DOI: 10.24996/ijs.2022.63.9.26

Iraqi Journal of Science

ISSN: 0067-2904

Evaluation of Feature Ranking for Optimization of Random Forest Model for Heart Disease Diagnosis

Tsehay Admassu Assegie[1]*, Dr. Tamilarasi[2], N.Komal Kumar[3]

[1]*Department of Computer Science, Faculty of Computing Technology, Aksum Institute of technology, Aksum University, Axum, Ethiopia*

[2]*Department, of Information Technology, St. Peter's Institute of Higher Education and Research, Chennai, India*

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

Received: 17/4/2021 Accepted: 2/8/2021 Published: 30/9/2022

Abstract

Diagnosing heart disease is one of the most challenging tasks that require highly experienced cardiologists. However, in developing nations such as Ethiopia, there are few cardiologists, and heart disease diagnosis is challenging. As an alternative solution to cardiologists, this study proposed a more effective model for heart disease detection by employing random forest and feature ranking (FR). FR improves the performance of the random forest model on heart disease diagnosis. FR ranks features based on their relative importance to the heart disease diagnosis improving the performance of the random forest model on heart disease diagnosis. This study investigates the effectiveness of the FR approach for feature selection on a real-world clinical heart disease diagnosis dataset collected from the University of California at Irvine (UCI) data repository. The simulation on the validation set reveals that the FR is an effective approach for feature selection. The result shows the random forest model as an effective method for heart disease diagnosis giving 98.53% accuracy.

Keywords: Predictive analytics, predictive modeling, machine learning, automated diagnosis.

1. Introduction

Random forest classifier has become a core technology in the health care industry, largely due to the promising results and higher precision and cost-effective way of solving complex problems such as heart disease diagnosis [1]. Artificial intelligence methods are widely applied in the health care industry. However, optimization of the performance of decision support needs further study as medical decisions are highly risky and the cost of false negatives is higher. One of the major problems that affect the performance of a decision support model for heart disease diagnosis is the quality of the input features employed in training the model.

The feature ranking (FR) improves the performance of the automated model involving disease diagnosis for the heart disease dataset. Removing unrelated and unimportant features of the heart disease dataset improves the precision and diagnosis accuracy of a classification model. The goal of the FR is to rank input features used for training the heart disease diagnosis

*Email: tsehayadmassu2006@gmail.com

model, ultimately improving the precision and accuracy of the model for heart disease diagnosis.

Unrelated and redundant features in a real-world heart disease dataset propose relationships between irrelevant features and the target class rising by chance, and a strong relationship between irrelevant features and the target class tends to decrease the classification accuracy of a decision support model [2]. Furthermore, a larger number of input features results in substantially higher computational time complexity without resultant random forest model performance improvement. Consequently, training a random forest classifier with a smaller and only relevant input feature subset tends to improve the performance of a medical decision support model. Accordingly, this study investigates feature raking for selecting the appropriate and more related feature subset to the target class among the entire input feature set in a real-world clinical heart disease dataset for improved and more accurate heart disease detection compared to the existing decision support model in the scientific literature.

Dealing with a smaller number of input features brings us to a reduction in the dimensions of input features [3]. More features in the training set make the heart disease classifier model more difficult to learn and the interpretation of the random forest model more difficult. In addition to model complexity, more input features lead to model over-fitting. Model overfitting in turn leads to low performance, resulting in higher performance on the training dataset but poorer performance on new or unseen observations in the test set.

This study investigates feature ranking primarily focusing on model optimization for heart disease diagnosis by identifying more important features for heart disease diagnosis model development with a random forest algorithm. In addition to FR, model optimization approaches such as parameter tuning for selecting better combinations of parameter settings are employed for improving the performance of the random forest model on heart disease diagnosis.

2. Related Work

Several research studies have suggested that a larger volume of input features has an impact on the performance of decision support models in disease classification [4]. For instance, a study conducted in [5] claims that the most effective approach to decision support model optimization for improved model performance is feature selection. The authors employed feature selection and achieved 98.17% accuracy in heart disease detection. The study concluded that the application of feature selection and working with a reduced optimal set of input features significantly improves the performance of the decision support model for heart disease diagnosis.

Another study [6] proposed an information gain-based feature (IFSA) selection approach. Moreover, the researchers studied the effect of a higher dimension in the input feature on the performance of the decision support model for text classification. The experimental result shows that the Naive Bayes model performed with an accuracy of 87.3% when trained input features were selected with the IFSA feature selection method. There are two categories of feature selection approaches, namely, filter-based and wrapper-based [7]. The filter approach assigns weight to each feature, and based on the weight value assigned to each feature, the optimal feature set is selected for training a decision support model for disease diagnosis. The weight is determined based on distance and statistical approaches such as a correlation between a given feature and the target or class label. In the wrapper approach, features with higher weight are selected as the optimal feature subset and the medical decision support model is trained on the optimal feature subset. In contrast, the wrapper approach works with a heuristic search, calculating the predictive accuracy of a feature, and then a feature combination that produces the highest predictive accuracy among all possible combinations of features in the entire input feature is considered for training. Hence, the wrapper approach is computationally costly as compared to the filter approach because the wrapper approach requires more time to find the optimal feature subset by calculating performance on all of the possible combinations of input features in the dataset. Thus, if the original input feature is large, then the number of possible combinations of features expands and the time complexity for determining the optimal feature subset will increase.

In [8], the researchers presented a review of feature selection methods for improving the classification accuracy of decision support models for heart disease detection. The researchers suggested that designing and implementing medical decision support systems with higher classification accuracy for medical dataset classification is one of the major concerns of automated medical diagnosis systems. The study compared the performance of a decision support model with various feature selection approaches. The feature selection approaches employed in the study for the heart disease dataset are the principal component analysis (PCA) and the chi-squared test for selecting relevant features for the heart disease dataset. A principal component analysis is used to reduce the dimensionality of the input feature set before the model is trained on the dataset for heart disease detection. The experimental result appears to prove that more accurate and effective heart disease detection is achieved with feature selection. Heart disease detection accuracy of 85% is achieved when feature selection is applied to heart disease input features. Thus, the heart disease classification model is important to combat the problem of heart disease prediction. In recent years, the wider accessibility of large volumes of clinical heart disease datasets and better computational power in the computing industry has made the design and implementation of classification models promising results. However, due to the complexity and the need for highly accurate classification models for making highly precarious medical judgments, the application of machine learning to the problem of heart disease prediction still needs to be studied. Therefore, we are interested in studying the existing work and developing a more accurate and effective heart disease prediction model by employing FR and random forest algorithm.

3. Research Method

To conduct this research, we have followed the following steps. Firstly, the heart disease dataset was collected. Secondly, we have divided the dataset into a training set (70%) and a testing set (30%). Thirdly, FR is implemented to select optimal input features, and then a random forest algorithm is trained on the training set using the Python programming language with the Jupyter Notebook integrated development environment for implementation. Finally, the implemented method is evaluated using accuracy, confusion matrix, and receiver operating characteristic curve as performance measures.

3.1. Heart Disease Dataset

The heart disease dataset employed in this study for experimenting with the proposed FR classification model optimization is described in Table 1. As we observe from Table 1, each of the 1025 observations in the clinical heart disease dataset is characterized by 13 numeric, nominal, and predictive input features and the class or target variable shown in Table 2. To conduct the experiment on a random forest model with FR, 70% of the dataset, or 717 observations, is used for training, and 30% of the dataset, or 308 observations, is used for testing.

Table 1-Dataset characteristics

No. of instances	Heart disease negative	Heart disease positive	No. of classes
1025	499	526	2

The heart disease dataset used for characterization of each observation in the heart disease dataset is demonstrated in Table 2.

Table 2-Heart disease dataset numeric feature descriptive statistics

Feature	Standard deviation	Mean
age	9.07	54.43
Chest pain	1.02	0.94
Cholesterol	51.59	0.14
Fasting blood sugar	0.35	0.14
Maximum heart arte	23.00	149.11

3.2 Feature ranking (FR)

FR is used to reduce an original N-dimensional input feature subset to a d-dimensional feature set for d<N. The motivation behind FR is to automatically select a 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 a predictive model by removing irrelevant features or noise from a dataset. Let the entire heart disease input feature set be X defined as follows:

$$X = X1 + X2 \dots + XN \tag{1}$$

Where X = the original input feature; N= the number of input features or dimensions. Then the optimal input feature subset is defined as follows:

$$subset = x1 + x2 \dots + xM \mid M < N \tag{2}$$

Where $subset$ = the optimal input feature subset; M= the number of optimal input features or dimensions. The performance of the model trained on a subset of the model trained on the entire input feature X. The FR starts with the last input feature in the dataset, then one more feature is selected randomly, and the model performance is tested on the selected input feature subset. If the performance of the model improves, then the input feature is considered relevant and irrelevant otherwise. The process is repeated until M number of input features that possibly produce the highest classification accuracy is reached.

3.3 Performance Metrics

The goal of machine learning is to find an algorithm that produces a predictive model that produces a predictor y for a real observation y that minimizes the error rate, producing higher predictive accuracy. Predictive accuracy is defined as the ratio of correct predictions to total predictions [9].

$$A = T/N \tag{3}$$

Where A = Predictive accuracy; T= the number of correctly predicted observations; N= the number of observations in the test set. The predictive accuracy is higher when the value of T is higher.

4. Experimental Results and Discussions

As shown in Figure 1, the performance of the model varies for variations in the input feature subset used for training. The input features such as chest pain (cp), heart rate (thal), depression (old peak), maximum heart rate (thalach), and the number of vessels (ca) have higher importance for heart disease diagnosis. The accuracy of the random forest model with the original 13 features is 97.62% and the random forest achieves an accuracy of 98.6%, with the selected features of heart disease. The difference between the lowest and highest accuracy is 0.98%, which means the performance of random forest on the classification of heart failure improves by 17.49% with feature selection. Thus, the quality and number of input features have a significant effect on the performance of the random forest model on heart failure detection.

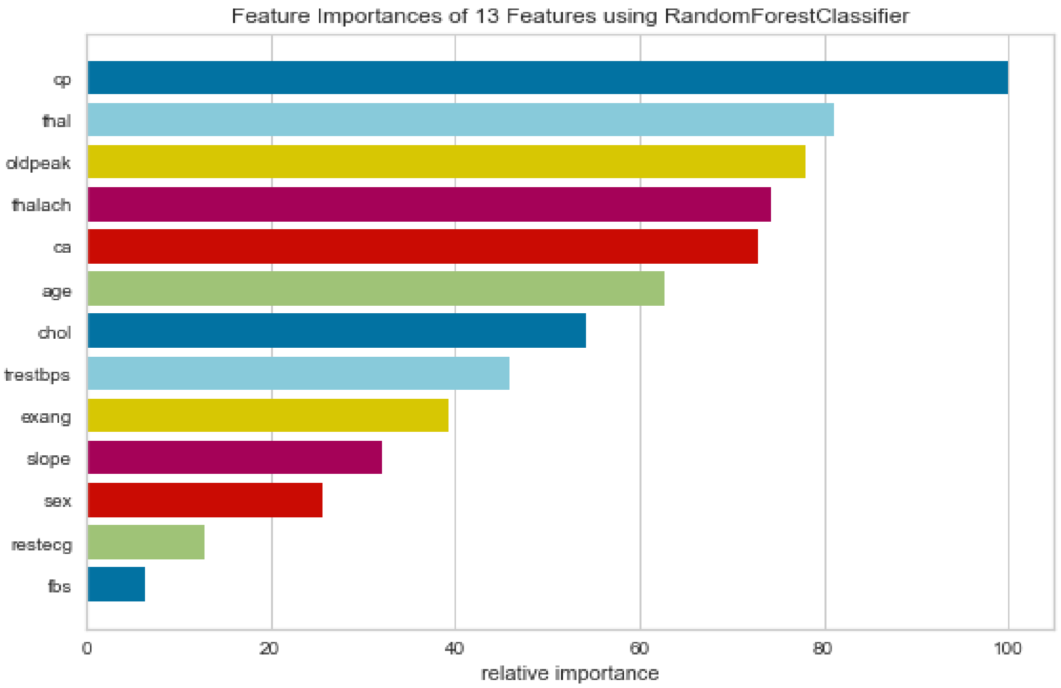

Figure 1-the performance of random forest model on different input features selected by feature ranking

As demonstrated in Figure 1, the performance of the random forest model is momentously affected by the number of input features used for training. The chest pain feature (cp), cholesterol, oldpeak, and maximum heart rate are ranked with higher importance.

Table 4-The optimal input features for heart disease detection

Feature index	Feature name
4	Cholesterol
7	Maximum heart rate achieved
9	Old peak
10	Slope
11	Cardio vascular
12	Thalassemia

In addition to prediction accuracy, we have used receiver operating characteristic curve (ROC) for evaluating the proposed classification model on heart disease prediction, shown in Figure 3.

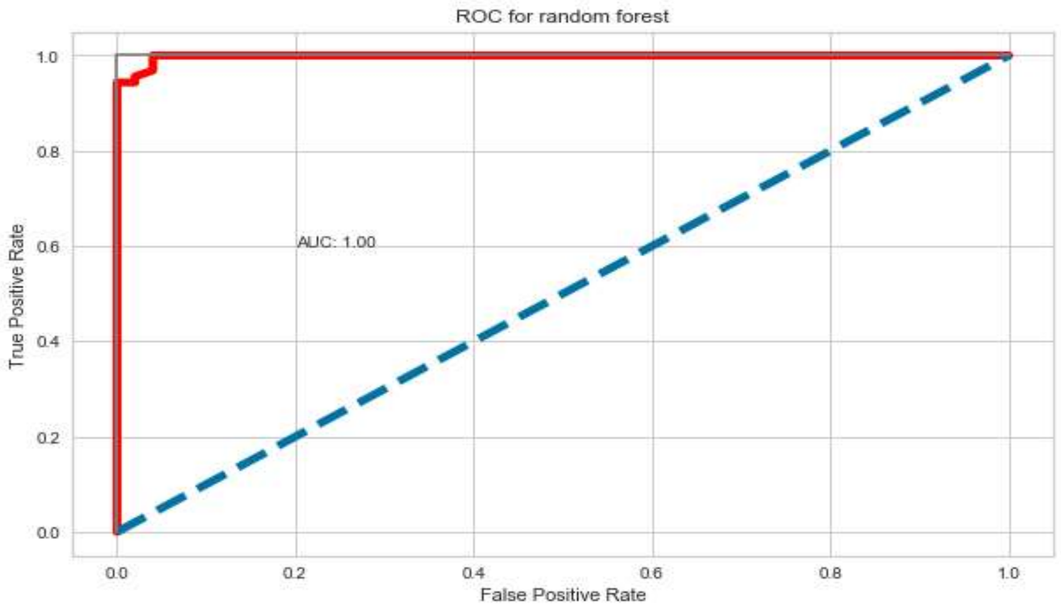

Figure 2-ROC curve of the random forest model on heart disease detection

As we observe from Figure 2, the area under the curve is 0.99, which is almost close to 1.00, revealing that the proposed classification model performs well when trained on the optimal input feature subset selected with the FR.

4.1 Comparison of the existing model

The performance of the existing heart disease prediction model is compared with the developed model with feature ranking. The comparative study of the existing and developed methods is summarized in Table 5. The comparison shows that this study improves the accuracy of existing work by 0.43%.

Table 5-Comparison of the existing and proposed method

Authors Name	Author	Algorithm	Accuracy in %
Sulekha Saxena et al.	[5]	Support vector machine	98.17%
Assegie, T.A et al.	[6]	Random forest	97.07%
RungChing Chen et. Al.	[7]	Naïve Bayes	83.11%
M.A. Jabbar et al.	[8]	Random forest	83.70%
Noor Basha et al.	[9]	Decision tree	84.00%
Assegie T.A et al.	This study	Random forest	98.6%

5. Conclusion

Recently, with an increased number of heart disease patients and the need for highly experienced cardiologists for accurate diagnosis, heart disease diagnosis has become challenging. With automated systems such as the random forest classifier, the problem of heart disease diagnosis is simplified. An automated model provides a faster and more cost-effective means of heart disease prediction. This study, investigated a random forest-based heart disease model feature ranking to investigate optimal features. The simulation result shows that the random forest classifier performed with 98.60% accuracy when trained on optimal input features. As for feature work, researchers are recommended to investigate the performance of feature ranking with different models such as decision trees, support vector machines, k-nearest neighbor algorithms, and deep learning.

Conflict of interest

The author declares that there are no conflicts of interest.

References

[1] K. Jayanthi, L. R. Sudha, ”Optimal Feature Subset Selection for Imbalanced Class Data using SMOTE and Binary ALO Algorithm,” *International Journal of Engineering and Advanced Technology*, pp.344-349, 2020.

[2] Wiharto, Esti Suryani, Vicka Cahyawati, ”The methods of duo output neural 3- network ensemble for prediction of coronary heart disease,” *Indonesian Journal of Electrical 4 Engineering and Informatics* (IJEEI), vol. 7, no. 1, March 2019.

[3] Assegie, T.A, Nair, P.S, ”The Performance of Different Machine Learning Models On Diabetes Prediction,” *International Journal of Scientific & Technology Research*, vol. 9, Issue 01, January 2020.

[4] Mary TS, Sebastian S, ”Predicting heart ailment in patients with varying number of features using data mining techniques,” *International Journal of Electrical and Computer Engineering*, 2019.

[5] Sulekha Saxena , Vijay Kumar Gupta, P. N. Hrisheekesha, ”Coronary Heart Disease Detection Using Nonlinear Features and Online Sequential Extreme Learning Machine,” *Biomedical Engineering: Applications, Basis and Communications*, vol. 31, no. 6, 2019.

[6] Assegie, T.A, S. J. Sushma, B. G. Bhavya, S. Padmashree, ”Correlation Analysis for Determining Effective Data in Machine Learning: Detection of Heart Failure,” *SN Computer Science*, vol. 2, no. 3, May 2021.

[7] Chen, RC., Dewi, C., Huang, SW. et al., "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 52 (2020). https: //doi.org /10 .11 86/s40537-020-00327-4

[8] M.A. Jabbar, B.L. Deekshatulu and Priti Chandra,” *Prediction of Heart Disease Using Random Forest and Feature Subset Selection,” Springer International Publishing Switzerland,* 2016.

[9] N. Basha, A. K. P.S., G. K. C. and V. P., "Early Detection of Heart Syndrome Using Machine Learning Technique," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), 2019, pp. 387-391, doi: 10.1109/ICEECCOT46775.2019.9114651.