Review: "Better-Than-Chance Classification for Signal Detection"

This paper is a comparative study for two classes of signal detection methods. One is the classic two-group testing, which tests between H0: F = G vs H1: F ≠ G, where F and G are the feature distributions in two groups of data respectively. The other is a machine learning (ML) based test: train a supervised model and estimate accuracy of the fitted model; if the accuracy is better (with statistical significance) than random guesses, then the features bear signal. The simulated data start with the basic model in Eq. (1), then extend to non-Gaussian noise, correlated noise, and heteroscedastic samples. In the end, the authors conclude that the two-group test should always be preferred over the ML-based accuracy-test.

My main concern is how the findings in this work generalize since the comparison is soly based on numerical results.

First, only linear SVM and LDA are compared for the accuracy-test. Both models can have limited predictive power for complicated dataset (thus lower test power for the signal detection problem). One example is mentioned by the authors in reference [11], where tree-type ML model forms better accuracy-test than two-group test for the sparse mean-shift signal.

This brings to my second question. Most empirical comparison is done for model (1) and $\mu = ce$, except in Section-III.J the mixture model is simulated. Thus, in effect, sparse signal is not studied. I would expect Section IV neuroimaging is a sparse signal example, but please provide the sparsity level in the experiment, i.e., the total number of regions that are monitored under vocal/non-vocal stimulus. For such a sparse signal, would the tree-type model based accuracy-test be preferred according to [11]? Without more details in Section IV, readers cannot tell what accuracy-test is compared in this experiment.

I find the discussion in Section V very interesting. My high-level take of this problem is as follows. The ML-based accuracy-test can fit more generic data ("learn a good test statistic" according to the authors), thus should be a better choice for wider range of datasets. On the contrary, two-group tests are good at certain dataset, where the given test statistic fits well to the data distribution. Therefore, drawing a universal conclusion as this paper did seems counter-intuitive to me.