Comments on "Better-Than-Chance Classification for Signal Detection"

I. Summary

This paper provides a simulation study to compare two different approaches to distinguish between two populations. The first approach involves developing a classifier to distinguish between the two populations and comparing the accuracy of the classifier to that of a purely guess-based classifier. The second is the more established practice of statistical tests for comparing two populations.

II. Comments

1) The quality of writing in the paper is poor. It is often hard to follow the arguments being made. Frequently in the paper unsubstantiated and questionable claims are presented as obvious facts. For instance, the following lines are on page 2: "Their argument is that the neural-net is able to learn the features that best separate the samples. Their examples, however, are lowdimensional (even if large-sample), and such feature learning is typically impossible in high-dimension" This statement is provided without any evidence and is also clearly incorrect. Neural nets have been shown to achieve state of the art performance in high-dimensional problems such as image classification.
2) In another place in section II B, it is stated that, "Because we focus on two-group testing under an independent sampling assumption, we know that a label-switching permutation test is valid." This statement is again made without a reference or arguments to support the claim. As this is the main tool used in the rest of arguments presented, the authors need to substantiate this claim.
   Moreover, the line "fix a test statistic T with a right tailed rejection region" is too general. The authors need to substantially improve clarity in their presentation.
3) The main technical limitation of this paper is that simulation studies cannot cover all scenarios. Moreover, the claims are all demonstrated exclusively on simulated data and not real data. The authors claim in the introduction, "we merely seek to demonstrate that accuracy-tests are never optimal, compared to high-dim two-group tests." It is impossible to conclude based on a finite number of simulations that accuracy-tests are never optimal.

4) Another concern with the paper is that the conclusions are expected and hence the evidence presented is not new. For population distinguishing tasks, tests designed for distinguishing hypothesis are expected to out-perform adhoc strategies that are based on comparing the accuracy of classifiers to accuracy of a purely chance-based guessing scheme.
5) Other issues:
    a) On page 1, column 2: It is hard to follow what is meant by “encode vocal/non-vocal information” and ”vocal / non-vocal stimulus”.
    b) The definition of V-fold CV accuracy on page 3 seems to be incorrect. Typically in each fold of the cross validation step, the predictor is trained on one part of the data and evaluated on the remaining part. However in the definition provided it seems like the predictor accuracy for each fold in evaluated on the same portion of the data as that on which it was trained.
    c) In the description of the permutation test on page 2, should (d) read “Repeat (b)-(c)” instead of “Repeat (a)-(c)”.

III. Recommendation

Given the poor presentation, the lack of any concrete technical results, and the insufficiency of the simulations to make a conclusive argument, I cannot recommend publication of this paper.