

Scalability and performance of decision tree for cardiovascular disease prediction

Tsehay Admassu Assegie¹, Komal Kumar Napa², Thiyagu Thulasi³, Angati Kalyan Kumar²,
Maran Jeyanthiran Thiruvarasu Vasantha Priya⁴, Vigneswari Dhamodaran⁵

¹School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, Republic of Korea

²Department of Computer Science and Engineering (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, India

³Department of Computer Science and Engineering (Cyber Security), Madanapalle Institute of Technology & Science, Madanapalle, India

⁴Department of Artificial Intelligence and Data Science, Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, India

⁵Department of Information Technology, KCG College of Technology, Karapakkam, Chennai, India

Article Info

Article history:

Received Oct 25, 2023

Revised Feb 16, 2024

Accepted Mar 14, 2024

Keywords:

Automated diagnostics

Computational model

Machine learning

Scalability in machine learning

ABSTRACT

As one of the most common types of disease, cardiovascular disease is a serious health concern worldwide. Early detection is crucial for successful treatment and improved survival rates. The decision tree is a robust classifier for predicting the risk of cardiovascular disease and getting insights that would assist in making clinical decisions. However, selecting a better model for cardiovascular disease could be challenging due to scalability issues. Hence, this study examines the scalability and performance of decision trees for cardiovascular disease prediction. The study evaluated the performance of a decision tree for predicting cardiovascular disease. The performance evaluation was carried out by employing a confusion matrix, cross-validation score, model complexity, and training score for varying sizes of training samples. The experiment depicted that, the decision tree model was 88.8% accurate in predicting the presence or absence of cardiovascular disease. Therefore, the implementation of the decision tree is beneficial for the prediction and early detection of heart disease events in patients.

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

School of Electronic and Electrical Engineering, Kyungpook National University

Daegu, Republic of Korea

Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

A recent study [1] suggested that the role of machine learning techniques has become significant in the clinical decision-making process. Machine learning systems save the time and money expenditure of cardiovascular patients with clinical assistance on a preliminary basis. Furthermore, machine-learning systems aid medical practitioners in their decision-making during the diagnosis of cardiovascular disease.

A prior study [2] has also demonstrated that machine learning-based heart disease diagnosis achieves an accuracy of 96%. A comparative analysis of different supervised learning models on logistic regression, random forest, support vector machine, artificial neural network, and k-nearest neighbor suggested that the support vector machine achieves a 96% accuracy score for heart disease prediction. Where the k-nearest neighbor achieves 91% accuracy, which is lower compared with the support vector machine.

Correspondingly, Mienye *et al.* [3] further improved the performance of classification and regression trees (CART) by introducing weights. The study has suggested that the proposed CART model

scored an accuracy of 93% outperforming other CART-based heart disease classification models in the literature. Even though, the proposed model achieved higher accuracy compared with the existing models; only accuracy, precision, and f-score were employed in comparison. Moreover, other ensemble models such as extreme gradient boosting (XGBoost), and random forest were not included in the comparison. Other metrics such as model complexity, scalability, and confusion matrix are important in the comparative analysis of machine learning models for cardiovascular disease prediction [4].

Additionally, several studies suggested that the ensemble model assists in cardiovascular disease prediction [5], [6]. Furthermore, the study emphasized that the prediction of cardiovascular disease has become significant for life-saving. The experimental result of multilayer perceptron in predicting cardiovascular disease demonstrated an accuracy of 87.23% showing promising results in assisting the cardiovascular disease prediction in the early stage.

Moreover, Jan *et al.* [7] developed an ensemble model by combining the predictive ability of different models to improve the accuracy of the machine-learning model for cardiovascular disease diagnosis. The proposed ensemble learning approach employed a support vector machine, artificial neural network, naïve Bayesian, regression analysis, and random forest, as a base classifier. The study employed cardiovascular datasets collected from Cleveland and Hungarian obtained from the UCI repository. The investigation established that an ensemble model is a superior approach in terms of predictive accuracy.

Similarly, Zaini and Awang [8] proposed an ensemble-stacking model for cardiovascular disease risk prediction. The study suggested that the performance of the ensemble-stacking model improves with hybrid feature selection. Chi-square and variance analysis of the cardiovascular disease feature improve the ensemble model for predicting the risk of heart disease in ten years. With the optimal features selected based on these methods, the performance of a stacking ensemble based on a logistic regression-based model yields 93.44% accuracy.

The predictive modeling of cardiovascular disease has attained significance in clinical research and patient care [9]. The accurate prediction of cardiovascular risk ensures that health implications and the risk of cardiovascular disease are timely considered with machine learning-based health risk assessment models. The recent healthcare sectors are data intensive. A huge quantity of patient data in the form of patient descriptions, clinical reports, and laboratory tests are being collected daily. With developments in data storage technologies, and with the growing use of digital health record systems in hospitals, it is now possible to process higher volumes of data and make decisions.

Among the various machine-learning techniques, a research paper [10] employed the K-nearest neighbor, decision tree, support vector machine, and naïve Bayes employed for the cardiovascular risk assessment. The support vector machine model scored 94.2% accuracy outperforming other models on the Cleveland dataset. Similarly, Molla *et al.* [11] proposed a predictive framework for heart disease prediction with machine learning techniques. The study employed the Cleveland heart disease dataset for training XGB, gradient boosting (XG), support vector machine, and decision tree. The study suggested that univariate feature selection improves the performance of the decision tree model. With 10 heart disease features selected by the univariate analysis, the decision tree model achieves higher accuracy 97.75% compared with other models.

In addition, Sarra *et al.* [12] enhanced the performance of the artificial neural network for heart disease prediction. The developed artificial neural networks-based heart disease prediction model achieved 93.44% accuracy. Compared to the support vector machine model the accuracy of artificial neural networks is 7.5%. The training time of the developed model demonstrated a training time of less than a minute. Machine learning and deep learning have been employed to develop a computational model for the angiographic disease status of a patient. Almulihi *et al.* [13] evaluated an ensemble-based computational intelligence model for early heart disease prediction. The computational intelligence model is developed by employing deep learning. The study improves the performance of the deep learning model by developing an ensemble model using deep learning as a based model. The proposed model achieves 98.42% accuracy.

Kumar and Vigneswari [14] developed a hard majority voting ensemble model for the assessment of early cases of heart disease. The study suggested that grid search improves cross-validation accuracy for the proposed model. In addition, the preprocessing method with scaling has improved the accuracy of the majority voting ensemble model. Overall, the developed model achieved a prediction accuracy of 90% with grid search and scaling as pre-processing methods. The performance ensemble-learning model improves with feature selection. According to Alim *et al.* [15], tree-based feature selection such as permutation feature importance with the random forest model improves the performance of the model by 1.51% accuracy. Moreover, pre-processing methods such as normalization and missing value removal improve the performance of the random forest model in predicting the presence of heart disease.

This study is motivated by the better performance achieved by machine learning models in the preliminary literature reviews presented in research articles [1]–[5]. Moreover, due to the higher performance of decision trees for cardiovascular disease prediction, the study aims to evaluate the performance of the

model. The objectives of the paper include i) providing a literature review of machine learning techniques employed to assist in the diagnosis of cardiovascular disease, ii) to study and evaluate the performance of different decision trees for cardiovascular disease prediction, and iii) to apply the different model evaluation performances such as model complexity, and confusion matrix in the comparative analysis, in addition to the accuracy measure which is usually employed for model comparison.

The organization of the study is given as follows: section 2 presents the methodology employed in the study which provides a discussion of the dataset, and evaluation approach. Section 3 discusses the result. Last, section 5 summarizes the work.

2. METHOD

The chronology for this research is discussed as follows. The first step involves data collection. The second step involves exploratory data analysis with the help of descriptive statistics such as correlation, and missing value analysis then feature scaling. The third step discusses model development with the help of ensemble learning algorithms. The fourth step involves the evaluation of the model performance with performance measures such as confusion matrix, cross-validation score, and model complexity. The final step presents the selection of a high-performing model among the candidate models and the recommendation of a better ensemble model for early diagnosis of heart disease.

The dataset for this study is collected from the Cleveland UCI heart disease data repository. The UCI data repository is a benchmark dataset. It is widely used for machine learning research, previously verified by several studies [16]–[18]. In exploratory data analysis, the study employed descriptive statistics such as mean, standard deviation, maximum, and count to analyze the collected dataset. This study employed the K-fold cross-validation to evaluate the performance of the ensemble learning model for heart disease prediction. The K-fold cross-validation refers to separating the dataset into a subset of K-folds for training and testing [19]–[21]. This method is often used for prediction models to determine how the built model implements unobserved instances in the training set [22]–[25]. After the data is collected, we explore the data for missing values, and the feature is scaled. Thirdly, the classifier model was developed using a decision tree algorithm. Then, the final step involved the evaluation of the performance of the decision tree algorithm with various metrics such as confusion matrix, and cross-validation accuracy.

3. RESULTS AND DISCUSSION

This section presents the results achieved with ensemble learning models for heart disease prediction. The effectiveness of the ensemble-learning model is evaluated against prediction accuracy, training time complexity, and confusion matrix. In addition, this section presents the result obtained from standard performance measures such as validation score and the scalability of the model over a larger training sample for each of the models.

Figure 1 indicates the confusion matrix for the decision tree model for predicting the presence or absence of heart disease. As indicated in Figure 1 the decision tree model incorrectly predicted 41 observations among the total number of observations used in the testing. Figure 1 also indicates that the decision tree model achieves a training accuracy of 88.8%, and 73.3% validation accuracy in predicting the true positive class (heart disease patient observations).

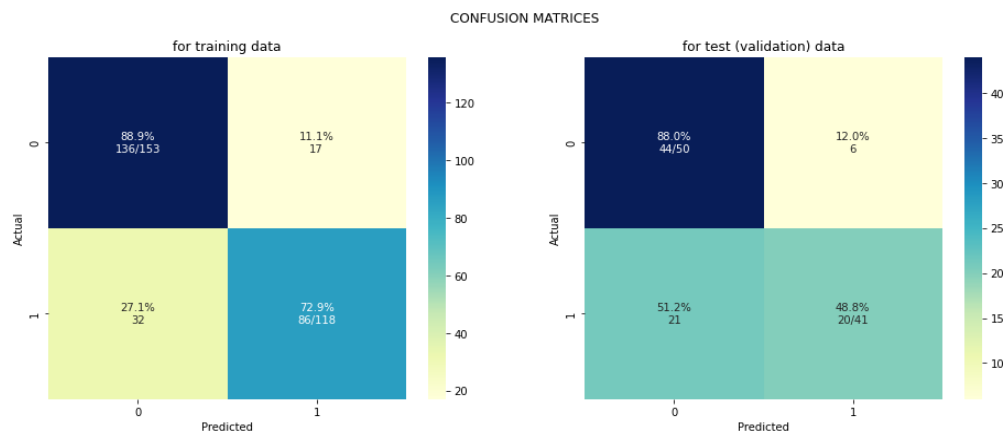


Figure 1. The confusion matrix for the decision tree model

Figure 2 demonstrates the training, cross-validation score, and scalability of the model with the time required to fit on varying the size of the training sample. The training score of the decision tree achieves higher results compared to the cross-validation score. Moreover, the training score is higher for samples of size 101, and the cross-validation accuracy is high for approximately 176 training samples. The decision tree is also scalable as the fit time is less than 0.18 seconds for a training sample of 176 observations.

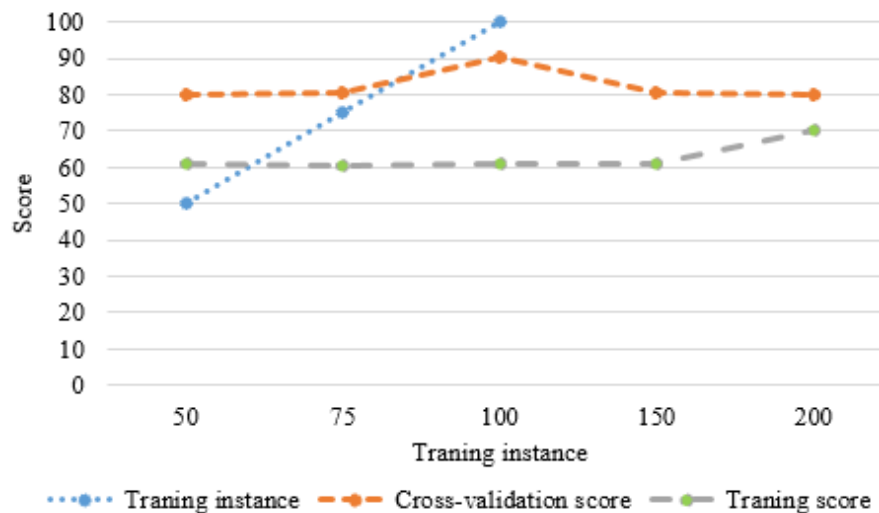


Figure 2. Scalability, training, and validation score of decision tree model

4. CONCLUSION

This study investigated the performance and scalability trade-off in terms of time complexity of ensemble learners. The paper evaluated the performance of the decision tree on the heart disease dataset. The experiment suggests that the decision tree achieves a higher cross-validation score of 83.3%. In future work, the researchers recommend further investigation of different models such as support vector machine, K-nearest neighbor, and the ensemble learning methods on different datasets such as hypothyroid and breast cancer. Moreover, the problem of the trade-off between cross-validation score and model complexity needs much research effort for investigating feature scaling, and feature selection methods as a solution to model complexity and cross-validation score trade-off. However, our study has some limitations. First, we used only one dataset for our study, which may not be representative of all cardiovascular disease cases. Second, we used only decision tree techniques, and other techniques may yield different results.




REFERENCES

- [1] P. Guleria, P. Naga Srinivasu, S. Ahmed, N. Almusallam, and F. K. Alarfaj, "XAI framework for cardiovascular disease prediction using classification techniques," *Electronics*, vol. 11, no. 24, 2022, doi: 10.3390/electronics11244086.
- [2] B. S. Shukur and M. M. Mijwil, "Involving machine learning techniques in heart disease diagnosis: a performance analysis," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, pp. 2177-2185, Apr. 2023, doi: 10.11591/ijece.v13i2.pp2177-2185.
- [3] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, 2020, doi: 10.1016/j.imu.2020.100402.
- [4] Y. Jiang *et al.*, "Cardiovascular disease prediction by machine learning algorithms based on cytokines in kazakhs of china," *Clinical Epidemiology*, vol. 13, pp. 417-428, 2021, doi: 10.2147/CLEP.S313343.
- [5] A. Alfaidi, R. Aljuhani, B. Alshehri, H. Alwadei, and S. Sabbeh, "Machine learning: assisted cardiovascular diseases diagnosis," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 2, pp. 135-141, 2022, doi: 10.14569/IJACSA.2022.0130216.
- [6] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, pp. 261-268, 2019, doi: 10.14569/ijacsa.2019.0100637.
- [7] M. Jan, A. A. Awan, M. S. Khalid, and S. Nisar, "Ensemble approach for developing a smart heart disease prediction system using classification algorithms," *Research Reports in Clinical Cardiology*, vol. 9, pp. 33-45, 2018, doi: 10.2147/rcc.s172035.
- [8] N. A. M. Zaini and M. K. Awang, "Hybrid feature selection algorithm and ensemble stacking for heart disease prediction," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, pp. 158-165, 2023, doi: 10.14569/IJACSA.2023.0140220.
- [9] T. A. Assegie, A. O. Salau, C. O. Omeje, and S. L. Braide, "Multivariate sample similarity measure for feature selection with a resemblance model," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 3, pp. 3359-3366, 2023, doi: 10.11591/ijece.v13i3.pp3359-3366.




- [10] 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.
- [11] S. Molla *et al.*, "A predictive analysis framework of heart disease using machine learning approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 5, pp. 2705–2716, 2022, doi: 10.11591/eei.v11i5.3942.
- [12] R. R. Sarra, A. M. Dinar, and M. A. Mohammed, "Enhanced accuracy for heart disease prediction using artificial neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 375–383, 2023, doi: 10.11591/ijeecs.v29.i1.pp375-383.
- [13] A. Almulihi *et al.*, "Ensemble learning based on hybrid deep learning model for heart disease early prediction," *Diagnostics*, vol. 12, no. 12, Dec. 2022, doi: 10.3390/diagnostics12123215.
- [14] N. K. Kumar and D. Vigneswari, "A drug recommendation system for multi-disease in health care using machine learning," *Lecture Notes in Electrical Engineering*, vol. 668, pp. 1–12, 2021, doi: 10.1007/978-981-15-5341-7_1.
- [15] M. A. Alim, S. Habib, Y. Farooq, and A. Rafay, "Robust heart disease prediction: a novel approach based on significant feature and ensemble learning model," *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies: Idea to Innovation for Building the Knowledge Economy, iCoMET 2020*, 2020, doi: 10.1109/iCoMET48670.2020.9074135.
- [16] N. Rajinikanth and L. Pavithra, "Heart diseases prediction for optimization based feature selection and classification using machine learning methods," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021, doi: 10.14569/ijacsa.2021.0120280.
- [17] C. A. U. Hassan *et al.*, "Effectively predicting the presence of coronary heart disease using machine learning classifiers," *Sensors*, vol. 22, no. 19, 2022, doi: 10.3390/s22197227.
- [18] D. Swain, S. K. Pani, and D. Swain, "A metaphoric investigation on prediction of heart disease using machine learning," *International Conference on Advanced Computation and Telecommunication (ICACAT 2018)*, pp. 1–6, 2018, doi: 10.1109/ICACAT.2018.8933603.
- [19] K. Yuan, L. Yang, Y. Huang, and Z. Li, "Heart disease prediction algorithm based on ensemble learning," *Proceedings - 2020 7th International Conference on Dependable Systems and Their Applications, DSA 2020*, pp. 293–298, 2020, doi: 10.1109/DSA51864.2020.00052.
- [20] T. A. Assegie, R. Subhashni, N. K. Kumar, J. P. Manivannan, P. Duraisamy, and M. F. Engidaye, "Random forest and support vector machine-based hybrid liver disease detection," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 3, pp. 1650–1656, 2022, doi: 10.11591/eei.v11i3.3787.
- [21] S. Chaitusaney and A. Yokoyama, "An appropriate distributed generation sizing considering recloser-fuse coordination," *Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference*, pp. 1–6, 2005, doi: 10.1109/TDC.2005.1546838.
- [22] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system," *World Journal of Engineering and Technology*, vol. 6, no. 4, pp. 854–873, 2018, doi: 10.4236/wjet.2018.64057.
- [23] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A method for improving prediction of human heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/1410169.
- [24] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, no. 10, pp. 685–693, 2018, doi: 10.1007/s00521-016-2604-1.
- [25] T. A. Assegie, P. K. Rangarajan, N. K. Kumar, and D. Vigneswari, "An empirical study on machine learning algorithms for heart disease prediction," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, pp. 1066–1073, 2022, doi: 10.11591/ijai.v11i3.pp1066-1073.

BIOGRAPHIES OF AUTHORS






Tsehay Admassu Assegie    received his M.Sc. in Computer Science from Andhra University, India 2016. He received his B.Sc. in Computer Science from Dilla University, Ethiopia, in 2013. He is currently a Ph.D. student in the School of Electronic and Electrical Engineering, Kyungpook National University Daegu, Republic of Korea. His research includes machine learning, the application of machine learning in healthcare, network security, and software-defined networking. His research has been published in many reputable international journals and international conferences. He is a member of the International Association of Engineers (IAENG). He has reviewed many papers published in different scientific journals. He is an active reviewer of different reputed journals. Recently, Web of Science has verified 9 peer reviews by him, published in multi-disciplinary digital publishing institute (MDPI) journals. He can be contacted at email: tsehayadmassu2006@gmail.com.






Dr. Komal Kumar Napa    is currently working as an Assistant Professor in the Department of Computer Science and Engineering (Data Science) at Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India. His research interests include machine learning, data mining, and cloud computing. He can be contacted at email: komalkumarnapa@gmail.com.






Thiyagu Thulasi    is currently working as an Assistant Professor in the Department of Computer Science & Engineering (Cyber Security) at Madanapalle Institute of Technology & Science, Madanapalle, India. His research interests include cyber security, machine learning, and deep learning. He can be contacted at email: tgiyagu.57@gmail.com.






Angati Kalyan Kumar    is currently working as an Assistant Professor in the Department of Computer Science and Engineering (Data Science) at Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India. His research interests include machine learning and data mining. He can be contacted at email: kalyankumara@mits.ac.in.



Mrs. Maran Jeyanthiran Thiruvarasu Vasantha Priya    is currently working as an Assistant Professor in the Department of Artificial Intelligence and Data Science at Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Chennai, India. Her research interests include data mining and machine learning. She can be contacted at email: vasanthapriya@velhightech.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.