ν_e identification in the Near Detector

Ciro Riccio
Fermilab
(Dated: September 24, 2014)

This note describes the ν_e identification in the NOνA Near Detector (ND) using Boosted Decision Trees technique to combine various reconstructed variables.

Contents

I. Introduction 3

II. BDT Training and Testing 3
A. Input variables 3
B. Training and Testing 7
C. Performance Evaluation 15

III. Conclusion 18

References 18

I. INTRODUCTION

The NOνA Near Detector (ND) is used to study the background compositions and contributions for $\nu_\mu \rightarrow \nu_e$ oscillation analysis. One source of background is due to the beam ν_e events, therefore it is very important to identify this events. For this study, the Boosted Decision Trees (BDT) has been used to produce a particle identification technique (PID). BDT is a multivariate classification algorithms implemented in the Toolkit for Multivariate Data Analysis (TMVA) [1]. This Toolkit has been used to extract the smaller signal from large data and maximize the separation power between signal and background. Thus BDT based PID was applied on 1779 ND Monte Carlo (MC) files, for a total of 8.90×10^{19} Proton On Target (POT),using the same selection as for Far Detector (FD) events[2]. The events are normalized to an exposure of 8.64×10^{19} POT, which is about 3 months of ND data taking.

II. BDT TRAINING AND TESTING

A. Input variables

The BDT was initially trained with 15 variables. Since the BDT provides as output the ranking of input variable it is possible to reduce their number finding the better choice of variables. The variables used are:

- the ratio of number of cells associated to the longest track over number of cells N_{cells} $\left(\frac{N^{track}_{cells}}{N^{slice}_{cells}}\right)$;
- MIP fraction $\frac{N^{mip}_{cells}}{N^{slice}_{cells}}$;
- fraction of energy in 2σ road ($Efrac_in2sigmaHOR$).
  The ν_e should have relatively narrow transverse shower than the π^0. Thus, to calculate the transverse shower width, was used the idea of the "Hits In Road" [2]. For each plane, the energy-weighted center is used as the centroid position $\left(\frac{\sum E^i_{cell} \cdot X^i_{cell}}{\sum E^i_{cell}}\right)$, then sum energy for the cells around the center by $2 \cdot \sigma$ $\left(\sum E^{\Delta X(\Delta Y)<2\cdot\sigma}_{cells}\right)$, where the width σ is 2 cm. Fraction of energy in 2σ road is the ratio of $\sum E^{\Delta X(\Delta Y)<2\cdot\sigma}_{cells}$ over $\sum E_{cells}$.
- maximal fraction of energy in 2 planes ($Efrac_2slides$).
  The maximal fraction of energy deposited in these 2 planes was found sliding the window of 2 planes along the beam direction. This variable reflects the condensity of the longitudinal shower;
- number of MIP cells (N_{mip}). The MIP cells are defined by requiring $100 <$ PECorr < 245, where PECorr is the corrected photo-electrons;
- recostructed energy $E_{reco} = \sum E_{cells}$;
- maximal fraction of energy in 6 planes ($Efrac_6slides$);
  Same as $Efrac_2slides$ but using 6 planes as a window;
- number of cells N_{cells};
- energy balance between 2 most energetic prongs eveluated for 2D prongs ($E_{balance2D}$) using this formula

$$E_{balance} = \frac{E_{prong1} - E_{prong2}}{E_{prong1} - E_{prong2}} \tag{1}$$

  where $E_{prong2} = 0$ for events having only one prong.
- energy balance between 2 most energetic prongs eveluated for 3D prongs ($E_{balance3D}$).
  Same as $E_{balance2D}$ but in 3 dimensions;
- fraction of energy out 3σ road ($Efrac_out3sigmaHOR$);
- Length of track L_{track};
- fraction of energy in first 20 planes ($Efrac_20slides$);
- number of 2D prongs ($N_{2Dprongs}$);
- number of 3D prongs ($N_{3Dprongs}$);

The distributions of these 15 input variables for beam ν_e signal, CC bkg and NC bkg are shown in figs. $1 - 3$.

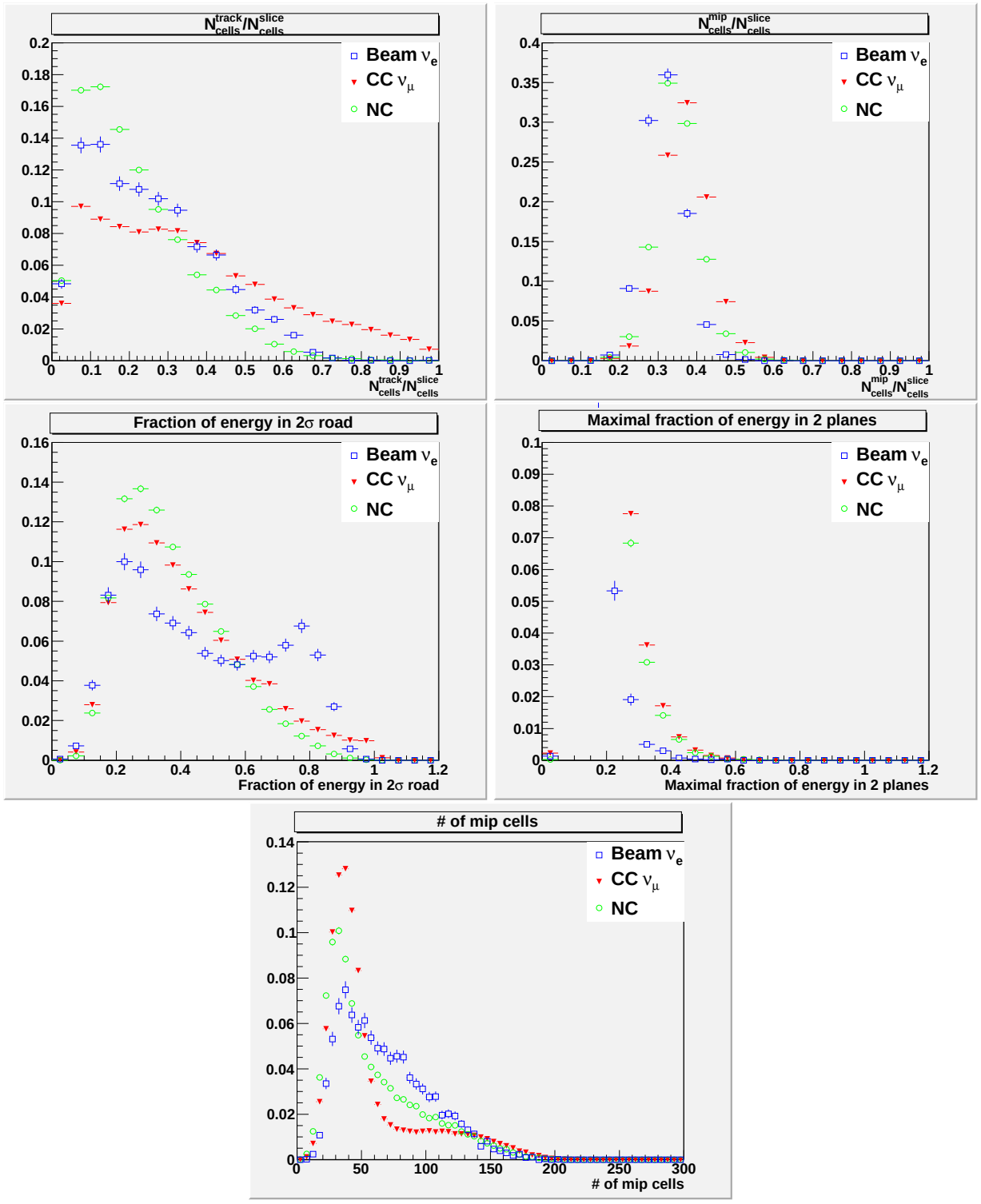


FIG. 1: The distributions of BDT input variables for beam ν_e, CC bkg and NC bkg.

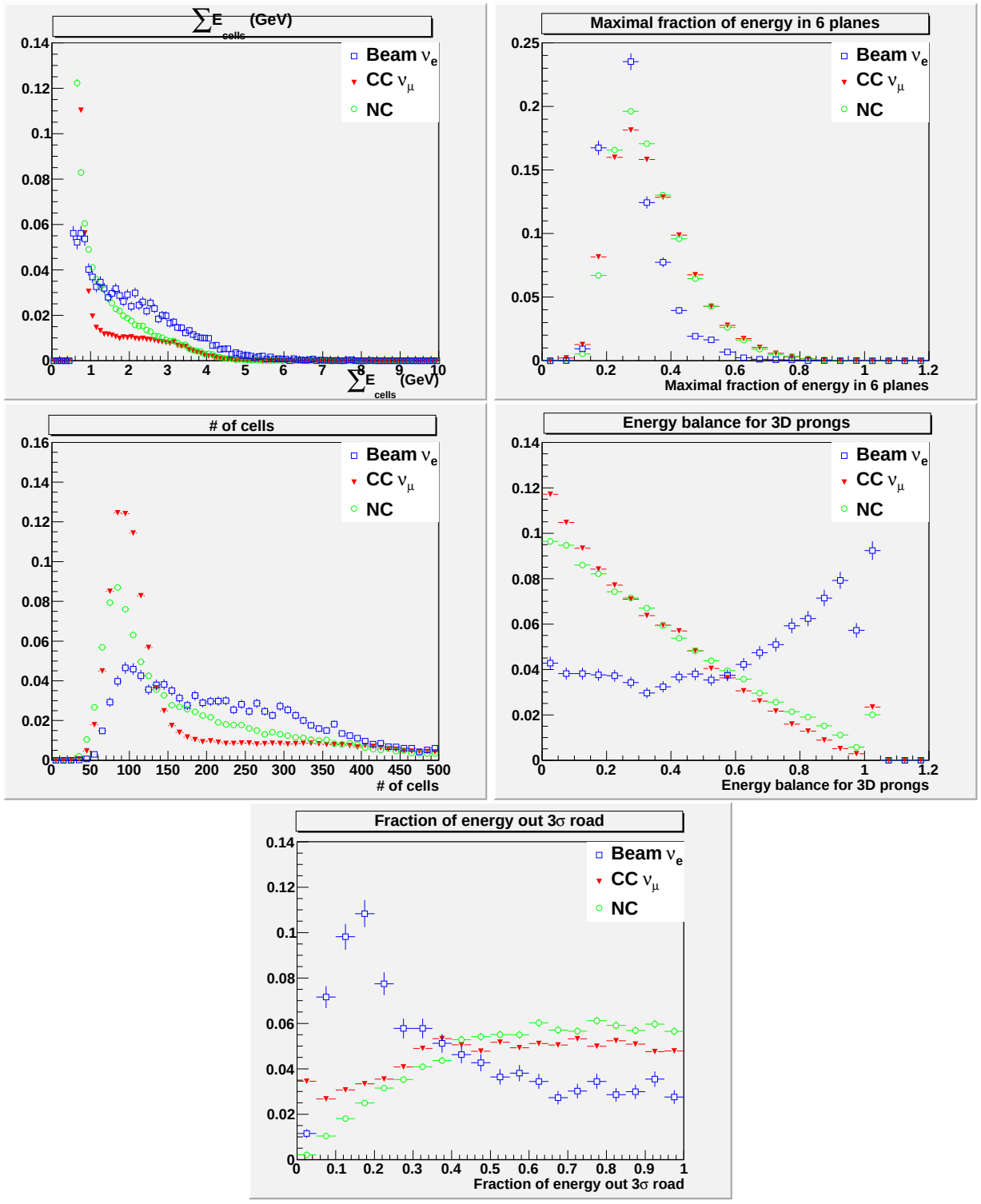


FIG. 2: The distributions of BDT input variables for beam ν_e, CC bkg and NC bkg.

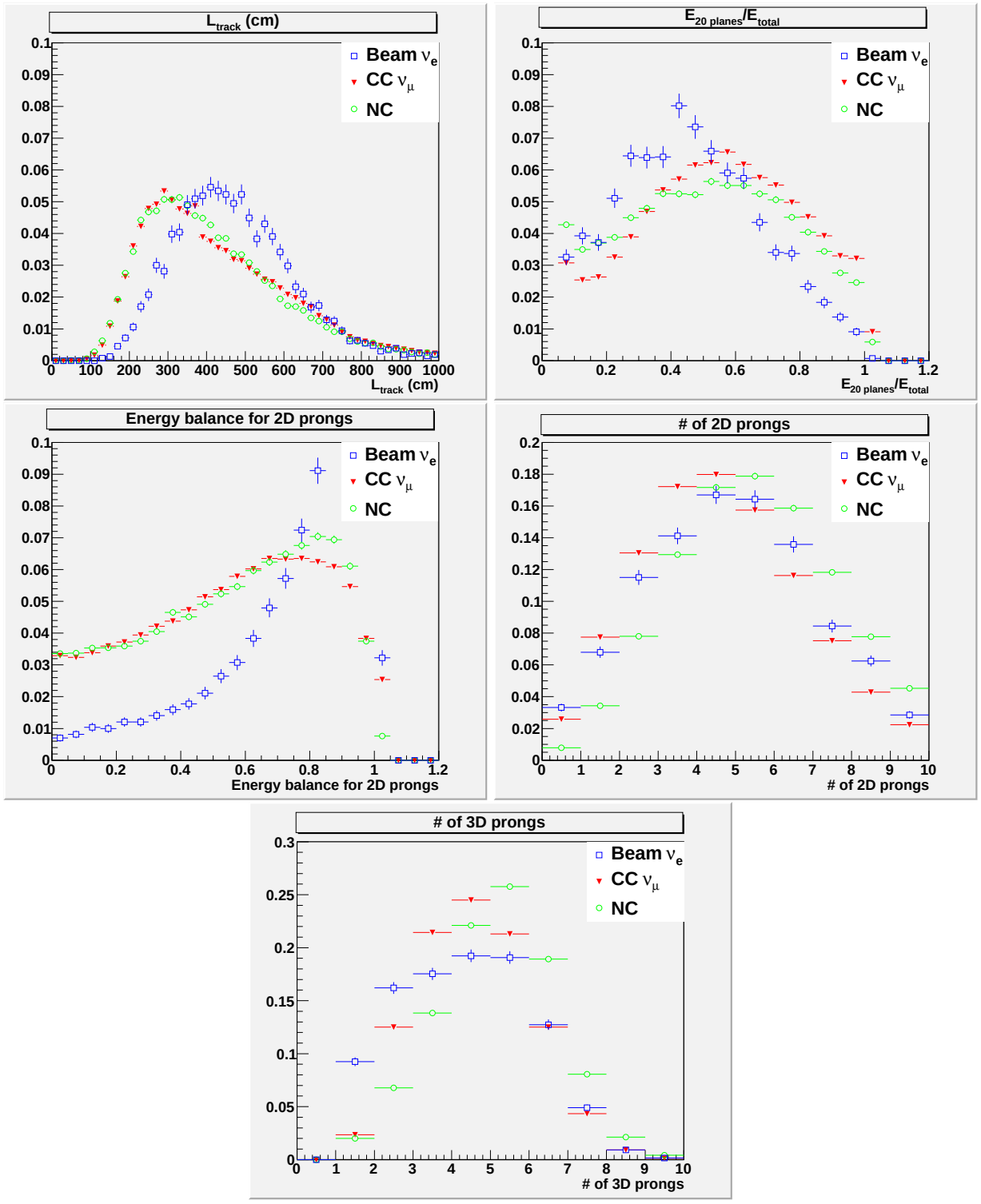


FIG. 3: The distributions of BDT input variables for beam ν_e, CC bkg and NC bkg.

B. Training and Testing

The TMVA of version 4.1.3 has been used to do the training, the number of used signal and background events are summarized in the Table I.

	signal	background
Training	4177	4177
Testing	57575	57575

TABLE I: Number of signal and background events used for BDT training and testing. These are raw number of MC events without oscillation probability and POT normalization.

In fig. 4 are shown:

- The overtraining check plot for BDT (top left). The Kolmogorov-Smirnov test probability between the training and testing samples is measured to be 0.0.73 for signal and 0.03 for background.
- The signal efficiency versus the background rejection (top right). More than 90% of the background can be rejected obtaining 60% signal efficiency.
- The linear correlation coefficients for signal (bottom left) and background (bottom right).

The ranking for the contributions of the input variables is shown in Table II.

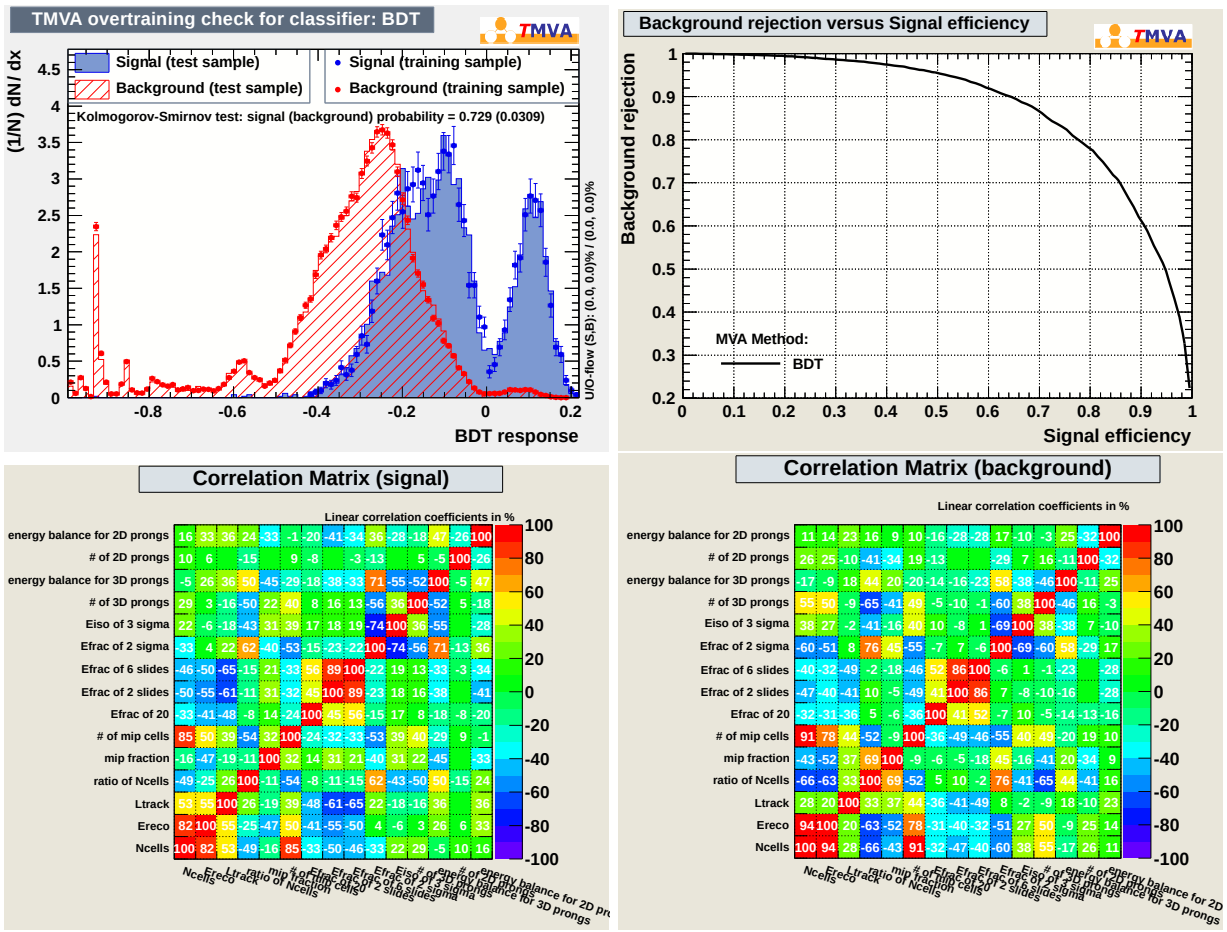


FIG. 4: Overtraining check for BDT (top left), background rejection versus signal efficiency rate (top right), linear correlation coefficients for signal (bottom left) and linear correlation coefficients for background (bottom right) in case of 15 Variables.

At this stage referring to variables correlation and ranking the number of variables was reduced starting from the most correlated and the less important. Therefore the case in which 9, 10, 11 and 12 variable was analyzed.

Variable	Importance (%)
$\frac{N_{cells}^{track}}{N_{cells}^{slice}}$	79
$\frac{N_{cells}^{mip}}{N_{cells}^{slice}}$	9.3
Efrac_in2sigmaHOR	3.9
Efrac_2slides	1.7
N_{mip}	1.4
$\sum E_{cells}$	0.98
Efrac_6slides	0.83
N_{cells}	0.76
$E_{balance3D}$	0.46
Efrac_out3sigmaHOR	0.42
L_{track}	0.30
Efrac_20slides	0.19
$E_{balance2D}$	0.14
$N_{2Dprongs}$	0.10
$N_{3Dprongs}$	0.07

TABLE II: Ranking in case of 15 input variables for BDT.

In figs. $5-8$ are shown the overtraining check plot and the linear correlation coefficients for signal and background in case of 9, 10, 11 and 12 variables.
The ranking for the contributions of the input variables in case of 9, 10, 11 and 12 variables are shown in Tables III $-$ VI.

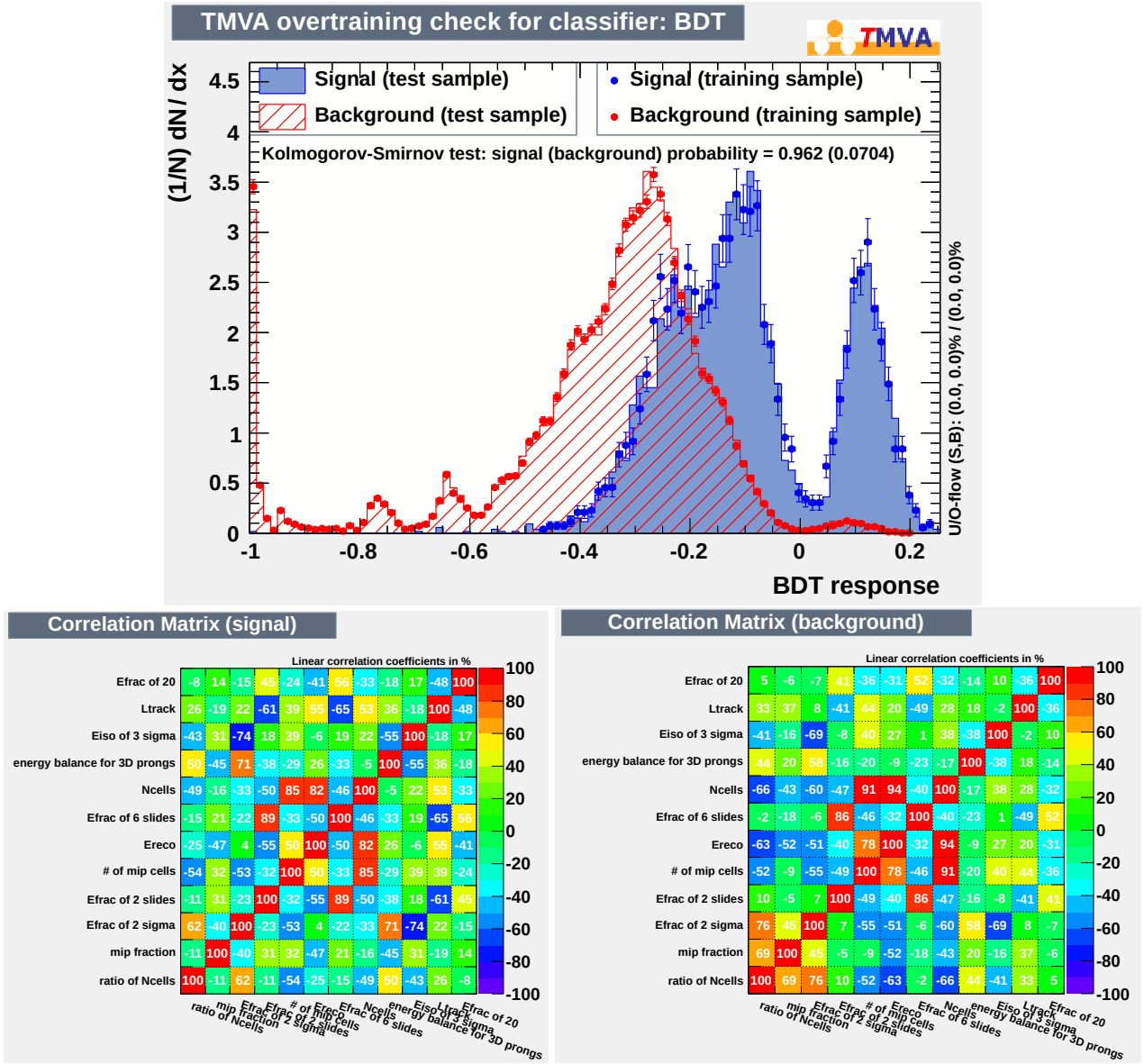


FIG. 5: Overtraining check plot (top center) and linear correlation coefficients for signal (bottom right) and background (bottom left) in case of 12 Variables.

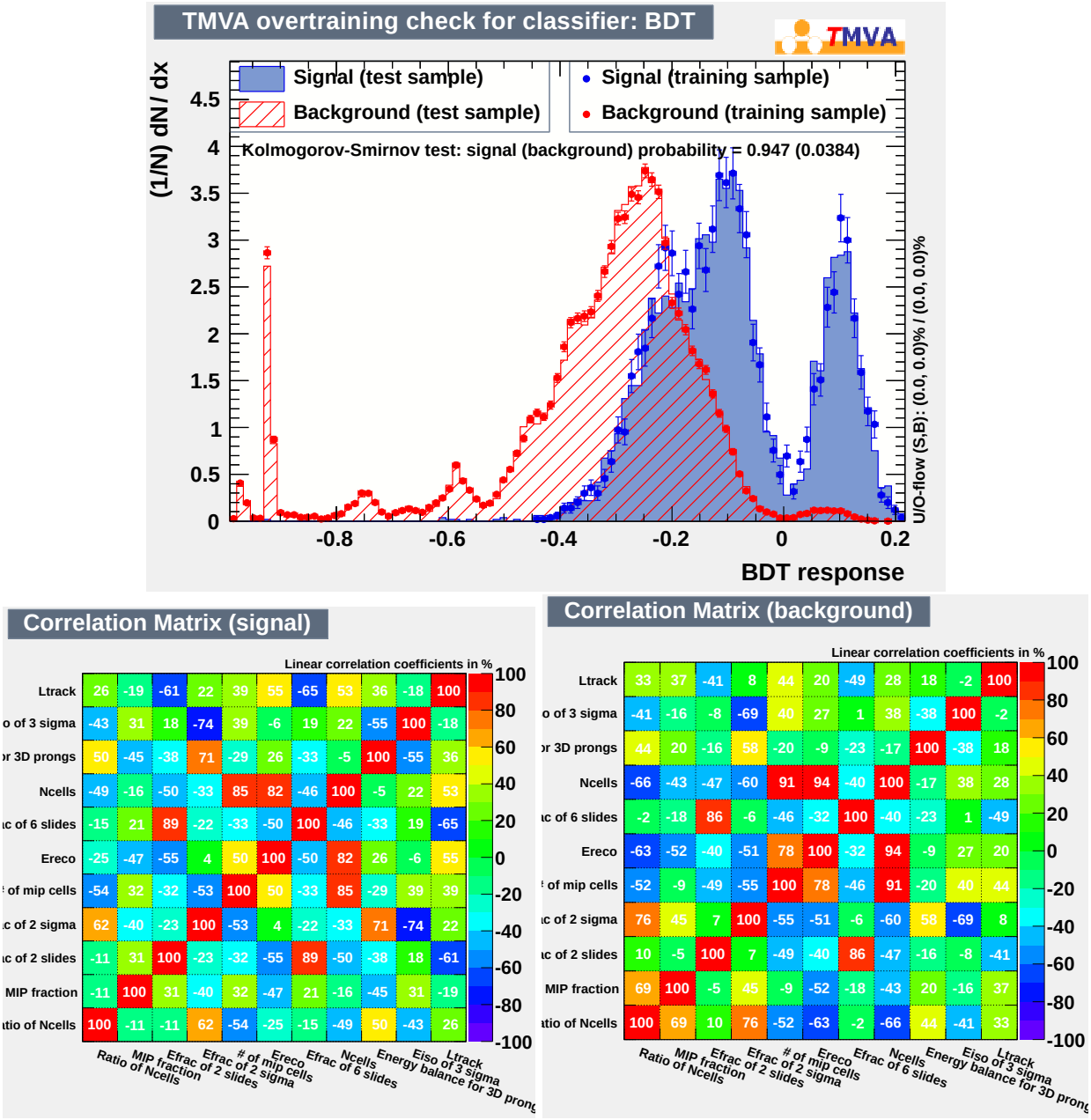


FIG. 6: Overtraining check plot (top center) and linear correlation coefficients for signal (bottom right) and background (bottom left) in case of 11 Variables.

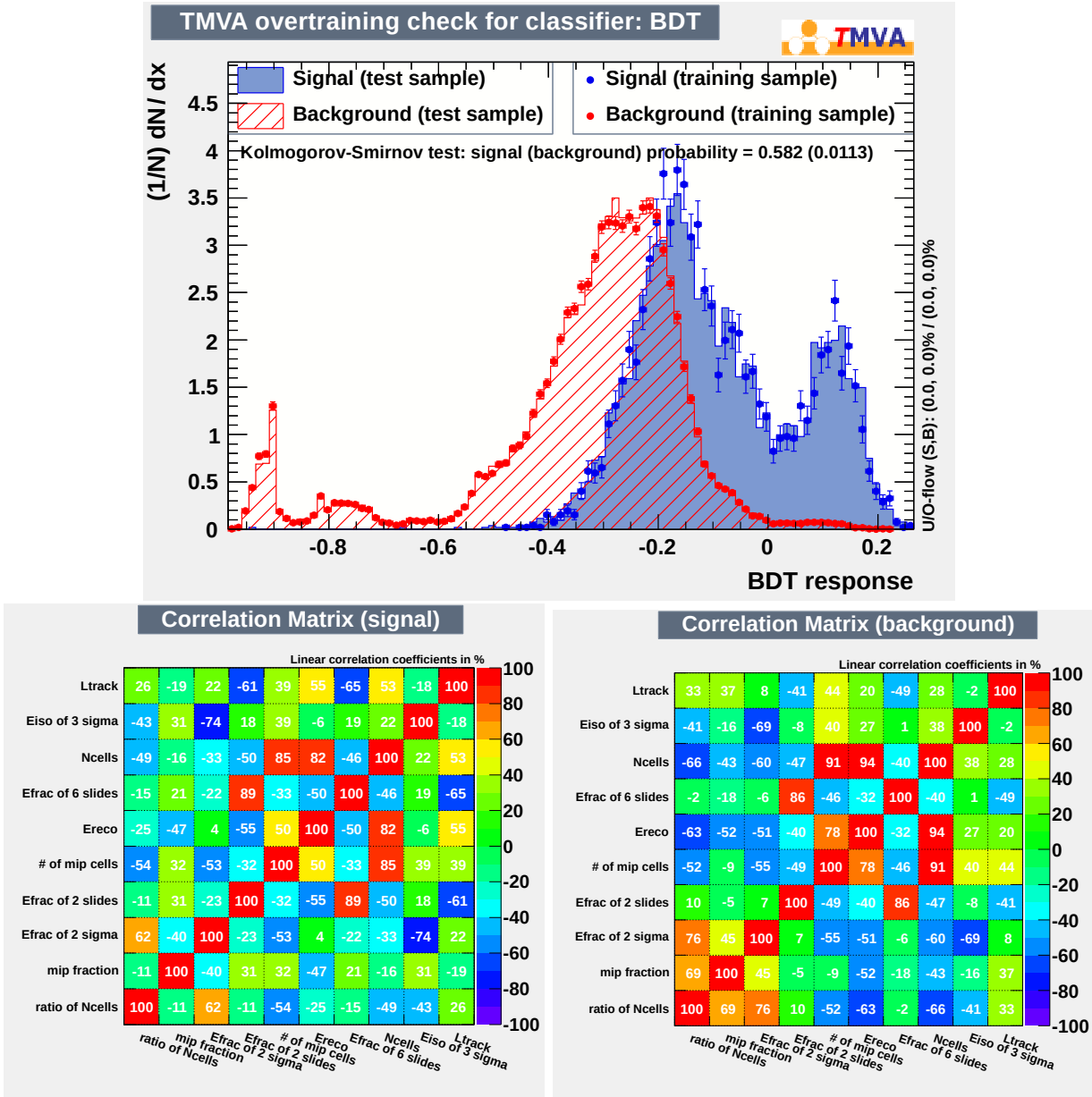
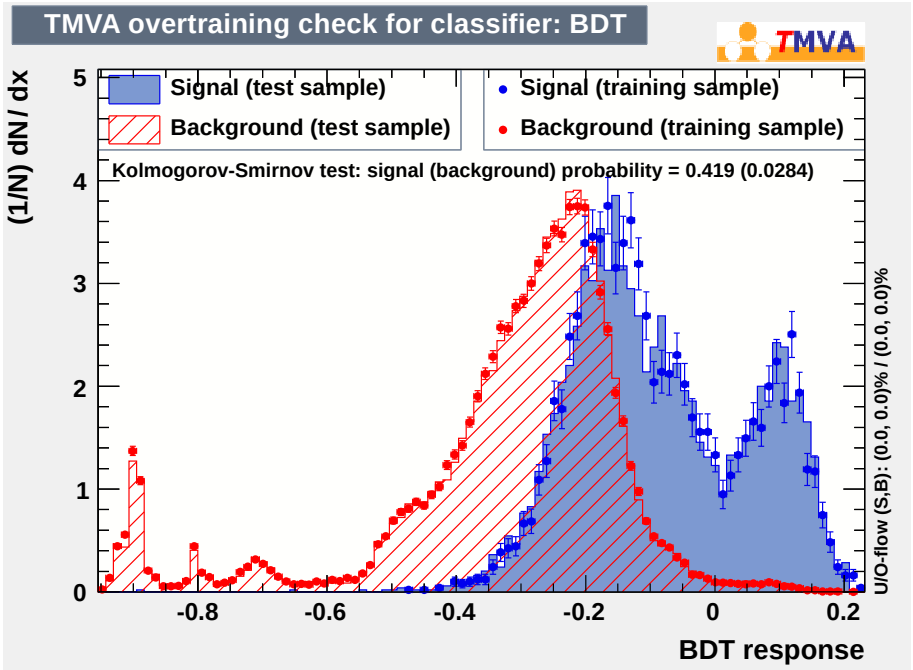
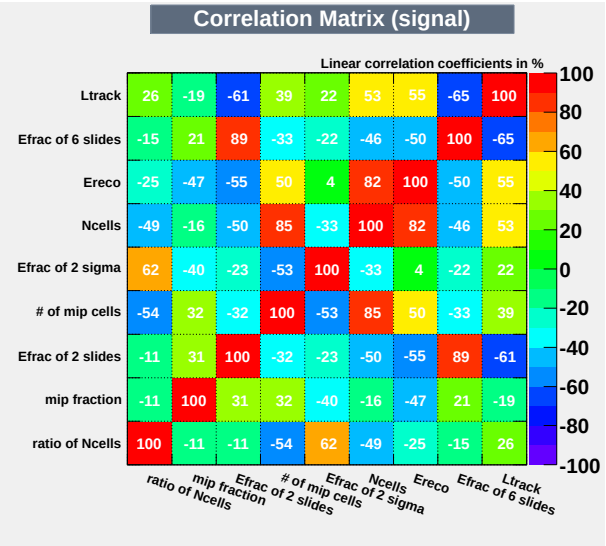


FIG. 7: Overtraining check plot (top center) and linear correlation coefficients for signal (bottom right) and background (bottom left) in case of 10 Variables.





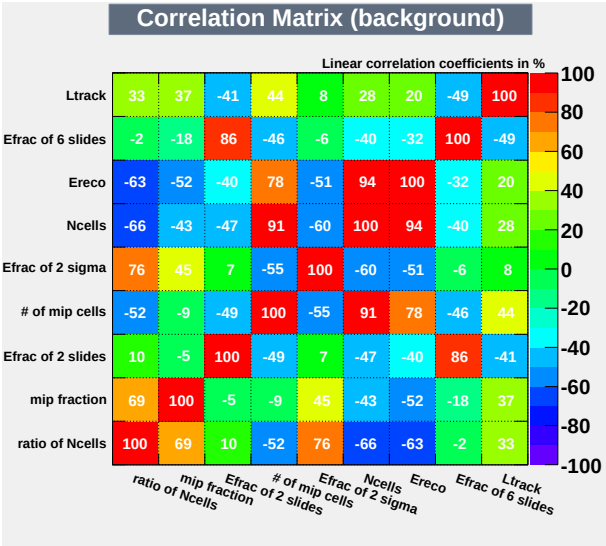


FIG. 8: Overtraining check plot (top center) and linear correlation coefficients for signal (bottom right) and background (bottom left) in case of 9 Variables.

Variable	Importance (%)
$\frac{N_{cells}^{track}}{N_{cells}^{slice}}$	79
$\frac{N_{cells}^{mip}}{N_{cells}^{slice}}$	14
Efrac_2slides	1.6
Efrac_in2sigmaHOR	1.5
N_{mip}	1
$\sum E_{cells}$	0.9
Efrac_6slides	0.8
N_{cells}	0.4
Efrac_out3sigmaHOR	0.3
L_{track}	0.3
$E_{balance3D}$	0.2
Efrac_20slides	0.07

TABLE III: Ranking in case of 12 input variables for BDT.

Variable	Importance (%)
$\frac{N_{cells}^{track}}{N_{cells}^{slice}}$	72
$\frac{N_{cells}^{mip}}{N_{cells}^{slice}}$	14
Efrac_in2sigmaHOR	3.7
N_{mip}	2
Efrac_6slides	1.9
Efrac_2slides	1.8
$\sum E_{cells}$	1.3
N_{cells}	1.1
Efrac_out3sigmaHOR	0.7
$E_{balance3D}$	0.6
L_{track}	0.4

TABLE IV: Ranking in case of 11 input variables for BDT.

Variable	Importance (%)
$\frac{N_{cells}^{track}}{N_{cells}^{slice}}$	79
$\frac{N_{cells}^{mip}}{N_{cells}^{slice}}$	10
Efrac_2slides	2.6
N_{mip}	1.8
Efrac_in2sigmaHOR	1.5
N_{cells}	1.5
$\sum E_{cells}$	1.3
Efrac_6slides	1.12
L_{track}	1.1
Efrac_out3sigmaHOR	0.3

TABLE V: Ranking in case of 10 input variables for BDT.

Variable	Importance (%)
$\frac{N_{cells}^{track}}{N_{cells}^{slice}}$	75
$\frac{N_{cells}^{mip}}{N_{cells}^{slice}}$	10
Efrac_2slides	3.7
N_{mip}	2.4
$\sum E_{cells}$	2.3
N_{cells}	2.3
Efrac_in2sigmaHOR	1.8
Efrac_6slides	1.4
L_{track}	0.6

TABLE VI: Ranking in case of 9 input variables for BDT.

C. Performance Evaluation

BDT obtained for different number of variables was applied to the ND MC sample. The BDT output distributions for signal and various background are shown in Fig. 9.

In order to find the best choice of variable to obtain the better ID, statistical significances $\left(\frac{S}{\sqrt{S+B}}\right)$ and $\left(\frac{S}{\sqrt{B}}\right)$ were evaluated varying the BDT output beetween -1 and $+1$. The final plots are shown in fig. 10 and 11. The best choice of variables in order to obtain the better PID is 11. Indeed in this case $\frac{S}{\sqrt{S+B}} = 26\%$ and $\frac{S}{\sqrt{B}} = 39\%$ requiring BDT output largest than -0.05.

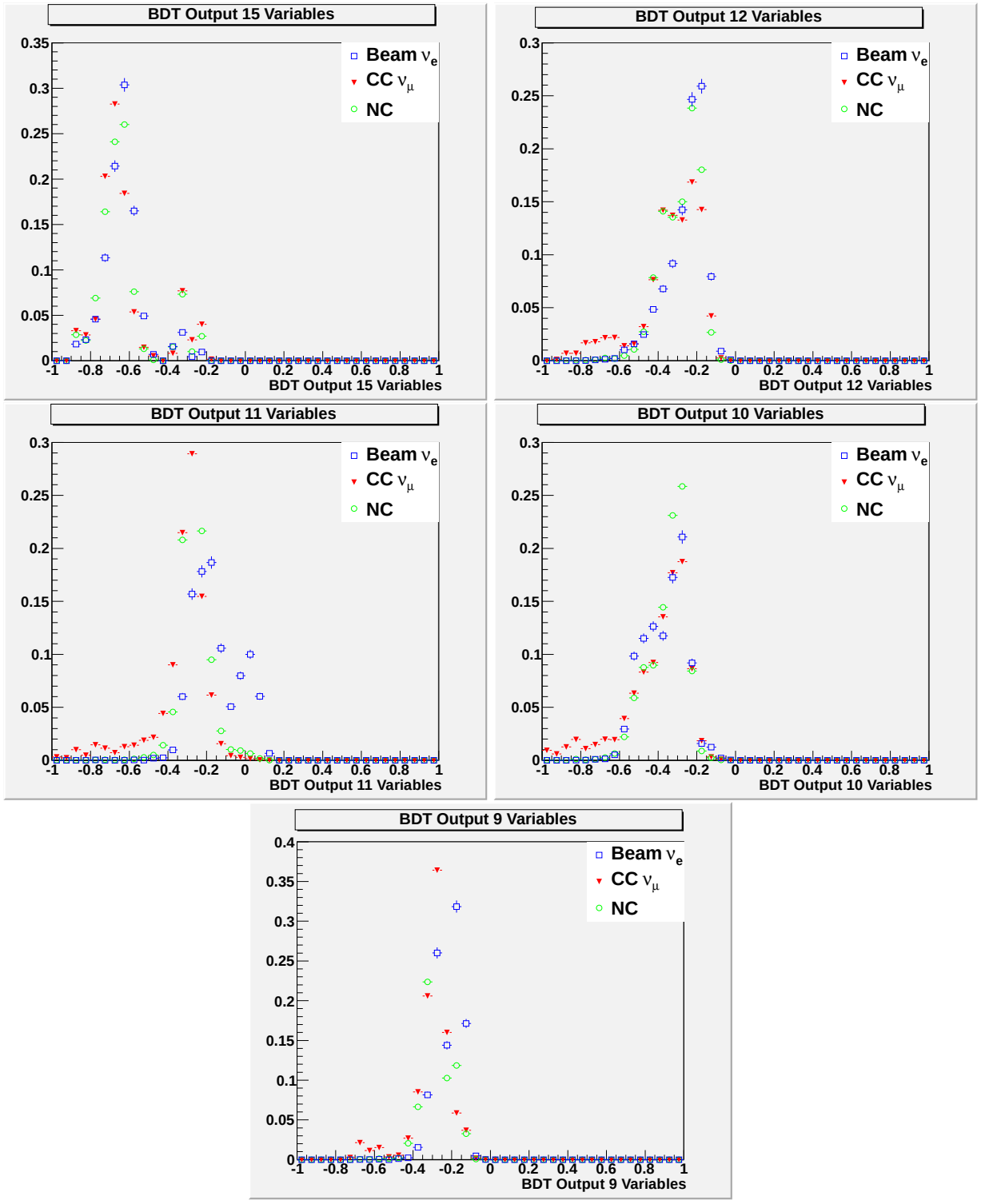


FIG. 9: The BDT output distributions in case of 15 (top left), 12 (top right), 11 (bottom left), 10 (bottom right) and 9 (bottom center) variables for beam ν_e, CC bkg and NC bkg. All distributions are normalized to unit.

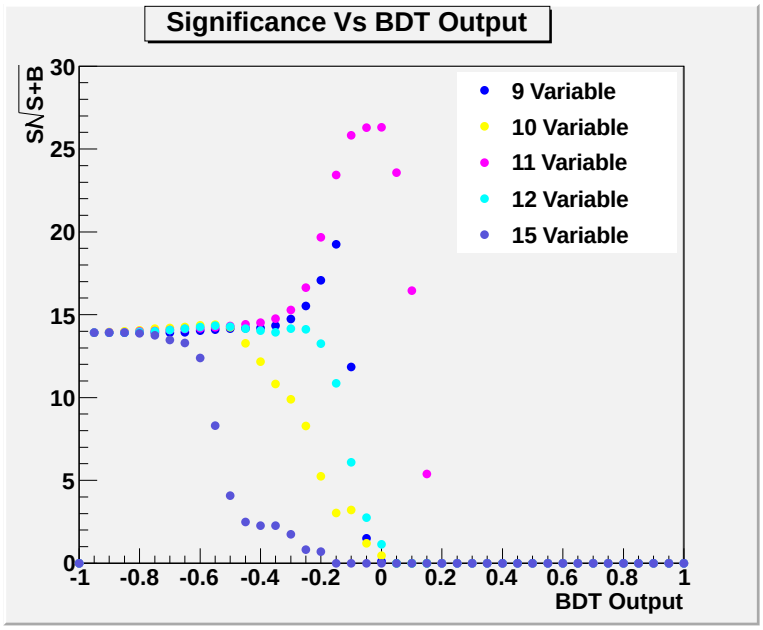


FIG. 10: Statistical significance $\frac{S}{\sqrt{S+B}}$ versus BDT output.

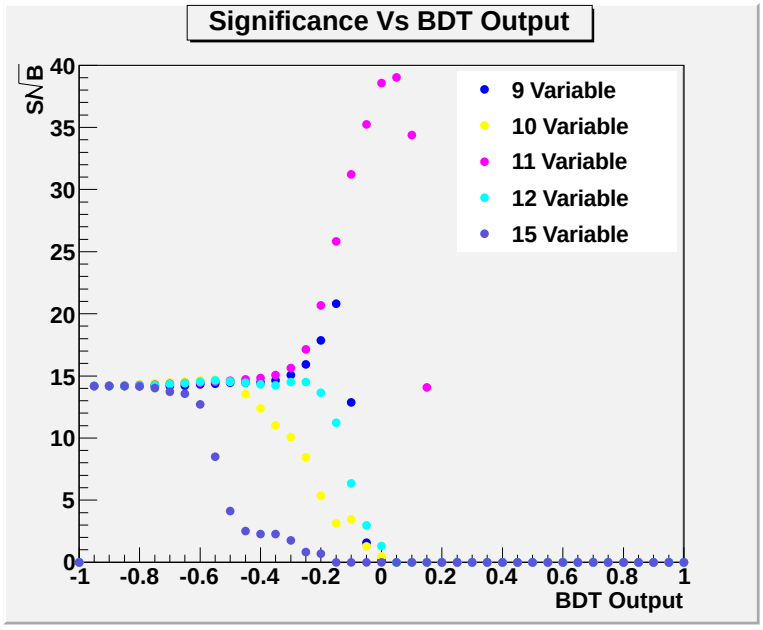


FIG. 11: Statistical significance $\frac{S}{\sqrt{B}}$ versus BDT output.

III. CONCLUSION

The ν_e identification in ND is a most important strategy directed towards the total understanding of the beam and the ν_e measurement at FD. Indeed for this measure the ν_e coming from kaons and muons decay in the decay pipe are background for the measuremet of oscillation probability $P(\nu_\mu \to \nu_e)$. In order to achieve this goal BDT has been used to produce a particle identification technique. BDT based PID was applied on ND MC sample to identify ν_e component in ν_μ beam. To achieve this goal 15 input variables was initially trained. Then the number of variables was reduced starting from the less important and statistical significances $\left(\frac{S}{\sqrt{S+B}}\right)$ and $\left(\frac{S}{\sqrt{B}}\right)$ were evaluated in case of 15, 12, 11, 10 and 9 variables varying the cut on BDT output beetween -1 and $+1$. Thanks to this method results that the best choice of variables in order to obtain the better PID is 11 since $\frac{S}{\sqrt{S+B}} = 26\%$ and $\frac{S}{\sqrt{B}} = 39\%$ using only 11 variables and requiring BDT output largest than -0.05.

[1] A. Hoecher et al., arXiv:0703039 (2009).
[2] Xuebing Bu, NOVA-doc-8381 (2012).