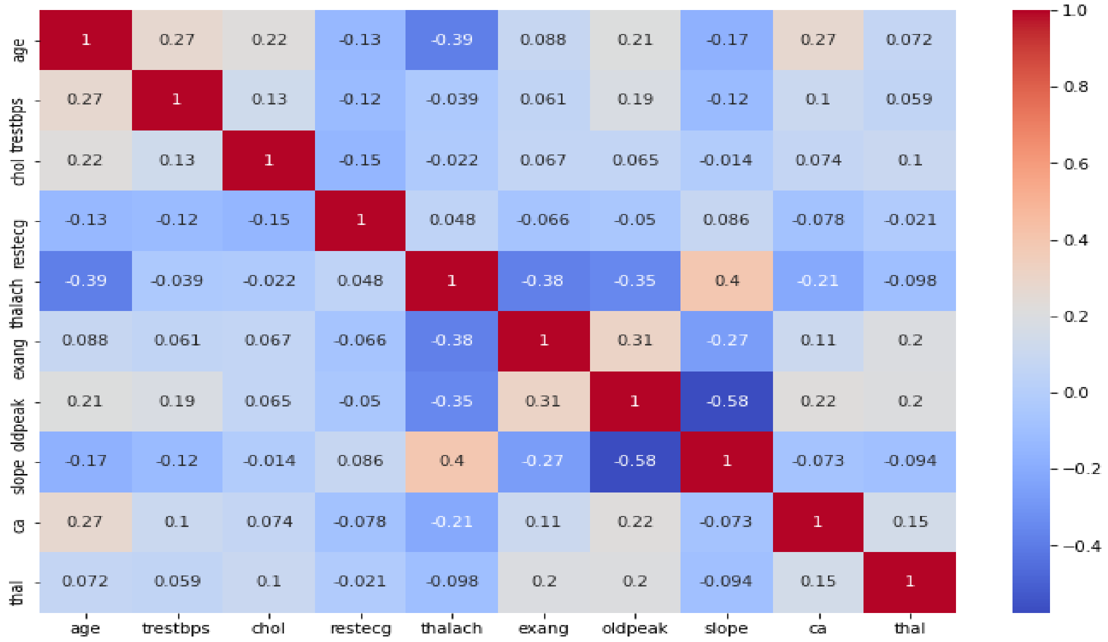


Figure 2. Slope vs heart disease positive and negative class

3.2. Feature correlation model

The author has employed Pearson’s correlation analysis for visualization of the relationship between each feature. This helps to identify the feature that is strongly related to the class feature in the data repository. The Pearson’s correlation matrix for each feature of the breast cancer dataset is shown in Figure 3. As illustrated in Figure 3, some of the features are highly correlated. For instance, age and total resting blood pressures (trestbps) has correlation value 0.27. Similarity cholesterol is highly correlated to age with correlation coefficient 0.22. In addition, number of major vessels has high correlation with age with correlation coefficient of 0.25. Slope and maximum heart rate achieved has high correlation value of 0.4. In contrast, features such as resting electrocardiogram and exercise-induced angina has negative correlation value with age feature.



	age	trestbps	chol	restecg	thalach	exang	oldpeak	slope	ca	thal
age	1	0.27	0.22	-0.13	-0.39	0.088	0.21	-0.17	0.27	0.072
trestbps	0.27	1	0.13	-0.12	-0.039	0.061	0.19	-0.12	0.1	0.059
chol	0.22	0.13	1	-0.15	-0.022	0.067	0.065	-0.014	0.074	0.1
restecg	-0.13	-0.12	-0.15	1	0.048	-0.066	-0.05	0.086	-0.078	-0.021
thalach	-0.39	-0.039	-0.022	0.048	1	-0.38	-0.35	0.4	-0.21	-0.098
exang	0.088	0.061	0.067	-0.066	-0.38	1	0.31	-0.27	0.11	0.2
oldpeak	0.21	0.19	0.065	-0.05	-0.35	0.31	1	-0.58	0.22	0.2
slope	-0.17	-0.12	-0.014	0.086	0.4	-0.27	-0.58	1	-0.073	-0.094
ca	0.27	0.1	0.074	-0.078	-0.21	0.11	0.22	-0.073	1	0.15
thal	0.072	0.059	0.1	-0.021	-0.098	0.2	0.2	-0.094	0.15	1

Figure 3. Heart disease feature correlation matrix

4. RESULT AND DISCUSSION

In this section, the experimental test results on the proposed model are explained. The predictive performance of decision tree and adaptive boosting algorithm is analyzed by employing the performance metrics such as accuracy and confusion matrix along with learning curve of the algorithms. Table 2 illustrates the accuracy of the proposed KNN model on five random tests.

As demonstrated in Table 1, the highest accuracy score on five random test is 92.68% with average accuracy of x%. The predictive performance of the proposed model is experimented on the training set. The predictive accuracy of the proposed model is shown in Figure 4.

Table 2. Accuracy of KNN

Experimental test	Accuracy in %
1	91.70
2	89.26
3	96.09
4	90.24
5	92.68

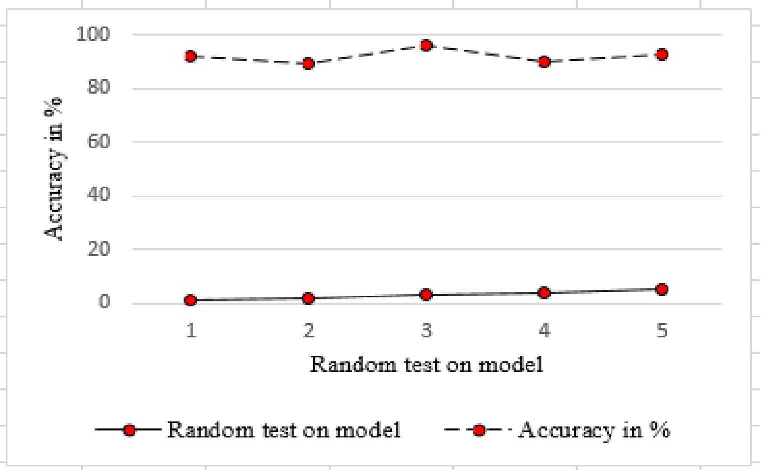


Figure 4. Heart disease feature correlation matrix

4.1. Confusion matrix

A confusion matrix is a measure the predictive performance of the proposed models in terms of the number of correct and incorrect predictions on the test set by the decision tree and adaptive boosting algorithm. The confusion matrix of the decision tree and adaptive boosting algorithm is shown in Figure 5.

4.2. Training and test accuracy vs k-values

Learning curves of the proposed model shows the performance of the model on training set for different k-values as demonstrated in Figure 6. The Figure 6 demonstrates the training and test set accuracy on the y-axis against the k-neighbors on the y-axis. The worst performance of the model is approximately 72.25%, which is still acceptable. And the model's best performance is at 1 neighbor ($K = 1$) and drops with higher values of neighbors.

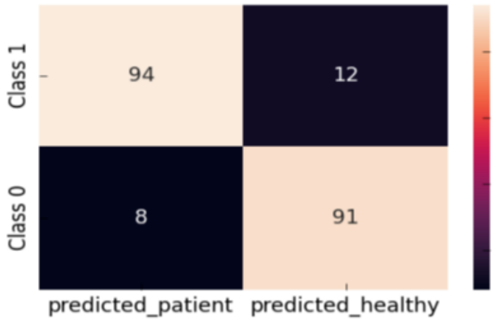


Figure 5. Confusion matrix of KNN model

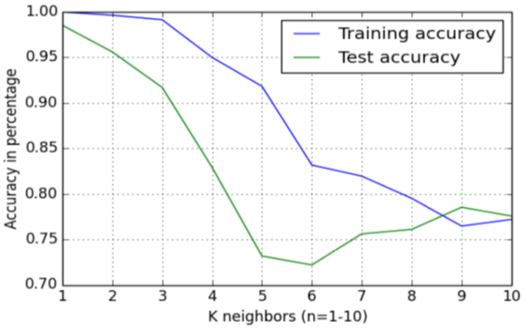


Figure 6. K value vs accuracy for KNN model

5. CONCLUSION

In this research, author proposed KNN based model for heart disease prediction by using dataset obtained from Kaggle machine learning data repository. The proposed model solves the problem of biased classification on imbalanced observation by non-ensemble algorithm through ensemble classifier namely the adaptive boosting. The predictive performance of the proposed model is evaluated by employing different performance metrics such as accuracy and confusion matrix on the test set. The result of performance analysis shows that the adaptive boosting algorithm has better performance than the decision tree. Hence, the adaptive boosting algorithm is a better classifier for imbalanced observations where the use of non-ensemble algorithm such as decision tree, results in biased prediction towards the majority class yielding better performance on prediction of the majority class and poor performance on the minority class.

ACKNOWLEDGEMENTS

The author would like to thank Injibara University for providing an internet and laboratory facilitates for conducting this work.

REFERENCES

[1] A. M. Yusof, N. A. M. Ghani, K. A. M. Ghani, and K. I. M. Ghani, "A predictive model for prediction of heart surgery procedure," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 3, pp. 1615–1620, Sep. 2019, doi: 10.11591/ijeecs.v15.i3.pp1615-1620.

[2] T. Suresh, T. A. Assegie, S. Rajkumar, and N. K. Kumar, "A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, pp. 1831–1838, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1831-1838.

[3] A. A. Hussein, "Improve the performance of K-means by using genetic algorithm for classification heart attack," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 2, p. 1256, Apr. 2018, doi: 10.11591/ijece.v8i2.pp1256-1261.

[4] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," Mar. 2018, doi: 10.1109/iceca.2018.8474922.

[5] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, and M. A. Hossain, "Comparative analysis of classification approaches for heart disease prediction," in *International Conference on Computer, Communication, Chemical, Material and Electronic Engineering, IC4ME2 2018*, Feb. 2018, pp. 1–4, doi: 10.1109/IC4ME2.2018.8465594.

[6] S. Ambekar and R. Phalnikar, "Disease risk prediction by using convolutional neural network," in *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, Aug. 2018, pp. 1–5, doi: 10.1109/ICCUBEA.2018.8697423.

[7] A. P. Pawlovsky, "An ensemble based on distances for a kNN method for heart disease diagnosis," in *International Conference on Electronics, Information and Communication, ICEIC 2018*, Jan. 2018, vol. 2018-Janua, pp. 1–4, doi: 10.23919/ELINFOCOM.2018.8330570.

[8] W. H. A. W. Zunaidi, R. D. R. Saedudin, Z. A. Shah, S. Kasim, C. Sen Seah, and M. Abdurohman, "Performances Analysis of Heart Disease Dataset using Different Data Mining Classifications," *International Journal on Advanced*

Science, Engineering and Information Technology, vol. 8, no. 6, p. 2677, Dec. 2018, doi: 10.18517/ijaseit.8.6.5042.

[9] N. Jothi, W. Husain, N. A. Rashid, and S. M. Syed-Mohamad, "Feature selection method using genetic algorithm for medical dataset," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 9, no. 6, pp. 1907–1912, Dec. 2019, doi: 10.18517/ijaseit.9.6.10226.

[10] M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier," Oct. 2016, doi: 10.1109/cimca.2016.8053261.

[11] T. A. Assegie, "a Support Vector Machine Based Heart Disease Prediction," *Journal of Software Engineering & Intelligent Systems Issn 2518-8739*, vol. 4, no. 3, pp. 111–116, 2019.

[12] T. A. Assegie and P. S. Nair, "The performance of different machine learning models on diabetes prediction," *International Journal of Scientific and Technology Research*, vol. 9, no. 1, pp. 2491–2494, 2020.

[13] T. A. Assegie, S. J. Sushma, and S. C. Prasanna, "Weighted decision tree model for breast cancer detection," *Technology reports of Kansai university*, vol. 62, no. 3, 2020.

[14] S. J. Sushma, T. A. Assegie, D. C. Vinutha, and S. Padmashree, "An improved feature selection approach for chronic heart disease detection," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3501–3506, Dec. 2021, doi: 10.11591/eei.v10i6.3001.

[15] K. Wang *et al.*, "Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and (SHAP)," *Computers in Biology and Medicine*, vol. 137, p. 104813, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104813.

[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*, Oct. 2020, doi: 10.1016/j.jksuci.2020.10.013.

[17] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," *Journal of King Saud University-Computer and Information Sciences*, vol. 24, no. 1, pp. 27–40, Jan. 2012, doi: 10.1016/j.jksuci.2011.09.002.

[18] K. Bahani, M. Moujabbir, and M. Ramdani, "An accurate fuzzy rule-based classification systems for heart disease diagnosis," *Scientific African*, vol. 14, p. e01019, Nov. 2021, doi: 10.1016/j.sciaf.2021.e01019.

[19] S. P. Patro, G. S. Nayak, and N. Padhy, "Heart disease prediction by using novel optimization algorithm: A supervised learning prospective," *Informatics in Medicine Unlocked*, vol. 26, p. 100696, 2021, doi: 10.1016/j.imu.2021.100696.

[20] T. A. Assegie, S. J. Sushma, B. G. Bhavya, and S. Padmashree, "Correlation Analysis for Determining Effective Data in Machine Learning: Detection of Heart Failure," *Computer Science*, vol. 2, no. 3, Apr. 2021, doi: 10.1007/s42979-021-00617-5.

[21] T. R. S. Mary and S. Sebastian, "Predicting Heart Ailment in Patients with Varying number of Features using Data Mining Techniques," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, p. 2675, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2675-2681.

[22] T. A. Assegie, "An optimized K-Nearest Neighbor based breast cancer detection," *Journal of Robotics and Control (JRC)*, vol. 2, no. 3, 2021, doi: 10.18196/jrc.2363.

[23] Y. K. Singh, N. Sinha, and S. K. Singh, "Heart Disease Prediction System Using Random Forest," in *Communications in Computer and Information Science*, Springer Singapore, 2017, pp. 613–623.

[24] M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," *Journal of Physics: Conference Series*, vol. 1817, no. 1, p. 12009, Mar. 2021, doi: 10.1088/1742-6596/1817/1/012009.

[25] L. Ali *et al.*, "A Feature-Driven Decision Support System for Heart Failure Prediction Based on χ^2 Statistical Model and Gaussian Naive Bayes," *Computational and Mathematical Methods in Medicine*, vol. 2019, pp. 1–8, Nov. 2019, doi: 10.1155/2019/6314328.

BIOGRAPHY OF AUTHOR

Tsehay Admassu Assegie received his first degree in Computer Science from Dilla University, Ethiopia, in 2013. He has also Master degree in Computer Science from Andhra University, India, in 2016. He is currently Computer Science researcher working as Lectured in Department of Computer Science, Injibara University, Ethiopia. His main research interests focus on Machine learning, Medical data Analysis and Data Mining. He can be contacted at email: tsehayadmassu2006@gmail.com

.