

An empirical study on machine learning algorithms for heart disease prediction

Tsehay Admassu Assegie¹, Prasanna Kumar Rangarajan², Napa Komal Kumar³,
Dhamodaran Vigneswari⁴

¹Department of Computer Science, College of Natural and Computational Science, Injibara University, Injibara, Ethiopia

²Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Chennai, India

³Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, Chennai, India

⁴Department of Information Technology, KCG College of Technology, Chennai, India

Article Info

Article history:

Received Apr 25, 2021

Revised May 24, 2022

Accepted Jun 12, 2022

Keywords:

Decision tree

Heart disease prediction

Random forest

Recursive feature elimination

Support vector machine

ABSTRACT

In recent years, machine learning is attaining higher precision and accuracy in clinical heart disease dataset classification. However, literature shows that the quality of heart disease feature used for the training model has a significant impact on the outcome of the predictive model. Thus, this study focuses on exploring the impact of the quality of heart disease features on the performance of the machine learning model on heart disease prediction by employing recursive feature elimination with cross-validation (RFECV). Furthermore, the study explores heart disease features with a significant effect on model output. The dataset for experimentation is obtained from the University of California Irvine (UCI) machine learning dataset. The experiment is implemented using a support vector machine (SVM), logistic regression (LR), decision tree (DT), and random forest (RF) are employed. The performance of the SVM, LR, DT, and RF models. The result appears to prove that the quality of the feature significantly affects the performance of the model. Overall, the experiment proves that RF outperforms as compared to other algorithms. In conclusion, the predictive accuracy of 99.7% is achieved with RF.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Tsehay Admassu Assegie

Department of Computer Science, Injibara University

P.O.B: 40, Injibara, Ethiopia

Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

In the last few years, the implementation and adoption of a machine learning algorithm for heart disease diagnosis have been the major focus of researchers [1]. The reason behind the wider adoption and application of machine learning and predictive model to heart disease prediction include the promising accuracy of the learning model compared to a human expert, the speed, and the cost expenditure spent for heart disease prediction or detection. Despite the wider adoption and application of the predictive model to heart disease diagnosis and their promising result on heart disease prediction, the performance of the machine learning model has still scope for improvement. In the literature, the impact of heart disease feature quality on the learning model is largely focused on. Hence, this study is aimed to further investigate the impact of heart disease feature quality and explores the most important or informative heart disease features that represent the heart disease patient resulting in better predictive outcomes.

This research focuses on the application of recursive feature elimination with the cross-validation (RFECV) method to process heart disease data before the model is trained using a support vector machine

(SVM), logistic regression (LR), decision tree (DT), and random forest (RF). The RFECV method is used to determine the most relevant risk factor heart disease features that are important for improving the prediction outcome of SVM, LR, DT, and RF. Generally, the goal of this research is to investigate the number of heart disease features required to develop a more accurate and computationally efficient model for heart disease prediction. In addition, the variability of the performance of SVM, LR, DT, and RF models for heart disease prediction is explored using the implemented RFECV method. This research follows an empirical methodology to experiment using RFECV for feature selection and SVM, LR, DT, and RF for model development using real-world data obtained from the University of California Irvine (UCI) heart disease data repository. The objectives of this research are to answer the questions shortlisted: i) what is the optimal number of heart disease feature that maximizes the performance of the SVM, DT, RF and LR model for heart disease prediction?; ii) what is the impact of varying the number of features, on the performance of SVM, DT, RF and LR model for heart disease prediction?; and iii) among SVM, LR, DT and RF which predictive model has high variability of cross-validation score for varying number of heart disease features?.

2. LITERATURE REVIEW

Heart disease is a cardiovascular disease that causes death all over the world [2]. The identification of heart disease is difficult using common heart disease risk factors such as high blood pressure, high cholesterol level, age, sex, and serum cholesterol. Overall, the characteristics of heart disease are complex and some of the heart diseases features overlap with other diseases such as chronic kidney disease. Thus, the identification of heart disease requires caution and heart disease treatment requires highly experienced cardiologists which is usually costly and requires much time and human effort.

Recently, machine learning is gaining importance in the health care industry as one of the means to combat the long-term effect of heart disease on society [3]–[5]. The higher precision, high performance, and cost-effectiveness is the major advantage of the predictive model in heart disease identification. With more and more patients admitted to hospitals, the diagnosis of heart disease is becoming more challenging. One of the major challenges in heart disease identification is that highly experienced cardiologists are required to identify heart disease accurately. However, training humans requires much effort and time usually, many years for healthcare clinicians to gain the necessary skill and experience in heart disease identification. Thus, machine learning has become not only an alternative solution to replace human experts in heart disease identification, but also a necessity to aid the decision-making process during heart disease identification. In this study, the authors developed a predictive model for heart disease diagnosis using supervised learning algorithms specifically, SVM, LR, DT, and RF. The authors have also experimented on the developed model using the heart disease dataset collected from the UCI data repository.

In [6], the researchers evaluated the performance of DT, RF, and artificial neural network (ANN) for heart disease diagnosis using the UCI heart disease data repository. The experimentation on DT, RF, and ANN shows that artificial neural network outperforms as compared to DT and RF model for heart disease detection. Overall, the predictive performance of 85.03%, 79.93%, and 79.93% is achieved with an artificial neural network, DT, and RF model respectively. Thus, the performance of an artificial neural network is better compared to DT and RF models. Moreover, in another study [7], the researchers applied RF to develop a predictive model that predicts heart disease. In addition, the authors experimented on the model with a heart disease test-set and the result shows that the performance of the model achieved an accuracy of 94.03%.

Despite the wide application of supervised machine learning algorithms to heart disease datasets for implementation of an automated intelligent model for heart disease prediction [8]–[10], literature shows that heart disease feature has an impact on heart disease prediction performance of the predictive model and feature selection also reduces the computational cost such as time and memory space. Thus, heart disease symptom or feature, which is important to represent a heart disease sample, has to be determined to improve the performance of the machine learning model for heart disease prediction. Thus, this research focused on the implementation automation of heart disease diagnosis with the RFECV method to obtain a better result on heart disease prediction using SVM, LR, DT, and RF models.

3. METHOD

To conduct this study, the authors reviewed recently published articles in reputed international scholarly journals indexed in Scopus. Then we collected clinical heart disease records and conducted exploratory data analysis using statistical methods such as correlation analysis and descriptive statistics. Much of the research work in the literature [11]–[16] has employed RF, SVM, LR, and DT to predict the framework of heart disease diagnosis. Based on the literature survey, we have selected the four most popular supervised machine learning algorithms namely, SVM, DT, RF, and LR to conduct experimental research on the performance of SVM, LR, DT, and RF. The heart disease dataset is collected from the UCI machine

learning data repository, which is one of the most popular machine learning data repositories for conducting experimental research in machine learning research [17]–[21]. The heart disease dataset employed in this study is demonstrated in Table 1.

Table 1. Heart disease dataset characteristics

Data source	No. of instances	No. of patients	No. of non-patients	No. of classes
UCI data repository	1025	526	498	2 (patient and healthy)

3.1. UCI heart disease feature description

The UCI heart disease dataset consists of 1,025 sample data points, each data point or sample is described by 13 heart disease features described in Table 2. The authors have considered 70% of the dataset or 717 samples and the remaining 30% or 308 data samples are used for testing. In addition, the dataset consists of balanced observations of the patient and non-patient class distribution.

Table 2. Heart disease dataset description

Feature	Data type	Description	Value
Age	Numeric	Age of patient	Mean=54, Max=77, Min=29
Sex	Nominal	Patient's gender (1=male,0=female)	1=Male, 0=Female
restecg	Numeric	Blood pressure in mmHg	0=Normal, 1=Having ST-T, 2=Showing probability
Cholesterol (chol)	Numeric	Continuous value in mm/dl	Mean=246, Max=564, Min=126
Fasting blood pressure (fbs)	Nominal	Level of sugar in blood >120 mg/dl (1=yes, 0=no)	>120 mg/dl, 1=Yes, 0=No
Heart rate achieved (thalach)	Numerical	Heart rate in mmHg	Mean=149, Max=202, Min=71
Thallium scan(thal)	Nominal	Nominal (3=Normal,6=fixed defect,7=Reversible defect)	
Exercise induced angina(exang)	Nominal	Nominal (presence of exercise induced angina,1=present,0=absent)	1=Yes, 0=No
Slope (slope)	Nominal	Nominal (1=Up slopping,2=Flat,3=Down slopping)	1=Up slopping, 2=Flat, 3=Down slopping
Status of fluoroscopy (ca)	Nominal	Nominal (number of vessels colored through X-ray (0 to 3))	Continuous values (0-3)
Chest pain (cp)	Nominal	Nominal (With chest pain=1, no chest pain=0)	1=Typical, 2=Atypical, 3=Non-angina, 4=Asymptomatic
oldpeak		S-T depression induced by exercise relative to rest nominal (0 to 6)	Mean=1.07, Max=6.2, Min=0
Resting blood pressure (tresbps)	Nominal	Blood pressure at rest	Mean=131, Max=200, Min=94
Target	Nominal	Predicted class	1=Patient, 0=Healthy or not patient

3.2. Correlation model

To get insight into the heart disease dataset and explore the dependency or collinearity that exists among heart disease features we have employed Pearson correlation to the heart disease dataset for exploratory data analysis. Figure 1 demonstrates the correlation circle for heart disease features. Pearson correlation among each heart disease feature is determined by using the Pearson's correlation formula given in (1) [22]–[24]:

$$r = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\sum (xi - \bar{x})^2 \sum (yi - \bar{y})^2}} \quad (1)$$

where r denotes Pearson correlation coefficient, xi denotes values of the variable x in the heart disease dataset, Yi denotes values of the variable y in the heart disease dataset, \bar{x} denotes the mean of variable x and \bar{y} denotes the mean of the values of variable y in the heart disease dataset.

Figure 1, shows the correlation of heart disease dataset features. Total resting blood pressure, age, fasting blood sugar, cholesterol, and quantity of main vessels colored by fluoroscopy have a positive correlation. In addition, sex, heart rate, exercise-induced angina, oldpeak, and heart rate have a positive correlation to each other. Similarly, chest pain, slope and maximum heart rate achieved has a positive correlation to each other. In contrast, chest pain, maximum heart rate, and slope are negatively correlated to sex, heart rate, exercise-induced angina, oldpeak, and heart rate. Similarly, resting electrocardiography has a

negative correlation to total resting blood pressure, age, fasting blood sugar, cholesterol, and quantity of main vessels colored X-ray.

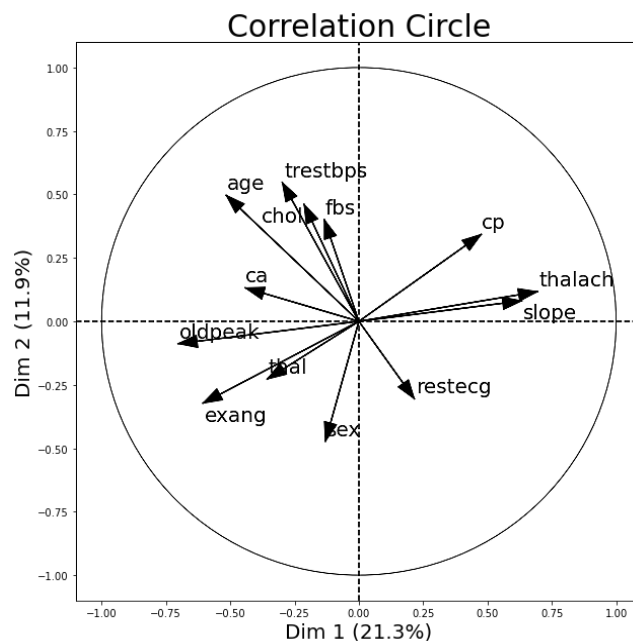


Figure 1. Correlation graph for heart disease features

3.3. RFECV

Recursive feature elimination (RFE) is a feature selection technique that fits the model and eliminates less important features (or features) until the specified number of features is selected. Features are ranked by their important characteristics, and by recursively removing a small number of features per loop, RFE eliminates dependencies or collinearity that exists between the features [25]. To determine the relevant number of heart disease features, we have employed RFE with cross-validation or RFECV to compute the cross-validation score on the selected heart disease feature.

4. RESULTS AND DISCUSSION

This section presents the results such as accuracy and cross-validation variability between SVM, LR, DT, and RF for varying features. The performance of SVM, DT, RF, and LR is evaluated using accuracy, receiver operating characteristics the area under the curve (AUC), and average precision. In experimentation, the model is tested against a varying number of input features. Cross-validated accuracy is computed for different input feature sizes and the result is compared.

4.1. The effect of heart disease input feature size on classifier performance

To determine the optimal number of heart disease features, cross-validation is used with RFE to score different feature subsets and select the best performing collection of heart disease features. The RFECV is demonstrated in Figures 2(a) and 2(b) and Figures 3(a) and 3(b) show the number of heart disease features in the SVM, R, DT, and LR models along with their cross-validated test score and variability respectively. Moreover, Figures 2(a) and 2(b) and Figures 3(a) and 3(b) demonstrate the selected number of heart disease features for each model.

Figure 2 proves that the performance of the SVM and RF model is highly affected by the heart disease feature used for model training. Figures 2(a) and 2(b) and Figures 3(a) and 3(b) show the RFECV curve of the proposed model, the SVM achieves a higher accuracy when 11 informative heart disease features are used as shown in Figures 2(a), then gradually decreases the accuracy as the non-informative heart disease features are added into the model. Similarly, RF achieves a higher accuracy when 11 informative heart disease features are used as shown in Figure 2(b). In addition, the shaded region represents the variability of cross-validation or standard deviation above and below the mean accuracy score drawn by the cross-validation curve. We see from Figures 2(a) and 2(b) that there is a high variability of cross-

validation scores with varying heart disease features using SVM as compared to the RF model on heart disease prediction.

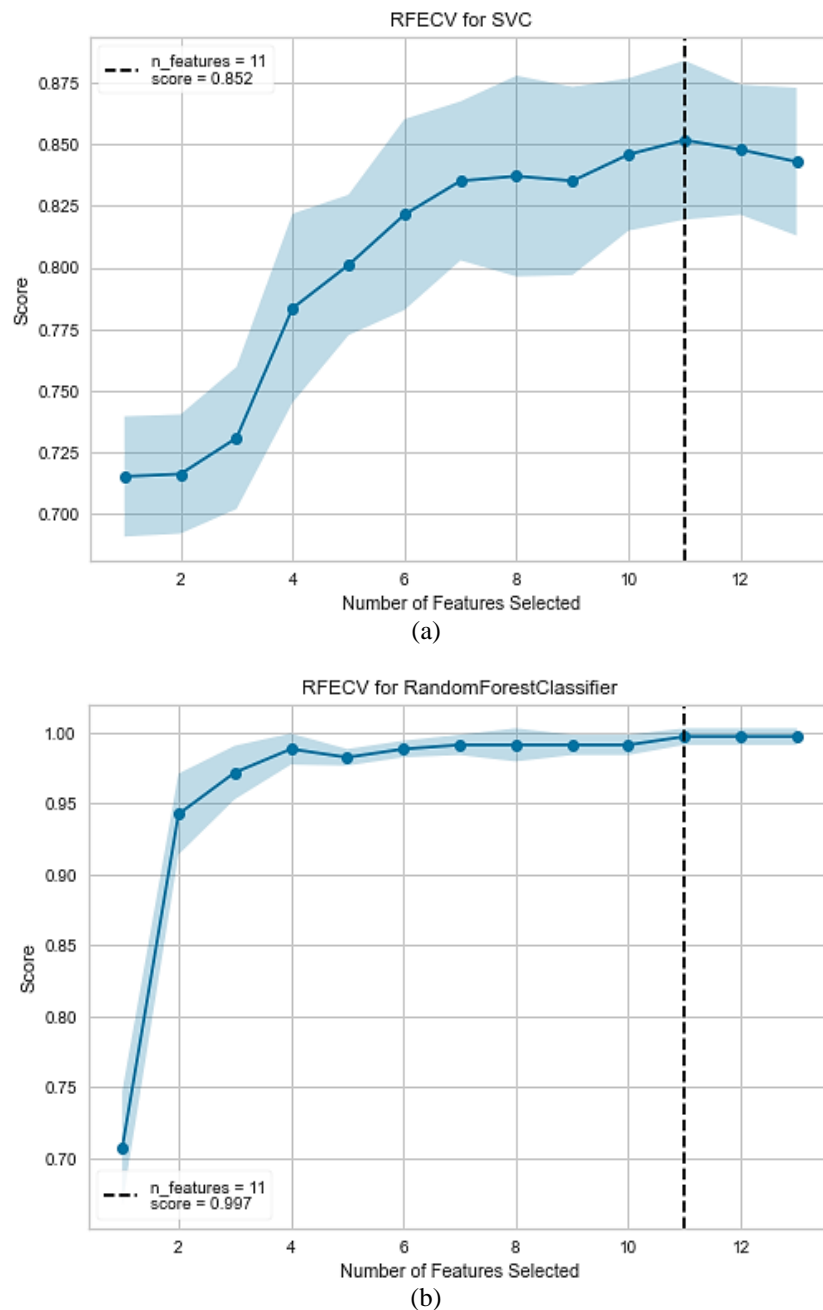


Figure 2. Effects of selecting heart disease feature on the performance of classifier (a) SVM and (b) RF

We see from Figures 3(a) and 3(b) that DT achieves a higher accuracy when 8 informative heart disease features are used as shown in Figure 3(a), then remained constant accuracy as the non-informative heart disease features are added into the model. Similarly, LR achieves a higher accuracy when 10 informative heart disease features are used as shown in Figure 3(b). In addition, we see from Figures 3(a) and Figure 3(b) that there is a high variability of cross-validation scores with varying heart disease features using LR as compared to the DT model on heart disease prediction.

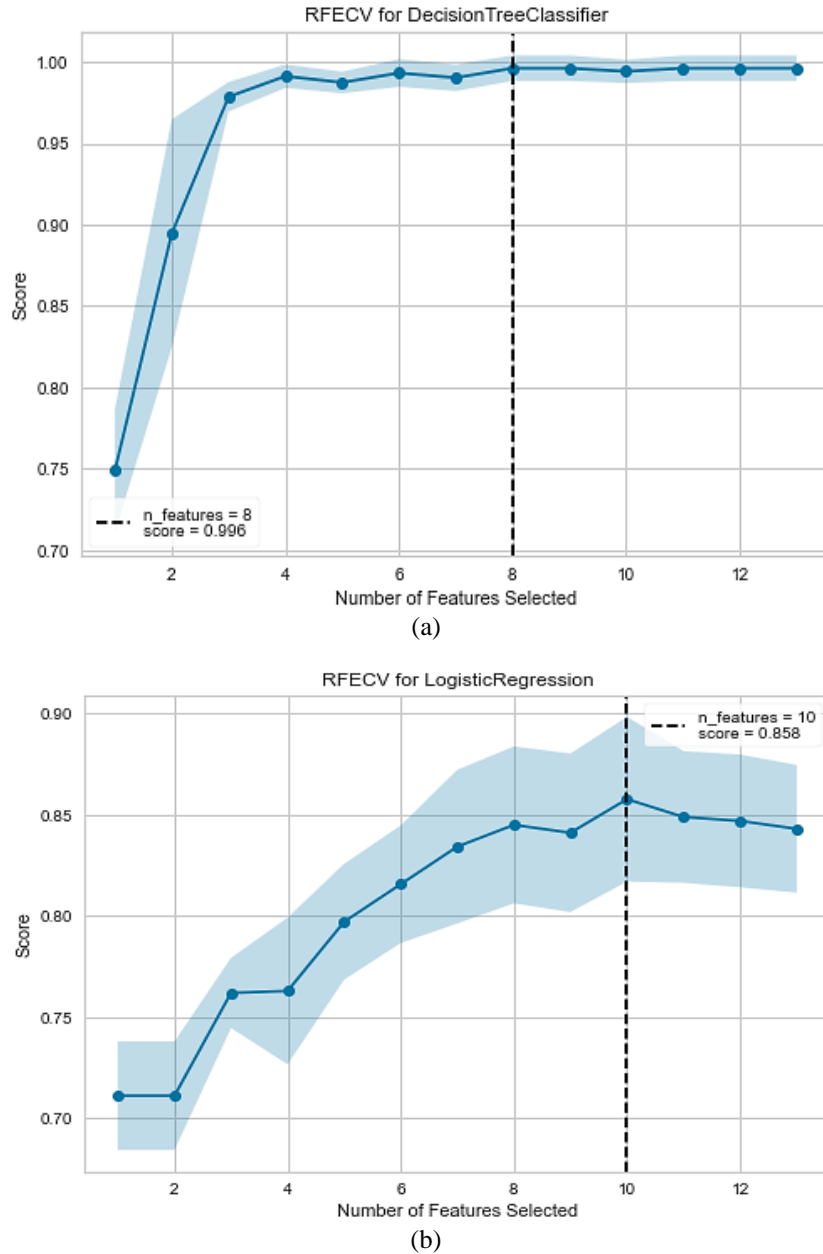


Figure 3. Effects of selecting heart disease feature on the performance of classier (a) DT and (b) LR

4.2. Comparison between the accuracy of the model

The authors employed accuracy as a performance metric to evaluate and compare the performance of SVM, DT, LR, and RF models. In comparison, the highest accuracy achieved by each model is different. In addition to accuracy variation across the different models, the experimental result appears to prove that the model performance varies for different features. Table 3 illustrates the variation in the performance of the model for a varying input feature.

Table 3. Heart disease dataset feature description

Model	No. of features	Highest accuracy	Avg. AUC	Avg. precision
LR	10	85.8%	0.88	0.94
SVM	11	85.2%	0.84	0.86
DT	8	99.6%	0.96	1.00
RF	11	99.7%	1.00	1.00

5. CONCLUSION

In this study, the authors conducted an empirical study on the performance of machine learning for heart disease prediction using SVM, LR, DT, and RF. Furthermore, we have employed RFECV to select optimal input features to obtain better heart disease diagnosis outcomes. With RFECV, we have determined the optimal number of heart disease features that maximize the heart disease diagnosis outcome of the proposed model. In addition, the proposed model is compared to the existing model and the experimental result shows that the RF model outperforms as compared to DT, SVM, and LR. Overall, a random forest model was performed with a classification accuracy of 99.7%. The performance of RF, DT, SVM, and LR models is 99.7% and 99.6%. 85.2% and 85.8% respectively.

ACKNOWLEDGEMENTS

The authors would like to thank Inijbara University for providing internet facilities and laboratory equipment (Laptop) for conducting this work.





REFERENCES

- [1] T. A. Assegie, R. L. Tulasi, V. Elanangai, and N. K. Kumar, "Exploring the performance of feature selection method using breast cancer dataset," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 25, no. 1, pp. 232–237, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp232-237.
- [2] T. A. Assegie, "An optimized k-nearest neighbor based breast cancer detection," *J. Robot. Control*, vol. 2, no. 3, pp. 115–118, 2021, doi: 10.18196/jrc.2363.
- [3] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digit. Heal.*, vol. 6, p. 205520762091477, Jan. 2020, doi: 10.1177/2055207620914777.
- [4] S. Sun, Z. Cao, H. Zhu, and J. Zhao, "A survey of optimization methods from a machine learning perspective," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3668–3681, Aug. 2020, doi: 10.1109/TCYB.2019.2950779.
- [5] R. Aggrawal and S. Pal, "Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease," *SN Comput. Sci.*, vol. 1, no. 6, Nov. 2020, doi: 10.1007/s42979-020-00370-1.
- [6] A. Niakouei, M. Tehrani, and L. Fulton, "Health disparities and cardiovascular disease," *Healthcare*, vol. 8, no. 1, Mar. 2020, doi: 10.3390/healthcare8010065.
- [7] L. J. Muhammad, I. Al-Shourbaji, A. A. Haruna, I. A. Mohammed, A. Ahmad, and M. B. Jibrin, "Machine learning predictive models for coronary artery disease," *SN Comput. Sci.*, vol. 2, no. 5, Sep. 2021, doi: 10.1007/s42979-021-00731-4.
- [8] J. H. Joloudari *et al.*, "Coronary artery disease diagnosis; ranking the significant features using a random trees model," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, p. 731, Jan. 2020, doi: 10.3390/ijerph17030731.
- [9] T. R. S. Mary and S. Sebastian, "Predicting heart ailment in patients with varying number of features using data mining techniques," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 4, p. 2675, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2675-2681.
- [10] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00327-4.
- [11] R. Sigit, A. Basuki, and - Anwar, "A new feature extraction method for classifying heart wall from left ventricle cavity," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 10, no. 3, p. 964, Jun. 2020, doi: 10.18517/ijaseit.10.3.12152.
- [12] C. R., "Heart disease prediction system using supervised learning classifier," *Bonfring Int. J. Softw. Eng. Soft Comput.*, vol. 3, no. 1, pp. 1–7, Mar. 2013, doi: 10.9756/BIJSESC.4336.
- [13] D. Khanna, R. Sahu, V. Baths, and B. Deshpande, "Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease," *Int. J. Mach. Learn. Comput.*, vol. 5, no. 5, pp. 414–419, Oct. 2015, doi: 10.7763/IJMLC.2015.V5.544.
- [14] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *IAES Int. J. Artif. Intell.*, vol. 10, no. 1, pp. 184–190, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp184-190.
- [15] C.-A. Cheng and H.-W. Chiu, "An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2017, pp. 2566–2569, doi: 10.1109/EMBC.2017.8037381.
- [16] E. Nikookar and E. Naderi, "Hybrid ensemble framework for heart disease detection and prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 5, pp. 243–248, 2018, doi: 10.14569/IJACSA.2018.090533.
- [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, doi: 10.14569/IJACSA.2019.0100637.
- [18] A. M. Alqudah, M. AlTantawi, and A. Alqudah, "Artificial intelligence hybrid system for enhancing retinal diseases classification using automated deep features extracted from OCT images," *Int. J. Intell. Syst. Appl. Eng.*, vol. 9, no. 3, pp. 91–100, Sep. 2021, doi: 10.18201/ijisae.2021.236.
- [19] Z. Khandezamin, M. Naderan, and M. J. Rashti, "Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier," *J. Biomed. Inform.*, vol. 111, Nov. 2020, doi: 10.1016/j.jbi.2020.103591.
- [20] S. J. Sushma, T. A. Assegie, D. C. Vinutha, and S. Padmashree, "An improved feature selection approach for chronic heart disease detection," *Bull. Electr. Eng. Informatics*, vol. 10, no. 6, pp. 3501–3506, Dec. 2021, doi: 10.11591/eei.v10i6.3001.
- [21] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *J. King Saud Univ. - Comput. Inf. Sci.*, Oct. 2020, doi: 10.1016/j.jksuci.2020.10.013.
- [22] T. Suresh, T. A. Assegie, S. Rajkumar, and N. Komal Kumar, "A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 2, pp. 1831–1838, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1831-1838.
- [23] I. Javid, A. Khalaf, 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, doi: 10.14569/IJACSA.2020.0110369.





- [24] P. C. Kaur, "A study on role of machine learning in detectin heart diseases," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Mar. 2020, pp. 188–193, doi: 10.1109/ICCMC48092.2020.ICCMC-00037.
- [25] R. Naseem *et al.*, "Performance assessment of classification algorithms on early detection of liver syndrome," *J. Healthc. Eng.*, vol. 2020, pp. 1–13, Dec. 2020, doi: 10.1155/2020/6680002.

BIOGRAPHIES OF AUTHORS







Mr. Tsehay Admassu Assegie     holds a Master of Science degree from Andhra University, India 2016. He also received his B.Sc. (Computer Science) from Dilla University, Ethiopia in 2013. His research includes machine learning, data mining, bioinformatics, network security and software defined networking. He has published over 33 papers in international journals and conferences. He can be contacted at email: tsehayadmassu2006@gmail.com.







Mr. Prasanna Kumar Rangarajan     is presently serving as Faculty in the Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham Chennai. He has 20 years of teaching experience. His area of interest includes the theory of computation, compiler design, machine learning and data science. He can be contacted at email: r_prasannakumar@ch.amrita.edu.



Mr. Napa Komal Kumar     is currently working as an Assistant Professor in the Department of Computer Science and Engineering at St. Peter's Institute of Higher Education and Research, Avadi, Chennai. His research interests include Machine Learning, Data Mining, and Cloud Computing. He can be contacted at email: komalkumarnapa@gmail.com.



Vigneswari Dhamodaran     is currently working as an Assistant Professor in the Department of Information Technology, KCG College of Technology, Chennai, Tamil Nadu, India. Her research interests include data mining, and machine learning. She can be contacted at email: vigneswari121192@gmail.com.