Comments to the Authors,

This manuscript provided a specific methylation-based classification signature for distinguishing different diffuse glioma subtypes with machine learning method based on TCGA dataset. The idea and the strategy is great and all the analysis was based on public dataset which provided a great way to redefined the molecular subtype of complex disease, meanwhile, the prediction model based on pre-selected features was quite promising according to the prediction result provided by the authors. The study was performed rigorously and the findings are interesting. I am not specialist on the molecular characteristic of Glioma. However, I still have several concerns on the statistic and machine learning process.

**Major Compulsory Revisions**

1, Since it is a bioinformatics or biostatistics paper, it is most suitable to published to the journal of bioinformatics or biostatistics. Now the authors would like to publish it in clinical epigenetics, I hope the author could provide completed R script or Rmd procedures, or the trimed dataset and script would be reorganized and upload to github so that everyone could validate the manuscript. Meanwhile, I don’t know how the authors get the mutation data from TCGA and how to defined negative mutation patients? The process how to get the raw data should be provded. It is easy to understand that the patents were defined to mutated when the mutation were found, however, it is hard to defined the patients to negative mutated since maybe the probed or sequencing was not covered by the technique, right?

2, The current model was highly suspected to be overfitted since the differential methylation analysis was conducted with all the samples before the fold-cross-sampling. Please make sure that the differential methylation analysis should be conducted only based on train-set samples in each fold, rather than the whole samples.

3, Could the author try to find a uniform methylation signature for multi-class classification and predict all these different subtype at the same time?

4, Could the author validate these biomarkers in different dataset which could be collected from GEO or some other HM450K database?

**Minor Essential Revisions**

1, Reference of Prediction Analysis of Microarrays (PAM) should be provided not only in the method section but also the first position it is occurred.

2, Page 17, Line 421, the threshold P-value should be provided.

3, Line 429-Line 430, the current feature selection pipeline was not quite explicit as present description, please make it more clearly. Since cross-fold sampling was applied in the PAM, how to deal with the features in each fold? And how to select the features eventually?

4, Between Line 422 and 437, why different threshold was applied?

5, *pamr* package should be cited in the reference list.

6, Line 480, PAM should be SVM? It is not quite clear the meaning for the sentence.

7, Line 478-479, whether this kind of feature selection will cause over-fitting?