**International Journal of Science Academic Research**
Vol. 01, Issue 06, pp.343-345, September, 2020
Available online at http://www.scienceijsar.com

ISSN: 2582-6425

**Research Article**

# PRESENTING A PREDICTION ALGORITHM FOR HEART PATIENTS USING DATA MINING

**[1]Ebrahim Salimi Tork, [2]Sima Dehnavi and [3,*]Amin Golabpour**

[1]Faculty member of Karoon Institute of Higher Education, Ahvaz, Iran,
[2]Sobhan Private Higher Education Institute, Neishabour, Iran,
[3]Assistant Professor of Shahroud University of Medical Sciences, Shahroud, Iran

**Received** 17th July 2020; **Accepted** 10th August 2020; **Published online** 30th September 2020

**Abstract**

Heart disease is one of the most common human health issues in the world, which there is not a definite method to predict it. The only hopeful way for these patients to continue living is through proper care. The accurate prediction of a patient's heart condition is very critical to prevent any side effects caused by disease. Therefore, in this article, a new method based on a combination of the random forest algorithm used to predict the condition of the patient's heart. The proposed structure can calculate the output corresponding to each learning pattern obtained from performing various experiments on several experimental items and compare them with the desired output of each pattern. For evaluation, the standard UCI data is used. The results show that the obtained accuracy of our method outputs in comparison to similar algorithms is up to 93.3%, which indicates the suitability of the proposed algorithm.

**Keywords:** Prediction, Data Mining, Heart disease, Random Forest.

## INTRODUCTION

The circulatory system has the main task of preparing and supplying the necessary materials for the continuation of the biological processes of all cells of the body. In addition, it performs the excretion of waste materials from cellular metabolism in the appropriate excretory pathways .The circulatory system consists of the heart and arteries.The heart is responsible for circulating blood throughout the body. For this purpose, it produces the main strength for moving blood through its muscle contractions (Mozaffarian *et al.,* 2016). Heart failure is one of the most common issues in Heart diseases, which has a very adverse effect on patient's life quality.Heart failure means that the heart cannot pump enough blood to meet the body's needs and send it to other parts of the body.In some cases, the heart cannot fill with blood, and in other cases, it cannot pump blood well to other tissues.Some people may have both problems at the same time. The result is the insufficient blood supply to the tissues and heart failure (Lewis, 2019). About 610,000 people die of Heart disease in the United States each year, according to the World Health Organization. Heart disease is the leading. More than half of all deaths from Heart disease in 2009 were in men (Benjamin *et al.,* 2018). In Iran, official statistics from the Ministry of Health and Medical Education show that 33 to 39.3% of deaths in the country are due to cardiovascular disease so that Iran has a record of the highest heart death rate in the world. According to this report, 300 people die every day in the country due to Heart complications (Davoudi *et al.,* 2016). This study presented a Heart disease predictive model using a random forest algorithm.The data for this study is standard Swiss data from the UCI Usage Database to determine if the model can predict the previous disease.In the following, first, we look at the background of the research. Then the random forest algorithm briefly reviewed. Finally, the proposed algorithm discussed and evaluated.

***Corresponding Author:** ***Amin Golabpour***
*Assistant Professor of Shahroud University of Medical Sciences, Shahroud, Iran.*

### Research background

There is a lot of research has been done about Heart diseases. In the following, briefly reviewed several of them. In 2017, Davari et al. published an article predicting previous illness . They used the optimal Support Sector Machine algorithm for predicting the model. The research used standard data. This study shows the sensitivity rate is 98.43%, the specificity is 100%, and the accuracy is 99.21%, which makes it a tolerable algorithm (Dolatabadi *et al.,* 2017). In 2018, Mathan *et al.* presented a model for predicting Heart failure. In this study, they presented a predictive model of the combination of decision tree and artificial neural network. First, the decision tree calculates the entropy of the variables. Next, the variables with higher entropy selected, and then the variables with lower entropy removed.The reduced variables then given to the leading artificial neural network, and then the proposed model is presented.In this study, 10 independent variables and 1 dependent variable are used. This model has been reached to an accuracy of 95.5%, which shows that the proposed model is suitable (Mathan *et al.,* 2018). In 2019, Nagappan *et al.* evaluated heart attack prediction models in a clinical article. For evaluation, the opinions of physicians were used, and it indicated that the machine learning algorithms are a suitable method for Heart attack prediction (Nagappan *et al.,* 2019).

### Random Forest

The random forest algorithm is an easy-to-use machine-learning algorithm that often provides excellent results even without adjusting its meta-parameters.This algorithm is one of the most widely used machine learning algorithms for both classification and regression due to its simplicity and usability (Pavlov, 2000). To understand how a random forest works, first must learn the decision tree algorithm, which is the building block of the random forest.Humans use the decision tree every day for their decisions and choices; even if they do not know that, what they are using is a kind of machine learning algorithm .However, after adequate training with

qualitative data, the decision tree's ability to make predictions exceeds human ability.It should be consider that the decision tree has no conceptual understanding of the problem, and even after training, such an understanding not achieved. From the model's perspective, it receives numbers from the input for the training and gives to the output different numbers from what it gets. So far, the high-level concept of a decision tree has explained. Figure 1 shows an example of a random forest.

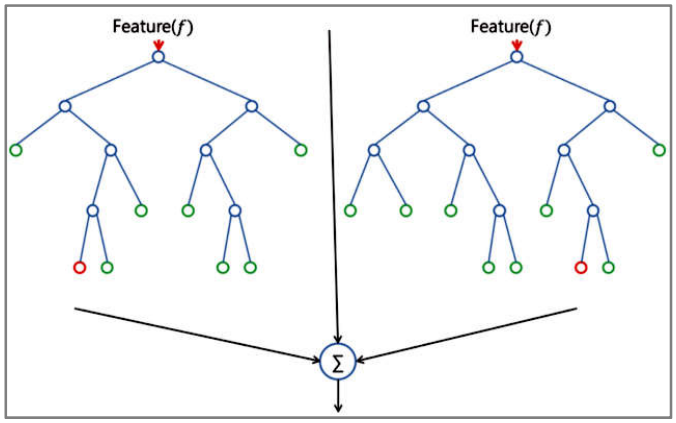


**Figure 1. An example of a random forest**

## Study Data

This study uses the heart disease data set in the tasks section related to the classification of the University of California UCI machine-learning repository, which includes Switzerland with 123 samples (Janosi *et al.,* 2019). The main dataset contains 76 features;but in published experimental researches, only a subset of 14 members of these features have used.This study also uses 14 reduced features;So that the last attribute is introduced as the target attribute. The target feature has values of 0-4, which 0 indicates the absence of disease, and values of 4-1 indicate a degree of heart disease.

Due to the concise and usefulness of these features in this study, no other feature reduction method was used.Table 1 shows the specifications of the standard data set used in this research by the number of samples in each class, and Table 2 shows the specifications of the existing data.

## Pre-processing

The study database has missing values. These missing values must bereplacing. The Switzerland database contains 1722 data and 273 missing data. The missing values replaced using the EM[1]method (Bowers, 2008). SPSS software used for this purpose. The Table 3 shows the percentage of missing.

## Proposed model

Random tree forest produces many decision trees.To classify a new object, it places an input vector at the end of each random forest tree.Each tree gives us a classification, and we say that this tree votes for that class.The forest with the highest number of votes (among all forest trees) is selected one.Each tree is forming as follows:

- If N is the number of states in the train data set, N randomly samples the positions by substituting the original data.This example is a collection of works for this tree.
- If M has a variable and m considered smaller than M, so that in each node, the m variable being randomly selected from M, and the best separation on this m variable is used to separate the node, the value of m is considered constant during forest construction.

Each tree grows as big as possible. There is no pruning.

The forest error rate depends on the following two things:

- The correlation between the two trees; in the forest increases the correlation between the forest error rates.

**Table 1. Summary of the Switzerland dataset**

| Data | No disease | disease | | | | Count |
|---|---|---|---|---|---|---|
| | Class 0 | Class 1 | Class 2 | Class 3 | Class 4 | |
| Switzerland | 8 | 48 | 32 | 30 | 5 | 123 |

**Table 2. Details of standard heart disease data**

| Id | Feature name | description |
|---|---|---|
| 1 | Age | Age |
| 2 | Sex | Sex |
| 3 | CP | Types of chest pain include four values: normal angina, unusual angina, non-angina pain and no pain |
| 4 | Trestbps | Blood pressure at rest |
| 5 | Chol | blood cholesterol |
| 6 | FBS | Fasting blood sugar |
| 7 | Restecg | Resting ECG results include three values: normal, abnormal ST-T wave, cross-sectional increase |
| 8 | Thalach | Maximum heart rate obtained |
| 9 | Exang | Exercise-induced angina |
| 10 | Oldpeak | ST falls due to exercise |
| 11 | Slope | The slope of the ST area at the peak of training includes three values: high slope, smooth slope, and low slope |
| 12 | Ca | Number of major vessels identified by fluoroscopy |
| 13 | Thal | Thallium scan consists of three values: normal, permanent lesion and reversible lesion |
| 14 | Diagnosis of heart disease | Diagnosis of heart disease |

**Table 3. Total number of existing and missing records of standard database**

| Database name | All Data | | Missing values count | Percentage of missing |
|---|---|---|---|---|
| | Records | Column | | |
| Switzerland | 123 | 14 | 273 | 15.85% |

- The power of each tree in the forest; each tree with a low error rate is a strong classifier.Increasing the strength of each tree reduces the forest error rate.

Decreasing m reduces both correlation and power; and increasing it increases both .Initially, for each database, the data being divided into two sets of training and testing. How to divide the data is based on the K-Fold method, the value of K in this study is 10The random forest algorithm is then implemented on the training data and generates a predictive model. This model evaluated by both test data and training data, which is reported in section 6.Due to the division of data into two sets of training and testing, in every execution, a different answer will being generated each time, so the algorithm is executed 100 times and the average of this is reported 100 times. This algorithm was performed for each data in two cases of disease or non-disease and type of disease. In the random forest algorithm, the important parameter is the number of decision trees. The algorithm was executed with a random number of forests of 10 to 100 in steps of 10 units to be able to calculate its best state. All of these reports are presented in Section 6. Figure 2 shows the flowchart of a random forest algorithm.

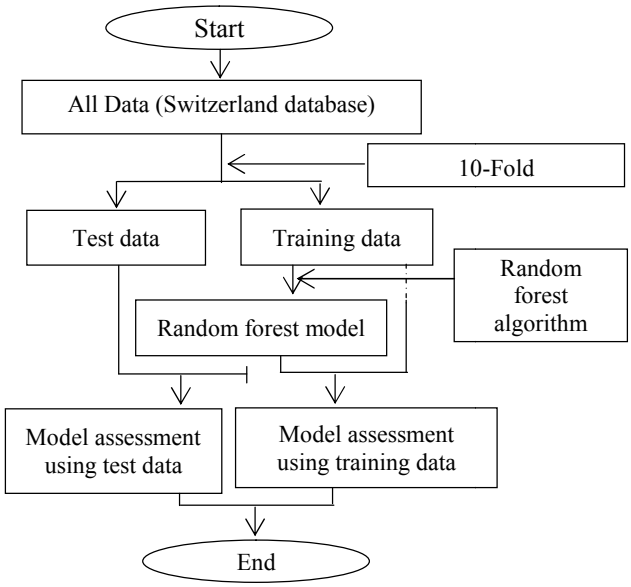


**Figure 2. Flowchart of the proposed algorithm**

### Assessment

This section describes the evaluation of the proposed model. Figure 3 shows the results of implementing a random forest algorithm for the Switzerland database run ten times by changing the number of trees.As can be seen, the average accuracy for 100 decision trees in educational data has the highest value of 99.99%, and test data with 40 trees have the highest value of 93.13%.At the end of this result, it observed that with increasing the number of trees, the average accuracy for the training data increases, and in the test data, the outputs are almost convergent.Figure 3 shows the correct output of the random forest algorithm with different number of trees.

## RESULTS

In this study, using different algorithms, we tried to provide better results for the problem of predicting and diagnosing heart disease.The algorithms used in this research include the random forest algorithm .The reason for choosing this algorithm is the high accuracy of the random forest algorithm . The results of this algorithm show the high accuracy of the algorithm.Using the double run method, the outputs of the proposed model can be converged and increasing the reliability of the proposed model.The high volume of the data set slows down the algorithm.Finally, should note that the combination of these algorithms, the random forest, and the artificial neural networks, increases the accuracy of the proposed model and increases the reliability and trust in the model.

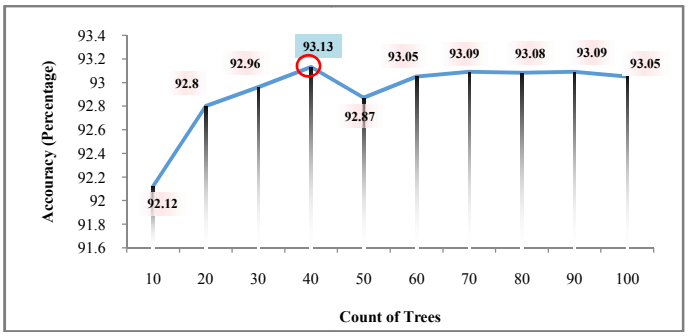


**Figure 3. Calculation of the optimal number of trees in terms of accuracy**

## REFERENCES

Mozaffarian, D., *et al.* 2016. Heart disease and stroke statistics-2016 update a report from the American Heart Association. *Circulation*, 133(4): p. e38-e48.

Lewis, T., The Heart. 2019: Univ of California Press.

Benjamin, E.J., *et al.,* 2018. Heart disease and stroke statistics-2018 update: a report from the American Heart Association. *Circulation*, 137(12): p. e67.

Davoudi, A., *et al.,* 2016. Resistance pattern of antibiotics in patient underwent open heart surgery with nosocomial infection in North of Iran. *Global journal of health science,* 8(2): p. 288.

Dolatabadi, A.D., S.E.Z. Khadem, and B.M. Asl, 2017. Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. *Computer methods and programs in biomedicine,* 138: p. 117-126.

Mathan, K. *et al.* 2018. A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. *Design Automation for Embedded Systems,* 2018. 22(3): p. 225-242.

Nagappan, M., *et al.,* 2019. Heart Disease Prediction Using Data Mining Technique. 2019.

Pavlov, Y.L., Random Forests. 2000: VSP.

Janosi, A., et al. Heart Disease Data Set. 1988 [cited 2019 10/10/2019]; Available from: https://archive.ics.uci.edu/ml/datasets/heart+Disease.

Bowers, D. Medical Statistics from Scratch: An Introduction for Health Professionals. 2008: Wiley.

*******