Response to Reviewer 3 Comments

**Point 1:** What is the meaning of BDCEFS authors mentioned in the abstract? It is not the same the version placed at the manuscript against the one placed at the Review Report Form.

**Response 1:** Thanks for the comments. This was a mistake when we submitted the first version of the manuscript. We now have corrected it in the manuscript.

**Point 2:** It is vague to say that algorithm implementation takes advantage of multiple CPU cores to significantly accelerate the training process. Which is the minimum of computing power required to have good results with technique reported in the manuscript? How this can be measured against other wrapper methods?

**Response 2:** Thanks for the comments. The minimum of computing power is one CPU and enough memory space for the analysed dataset to reproduce the results reported in the manuscript. However, more CPUs will shorten the time of training. We have added the following description in the manuscript: “Our server had 80 Xeon E7-8870 CPUs cores (2.10GHz) and 120 GB of memory. The training time of BetaDCE for the leukemia dataset was 149.7 minutes on single CPU, and it was shortened to 4.1 minutes when 80 CPUs was used. As a comparison, the training time of RF+SFFS for the leukemia dataset was 380.3 minutes.” (line 140-143)

**Point 3:** Bulls-eye diagram with both bias and variance can help see the results of the algorithm implementation over the datasets involved, so it is suggested to do that in order to show the bias and variance trade-off complexity of the algorithm applied to the datasets involved.

**Response 3:** Thanks for the suggestion. We now have added the diagram (Figure 7) to show the bias and variance trade-off of the algorithm on leukemia dataset. The description is on line 238-242: “Figure 7 shown the bias and variance trade-off for BetaDCE on the leukemia dataset. As can be seen, the variance was dominant when number of nearest neighbors was small, and the bias became the main contribution to predictive results for big number of nearest neighbors. The cross entropy by BetaDCE is a trade-off between bias and variance.”

**Point 4:** It is encouraged to present a benchmark of wrapper methods for feature selection to demonstrate the outperformance described in the manuscript.

**Response 4:** Thanks for the suggestion. We now have added another two methods for comparisons in analysis of leukemia dataset: random forest with sequential floating forward selection, and LASSO. (as revised in Table 2, line 261-264). The corresponding description is revised: “The AUC of BetaDCE on the test set was 0.93, which was larger than that of Golub’s method (0.83), RF+SFFS (0.81) and LASSO (0.81).” (line 242-244), and “The performance of BetaDCE was compared with Golub’s method, as well as least absolute shrinkage and selection operator (LASSO) and random forest (RF) with sequential floating forward selection (SFFS).” (line 229-231). Meanwhile, SVM has been added in the analysis of XOR-like dataset and a description about it has been added in the manuscript: “To compare the performance, the algorithms were firstly evaluated on the raw training set. Then,
they were compared on the datasets with feature of noise by the percentage of noises detected. For KNN, the accuracy of leave-one-out cross validation was used to evaluate the performance of features, and the parameter K was optimized automatically in cross validation. For SVM, the radial basis function kernel was used as kernel function. The penalty parameter C and kernel coefficient were optimized simultaneously in cross validation. The accuracy of five-fold cross validation was used to evaluate the performance of features.” (line 174-180). Table 1 and Figure 5 were also updated