Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 6, December 2021, pp. 3501~3506
ISSN: 2302-9285, DOI: 10.11591/eei.v10i6.3001

An improved feature selection approach for chronic heart disease detection

S. J. Sushma[1], Tsehay Admassu Assegie[2], D. C. Vinutha [3], S. Padmashree[4]
[1,4]GSSS Institute of Engineering and Technology for Women, Visvesvaraya Technological University, Belgaum, Karnataka, India
[2]Department of Computer Science, College of Natural and Computational Science, Injibara University, Injibara, Ethiopia
[3]Vidyavardhaka College of Engineering, Mysuru Visvesvaraya Technological University, Belgaum, Karnataka, India

Article Info

Article history:

Received Mar 25, 2021
Revised Sep 5, 2021
Accepted Oct 16, 2021

Keywords:

Chronic heart disease
Feature selection
Heart disease diagnosis
Model optimization
Random forest model

ABSTRACT

Irrelevant feature in heart disease dataset affects the performance of binary classification model. Consequently, eliminating irrelevant and redundant feature (s) from training set with feature selection algorithm significantly improves the performance of classification model on heart disease detection. Sequential feature selection (SFS) is successful algorithm to improve the performance of classification model on heart disease detection and reduces the computational time complexity. In this study, sequential feature selection (SFS) algorithm is implemented for improving the classifier performance on heart disease detection by removing irrelevant features and training a model on optimal features. Furthermore, exhaustive and permutation based feature selection algorithm are implemented and compared with SFS algorithm. The implemented and existing feature selection algorithms are evaluated using real world Pima Indian heart disease dataset and result appears to prove that the SFS algorithm outperforms as compared to exhaustive and permutation based feature selection algorithm. Overall, the result looks promising and more effective heart disease detection model is developed with accuracy of 99.3%.

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 diagnosis with supervised machine learning model involves identifying the class of a given observation based on previous experience without explicitly programmed [1]. The input features represent the class to be determined. However, the complete set of input features in the medical dataset does not determine the class of an observation. Irrelevant input features tend to mislead the machine-learning algorithm and result in low performance on heart disease classification. Thus, sequential feature selection (SFS) algorithm is a novel approach to input feature selection to improve the performance and computational time complexity for classification problem involving disease diagnosis using medical diagnostic datasets. Moreover, eliminating irrelevant and less important features from the medical dataset tends to improve the precision and classification accuracy of machine learning models in heart disease diagnosis. The goal of sequential feature selection (SFS) algorithm is to ensure that the best possible subset of features are used for training a model on medical dataset for classification ultimately improving the precision and classification accuracy of a model on the classification task.

Irrelevant features in a real world medical dataset such as heart disease dataset suggest strong correlations between features and the target class label arising by chance and strong correlation of between

features tends to deteriorate the classification accuracy of model [2]-[4]. Moreover, large number of features in a dataset significantly increases computational time complexity without corresponding model performance improvement. Consequently, training classification model with smaller and optimal feature subset tends to improve classification performance. Thus, we have proposed an efficient sequential feature selection algorithm for selecting the relevant and more important feature subset among the larger input feature in a real world dataset for improving the performance of machine learning model for classification task. Dealing with a large number of features brings us to reduction of the dimensionality of dataset features [2]. More features in training set tend to make models more complex to learn and difficult to interpret the classification performance. In addition to model complexity, more features tend to lead model overfitting.

In this study, we primarily focused on heart disease classification model optimization with input feature selection. The problem of supervised learning is to approximate the functional relationship f() between an input $X=\{X_1, X_2, \ldots, X_N\}$ and output Y called the class label on a memory of data points $\{X_i, Y_{i,}\}$ for $i=\{1,\ldots,N\}$ where X_i is input vector and d $Y_{i,1}$ is a real number. However, some of the input features are irrelevant in medical diagnosis datasets. For example, patient ID is not relevant for machine learning model. Moreover, using all input features requires sufficient time and irrelevant features introduce overhead time complexity and result in lower classification accuracy. Thus, this study, introduces sequential feature selection algorithm for optimizing heart disease classification accuracy of random forest model. We implemented exhaustive, correlation and permutation based features selection algorithm to compare with the proposed feature selection algorithm by employing real world Pima Indian heart disease dataset data repository for testing the classifier model performance.

2. LITREATURE REBVIEW

Numerous research works have shown that large number of feature have impact on performance of supervised machine learning model for classification. Recent literature review by [5]-[7] summarized the current optimization approaches employed for improving the performance of supervised machine-learning algorithms on medical dataset classification tasks. In the study, the authors suggested that the application of model optimization methods such as parameter tuning, correlation based feature selection and dimensionality reduction with principal component analysis (PCA) have significant importance for improving the performance of supervised machine learning model on medical dataset classification task. Moreover, irrelevant input features induce extra computational cost such as processor time and memory space. Moreover, irrelevant feature lead to model overfitting, where the learning model performs good on training set as compared to the test set. Thus, irrelevant feature not only incur additional computational cost, but also mislead the model and ultimately results in low performance on disease classification.

The researchers studied the effect of high dimensional dataset on the performance of supervised classification model [8], [9]. The authors proposed an information gain based (IFG) feature selection algorithm for reducing a high dimensional input feature for improving classification performance of Naïve Bayes classification algorithm on text data classification. Moreover, the authors carried out an extensive experiment test on the classification performance of the proposed model and the experimental result appears to prove that information gain based feature selection improved the classification performance of Naïve Bayes model on text dataset classification. Moreover, the computational time and storage space required for training and testing the proposed text classification model is lower as compared to the complete high dimensional input feature. In heart disease diagnosis, our purpose is to infer the relationship between the symptoms, input features and their corresponding class label (heart disease positive or heart disease negative). If mistakenly one feature is included in model training, this the learning model comes to false conclusion due to mistakenly included feature for training.

Feature selection significantly reduces the computational costs such as computational time complexity and memory space requirement. A review on the effect of input features in [10]-[12] suggested that feature selection algorithms are categorized into two categories namely, filter and wrapper approach. The filter approach assigns weight to each feature subset and based on the weight value assigned to each feature subset, the optimal feature set is selected for training a classification model. The weight is determined based on distance, statistical method such as correlation between a given feature set and the target or class label. The features with higher weight are selected as optimal feature set and the classification model is trained on the selected feature set. The motivation for feature selection in medical dataset classification is that, since the goal of medical diagnosis model is to approximate the underlying relationship between the input features or symptoms and the class label, ignoring those input features with little effect on the class label leads to better performance.

A current literature review in [13], [14], related to the dimensionality reduction problem in medical dataset classification showed that the issue of irrelevant feature on the accuracy of classification model is still an ongoing and open research issue. Developing classification model with higher accuracy on medical

dataset is one of the major concern of automated medical diagnosis systems. In [15]-[18], the researchers further investigated the performance of machine learning model with various feature selection methods. The methods employed for feature selection include the PCA, statistical method such as chi-squared test for selecting relevant features in a medical dataset. The PCA is employed for reducing the dimensionality of the feature before training model on a medical dataset for classification. The authors carried out extensive experiment, experimental result appears to prove that more accurate, and effective heart disease classification model is achieved with feature selection. Overall, the classification accuracy 85% is achieved when feature selection is applied on heart disease dataset.

Dimensionality reduction is the most important and popular approach for noise reduction (removing irrelevant features) and redundant features [19]-[25]. Moreover, input feature extraction method such as PCA reduces the dimensionality of the original or the complete input feature by projecting the original input feature space into a new constructed feature space, preserving the combinations in original feature space. Thus, principal component analysis is important to visualize a high dimensional dataset and investigate the relationship among input feature subset. However, finding overall, optimal input feature subset incurs additional computational overhead, because we employ exhaustive search to find the optimal feature subset. The accuracy and efficiency of the feature selection process depends on the type of feature selection algorithm employed for searching optimal feature subset. Overall, feature selection algorithm leads to great accuracy in classification but rarely applied to medical diagnosis and optimal feature selection may incur additional time complexity. However, in medical diagnosis, accuracy is more important than efficiency.

In this study, the existing feature selection algorithms, namely exhaustive, permutation and correlational-based feature selection algorithms are critically reviewed. Moreover, we have proposed sequential feature selection (SFS) algorithm to improve the performance of random forest model for heart disease classification. In addition to that, we have implemented random forest model for heart disease detection to evaluate the efficiency and accuracy of the proposed approach. For comparison with the existing feature selection algorithm we have implemented exhaustive, permutation and correlation based feature selection well known algorithms for feature selection. An extensive experiment is conducted on the proposed approach and the existing methods for feature selection. In the experiment, we have used 5-fold cross validation to test the efficiency and accuracy of the proposed and existing approaches. Result appears to prove that sequential feature selection is better in terms of efficiency and the exhaustive feature selection is better in terms of accuracy but computationally expensive.

3. METHODS AND MATERIALS

In this study, we have conducted experiment and comparisons among feature selection algorithms. At the end, we have suggested an efficient sequential feature selection method for selecting optimal feature subset. To conduct our study, we followed the flowing procedure. First, we have conducted a preliminary review of previous related work in section 2 and then we have collected dataset. Finally, we have conducted an experiment using the collected dataset with different feature selection algorithms discussed in previous section. The data is collected from Kaggle and contains 1025 observations. In Kaggle dataset, each observation belongs to either the heart disease patient (positive class) or not patient class (negative class). Hence, the problem of heart disease detection is binary classification problem where the class of a particular observation belongs to the positive or negative class.

3.1. Dataset characteristics

The descriptive statistics, the mean, standard deviation, maximum and minimum values for the numeric features in the dataset is summarized in Table 1.

3.2. Sequential feature selection

Sequential feature selection algorithm is used to reduce an initial *d*-dimensional feature subset to a *k*-dimensional feature set for $k<d$. The motivation behind sequential feature selection algorithms that automatically selects feature subset that is most relevant to the problem. The goal of feature selection is to improve the computational efficiency and reduce the classification error of predictive model by removing irrelevant features or noise from a dataset. Sequential feature selection algorithm removes or adds a feature at a time based on the relevance of the feature to the classifier performance. Let us consider the (min_f, max_f) is a tuple representing the minimum and maximum feature in the range min_f to max_f in the feature set. The best feature combination that produces optimal performance for a classification model is obtained by iteratively testing the performance of the classification model on feature subset on 1 to max_f (forward) or max_f to min_f (backward). The size of the returned feature subset within max_f to min_f depending on which combination scored higher classification accuracy during cross validation is selected as best combination of features.

Table 1. Descriptive statistics for high risk factor in heart disease dataset

Feature	Mean	Std.Dev	Min	max
Age	54.43	9.07	29.00	77.00
Cholesterol	246.00	51.59	126.0	546.00
Exercise induced angina	0.59	0.52	0.00	2.00
Maximum heart rate achieved	149.11	23.00	71.00	202.00
Exercise induced angina	0.33	0.47	0.00	1.00
Resting elector cardio graphic	1.07	1.17	0.00	9.20
Cardio vascular disease	0.75	1.03	0.00	4.00
Thalassemia	2.32	0.62	0.00	3.00
Fasting blood sugar	0.14	0.35	0.00	1.00
Total resting blood pressure	131.61	17.51	94.00	200.00
Slope	1.38	0.61	0.00	2.00

Algorithm 1. Sequential feature selection

Input: *$Y=\{y_1, y_2, ...y_d\}$, original d-dimensional feature set*
Output: *$X_k = \{x_j | j=1,2,3...,Ik, x_j \in Y\}$, where $k=(0,1,2.....d\}$*
Begin *:*

1. *$X_0=0=k$*
2. *$X^a=arg\ max\ J(X_k+X)$ where $X \in X_k$*
3. *6. $X_k+1=X_k+X^a$*
4. *7. $k=k+1$*
5. *Go to Step 3*

Stop if k=p, where p is a priori defined number of optimal features to be selected

4. RESULTS AND DISCUSSION

The features selecting by sequential feature selection algorithm among the 13 heart disease dataset features in the heart disease dataset are the following: The better combination of features selected by the proposed approach are as: Best combination (highest accuracy achieved: 0.971): (0, 1, 2, 4, 6, 7, 8, 9, 11, 12).

As demonstrated in the output, the classification accuracy achieved by the random forest model with the sequential feature selection algorithm is 97.1% on the heart disease detection. The features used for training the model are discussed in Table 2. The most representative features among the 13 features in the original dataset selected by the proposed method are discussed as follows. indexes, index 0 age feature index 1 sex, index 2 chest pain, index 4 cholesterol, index 6 exercise induced angina compared to rest, index 7 maximum heart rate achieved, index 8 exercise induced angina, index 9 old peak, index 11 cardio vascular disease and index 12 thalassemia. Highest classification accuracy achieved with the selected feature using the proposed sequential feature selection algorithm is 97.1% as demonstrated in Figure 1.

Table 2. Optimal features selected by sequential feature selection algorithm

No	Feature Index	Selected feature name
1	0	Age
2	1	Sex
3	2	Chest pain
4	4	Cholesterol
5	6	Exercise induced angina relative to rest
6	7	Maximum heart rate achieved
7	8	Exercise induced angina
8	9	Resting elector cardio graphic
9	11	Cardio vascular disease

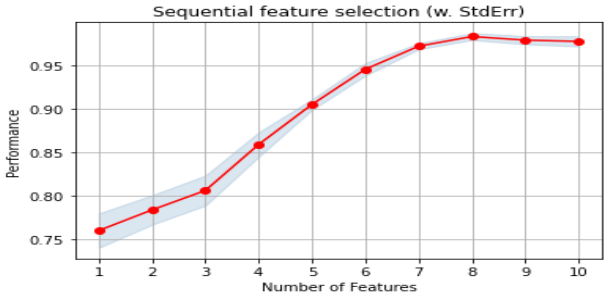

Figure 1. Effects of selecting different switching under dynamic condition

Figure 1 demonstrates the performance of random forest model using sequential feature selection. As illustrated in Figure 1, 10 features among the 13 features, characterizing heart disease dataset are selected as optimal feature subset by sequential feature selection as shown in Table 1. We observe from Figure 1 that the highest accuracy is archived with 10 features. As demonstrated in Figure 1, the performance the random forest model improves with an increase in feature space. However, increasing irrelevant feature space will degrade the performance of the classification model. As shown in Figure 1, an increase in only selected features with sequential feature selection have positive effect on the performance of random forest model.

4.1. Comparison of the proposed and existing feature selection algorithms

Our first experiment demonstrated the sequential feature selection algorithm. The second experiment-demonstrated permutation based feature selection algorithm and the third experiment is conducted on the exhaustive feature selection algorithm. Experimental results on the three feature selection algorithm appears to prove that the exhaustive feature selection algorithm is time consuming, computationally costly but performs well on selecting optimal features set. The sequential feature selection algorithm has lower computational overhead as compared to the exhaustive feature selection algorithm. The optimal feature selected by sequential, exhaustive, permutation and correlation based feature selection is summarized in Table 3.

Table 3. Optimal features selected by sequential feature selection algorithm

Feature selection algorithm	Time cost	Selected features	Performance
Sequential	146 Sec	(0, 1, 2, 4, 6, 7, 8, 9, 11, 12)	97.1%
Permutation based	100 Sec	(0, 1, 2, 5, 6, 8, 9, 10, 11, 12)	94.3%
Correlation based	76 Sec	(1, 2, 3, 4, 6, 7, 8, 9, 11, 12)	94.1%
Exhaustive	349 Sec	(0, 1, 2, 6, 7, 8, 9, 11, 12)	94%

As we observe in Table 3, the permutation based feature selection algorithm removes irrelevant features as compared to the sequential feature selection and exhaustive feature selection algorithm. However, the model does not perform well as compared to the sequential and exhaustive feature selection algorithm. We also realize that different feature selection algorithms find different features set as optimal features. To measure the goodness of these algorithms we evaluate the mean fold score with different feature selection algorithms, namely the sequential, exhaustive, correlation and permutation based feature selection. The performance is higher for sequential feature selection method as compared to exhaustive, permutation and correlation based feature selection method.

5. CONCLUSION

In this study, we have proposed sequential feature selection algorithm for feature selection. Moreover, we have studied different types of feature selection algorithms along with their practical 23 implementations. The goal of the proposed sequential feature selection algorithm is to improve model performance for machine learning classifier. We have implemented a number of feature selection algorithms such as permutation based, exhaustive and correlation based feature selection method and compared with the proposed algorithm. We employed random forest model to test the proposed algorithm on heart disease classification. We employed real world Pima Indian heart disease dataset for evaluating the performance of the proposed and exiting feature selection methods. The exhaustive features selection algorithm produces better performance result. However, the computational time is higher for the exhaustive feature selection algorithm as compared with sequential feature selection algorithm. Exhaustive algorithm has advantage of fitting to specific machine learning algorithm. The sequential feature selection algorithm is preferred for large datasets. Moreover, the computational time complexity for sequential feature selection algorithm is less as compared to the exhaustive feature selection algorithm. We have conducted experiment on random forest model for testing the performance of selected feature subset.

As a future work, we recommend the researchers to extend this work by using different feature selection algorithms such as filter based feature selection algorithm and compare performance results with the current work to optimize the model to more robust and effective level. Moreover, we recommend researchers to conduct empirical study with the existing feature selection algorithms with different high dimensional high input space datasets such as text dataset.

ACKNOWLEDGEMENTS

I would like to express my special thanks of gratitude to my wife Aster Belay who helped me a lot in typesetting and writing this manuscript.

REFERENCES

[1] G. Nguyen *et al.*, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artif Intell Rev.*, vol. 52, no. 1, pp. 77-124, 2019, doi: https://doi.org/10.1007/s10462-018-09679-z.

[2] Z. M. Yusof, Md. M. Billah, K. Kadir, N. Armanina, N. Hidayah and H. Nasir, "Design and fabrication of cost-effective heart rate pulse monitoring sensor system," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 17, no. 5, pp. 2497-2504, 2019, doi: 10.12928/TELKOMNIKA.v17i5.9926.

[3] T. A. Assegie, S. J. Sushma, B. G. Bhava and S. Padmashree, "Correlation Analysis for Determining Effective Data in Machine Learning: Detection of Heart Failure," *SN Computer Science*, vol. 2, no. 213, 2021, doi: https://doi.org/10.1007/s42979-021-00617-5.

[4] 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*, vol. 9, no. 4, pp. 2675-2681, 2019, doi: 10.11591/ijece.v9i4.pp2675-2681.

[5] 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.

[6] T. A. Assegie, R. L. Tulasi and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *International Journal of Artificial Intelligence*, vol. 10, no. 1, pp. 184-190, 2021, doi: 10.11591/ijai.v10.i1.pp184-190.

[7] X. Gao, A. A. Ali, H. S. Hassan and E. M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method," *Hindawi Complexity*, pp.1-10, 2021, doi: 10.1155/2021/6663455.

[8] A. Yazdani, K. D. Varathan, Y. K. Chiam, A. W. Malik and W. A. W. Ahmad, "A novel approach for heart disease prediction using strength scores with significant predictors," *BMC Med Inform Deci Mak*, vol. 21, no. 194, pp. 2-16, 2021, doi: 10.1186/s12911-021-01527-5.

[9] H. Meshref, "Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, pp. 258-269 2019, doi: 10.14569/IJACSA.2019.0101236.

[10] T. A. Assegie and P. S. Nair, "The Performance of Different Machine Learning Models on Diabetes Prediction," *International journal of scientific & technology research*, vol. 9, no. 1, pp. 2491-2494, 2020.

[11] A. W. Sugiyarto, A. M. Abadi and Sumarna, "Classification of heart disease based on PCG signal using CNN," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 19, no. 5, pp. 1697~1706, 2021, doi: 10.12928/TELKOMNIKA.v19i5.20486.

[12] S. Budiyanto *et al*., "Design and monitoring body temperature and heart rate in humans based on WSN using star topology," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 1, pp. 326-334, 2021, doi: 10.11591/ijeecs.v22.i1.pp326-334.

[13] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Hindawi Computational Intelligence and Neuroscience*, vol. 2021, 2021, doi:10.1155/2021/8387680.

[14] A. Alaiad, H. Najadat, B. Mohsen and K. Balhef, "Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease," *Journal of Information & Knowledge Management*, vol. 19, no.1, 2020, doi: 10.1142/S0219649220400158.

[15] Y. T. Lo, H. Fujita and T. W. Pai, "Prediction of Coronary Artery Disease Based on Ensemble Learning Approaches and Co-Expressed Observations," *Journal of Mechanics in Medicine and Biology*, vol. 16, no. 1, 2016, doi: 10.1142/S0219519416400108.

[16] G. T. Reddy and N. Khare, "An Efficient System for Heart Disease Prediction using Hybrid OFBAT with Rule-Based Fuzzy Logic Model," *Journal of Circuits, Systems and Computers*, vol. 26, no. 4, 2017, doi: 10.1142/S021812661750061X.

[17] S. Sreejith, S. Rahul and R. C. Jisha, "A Real Time Patient Monitoring System for Heart Disease Prediction Using Random Forest Algorithm," *Advances in Signal Processing and Intelligent Recognition Systems*, pp. 485-500, 2016, doi: 10.1007/978-3-319-28658-7_41.

[18] R. Chadha, S. Mayank, A. Vardhan and T. Pradhan, "Application of Data Mining Techniques on Heart Disease Prediction: A Survey," *Emerging Research in Computing, Information, Communication and Applications*, pp. 413-426, 2016, doi: 10.1007/978-81-322-2553-9_38.

[19] T. N. Nguyen, T. H. Nguyen and V. T. Ngo, "Artifact elimination in ECG signal using wavelet transform," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 18, no. 2, pp. 936-944, 2020, doi: 10.12928/TELKOMNIKA.v18i2.14403.

[20] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir and R. Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," *Hindawi Mobile Information Systems*, vol. 2018, 2018, doi: 10.1155/2018/3860146.

[21] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[22] T. A. Assegie and P. S. Nair, "Handwritten digits recognition with decision tree classification: a machine learning approach," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 5, pp. 4446-4451, 2019, doi: 10.11591/ijece.v9i5.pp4446-4451.

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

[24] I. Tougui, A. Jilbab and J. E. Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," *Health and Technology*, vol. 10, no. 284, 2020, doi:10.1007/s12553-020-00438-1.

[25] J. N. Khiarak, M. R. F. Derakhshi, K. Behrouzi, S. Mazaheri, Y. Z. Harghalani and R. M. Tayebi, "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection," *Health and Technology*, vol. 10, no. 1, 2020, doi: 10.1007/s12553-019-00396-3.