Shiraz E-Medical Journal
Volume 23 · Issue 7 · e120140

OPEN PEER REVIEW

Comparing Data Mining Algorithms for Breast Cancer Diagnosis

Mostafa Shanbezadeh, Raoof Nopour, Leila Erfannia, Morteza Amraei, Nahid Mehrabi, Mehrnaz Mashoufi

SUBMITTED	REVISED	ACCEPTED	DOI
8 Oct 2021	12 Jan 2022	2 Feb 2022	10.5812/semj-120140

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 2
2 Revision 1 3
Authors' reply to reviewers 4
Reviewer and editor comments 6
3 Revision 2 7
Authors' reply to reviewers 8
Reviewer and editor comments 9
4 Revision 3 10
Authors' reply to reviewers 11
Reviewer and editor comments 13
5 Revision 4 Editor-in-Chief 14
Authors' reply to reviewers 15
Reviewer and editor comments 16
6 Revision 5 17
Authors' reply to reviewers 18
Reviewer and editor comments 19

1 Original submission

0 comments

No public comments in this round.

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.

Hi Dear Reviewer! We answered all your precious questions point to point as you wanted and journal policy,

Thank you for your very precious comments, your answers are:

1- Q: *Check the English proficiency and writing style.* A: Dear reviewer. Thanks for your good and constructive comments. Based on your eighth comment, the entire text of the article was re-edited by English language experts in terms of writing and spelling features. It is provable by comparison with the first text.
2- Q: *The research gap and research question, its importance to the literature is unclear.* A: Study question and gap has been brought in the Introduction section.
3- Q: *You have used 526 samples which is a low amount of data for machine learning. Moreover, data preparation and cleaning are not presented. How did you clean and verify data before running ML* A: Dear reviewer, the three steps of dataset dimensions reduction have been represented clearly.
4- Q: *After feature selection, you have finally selected 16 features out of 32. This method of feature selection is seemingly wrong and you had to run statistical analysis and check the dependency that might be seen among the selected features.* A: Dear reviewer, we have used the common statistical test at P for determining the dependency of each variable with the output class.
5- Q: *Please clearly state and compare the results and testing and training. This section is missing in the manuscript. Add the required details and corresponding metrics.* A: Dear reviewer, we didn't use the splitting process of testing and training in one step, and also the fold cross-validation has been utilized for measuring the performance of the algorithm randomly with 10 fold cross-validation, so we hadn't the testing and training performance indicators, separately in one step. But we brought the average of training and validation performance criteria like the total confusion matrix for evaluating a general algorithms performance.
6- Q: *You have just relied on F-score. Adopt more performance metrics and add post-analysis methods like cross-validation models.* A: The 10 fold cross-validation and also the PPV, NPV, Sensitivity, Specificity, and AUC of each data mining algorithm have been compared in this study in addition to F-Score.
7- Q: *It does not make sense at all to model 526 sample data using ML considering 16 features! Please revise the modeling procedure.* A: Dear reviewer, as you stated, the factors were analyzed using statistical analysis, and more variables were considered as the most important factors as you said, thank you.
8- Q: *Please clearly state and compare the results and testing and training. This section is missing in the manuscript. Add the required details and corresponding metrics.* A: Dear reviewer, in the previous version of the manuscript, we just brought the performance results of the training process, but, now we represented the average amounts of training and validation performance results as a general result about during and post-training process in

one table named total confusion matrix and then calculated the other performance criteria based on this matrix.

9- Q: *On page 12, I can see a short decision tree that conflicts with your modeling procedure and incurs 16 variables!!* A: Dear reviewer, as you wanted, we tried to bring all variables in the tree, but, some variables such as alcohol and … with low differences of info gain than other variables were removed during the tree pruning for preventing from overfitting.

10- Q: *The research gap and research question, its importance to the literature is unclear.* A: The importance of our research, based on the first comment.

11- Q: *Enrich the literature using recently published articles.* A: Dear Reviewer. Based on the given comments about updating research literature and prior background, older items (before 2016 and 2017) were removed and replaced with new contents (2018, 2019, 2020, etc.).

12- Q: *You have used 526 samples which is a low amount of data for machine learning. Moreover, data preparation and cleaning are not presented. How did you clean and verify data before running ML?* A: With respect. In many reviewed and prior researches in this study as study background, the sample population (as input to data mining software's) is less than 600, 700 cases. In our research, similarly, the research population is about 500 cases and this has been added as a limitation of the present research.

13- Q: *Before conclusion, please clearly state what is your new findings and contributions.* A: Dear reviewer, the capability of the performance of the proposed J-48 decision tree algorithms in this study using the various factors affecting the BC diagnosis has been described in the article text.

2 Revision 1

0 comments

No public comments in this round.

3

Revision 2

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.

Hi Dear Reviewer, We answered all your precious questions point to point as you wanted and journal policy,

Thank you for your very precious comments, your answers are:

Q7: ***It does not make sense at all to model 526 sample data using ML considering 16 features!!. please revise the modeling procedure.***

A7: Dear reviewer. We re-implemented the ML models based on more data which attained from the BC registry database. After this step, the decision tree was again identified as the better classifier.

Q8: ***Check the English proficiency and writing style.***

A8: Dear reviewer , regarding your comment about English language editing, in the first round of revision (prior round) , we actually delivered the whole article to an English language editor (non formal editor, DR HEIDARI) and he edited the whole of article by applying track change (editing documentation is available). However, if the present article is accepted by the journal reviwer and editorial, we will give the article to the official English editor if needed (for example RADAN EDIT ENGLISH), along with the proof of native editing documents.

Dear Reviewer, in fact, after the first round of review and the comment on language editing, we gave the entire manuscript to an English language editor, who edited the whole manuscript with track changes (the documentation of this editing is available from the editor and will be provided if needed). In any case, if the present manuscript is approved by the esteemed journal Shiraz E-Medical Journal and the respected reviewers, and if you, dear Reviewer, or the journal request it again, we will certainly submit the work to official editing organizations such as Radan Edit English and also provide a native-editing certificate. Thank you for your helpful comments.

Q9: ***Before conclusion, please clearly state what is your new findings and contributions.***

A9: Dear reviwer, according the announced comment, we added new finding and study contribution.

Q10: ***Enrich the literature using recently published articles.***

A10: Dear reviewer. According to your good comment, we have limited and updated all cases of research background to 2021. Previous research backgrounds (2017, 2018, 2019 and 2020) were deleted and updated items were added.

[Translated from Persian by the Brieflands editorial office.]

3 Revision 2

0 comments

No public comments in this round.

4

Revision 3

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.

In the name of god

Dear editor, thank you for considering our manuscript for publishing in your valuable journal; we answered all your precious questions point to point as you wanted and journal policy.

Hi Dear Reviewer,

Thank you for your very precious comments; your answers are:

Q1: In the abstract, "we weighted the features to achieve the best factors in BC diagnosis" is wage. Based on which methods and criteria?

A1: Dear reviewer, we didn't weigh the features as you stated but calculated the Phi coefficient, which as per your informative review, deleted the information about it; thanks.

Q2: In the abstract, "the data mining process was implemented using the selected algorithms" is wage. Which methods and algorithms?

A2: Dear reviewer, all the data mining algorithms are represented in the abstract section in detail, as you stated.

Q3: In the abstract, what do you mean by "different evaluation criteria"? Please declare.

A3: Dear reviewer, as you stated, we brought the confusion matrix and AUC as the performance criteria used in this research. Thank you.

Q4: What do you mean by "J-48"?

A4: Dear reviewer, The J-48 decision tree algorithm has been entirely mentioned in the abstract section.

Q5: Revise the whole abstract to give more information. There was no comparison among different methods.

A5: Dear reviewer, we compared the data mining algorithms using the AUC as the general performance criteria in the abstract section and reformed it with more informative materials.

Q6: Report descriptive statistics of understudy variables and compare groups using bivariate analysis.

A6: dear reviewer, we brought all variables; statistics description divided to the positive and negative BC cases in Table 1. We calculated the bivariate analysis using the chi-square test among two different groups.

Q7: Using any pre-processing methods for dimension reduction is not acceptable. Remove all the information and results about the phi coefficient.

A7: dear reviewer, we omitted the information regarding the phi coefficient correlation and retested the models using all study variables; thank you for your informative comment.

Q8: There was no information about the proposed classifiers.

A8: Dear reviewer, we described all algorithms definitions on one page highlighted in blue color. Thank you.

Q9: To assess prediction accuracy, you must divide the dataset into training and testing sets.

A9: Dear reviewer, we split our data into training and test modes (70% of the research data for training and 30% for the testing process). Then calculated all performance criteria, and the confusion matrix depicted in the article is the total confusion matrix, including the training and test confusion matrix accumulatively. Thank you for your precious comment.

Q10: Figure 2 is redundant.

A10: Dear reviewer, as you commented, we deleted the redundant table with its additive criteria.

Q11: Show all ROC diagrams in a figure.

A11: Dear reviewer, as you cited, we brought all the ROC curves in one diagram; thank you.

Q12: Figure 4 is not clear.

A12: dear reviewer, as you commented, we pruned our decision tree to make small our tree for more clarity by lowering the confidence factor.

4 Revision 3

0 comments

No public comments in this round.

5

Revision 4

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.

In the name of god

Dear editor, thank you for considering our manuscript for publishing in your valuable journal; we answered all the Associate Editor's precious questions point to point as you wanted and journal policy.

Hi Associate Editor,

Thank you for your very precious comments; your answers are:

Q1: In Table 1, remove columns 1, 3, and 6. Report number (%) (or mean± SD) in two separate columns for negative and positive cases.

A1: Dear associate editor, as you stated, we removed columns 1, 3, and 6 and reported the number (%) (or mean± SD) in two separate columns for negative and positive cases. Thank you for your highly precious comment.

Q2: Report exact P-values in the text.

A2: Dear associate editor, we cited the accurate values of P in the text, as you noted. Thank you.

Q3: Compare the training set and testing set in independent variables.

A3: Dear associate editor, as per your excellent comment, we compared our dataset's training and testing sets regarding the input class. Your idea was very precious, and we praised it.

Q4: Report Table 2 and Figure 2 based on the training set and also testing set.

Q4: Dear associate editor, as you said, we wrote the confusion matrix and AUC as our results regarding the training and testing modes of the dataset. Your comment was very beneficial; thank you, dear associate editor.

Q5: The discussion section is poor. There was no comparison between obtained results and previous studies and also no justification regarding the obtained results.

A5: Dear associate editor, as you stated, we enriched our discussion using more comparison of our study with other studies, then compared and analyzed our results with other researches and specified our results among different studies. Thank you for your precious comment.

5 Revision 4

1 comment · 9 Jan 2022

Editor-in-Chief

MINOR REVISION 9 Jan 2022

Minor revision

Thanks for sending the revised file.
I am not satisfied with the answers provided to comments #7, 8, 9, and 10.
Please work on them and provide the required responses.

6

Revision 5

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.

In the name of god

Dear editor, thank you for considering our manuscript for publishing in your valuable journal; we answered all the Associate Editor's precious questions point to point as you wanted and journal policy.

Hi Associate Editor,

Thank you for your very precious comments; your answers are:

Q1: In Table 1, use "P-value" rather than "P-level".

A1: Dear Associate Editor, as you stated, we changed the term of P-value to P-level in Table 1. Thank you.

Thanks for sending the revised file.

I am not satisfied with the answers provided to the comments # 7,8 , 9 and 10

please work on them and provide the required responses

Dear Editor, we answered to these questions in the Revision2, Please see the revision 2 file for more information. All of them are considered in current revision. Thank you.

6 Revision 5

0 comments

No public comments in this round.