IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 1, March 2024, pp. 730~736
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i1.pp730-736

Early prediction of chronic heart disease with recursive feature elimination and supervised learning techniques

Komal Kumar Napa[1], Angati Kalyan Kumar[1], Sangeetha Murugan[1], Kamaluru Mahammad[2], Tsehay Admassu Assegie[3]
[1]Department of Computer Science and Engineering, Madanapalle Institute of Technology and Science, Madanapalle, India
[2]Department of Computer and Engineering, Madanapalle Institute of Technology and Science, Madanapalle, India
[3]Department of Computer Science, College of Engineering and Technology, Injibara University, Injibara, Ethiopia

Article Info

Article history:

Received Jun 6, 2023
Revised Sep 22, 2023
Accepted Oct 2, 2023

Keywords:

Cardiovascular disease
Classification model
Feature selection
Healthcare data analytics
Machine learning

ABSTRACT

Chronic heart disease (CHD) is a common complication among patients suffering in the cardiological intensive care unit, often resulting in poor prognosis and high mortality. Early prediction of CHD can reduce mortality by preventing the severity of the disease. This study evaluated the efficacy of on recursive feature elimination for predicting CHD using supervised learning techniques for predicting CHD. The study employed 1190 Cleveland Hungarian CHD dataset. Different supervised learning techniques (support vector machine, decision tree, k-nearest neighbor, Naive Bayes, stochastic gradient descent, adaptive boosting, and multilayer perceptron) were used to study the efficacy of the recursive feature elimination. Chest pain type, sex, blood sugar level, angina, depression, and slope were associated with CHD occurrence. The accuracy of the K-nearest neighbor and decision tree model was 89.91% for the feature-selected dataset indicating good predictive ability. Ultimately, the support vector machine and logistic regression with the selected features exhibited good discriminatory ability for early prediction of CHD. Thus, the recursive feature elimination is a good approach to develop a a model with higher accuracy to predict CHD.

Corresponding Author:

Kennedy Okokpujie
Department of Electrical and Information Engineering, College of Engineering, Covenant University
Km. 10 Idiroko Road, Canaan Land, Ota, Ogun State, Nigeria
Email: kennedy.okokpujie@covenantuniversity.edu.nga

1. INTRODUCTION

Heart attack, commonly known as chronic heart disease (CHD), blocks the blood vessels in the heart and causes chest pain, and stroke [1]. A heart attack results in complications among patients suffering in the cardiological intensive care unit, often resulting in poor prognosis and high mortality. Early prediction of CHD with machine learning techniques is vital to save the life of the patient by reducing the mortality rate. However, the prediction of CHD with machine learning techniques is a critical challenge in the area of clinical data analysis [2]. Determining the optimal features for discrimination of patients from the healthy sample is one of the challenges in developing machine-learning techniques for the early prediction of CHD.

Machine learning (ML) techniques have become an important topic in the computer science field for assisting clinical decision-making [3]. ML techniques discover the pattern in the CHD dataset and try to generalize the relationship between the predictive outcome and the independent variables or features. Thus, by developing ML techniques the risk of getting CHD is predictedable at an early stage. Moreover, these techniques assist practitioners in clinical decision-making for accurate identification of CHD.

Krishnan *et al.* [4] developed a recurrent and agate hybrid neural network-based predictive model for heart dataset analysis. The discussion on the effectiveness of the developed method shows higher accuracy in predicting CHD. The effectiveness of the proposed method shows 98.69% accuracy. Even though higher overall accuracy shows promising results, the study does not show class-wise accuracy. Furthermore, the study does not suggest the effect of each CHD feature on the time complexity of fitting the proposed model on the training dataset.

Another research article by Masih *et al.* [5] discussed the multilayer perceptron-based deep neural network for predicting coronary CHD. The study suggested that pre-processing with missing value removal, and feature selection improved the performance of the proposed deep neural network by 3.36% achieving an overall accuracy of 96.50% after pre-processing for prediction of coronary CHD. Similarly, the sensitivity and specificity of the deep neural network improve after feature selection and pre-processing compared to the original dataset.

Correspondingly, another research by Assegie *et al.* [6] investigated that feature selection improves the effectiveness of the extreme boosting (XGBoost) model for matching patterns between the predictor and predicted feature in the CHD dataset. The result shows that the XGBoost model achieves 99.6% accuracy in generalizing the presence or absence of CHD. CHD predictors such as chest pain type, thallium scan, and the history of CHD are the most significant feature in discriminating the absence or presence of CHD. However, blood sugar, anginal pain, and cholesterol have less importance to the generalization of the XGBoost model. Although feature selection was the strength of the study, the performance of other supervised ML techniques is investigated.

Similarly, a study by Houssein *et al.* [7] applied a deep learning model that detects CHD risk factors. The result indicated that the proposed deep learning model scores 93.66% precision in detecting the risk factors of CHD. The result of the proposed model is promising. However, the study does not discuss the effect of features on the effectiveness of detecting the CHD.

Several studies [8]–[10] have discussed the efficacy of different ML techniques such as support vector machine (SVM), decision tree (DT), K-nearest neighbor (KNN), and artificial neural network (ANN) to predict CHD. The studies suggested that the efficacy of the ML techniques differs from study to study and the CHD dataset employed for developing the predictive ML techniques. Additionally, the study suggests that pre-processing with feature selection improves the efficiency of the ML model for detecting the CHD.

Likewise, Sarra *et al.* [11] investigated a chi-square (X2) statistical approach for improving the performance of ML techniques on CHD detection. The study employed chi-square statistics for selecting the most discriminative features of CHD. After selecting the discriminative feature of CHD, the ML techniques trained on the optimal feature set. Then the performance of the ML model compared to the original and feature-selected dataset. The result indicated that the performance of the SVM model improve by 5.26% achieving an overall accuracy of 89.47% on feature selected dataset.

Neurmious studies [12], [13] have investigated stacked ensemble learning models for predicting CHD. The stacked ensemble model is developed by employing different ML techniques such as logistic regression (LR), and the random forest (RF), DT, and KNN as base classifiers. The experimental result of the stacked ensemble model and the base ML techniques shows that the stacked model outperforms the base model with an accuracy of 88.71% for CHD prediction. Although the stacking of different base models improves the model prediction performance, the importance of feature selection to the improvement of model performance is not presented in the study.

The performance of ML techniques should be improved to obtain accurate results of the predictive outcome. The diagnosis of CHD largely depends on identifying the most important features with discriminative power between CHD positive and CHD negative class observations [14]. The study introduced the utilization of a hybrid model for improving the performance of ML techniques. The decision tree and random forest base hybrid model produces a predictive performance of 88.7% accuracy outperforming the individual decision tree and random forest model.

The use of chi-square ($\chi 2$) statistics for the elimination of the irrelevant features in the dataset helps in overcoming the overfitting and underfitting complexities of ML techniques employed for CHD diagnosis [15]. Furthermore, the performance of ML techniques improves by combining different models and producing hybrid ML models. The study of the neural network performance with an ensemble approach shows that the ensemble model achieves powerful predictive performance for CHD.

From the literature servey presented in this section, and other several studies [16]–[19], the researchers hypothesize that the important factor of CHD prediction is the risks that are associated with it. The RFE improves the predictive power of ML techniques by identifying the important features of the CHD. This article evaluates the importance of feature selection in improving the performance of different supervised ML techniques. The study also evaluates the predictive power of the ML techniques on the original and feature-selected dataset.

We discuss the contributions of this work as: i) To explore the efficacy of recursive feature selection for predicting CHD; ii) To apply the recursive feature elimination (RFE) and identify the most discriminative feature of CHD; iii) To improve the predictive accuracy by training the ML techniques on the selected features by the RFE; and iv) To discuss the supervised ML techniques on prediction of the presence or absence of CHD. The rest of the paper is structured as: i) Section 2 describes the method, the proposed approach to CHD prediction and research procedure is presented; ii) The simulation results are explained in section 3; and iii) Section 4 provides the conclusion and future scope.

2. METHOD

This section discusses the steps and research procedure followed to conduct this study. Firstly, the CHD dataset is collected from the Cleveland data repository. The Clevland CHD dataset is previously employed [20]–[24] for evaluation of the performance of machine learning in predicting CHD. Secondly, the dataset is pre-processed (at this steps the missing values are eliminated from the dataset, and the dataset is split into a training set (80%) and a testing set (20%). Secondly, different ML techniques (KNN, SVM, DT, stochastic gradient descent (SGD), adaptive booting (ADB), Naïve Bayes (NB), multilayer perceptron (MLP), and LR) model is trained on the training set and evaluated on the original dataset using accuracy, receiver operating characteristic, and fitting time of each model. Thirdly, the significant features of CHD are selected using the RFE feature selection technique. Finally, the models are trained on feature-selected datasets and their performance is measured by employing performance measures such as accuracy, receiver operating characteristic curve, fitting time of the models, and their predictive power in identifying CHD. Figure 1 indicates the flowchart for the study.

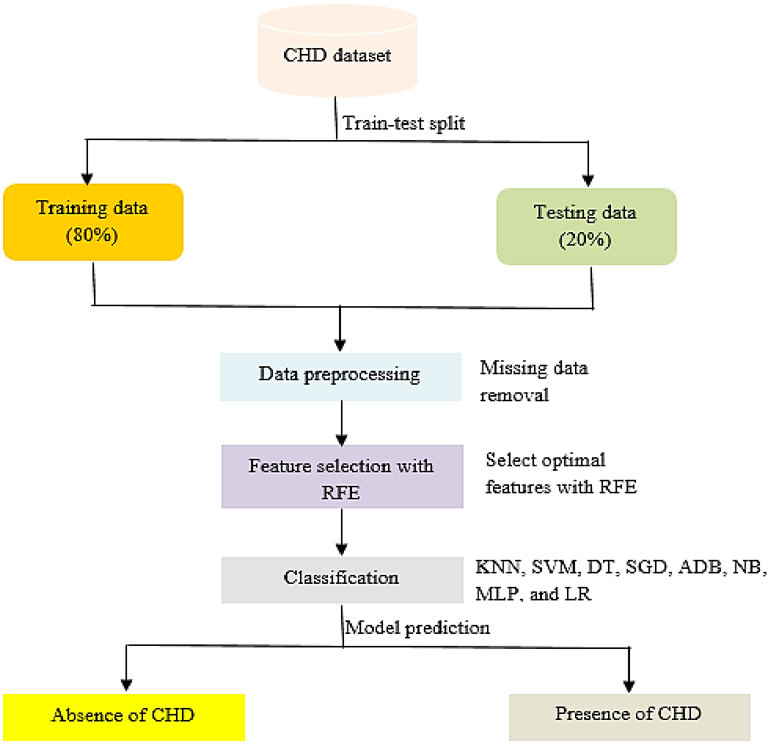
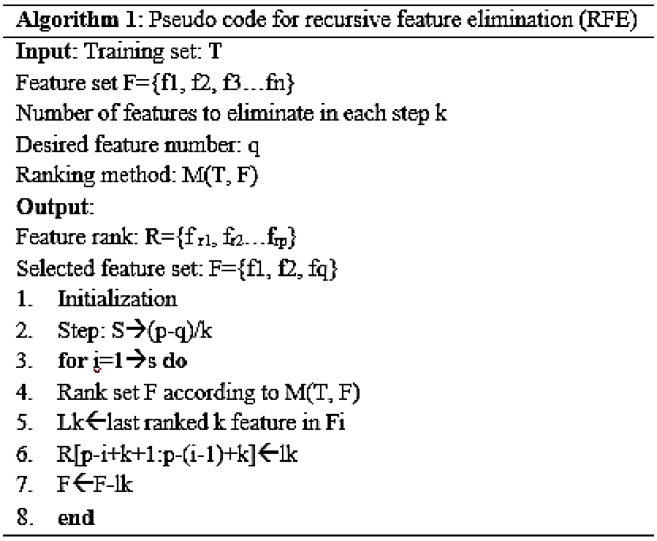


Figure 1. The flowchart of the study

2.1. Recursive feature elimination (RFE)

The recursive feature elimination improves the performance of the ML model. The RFE reduces the size of the feature for training the model. The feature elimination method reduces redundant features, which mislead model fitting and pattern identification processes during learning [25], [26]. Additionally, research articles [27], [28] investigated that the RFE improves the performance of gradient boosting, and KNN in predicting cardiovascular disease. Thus, we aimed to further investigate the effectiveness of RFE on other ML models.



```
Algorithm 1: Pseudo code for recursive feature elimination (RFE)
Input: Training set: T
Feature set F={f1, f2, f3...fn}
Number of features to eliminate in each step k
Desired feature number: q
Ranking method: M(T, F)
Output:
Feature rank: R={fr1, fr2...frp}
Selected feature set: F={f1, f2, fq}
Initialization
Step: S→(p-q)/k
for i=1→s do
    Rank set F according to M(T, F)
    Lk←last ranked k feature in Fi
    R[p-i+k+1:p-(i-1)+k]←lk
    F←F-lk
end
```

3. RESULTS AND DISCUSSION

This section presents the predictive power of different supervised ML techniques for detecting the absence or presence of CHD. The comparative analysis employed accuracy, the area under the receiver operating characteristics curve (ROC-AUC), and fitting time in comparing different models.

3.1. The performance of ML techniques

The performance of ML learning techniques is measured using an accuracy metric on the testing dataset. The performance of each ML model is evaluated on the original and the feature-selected dataset. Some of the models such as LR and SVM appear to improve with the feature-selected dataset. However, most of the models decrease in accuracy as demonstrated in Table 1.

Figure 2 shows the accuracies of each ML model on the original and feature-selected dataset. The MLP achieves the highest accuracy (93.69%) in predicting the presence of CHD on the original dataset. However, the accuracy of the MLP decreased on the feature-selected dataset having an accuracy value=87.39%. In contrast, the SVM and LR models scored higher accuracy on the feature-selected dataset than the original dataset. The DT and KNN model achieves the highest accuracy (89.91%) compared to other models on the feature-selected dataset, indicated in Figure 2.

Table 1. The performance of supervised ML techniues

ML model	Original dataset	Feature-selected dataset	ML model	Original dataset	Feature-selected dataset
LR	87.39%	88.65%	SGD	84.45%	81.09%
SVM	87.39%	88.65%	DT	93.69%	89.91%
NB	86.97%	87.39%	ADB	89.49%	86.55%
KNN	89.91%	89.91%	MLP	93.69%	87.39%

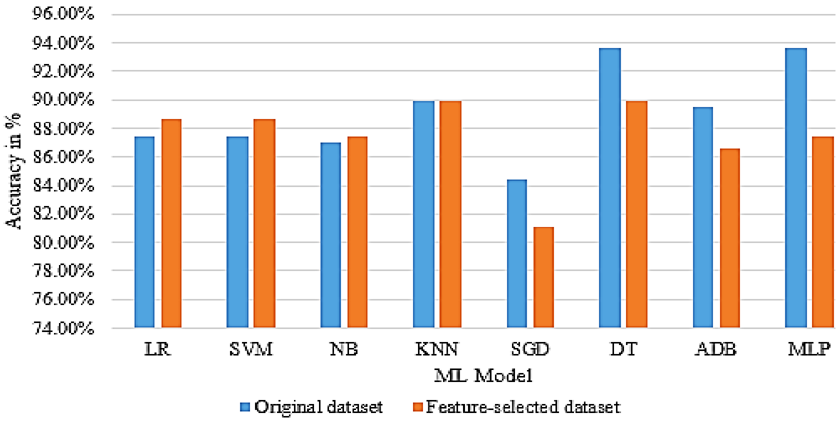


Figure 2. The perfromance of ML model on orginal and feature selected dataset

3.2. RFE and the fitting time of ML techniques

The fitting time complexity of the supervised ML model on the original and feature-selected dataset demonstrated the X model has a faster fitting time compared to the other models. Figure 3 indicates the fitting time complexity of the supervised model on the original and feature-selected dataset. As shown in Figure 3, the fitting time of LR, SVM, NB, KNN, SGD, DT, and ADB have lower fitting time. In contrast, the fitting time of the MLP has higher fitting time on the feature-selected dataset compared to the original dataset.

In addition to the fitting time demonstrated in Figure 3, the supervised ML model evaluation employed the receiver operating characteristics area (ROC-AUC). Figure 4 demonstrates the ROC-AUC for each of the supervised ML models. As revealed in Figure 4, recursive feature elimination improved the area under curve of NB, and KNN on predicting CHD. In contrast, the area under curve of LR, SVM, DT, and SGD remained roughly similar on the original and feature-selected dataset. The ADB and MLP models have a lower area under the curve for the feature-selected dataset than the original dataset.

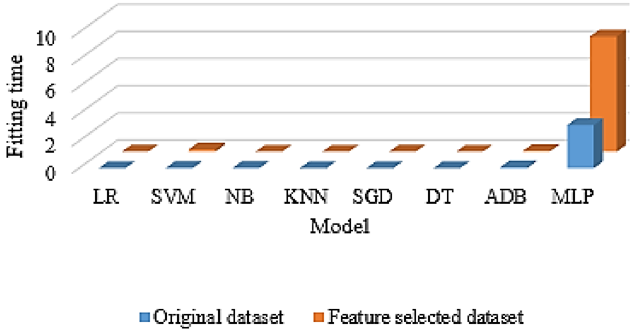


Figure 3. The cross-validation fitting time of the ML model on the original and feature-selected dataset

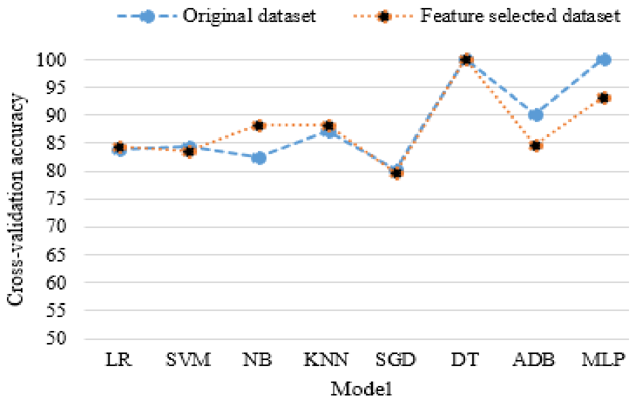


Figure 4. The ROC-AUC score of the ML model on the original and feature-selected dataset

4. CONCLUSION

This study investigated the efficiency of the recursive feature elimination (RFE) for selecting CHD features for predicting the presence of the disease using different ML models. The result indicated that RFE is important to select relevant features, and reducing training time. The RFE is effective at selecting those features in a training dataset that are more or most relevant in predicting the CHD. RFE preserves the feature importance of feature-selected data and is quite the same as the original data based on the observation above. With the feature-selected dataset, we obtained predictive accuracy of 89.91 % with the KNN and DT models. Thus, it implies that the ML models can be used in clinical decision-making and CHD risk analytics. The major contribution of this study is that it investigated how RFE influences the predictive power and the time complexity for fitting the ML models on datasets with different dimensions. However, the limitation of this study is that the ML models are not tested on various datasets. In future work, it is recommended to test the RFE on different datasets to validate the findings of this study.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Injibara University for internet and other facilities required to conduct this study.

REFERENCES

[1] 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, Sep. 2022, doi: 10.11591/ijai.v11.i3.pp1066-1073.

[2] A. Menshawi, M. M. Hassan, N. Allheeib, and G. Fortino, "A hybrid generic framework for heart problem diagnosis based on a machine learning paradigm," *Sensors*, vol. 23, no. 3, p. 1392, Jan. 2023, doi: 10.3390/s23031392.

[3] G. N. Ahamad *et al.*, "Influence of optimal hyperparameters on the performance of machine learning algorithms for predicting heart disease," *Processes*, vol. 11, no. 3, p. 734, Mar. 2023, doi: 10.3390/pr11030734.

[4] S. Krishnan, P. Magalingam, and R. Ibrahim, "Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5467–5476, 2021, doi: 10.11591/ijece.v11i6.pp5467-5476.

[5] N. Masih, H. Naz, and S. Ahuja, "Multilayer perceptron based deep neural network for early detection of coronary heart disease," *Health and Technology*, vol. 11, no. 1, pp. 127–138, Nov. 2021, doi: 10.1007/s12553-020-00509-3.

[6] 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, Jun. 2023, doi: 10.11591/ijece.v13i3.pp3359-3366.

[7] E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques," *Scientific Reports*, vol. 13, no. 1, May 2023, doi: 10.1038/s41598-023-34294-6.

[8] P. C. Bizimana, Z. Zhang, M. Asim, and A. A. Abd El-Latif, "An effective machine learning-based model for an early heart disease prediction," *BioMed Research International*, vol. 2023, pp. 1–11, Apr. 2023, doi: 10.1155/2023/3531420.

[9] K. Karthick, S. K. Aruna, R. Samikannu, R. Kuppusamy, Y. Teekaraman, and A. R. Thelkar, "Implementation of a heart disease risk prediction model using machine learning," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–14, May 2022, doi: 10.1155/2022/6517716.

[10] S. Haseena, S. K. Priya, S. Saroja, R. Madavan, M. Muhibbullah, and U. Subramaniam, "Moth-flame optimization for early prediction of heart diseases," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–10, Sep. 2022, doi: 10.1155/2022/9178302.

[11] R. R. Sarra, A. M. Dinar, M. A. Mohammed, and K. H. Abdulkareem, "Enhanced heart disease prediction based on machine learning and $\chi 2$ statistical optimal feature selection model," *Designs*, vol. 6, no. 5, p. 87, Sep. 2022, doi: 10.3390/designs6050087.

[12] A. Khan, A. Khan, M. M. Khan, K. Farid, M. M. Alam, and M. B. M. Su'ud, "Cardiovascular and diabetes diseases classification using ensemble stacking classifiers with SVM as a meta classifier," *Diagnostics*, vol. 12, no. 11, p. 2595, Oct. 2022, doi: 10.3390/diagnostics12112595.

[13] Gyanendra Kumar Pal and S. Gangwar, "Heart disease prediction by stacking ensemble models on multiple classifiers by applying feature selection methods," *Journal of Theoretical and Applied Information Technology*, vol. 100, p. 23, Jun. 2022.

[14] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in *Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021*, Jan. 2021, pp. 1329–1333, doi: 10.1109/ICICT50816.2021.9358597.

[15] P. Priyanga, V. V. Pattankar, and S. Sridevi, "A hybrid recurrent neural network-logistic chaos-based whale optimization framework for heart disease prediction with electronic health records," *Computational Intelligence*, vol. 37, no. 1, pp. 315–343, Oct. 2021, doi: 10.1111/coin.12405.

[16] P. Verma, V. K. Awasthi, and S. K. Sahu, "A novel design of classification of coronary artery disease using deep learning and data mining algorithms," *Revue d'Intelligence Artificielle*, vol. 35, no. 3, pp. 209–215, Jun. 2021, doi: 10.18280/ria.350304.

[17] E. Dritsas and M. Trigka, "Efficient data-driven machine learning models for cardiovascular diseases risk prediction," *Sensors*, vol. 23, no. 3, p. 1161, Jan. 2023, doi: 10.3390/s23031161.

[18] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," in *Proceedings - IEEE Symposium on Computers and Communications*, Jul. 2017, pp. 204–207, doi: 10.1109/ISCC.2017.8024530.

[19] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics in Medicine Unlocked*, vol. 26, p. 100655, 2021, doi: 10.1016/j.imu.2021.100655.

[20] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Analytics*, vol. 3, p. 100130, Nov. 2023, doi: 10.1016/j.health.2022.100130.

[21] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics*, vol. 2, p. 100060, Nov. 2022, doi: 10.1016/j.health.2022.100060.

[22] S. M. S. Shah, F. A. Shah, S. A. Hussain, and S. Batool, "Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods," *Computers and Electrical Engineering*, vol. 84, p. 106628, Jun. 2020, doi: 10.1016/j.compeleceng.2020.106628.

[23] H. Lin, Y. Xue, K. Chen, S. Zhong, and L. Chen, "Acute coronary syndrome risk prediction based on gradient boosted tree feature selection and recursive feature elimination: A dataset-specific modeling study," *PLoS ONE*, vol. 17, no. 11 November, p. e0278217, Nov. 2022, doi: 10.1371/journal.pone.0278217.

[24] X. Lin, X. Zhang, and X. Xu, "Efficient classification of hot spots and hub protein interfaces by recursive feature elimination and gradient boosting," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 5, pp. 1525–1534, Sep. 2020, doi: 10.1109/TCBB.2019.2931717.

[25] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," *International Journal of Pharmaceutical Research*, vol. 12, no. 4, pp. 56–66, 2020, doi: 10.31838/ijpr/2020.12.04.013.

[26] E. M. Senan *et al.*, "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–10, Jun. 2021, doi: 10.1155/2021/1004767.

[27] P. Theerthagiri, "Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique," *Intelligent Systems with Applications*, vol. 16, p. 200121, Nov. 2022, doi: 10.1016/j.iswa.2022.200121.

[28] V. P. C. Magboo and M. S. A. Magboo, "Cardiovascular disease prediction with imputation techniques and recursive feature

elimination," in *AIP Conference Proceedings*, 2023, vol. 2602, doi: 10.1063/5.0124079.

BIOGRAPHIES OF AUTHORS

Komal Kumar Napa 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.

Angati Kalyan Kumar is currently working as 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.

Sangeetha Murugan is currently working as Assistant Professor at the Department of Computer Science and Engineering at Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India. Her research interests include machine learning and data mining. She can be contacted at email: sangee525@gmail.com.

Kamaluru Mahammad is currently working as Assistant Professor in the Department of Computer Science and Engineering (artificial intelligence) at Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India. His research interests include data mining, computer networks and IoT. He can be contacted at email: mahammadk@mits.ac.in.

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 working as a lecturer at the Department of Computer Science, College of Engineering and Technology, Injibara University, Injibara, Ethiopia. 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 8 peer reviews by him, published in multi-disciplinary digital publishing institute (MDPI) journals. He can be contacted at email: tsehayadmassu2006@gmail.com.