Review of: "Implementing Machine Learning to predict the 10-year risk of Cardiovascular Disease"

Rui Tong

Potential competing interests: No potential competing interests to declare.

This research deprived 7 different machine learning algorithms for the prediction of 10-year risk of Cardiovascular and compared their performance from the aspect of discrimination. The task shows potential value in introduction the application of machine learning technology into clinical practice. However, a few aspects of the article might be improved via a revised vension:

1.In the section of related works, the authors reviewed the performance of previous ML models and mentioned the limitation in model generalizaion to different populations or settings. It is better to explain in the section of discussion why this research based on UCI datasets can overcome this limitation without external validation.

2.There are 76 variables in the original datasets and the researches selected 14 of these atrributes for analysis. However, readers might feel confused why these features are included but other features should be excluded. It might be better to present the details in feature selection whether based on the clinical knowledge of cardiologists or based on some feature selection algorithms.

3. All 13 variables included in prediction should be listed formally when they were mentioned for the first time, but in the section of methodoly the authors listed only 10 of them. The abbreviation of the variable might be presented in a more acadeical patterns. It is really confusion trying to understand what clinical items the variables “thal” “exang” and “ca” stand for.

4. It might be difficult for readers without background knowledge of ML to under the meaning of “SGD” “QDA” “EVCH” and “EVCS” in table 2. The full name of these algorithms should be listed.

5. Feature engineering are critically important for achieve better model perfomance. It is worth explaining how the researchers dealt with the outlets and missing data.

6. In the section of conclusion, the authors said"Our study provides compelling evidence that ML models surpass traditional models in CVD risk prediction". It might be better to calculate the accuracy, presicion, recall, F1 score and AUC-ROC of traditional predict models such as FRS, RRS, QRISK and other models the authors mentioned in the review section in the UCI dataset. It seems not so strict to reach the conclusion just based on the model performance reported in previous literature from other datasets.

7. A comprehensive evaluation of ML model includes the dicrimination and calibration of the model. AUC-ROC reflects the

discrimination of the model. What about the calibration of these models? If the brief scores of these 7 models can be reported, the article might be more interesting.

8. Were all algorhithms based on the same subset of variables(the selected 13 variables) in this article? Is it worth exploring whether different size of variable subset would bring better model performance?

9. The authors used different stragies to find out best combination of hyperparametres. What is the optimization target? The AUC-ROC or other indecis such as accuracy precision recall and f1-score

10. When accuracy precision recall and f1-score are reported the threhold of classification should also be reported. In some packages of sklearn, the default threshold for classification is 0.5. But in practice, in order to balance the accuracy and recall, different thresholds might be chosen.

11. As a more strict method to compare the AUC-ROCs of different models, the 95% confidence intervals of AUC-ROCs should be calculated to convince the readers that the diffenrence in model performance is/isn't statistically significant. Bootstrap might be used in the calculation of 95% CI.