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

Ramzi Guetari

Potential competing interests: No potential competing interests to declare.

The paper presents a comparative study of different Machine Learning models for predicting the 10-year risk of cardiovascular disease. The paper is well organized, but lacks a little depth, particularly in the state-of-the-art section. The authors should have presented the results of more work in the field before carrying out their own experiments. Moreover, their results are not compared with any other results. Nor is there any real novelty in what is presented. I encourage the authors to look at other work on the various algorithms presented and explain what they bring that's new. We also need a better choice of keywords, at least 5 in my opinion.

Sections 2.1 and 2.2, devoted to the prediction of CVDs by traditional models and by ML, are a little brief, but above all they lack a discussion highlighting the advantages of each of the two approaches and pointing out their disadvantages. If necessary, indicate when it is advantageous to use one or other of the two approaches, etc.

Paragraph 3.2 should explain how the choice of 14 features out of 76 was made, in other words, why the authors consider 62 features to be unsuitable for grading.

In the same paragraph, the idea of grouping classes 1, 2, 3 and 4 in the same class is an interesting one, but poses a problem. Diagnosis is generally made in order to choose a treatment, prescribe medication or carry out outright surgery. If we grouped the different degrees of severity into a single class, it would no longer be possible to guide a patient towards the right treatment.

Also in section 3.2, the authors constructed a correlation matrix to examine pairwise correlations between the selected variables. Variables with high positive or negative correlations (correlation>0.40 or correlation<-0.40) were considered potentially influential predictors of CVD risk. They then deemed these variables essential for their predictive model. The problem is that variables that are correlated (positively or negatively) provide the same information. Knowing one of the correlated variables, it is obvious to determine the effect of the other. In practice, the least correlated variables should be chosen, as they provide information on important criteria but communicate different information about the model. This seems to me to be the wrong choice.

An important question also arises: why was deep learning left out of this study? Models like RNN are powerful predictive tools. This would have added an extra dimension to the results, since it would have been possible to predict not only the risk, but also a time interval when the risk is major.

The idea of a comparative review in itself is a very good one, but it would be necessary to compare a good volume of existing work and draw up a balance sheet, or else contribute new approaches or new findings, which would have to be highlighted and compared with what already exists. On both counts, the paper has a few weaknesses.