Status report

Tokyo Institute of Technology

Satoshi Kawaguchi

Kuze lab. meeting, 3rd July (2017)

Table of contents

- Increase the number of input variables（6 →12）
- BDT optimization and overtraining check（NTrees, MaxDepth, MinNodeSize, AdaBoostBeta）
- Comparison of the results

Result by the default BDT in LHeC

We changed the number of input variables (6 → 12).

Score >	N_S	N_B	Br[%]
0	6004	66359	8.61
0.05	5301	35862	7.18
0.1	4137	15658	6.10
0.15	2692	5363	5.52
0.2	1335	1447	5.85
0.25	394	213	7.93
0.3	29	7	26.4
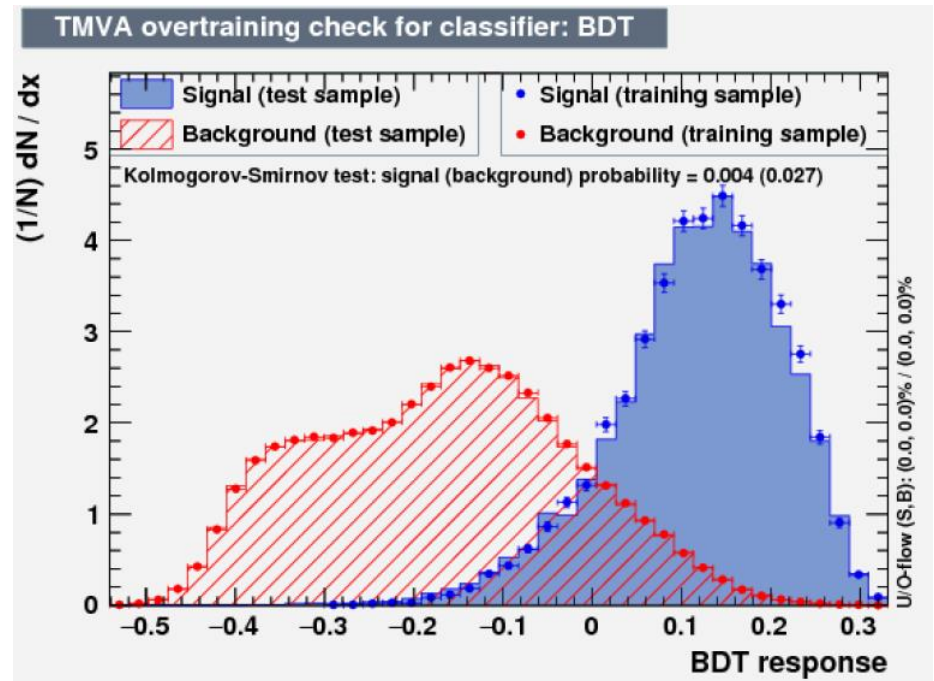
Branching ratio $(S/\sqrt{S+B})$:

$$Z = \frac{2692 \times \mathrm{Br}(h \to \not{E}_T)}{\sqrt{2692 \times \mathrm{Br}(h \to \not{E}_T) + 5363}}$$

In case of 2σ

⟹ $\mathrm{Br}(h \to \not{E}_T) \sim 5.52\%$

Variable ranking
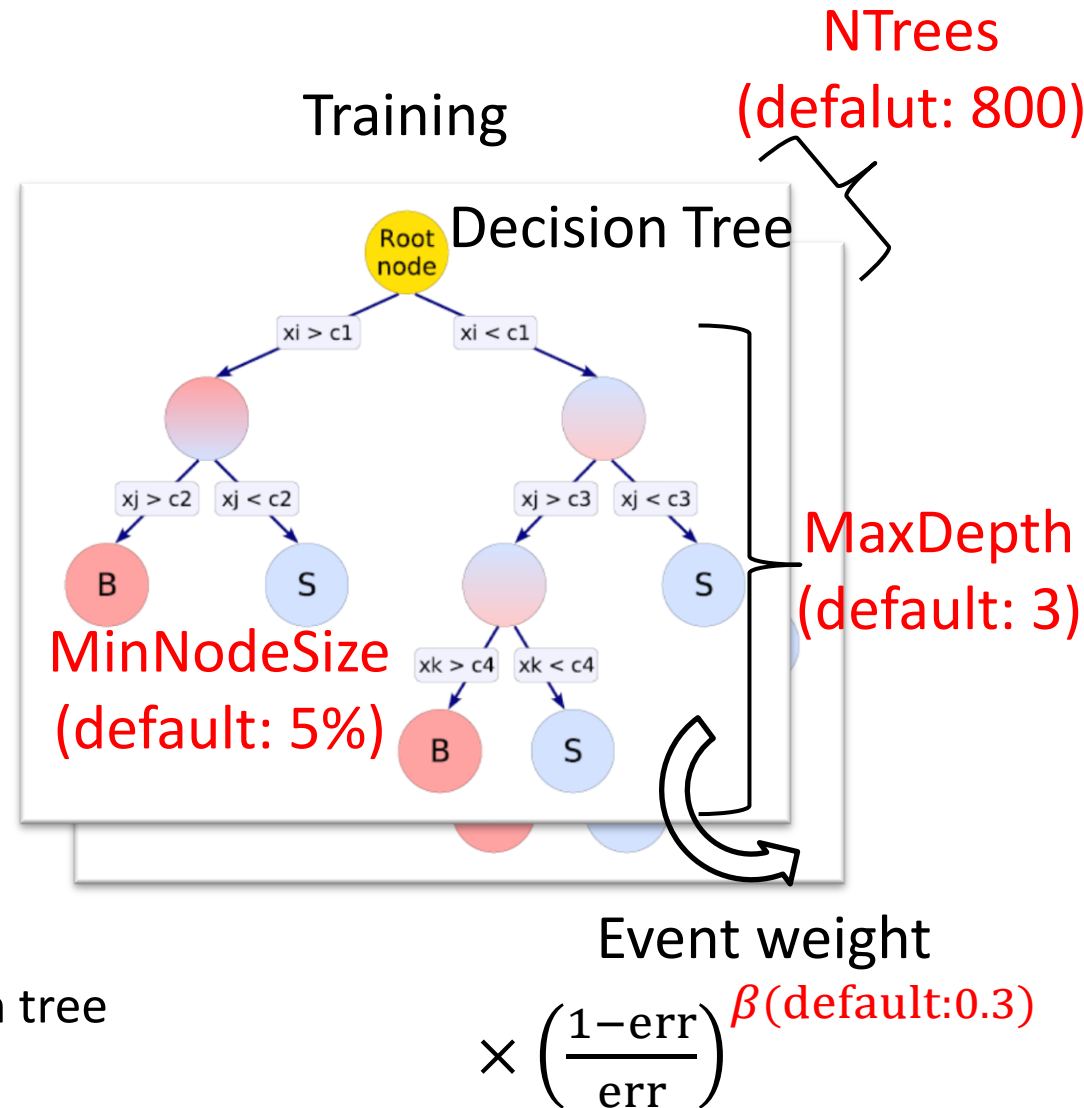
We compared the BDT performances when taking out each variable.
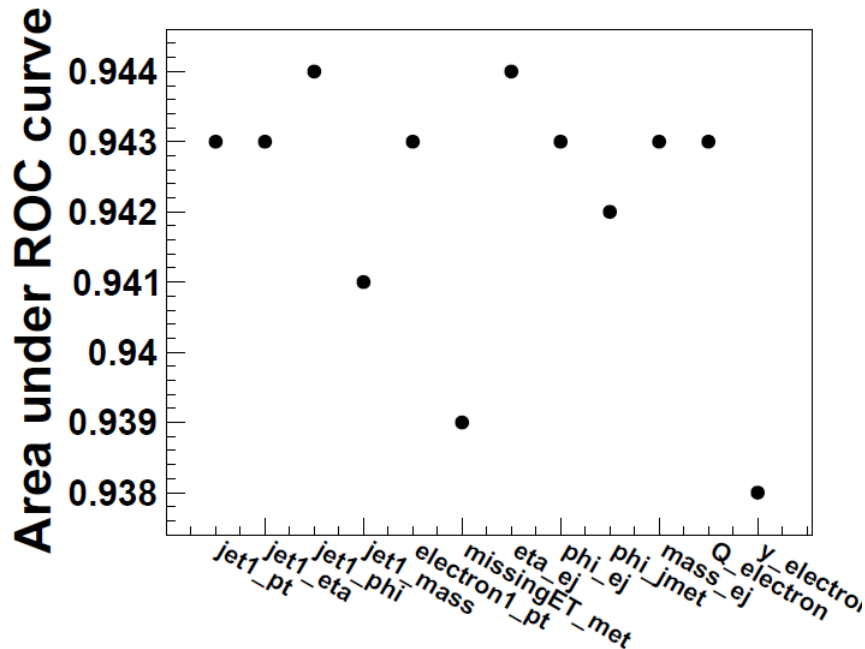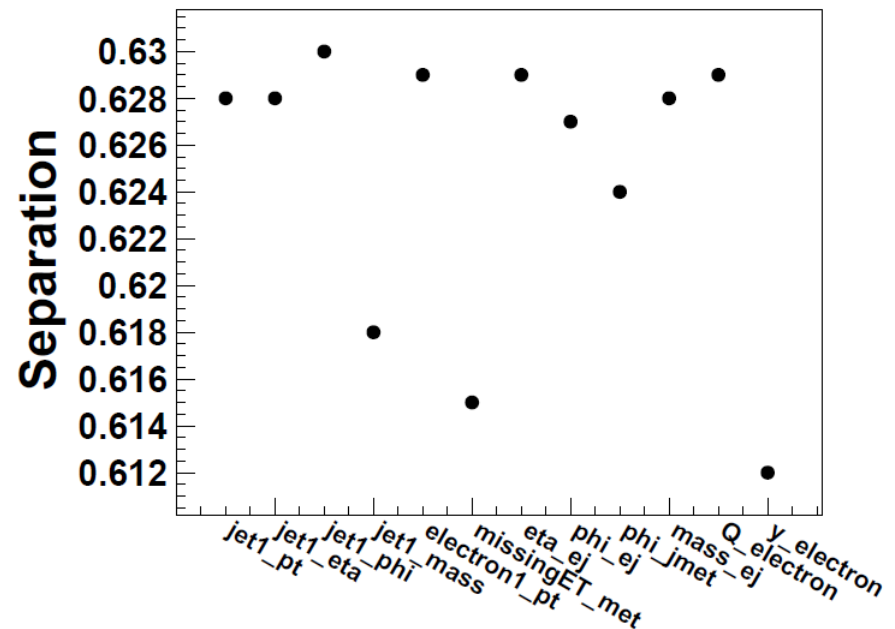
We didn't reduce the input variables in this analysis because the number of them wasn't so many.

BDT optimization

- NTrees(100-1600)
  →Number of trees in the forest

- MaxDepth(1-25)
  →Maximum depth of decision trees allowed

- MinNodeSize(0.5-9.5%)
  →Minimum percentage of training events in a leafnode

- AdaBoostBeta(0.1-2.5)
  →Learning rate for each decision tree

Training

Decision Tree

NTrees
(defalut: 800)

MaxDepth
(default: 3)

MinNodeSize
(default: 5%)

Event weight

$$\times \left(\frac{1-\text{err}}{\text{err}}\right)^{\beta(\text{default}:0.3)}$$
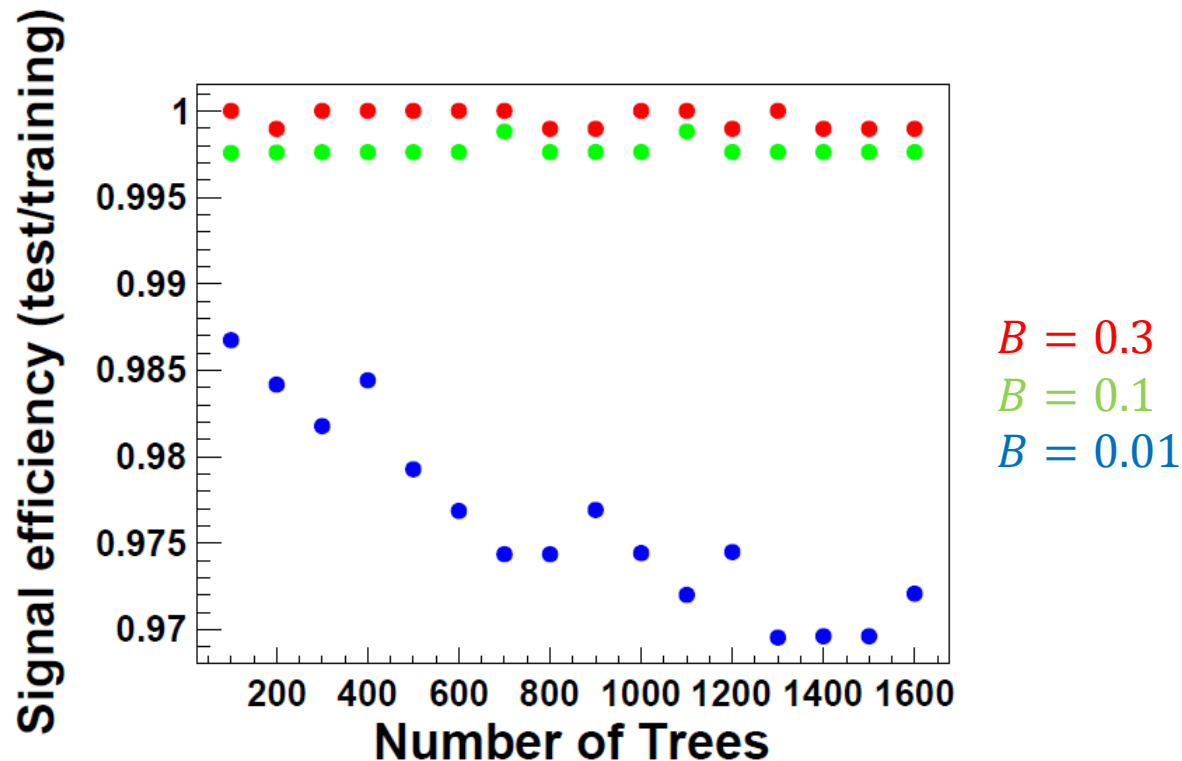
Optimization of 'NTrees'(default: 800)

Training MC events changing 'NTrees' between 100 and 1600
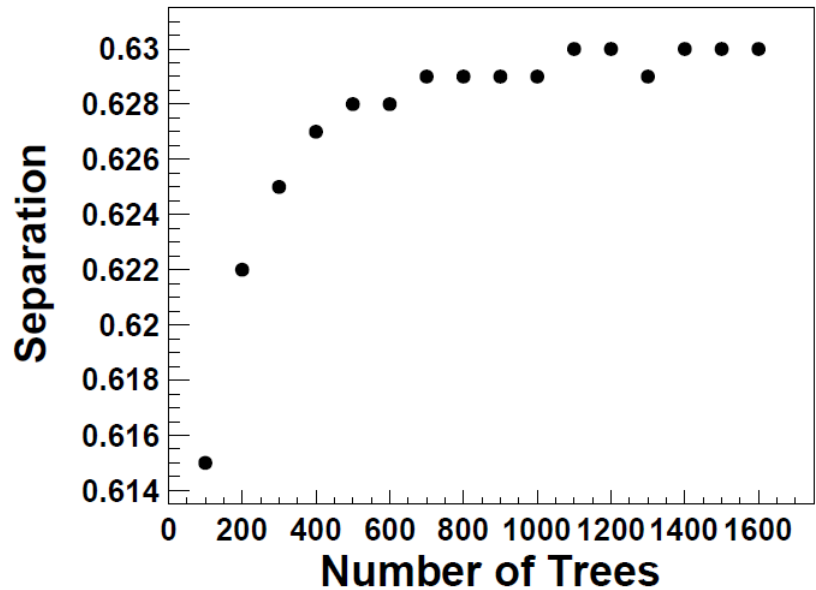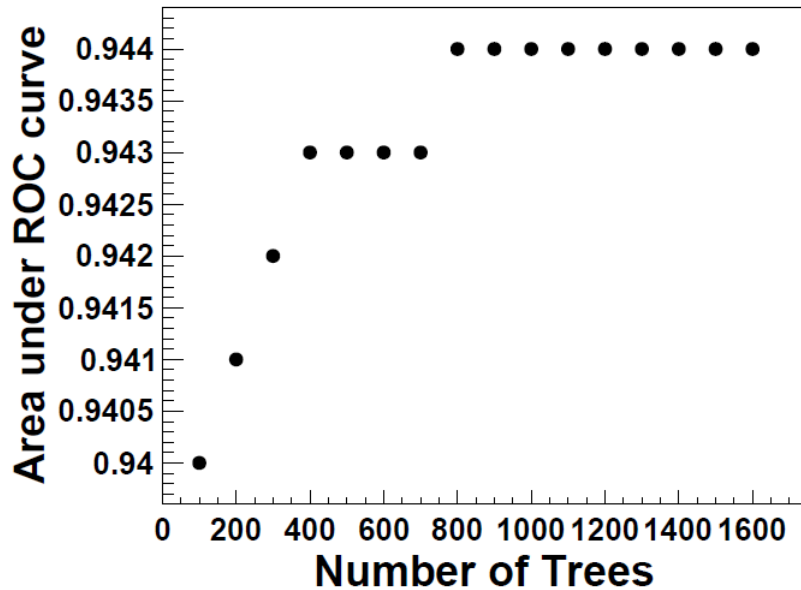
From these plots, we would use 'Ntrees' > 1000.
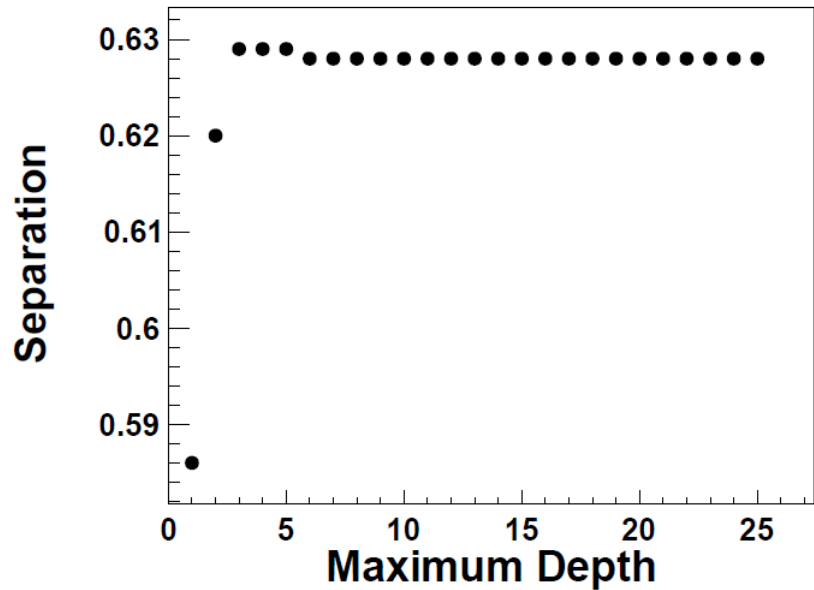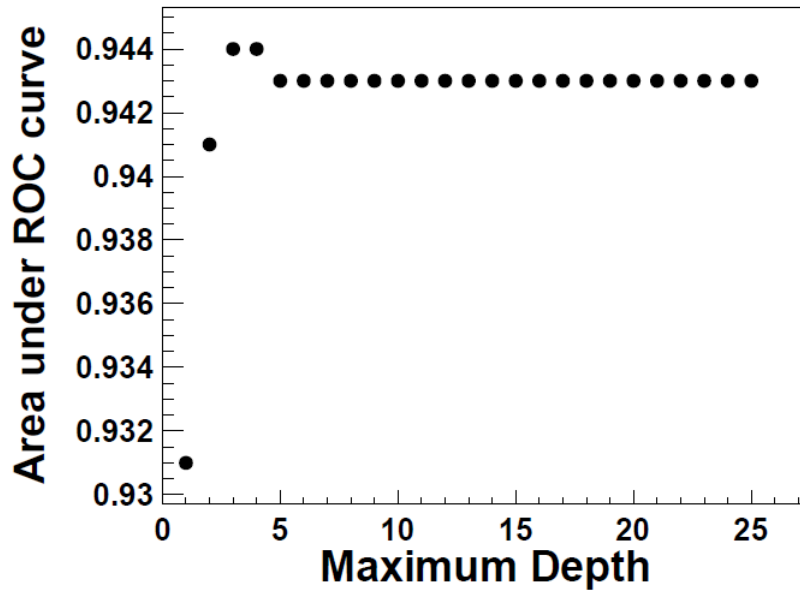
Optimization of 'NTrees'(default: 800)

We compared the ratios of the signal efficiency (test/training) at each 'Ntrees'.

$B = 0.3$
$B = 0.1$
$B = 0.01$

From these plots, we used 'NTrees' = 1200.
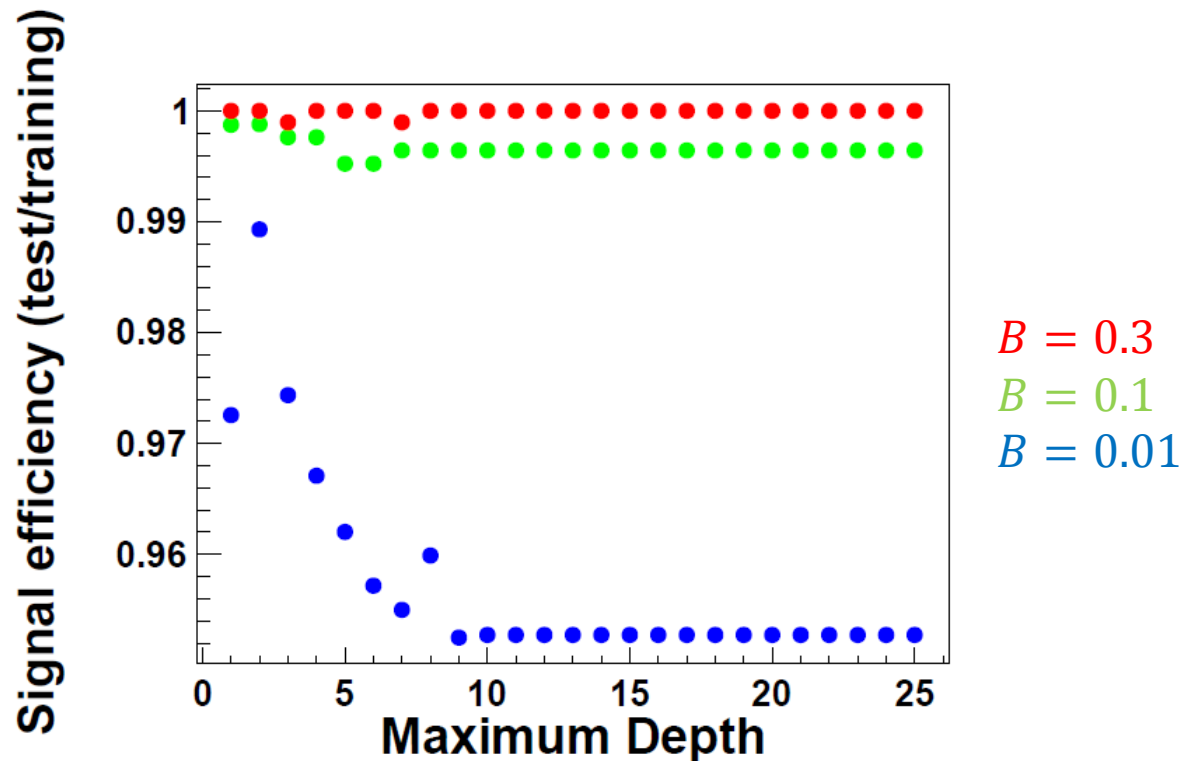
Optimization of 'MaxDepth'(default: 3)

Training MC events changing 'MaxDepth' between 1 and 25

From these plots, we would use 'MaxDepth' = 3 or 4.
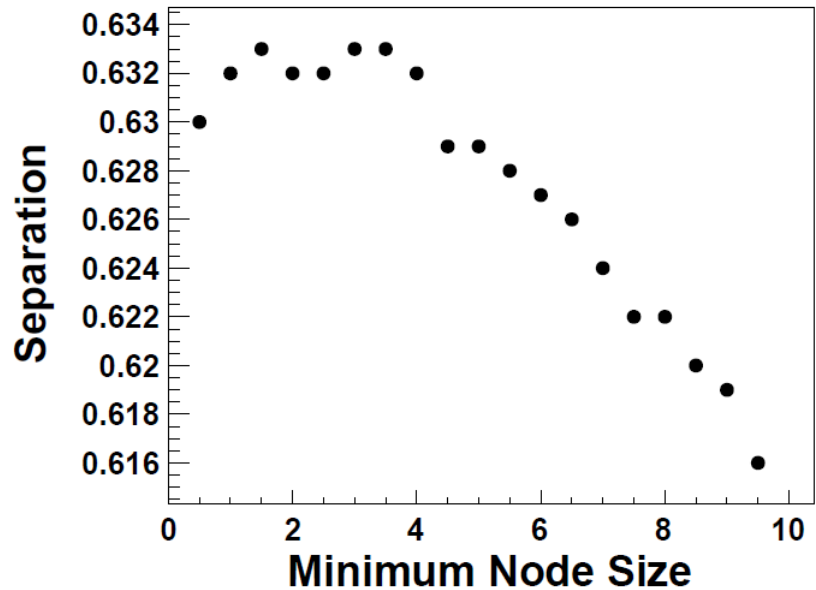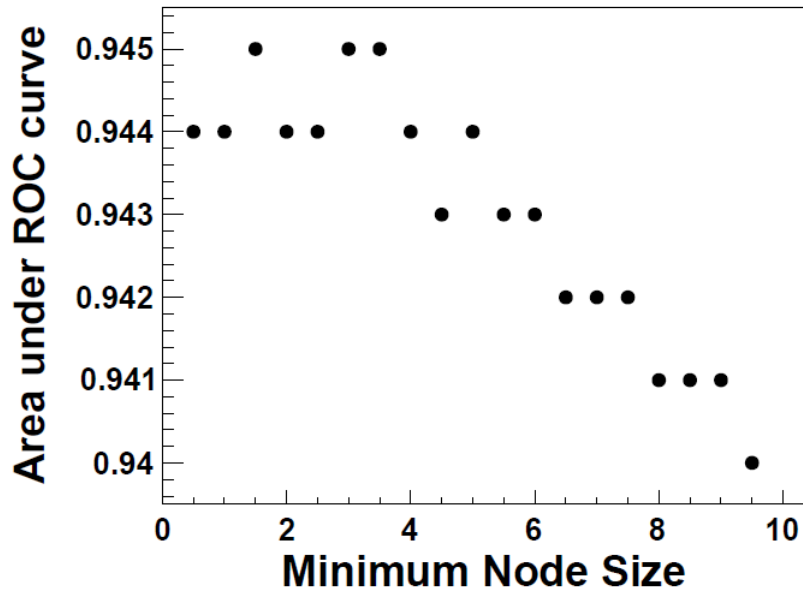
Optimization of 'MaxDepth'(default: 3)

We compared the ratios of the signal efficiency (test/training) at each 'MaxDepth'.

$B = 0.3$
$B = 0.1$
$B = 0.01$

From these plots, we used 'MaxDepth' = 3.

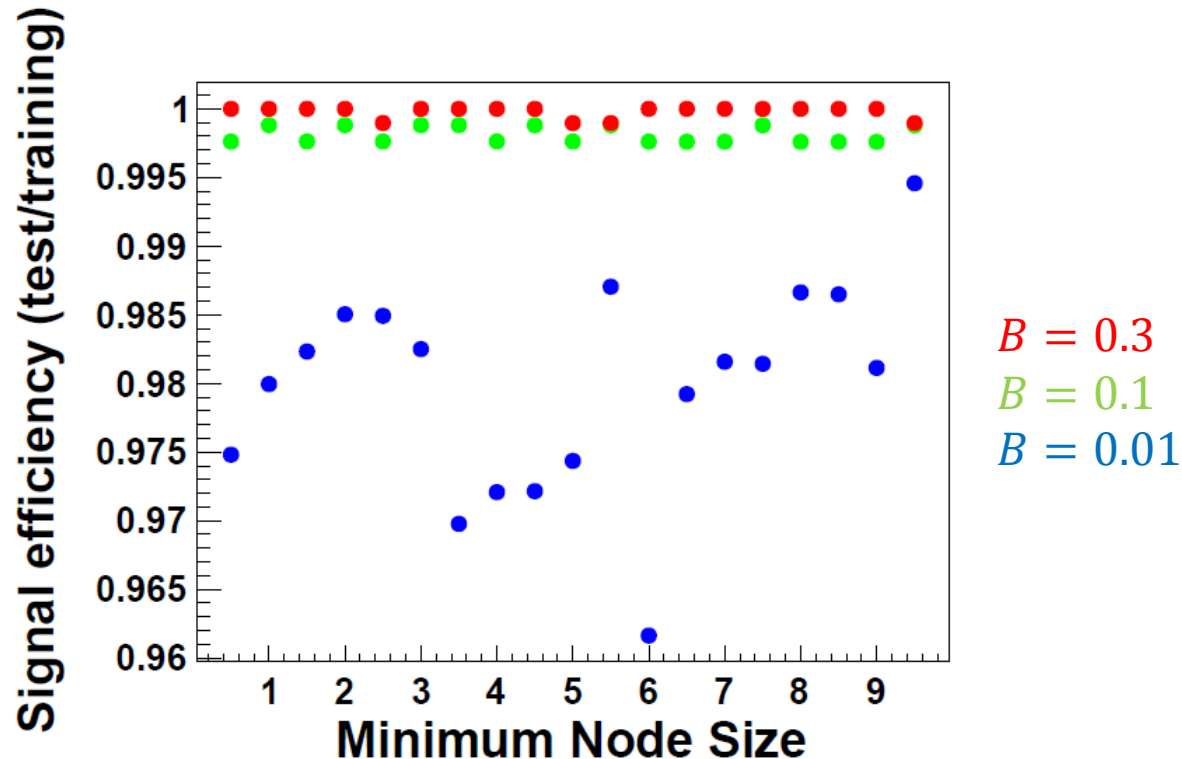# Optimization of 'MinNodeSize'(default: 5%)

Training MC events changing 'MinNodeSize' between 0.5 and 9.5%

From these plots, we would use 'MinNodeSize' = 1.5, 3.0 or 3.5.
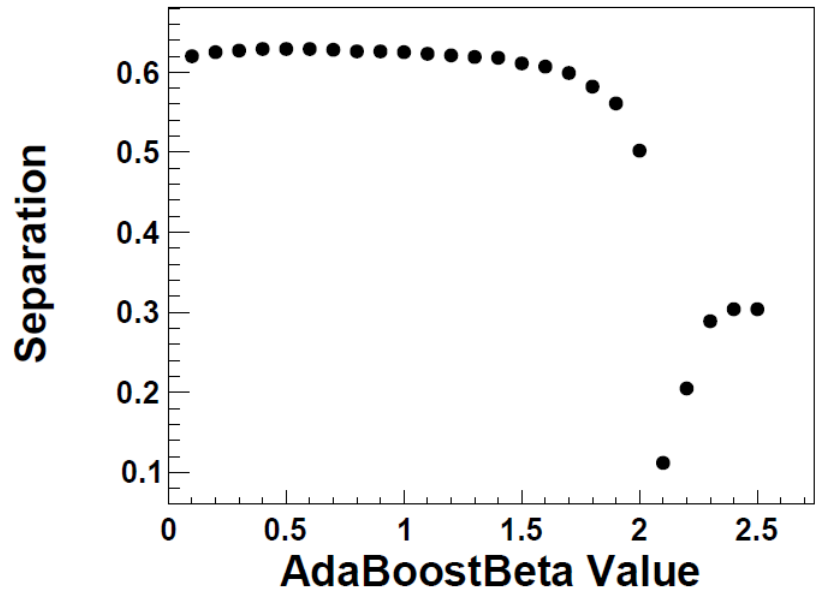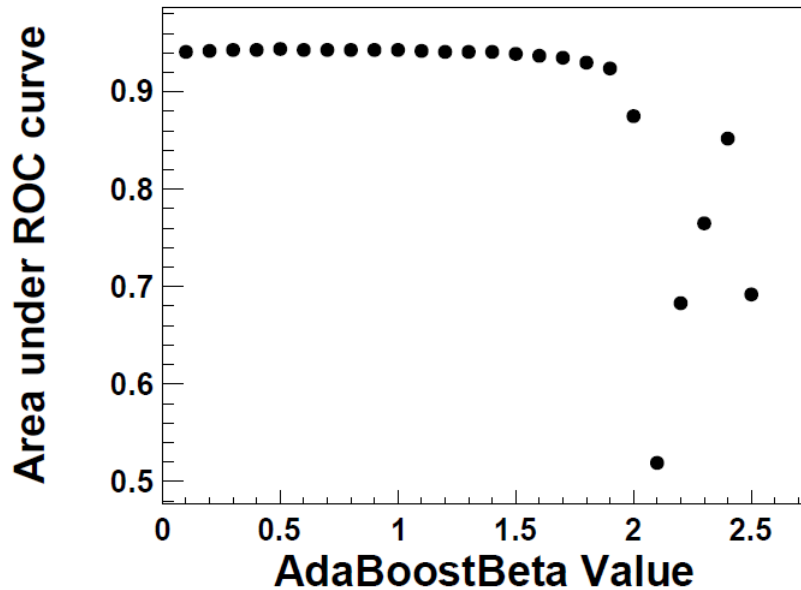
Optimization of 'MinNodeSize'(default: 5%)

We compared the ratios of the signal efficiency (test/training) at each 'MinNodeSize'.

$B = 0.3$
$B = 0.1$
$B = 0.01$

From these plots, we used 'MinNodeSize' = 3.

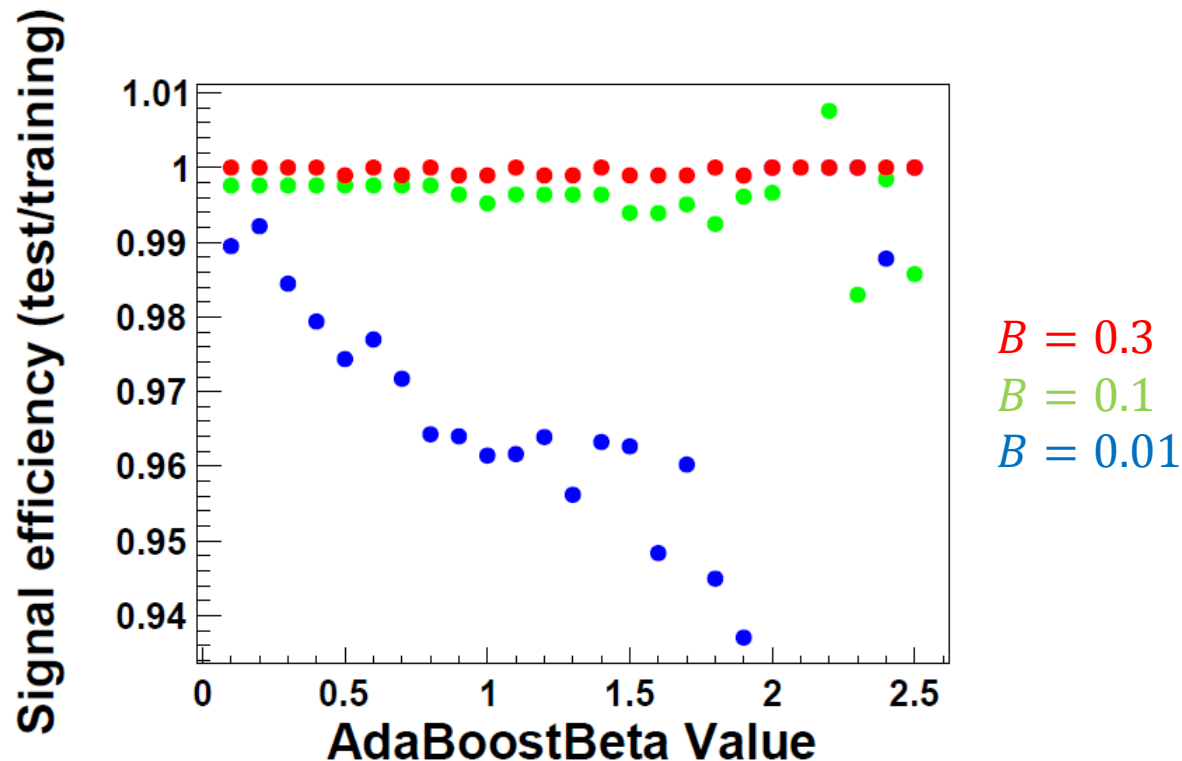# Optimization of 'AdaBoostBeta'(default: 0.5)

Training MC events changing 'AdaBoostBeta' between 0.1 and 2.5

From these plots, we would use 'AdaBoostBeta' < 2.0.

Optimization of 'AdaBoostBeta'(default: 0.5)

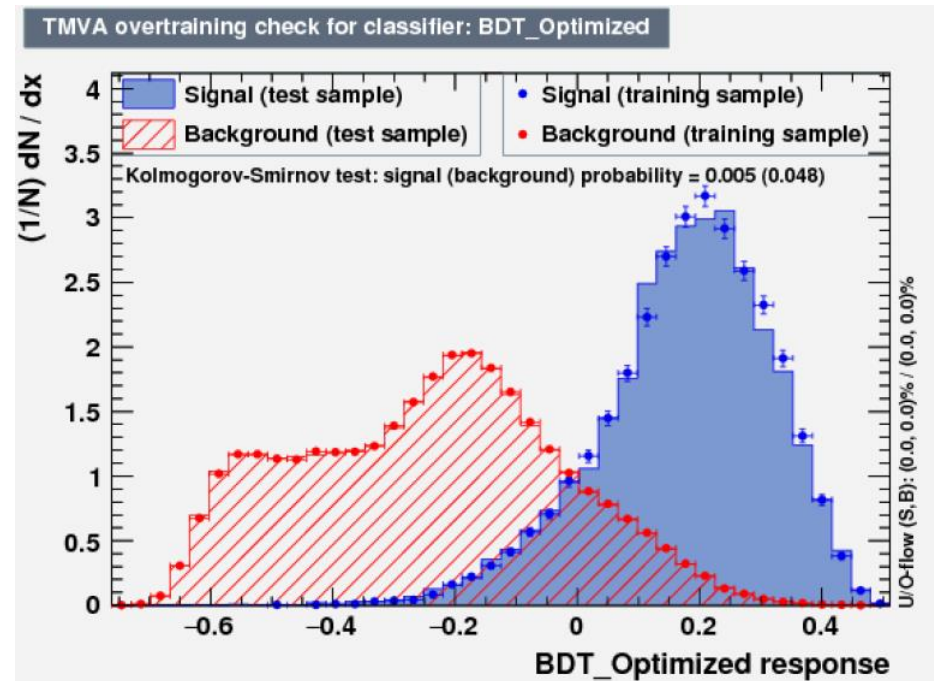We compared the ratios of the signal efficiency (test/training) at each 'AdaBoostBeta'.

$B = 0.3$
$B = 0.1$
$B = 0.01$

From these plots, we used 'MinNodeSize' = 0.2.

Result by the optimized BDT in LHeC

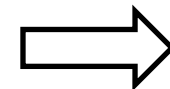We analyzed the signal and backgrounds by the optimized BDT.

Score >	N_S	N_B	Br[%]
0.1	5011	29054	6.84
0.15	4181	16162	6.13
0.2	3179	7788	5.62
0.25	2143	3286	5.44
0.3	1272	1267	5.76
0.35	598	399	7.02
0.4	185	91	11.5

Branching ratio $(S/\sqrt{S+B})$:

$$Z = \frac{2143 \times \mathrm{Br}(h \to \not{E}_T)}{\sqrt{2143 \times \mathrm{Br}(h \to \not{E}_T) + 3286}}$$

In case of 2σ

$$\Longrightarrow \quad \mathrm{Br}(h \to \not{E}_T) \sim 5.44\%$$

※Default BDT: 5.52%

Summary

- We increased the number of input variables（6 →12）.

- When we optimized the BDT configurations,
  $\mathrm{Br}(h \to \not{E}_T) = \color{red}{5.44}\%$ at 2σ level(default BDT: 5.52%)