Combining random forest and neural networks algorithms to diagnose heart disease

Sima Dehnavi¹, Majid Emamipour¹, Amin Golabpour^{2*}

¹Sobhan Nonprofit Higher Education Institution, Neyshabur, Iran

²Assistant Professor, Shahrood University of Medical Sciences, Shahrood, Iran

Article Info

Article type:
Research

Article History:
Received: 2020-03-04
Accepted: 2020-05-30
Published: 2020-06-23

* Corresponding author:
Amin Golabpour

Assistant Professor, Shahrood
University of Medical Sciences,
Shahrood, Iran

Email: a.golabpour@shmu.ac.ir

Keywords:
Heart Disease
Random Forest
Artificial Neural Network

ABSTRACT

Introduction: Heart disease is known as one of the most important causes of death in today's society and so far no definitive method has been found to predict it and several factors are effective in contracting this disease. Therefore, the aim of this study was to provide a data mining model for predicting heart disease.

Material and Methods: This study used standard data from UCI. These data include four Cleveland, Hungarian, Swiss and Long Beach VA databases. These data include 13 independent variables and one dependent variable. The data are missing, and the EM algorithm was used to control this loss, and at the end of the data, a suggestion algorithm was implemented that combined the two random forest algorithms and the artificial neural network.

Results: In this study, data was divided into two training sets and 10-Fold method was used. To evaluate the algorithms, three indicators of sensitivity, specificity, accuracy were used and the accuracy of the prediction algorithm for four data Cleveland, Hungarian, Switzerland and Long Beach VA reached 87.65%, 94.37%, 93.45% and 85%, respectively. Then, the proposed algorithm was compared with similar articles in this field, and it was found that this algorithm is more accurate than similar methods.

Conclusion: The results of this study showed that by combining the two algorithms of random forest and artificial neural network, a suitable model for predicting heart attacks can be provided.

Cite this paper as:

Dehnavi S, Emamipour M, Golabpour A. Combining Random Forest and Neural Networks Algorithms to Diagnose Heart Disease. Front Health Inform. 2020; 9: 34. DOI: [10.30699/fhi.v9i1.214](https://doi.org/10.30699/fhi.v9i1.214)

INTRODUCTION

The heart is an important part of the human body, and if you have a heart problem, you are more likely to die. The heart acts like a pump and is responsible for pumping blood in the arteries of the human body if there are problems with how a person lives, stress, stress, unhealthy nutrition, inactivity and physical activity, having a history of heart disease. Family and so on can cause heart disease. Diagnosis of heart disease in the early stages is very important and necessary. Early diagnosis of this disease plays an important role in a person's health and life [1]. Heart disease has negative effects on a person's life, making it impossible for them to do their daily activities. Mistakes in diagnosing the disease exacerbate the devastating effects and sometimes lead to death. The use of heart prevention and detection methods can be of great help to physicians in diagnosing this disease

[2].

Data mining is very important in medical data. In fact, the purpose of different stages of data mining is to discover knowledge and achieve results that can be used in the real world to improve efficiency. This extracted knowledge can be used as a decision-making system in the real world. Designing assistive devices to help physicians diagnose the type of disease or choose the right type of treatment with the help of research can be a great help in saving lives [3, 4]. The following are some of the articles from the field.

In 2019, in a paper entitled as predictive model for cardiovascular disease forecasting, Perpra presented seven data mining algorithms. Rapid Miner software was used in this study. It reached 91.67 percent accuracy. It was found that the vector machine algorithm has the best output between the seven

algorithms [5].

In 2019, Saklin et al. presented a paper in a paper entitled "Dimensions reduction for the diagnosis of heart disease with the help of a supportive vector machine". This article uses standard UCI data. In this study, first, the data were reduced to a dimensional scale, and then, with the help of a support vector machine with a radial base neural network kernel, a classification model was presented. And reached 81.19 percent accuracy for test data [6].

In 2019, Reddy et al. presented a paper to predict heart patients. This paper uses a combination of two genetic algorithms and a fuzzy logic algorithm. It is responsible for the treatment of heart patients. UCI standard data were used for evaluation and the algorithm reached 90% accuracy for test data [7].

In 2019, Prakash et al. designed a model to optimize the optimal criteria (OCFS) for the removal of inappropriate traits in an article titled an Optimal Criteria for Selection of Optimal Criteria Features for Predicting and Analyzing Effective Heart Disease Analysis (OCFS). Provides a number of patients for diagnosis and prognosis compared to RFS-IE and MRPS [8].

In 2016, Wei-Jia et al. published an article on optimizing particle density and supporting machine vectors optimized by association rules to identify the causes of heart disease using two particle density optimization (PSO) algorithms and SVM optimized support vector machine. And by the rules of the association (AR), they designed a proposed algorithm whose output has a total classification accuracy of 98.96% and the time spent for classification is 94.94 seconds [9].

In 2018, Yahyaie and colleagues designed an article in an article entitled Using the Internet of Things to provide a new model for predicting remote heart attack using the Internet of Things (IoT), the output of which is valid when using ECG data. 89.5% [10].

The proposed model uses a combination of two random forest artificial network algorithms that have the desired accuracy for diagnosing heart disease. With this model, it is possible to provide physicians with a decision-making system for diagnosing heart disease, which has the ability to differentiate the disease in the early stages and with high accuracy. The following is a description of the standard data, the proposed model, the model evaluation and the discussion and conclusion.

MATERIAL AND METHODS

The study looked at data from the Heart Disease Collection in the UCI Machine Learning Repository section of the University of California, which includes four Cleveland databases with 303 samples, Hungarian with 294 samples, Switzerland with 123

samples and Long Beach VA with 200. Samples are used. The main data set of heart disease includes 14 variables, so that the last variable is introduced as the dependent variable. The dependent variable has values of 4-0, with a value of 0 indicating no disease and values 1 to 4 indicating a degree of heart disease. Due to the conciseness and usefulness of other features in this research, reduction of dimensions has not been used. Table 1 shows the specifications of the data sets used in this study by the number of samples in each class, and Table 2 shows the data specifications [11].

Table 1: Summary of standard data set specifications [11]

Variable name	No disease	Disease				Number
	Class 0	Class1	Class 2	Class 3	Class 4	
Cleveland	164	55	36	35	13	303
Hungarian	188	37	26	28	15	294
Switzerland	8	48	32	30	5	123
Long Beach VA	51	56	41	42	10	200

Table 2: Standard heart disease variables [11]

Row	Variables
1	Age
2	Sex
3	CP
4	Trestbps
5	Chol
6	FBS
7	Restecg
8	Thalach
9	Exang
10	Oldpeak
11	Slope
12	Ca
13	Thal
14	Diagnosis of Heart Disease

In this study, a predictive model for the diagnosis of heart disease was obtained using random forest and artificial neural network algorithms. First, a random forest algorithm presents a model, the steps of which are described separately in Fig 1. In the following, using the artificial neural network, a model is presented, which is expressed in Fig 2. Then, using a combination of two random forest algorithms and an artificial neural network, a model with acceptable accuracy for the diagnosis of heart disease is presented. The study uses four databases: Cleveland, Hungarian, Switzerland and Long Beach VA. A total of 14 features have been used to evaluate the above algorithms, the results of which are discussed in the evaluation section. The following is a pre-processing and how to implement the random forest algorithm, then the artificial neural network algorithm, and finally the proposed hybrid algorithm using the combination of the two algorithms.

Preprocessing

The study database had missing values (Table 3). This missing values should be deleted. The Cleveland database has 4242 data and six missing data, the Swiss database has 1722 data and 273 missing data, the Long Beach VA database has 2800 data and 698 missing data and the Hungarian database have 4116 data and 782 missing data. The missing values are generated using the EM method [12] because the data has a completely randomized loss and the EM method has a better output for these missing data [13]. SPSS.V16.0 software was used to search for missing values.

Table 3: Total number of existing and missing records of four databases

Row	Database name	Records number	Missing value number	Missing value percent
1	Cleveland	303	6	0.141
2	Hungarian	123	273	15
3	Switzerland	200	698	24
4	Long Beach VA	294	782	19

Random forest algorithm

Initially, for all 4 databases, the data is divided into two sets of training and testing. The way the data is divided is based on the K-Fold method. The value of K in this study is 10, then the random forest algorithm [14] is executed on the training data and produces a predictive model. This model is evaluated by both test data and training data. According to the division of data into two sets of training and testing, each answer produces a different answer each time, so the algorithm is executed 100 times and the average of this 100 times is reported. This algorithm will be performed for each data in two states: disease or non-disease and disease type. In the random forest algorithm, the most important parameter is the number of decision trees. Fig 1 shows the flowchart of the classification algorithm. And in the figure, instead of the classification algorithm, a random forest is placed.

Artificial neural network algorithm

In all 4 databases, the data is divided into two sets of training and testing. The method of data division is based on the K-Fold method. Predicts. This model has been evaluated by both test data and training data. According to the division of data into two sets of training and testing, each answer produces a different answer each time, so the algorithm is executed 100 times and the average of this 100 times is reported. This algorithm is performed for each data in two states: disease or non-disease and disease type. In the neural network algorithm, the most important parameter is the number of neurons. Fig 1 shows the flowchart of the classification algorithm. And in the figure, instead of a classification algorithm,

an artificial neural network is placed.

Fig 1: Flowchart of classification algorithm

Suggested algorithm

In this step, the proposed algorithm is stated, the optimal random values of the decision trees are determined by the random forest algorithm. Then in each decision tree, the features and records that need to be selected are determined. The values are entered into the artificial neural network algorithm, and instead of each tree, the decision of a neural network with the optimal number of neurons is placed. The label is then determined by a majority vote. At the end, the results are evaluated. Fig 2 shows the structure of the proposed algorithm [5].

RESULTS

First, the random forest algorithm was executed with the number of different trees and it was determined that the best number of trees is 60 trees, then the artificial neural network algorithm was executed and it was determined that the best number of neurons is 10 neurons, then the proposed algorithm was presented. Three indicators of sensitivity, specificity, and accuracy are used to evaluate algorithms. To conduct research, the data can be divided into two parts, training and testing using the K-Fold method [15]. The value of K in this study is equal to 10 so that different states of educational data selection and testing can be expressed. Table 4 shows the output of the proposed algorithm. The proposed algorithm was performed for both the training data and the test data, and sensitivity, specificity, and accuracy were calculated. And its output is shown in Table 4. The highest accuracy for training data in these four databases is 99.07% and the highest accuracy for test data is 94.27%.

Fig 2: Structure of the proposed algorithm

Table 4: Evaluation of the proposed algorithm

Data Name	Training Data						Test Data					
	Confusion Matrix		Sensitivity	Specificity	Error Rate	Accuracy	Confusion Matrix		Sensitivity	Specificity	Error Rate	Accuracy
Cleveland	139	3	97.89	97	2.66	97.34	14	2	87.5	86	12.35	87.65
	4	117					2	12				
Hungarian	161	8	95.27	98	3.79	96.21	18	1	94.74	91	5.64	94.37
	2	93					1	10				
Switzerland	5	0	100	99	0.93	99.07	1	0	100	92	6.46	93.54
	1	102					1	11				
Long Beach VA	46	0	100	99	1.11	98.89	4	1	80	87	15	85
	2	132					2	13				

Evaluate the proposed hybrid algorithm with similar tasks

The proposed algorithm used in this study is compared with three articles (algorithm [5], algorithm [9] and algorithm [16]) in this study, the

results of which are evaluated on the test data in Table 5, respectively. Is observed. According to the table, the algorithm [9] with 98.96% accuracy for Cleveland database and the proposed algorithm in this study with 93.54% accuracy for Switzerland

database and algorithm [5] with 89.83% accuracy for Long Beach VA database and so on. The proposed hybrid algorithm with 94.37% accuracy for Hungarian database has the best output.

Table 5: Results of evaluating the proposed algorithm with similar tasks

Database name	Suggested Algorithm	Accuracy [5]	Accuracy [9]	Accuracy [16]
Cleveland	87.65	92.3	98.96	84.44
Switzerland	93.54	90.23	-	71.84
Long Beach VA	85	89.83	-	81.71
Hungarian	94.37	91.83	-	88.42

DISCUSSION

In this study, two different algorithms have been used to provide better results for predicting and diagnosing heart disease. The algorithms used in this study include the random forest algorithm and the artificial neural network. The proposed algorithm is a combination of a random forest algorithm and an artificial neural network. The reason for choosing these two algorithms is the high accuracy of random forest algorithms and artificial neural network. As shown in Table 3, the proposed hybrid algorithm has high accuracy in all four databases.

The proposed algorithm was run on all four databases. For Cleveland data, out of 303 records, 11 incorrect records were detected, which is 3% error. For Hungarian data, out of 294 records, 12 incorrect records were detected, which is 4% error. Record 2 miscalculated record of 1% error and also for Long Beach VA data out of 200 records 5 miscalculated

records represent 2% error and in total the predictive model has an error of less than 4% which indicates. The suitability of the proposed algorithm is appropriate.

CONCLUSION

The proposed algorithm for all four databases was also evaluated with two similar articles, and it was found that the proposed algorithm in both Swiss and Hungarian databases performed better than 3% in terms of accuracy, which indicates that the proposed algorithm is appropriate. But the proposed model for the other two databases did not work better.

Finally, it should be noted that the combination of these random forest and artificial neural network algorithms will greatly increase the accuracy of the proposed model and also increase the level of confidence and trust in the model.

AUTHOR'S CONTRIBUTION

The authors agree on this final form of the manuscript, and attested that all authors contributed in the final draft of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

REFERENCES

- DeSilva R. Heart disease. ABC-CLIO: Greenwood; 2013.
- Benjamin EJ, Muntner P, Bittencourt MS. Heart disease and stroke statistics-2019 update: A report from the American heart association. *Circulation*. 2019; 139(10): e56-e528. PMID: 30700139 DOI: 10.1161/CIR.0000000000000659 [PubMed]
- Sultana M, Haider A, Uddin MS. Analysis of data mining techniques for heart disease prediction. *International Conference on Electrical Engineering and Information Communication Technology*; IEEE:2016.
- Thomas J, Princy RT. Human heart disease prediction system using data mining techniques. *International Conference on Circuit, Power and Computing Technologies*; IEEE: 2016.
- Pereira N. Using machine learning classification methods to detect the presence of heart disease [Masters Dissertation]. Technological University Dublin; 2019.
- Saqlain SM, Sher M, Shah FA, Khan I, Ashraf MU, Awais M, et al. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. *Knowledge and Information Systems*. 2019; 58(1): 139-67.
- Reddy GT, Reddy MPK, Lakshmanan K, Rajput DS, Kaluri R, Srivastava G. Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. *Evolutionary Intelligence*. 2020; 13: 185-96.
- Prakash S, Sangeetha K, Ramkumar N. An optimal criterion feature selection method for prediction and effective analysis of heart disease. *Cluster Computing*. 2019; 22(5): 11957-63.
- Wei-Jia L, Liang M, Hao C. Particle swarm optimisation-support vector machine optimised by association rules for detecting factors inducing heart diseases. *Journal of Intelligent Systems*. 2017; 26(3): 573-83.
- Yahyaie M, Tarokh MJ, Mahmoodiyar MA. Use of Internet of things to provide a new model for remote heart attack prediction. *Telemed J E Health*. 2019; 25(6): 499-510. PMID: 30256729 DOI: 10.1089/tmj.2018.0076 [PubMed]

11. Aha DW. Heart disease data set [Internet]. 1988 [cited: 17 Dec 2019]. Available from: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>
12. McLachlan GJ, Krishnan T. The EM Algorithm and extensions. Wiley; 2007.
13. van Buuren S. Flexible imputation of missing data. Taylor & Francis; 2012.
14. Liaw A, Wiener M. Classification and regression by random forest. R News. 2002; 2(3): 18-22.
15. Hand DJ. Principles of data mining. Drug Saf. 2007; 30(7): 621-2. PMID: 17604416 DOI: 10.2165/00002018-200730070-00010 [[PubMed](#)]
16. Lahsasna A, Ainon RN, Zainuddin R, Bulgiba A. Design of a fuzzy-based decision support system for coronary heart disease diagnosis. J Med Syst. 2012; 36(5): 3293-306. PMID: 22252606 DOI: 10.1007/s10916-012-9821-7 [[PubMed](#)]