International Cardiovascular Research Journal
Volume 16 · Issue 2 · e122437
OPEN PEER REVIEW

Comparison of Random Forest and Artificial Neural Network Models to Evaluate Diagnostic Factors in the Necessity to Perform Angiography

Parastoo Golpour, Mohammad Tajfard, Majid Ghayour-Mobarhan, Mohsen Moohebati, Ali Taghipour, Habibollah Esmaily, Sara Sabbaghian Tousi

SUBMITTED	REVISED	ACCEPTED
5 Jan 2022	2 May 2022	29 May 2022

About this report

This open peer review (OPR) report shows the public comments made on this article during peer review, by the reviewers, the associate editor and the editor-in-chief, for every revision round, together with the authors' replies. Comments appear only with the consent of the people who wrote them. Reviewers who chose to sign their report are named; the others are shown as "Referee". Comments are edited for language only, and confidential remarks to the editors are not included.

More about open peer review: brieflands.com/knowledgebase/opr

Contents

1 **Original submission** Editor-in-Chief 2
2 **Revision 1** Editor-in-Chief 3
Authors' reply to reviewers 4
Reviewer and editor comments 7

1 Original submission

1 comment · 27 Feb 2022

Editor-in-Chief

MAJOR REVISION 27 Feb 2022

Needs revision.

2

Revision 1

AUTHORS' RESPONSE TO REVIEWERS

Authors' reply to reviewers

The next pages contain the authors' point-by-point reply to the comments of this revision round, as submitted with the revised manuscript. It is published as the authors wrote it.

Reviewer 1:

Q: The only comment I have is that the authors have considered an obstruction of more than 50% as "positive angiography" while we know that more than 75% is clinically significant and between 50-75% is clinically not significant. So I think the authors should analyze their data according to this new definition and then present their data again.

If they present, their data with this new definition and a biostatistician approves their data then it can be published.

R: Obstruction of more than 50% is considered "positive angiography" in articles performed on this dataset and published in prestigious journals with Q1 and Q2 (articles link attached). Our study is based on the same data.

https://pubmed.ncbi.nlm.nih.gov/24219895/

https://pubmed.ncbi.nlm.nih.gov/28534120/

https://pubmed.ncbi.nlm.nih.gov/30548615/

https://pubmed.ncbi.nlm.nih.gov/32160769/

Reviewer 2:

Q: Please remove the Python software form the title.

R: It was done.

Abstract:

Q: Statistical methods and machine learning cannot be a substitute for clinical methods for diagnosis. Without diagnostic tests, there is no data to be used by machine learning methods. Therefore, the introductory paragraph of the abstract is very erroneous.

R: It was done. Machine Learning methods have the ability to screen candidates for angiography. These methods use in health care to build models in order to accelerate the process of medical decision-making.

Q: Comparing statistical methods with clinical methods is completely wrong. The two must be together. You should refer to various statistical methods for diagnosis, such as decision tree, logistic regression, discriminant analysis, etc., and give the reason for using random forest, i.e., compare statistical methods with statistical methods, not diagnostic methods.

R: It was done. An artificial neural network works like a biological neural network and is a powerful way to predict a complex relationship between variables. Therefore, the performance of this method was compared with the random forest method.

Method:

Q: In this section, write which data was used instead of angiography.

R: In this study, the most important variables in the diagnosis of CAD were as follows: gender, age, smoking habit, body mass index (BMI), and fasting blood sugar (FBS).

Q: Add the sampling method and local of sampling.

R: The sample size calculated for the data set based on the first type error was 0.05 and the second type error was 0.1 and the odds ratio was 1.43, which were obtained for at least 964 angiography candidates. 20% was considered the outlier rate. The total sample size is equal to:

$$N = 964 + 20\% (964) = 1157$$

Finally, after 20 months, was collected data from 1187 patients.

Q: An important principle in modeling is the simplicity of the model. Here, 19 variables are to be used instead of angiography. Does this make sense and can this patient be managed by a doctor?

R: In this study, the most important variables in the diagnosis of CAD were as follows: gender, age, smoking habit, body mass index (BMI), and fasting blood sugar (FBS) which were selected based on previous similar research and our experts' opinion.

Q: New with 100 trees, the accuracy is 76%. So this model cannot be used instead of 100% angiography. It is useful as a screening test, not a diagnostic test.

R: The decision on whether a patient needs invasive angiography is complex and many risk factors must be considered. ML approaches are capable of constructing models to screen candidates for angiography and also accelerate the process of medical decision-making and are not used for diagnostic tests.

Q: The parts of modeling that are related to the results part are expressed in the method part. In the method section, the variables and the method of measurement, modeling and software used should be mentioned. Also significant level must be added.

R: It was done. The questioned parts were highlighted in the text.

Results:

Q: The variables mentioned are variables that are associated with cardiovascular disease and are determined by simple statistical tests that are more common in people with positive angiography, and this cannot show the ability of the random forest model.

R: In this study, our aim is to classify based on positive and negative angiographic results and cannot use regression statistical methods. In addition, this study compares Machine Learning classification algorithms that perform better than older classification methods in improving cardiovascular risk prediction.

Q: Most reported probability values are greater than 5% and are not significant. How important are these variables??

R: According to Table 1 in the manuscript, all P-values values are less than 5% and are significant.

Q: What method was used to check overfitting?

R: Due to the use of five important scaled parameters in modeling, overfitting has not occurred.

Q: The specificity index in your model is very low and this indicates that you had a lot of false negatives and this can be dangerous in a clinical decision.

R: It is important for us patients to be correctly diagnosed. Therefore, it is important sensitivity index be high because if it is low, it can cause patients to be incorrectly diagnosed as healthy people, and this may lead to a risk of death for the patient.

Reviewer 3:

Q: This manuscript is not acceptable for publishing because other important factors for predicting CAD and ischemia aren't included such as quality of chest pain, electrocardiographic findings and biomarkers and LV function.

R: The aim of this study was to compare machine learning methods with acceptable accuracy and AUC. We also tried to use the lowest parameters (using the important parameters available) to avoid over-fitting. Also, because a number of important variables in our dataset had missing values, they were omitted to increase the accuracy of the model.

2 Revision 1

1 comment · 29 May 2022

Editor-in-Chief

ACCEPT 29 May 2022

accept