Reviewer: 1  
  
Comments to the Author  
2) a) Please give more context to the study, have similar studies been done before? If not in medical domains, then what about in other domains.  
  
2) b)   
i) What is up with the text marked in red? please fix  
ii) Figures need larger labels, especially figure 2.  
iii) Colour legends for figs 1 & 2 seem incorrect  
iv) please provide reference for "no free lunch" theorem  
v) some citations appear only as a question mark, please fix  
vi) please make captions more informative, for example explaining classifier abbreviations  
  
Reviewer: 2  
  
Comments to the Author  
1) The study methodology relies on a comparison between classification performance with a range of algorithms but where such outcomes are compared, there are vague qualitative generalisations from figures or tables with no attempt to apply any statistical rigour to the comparisons. This is compounded by the poor tables and extremely poor figures. The result is that the study has generated a number of interesting "observations" which are presented with very little supporting evidence. For example, "does not perform catastrophically" (Pg4 line 47) is a very subjective comment which very little objective support and very little meaning. The result is that while some of the conclusions from these observations would be significant, the lack of evidence makes them questionable.  
  
The presentation of the work is poor and in particular the Figures are unacceptable (very busy graphics, poor legends, colour keys missing in Figure 1 and 2). The academic content of the paper also seems sparse which is reflected in the small number of references. In particular there is little exposition of the meta-learning approach, which might be unfamiliar to general bioinformaticians. The authors mention the Netflix and Kaggle prizes three times with little reference to studies more relevant to the medical domain. This adds to the rather unfocused, vague feeling of the text.   
   
2)a) It is not clear how model performances were estimated. A data splitting approach has been used but it is not clear if re-sampling iterations were used. A re-sampling approach seems sensible, as a single estimate of performance is misleading. Related to this, some statistical comparison of performance is needed. If data splitting was used, the same data splits could be used with all models and a Friedman type test could be employed to compare outcomes.   
Figures and table must be re-worked. Perhaps the information in Figs 1 to 3 could be contained in tables?  
The text needs to have more domain relevant machine learning references. A better explanation of the meta-learning approach is needed and perhaps an algorithmic explanation of this approach and the ensemble methods would be helpful.  
b) Pg3 line 21 Logistic misspelt  
Pg 5 line 26 "the their"?  
  
  
  
  
  
  
Reviewer: 3  
  
Comments to the Author  
Major Comments  
Page 3, 2.3 Base Classifiers Section, 1st Column: Please explain in detail all the flags and their meaning used by the base classification algorithms.  
Page 3, Ensembling Process Section: Please explain in detail all the flags and their meaning used by the ensemble classification algorithms.  
Page 5, Please correct figure 1,2, each algorithm should be represented by a different color in the legend of the figure.  
Page 6, Please use bold color for max AUC for each dataset.  
  
Minor Comments  
Page 2, 1 Introduction Section, 1 st Column, 2nd Par.: Replace the links with references.  
Page 2, 2 Methods Section, 1st Column, 4th Par.: “…ensemble are run in R.” change to “…ensemble were implemented in R.”  
----------------------------------------