Background Suppression via Multivariate Analysis

Chase Peterson

UNCW: Dr. Liping Gan

October 15, 2021

Jefferson Lab
EXPLORING THE NATURE OF MATTER

Introduction

A background free data sample allows for precisely testing low energy QCD.

$$\eta \to \pi^0 \gamma \gamma \to \gamma \gamma \gamma \gamma$$

Event Generator: 50million simulated eta decays
With 0.027 percent for $\pi^0 \gamma \gamma$, expecting ~13500

We are concerned with what combination of MVA Method, variable set, and manual cut will render the best result.

Eta decay branches

39.31%	$\gamma \gamma$
32.57%	$\pi^0 \pi^0 \pi^0$
22.74%	$\pi^- \pi^+ \pi^0$
4.6%	$\gamma \pi^- \pi^+$
0.7%	$\gamma e^+ e^-$
0.031%	$\gamma \mu^+ \mu^-$
0.027%	$\pi^0 \gamma \gamma$
0.02142%	$\pi^+ \pi^- e^+ e^-$
0.00048%	$\mu^+ \mu^-$
0.00010%	$\pi^0 e^+ e^-$

Method: Training

Using modified TMVAClassification and TMVAClassificationApplication macros:

MVA Methods Considered:
Boosted Decision Trees (BDT)
Fisher Discriminants (Fisher)
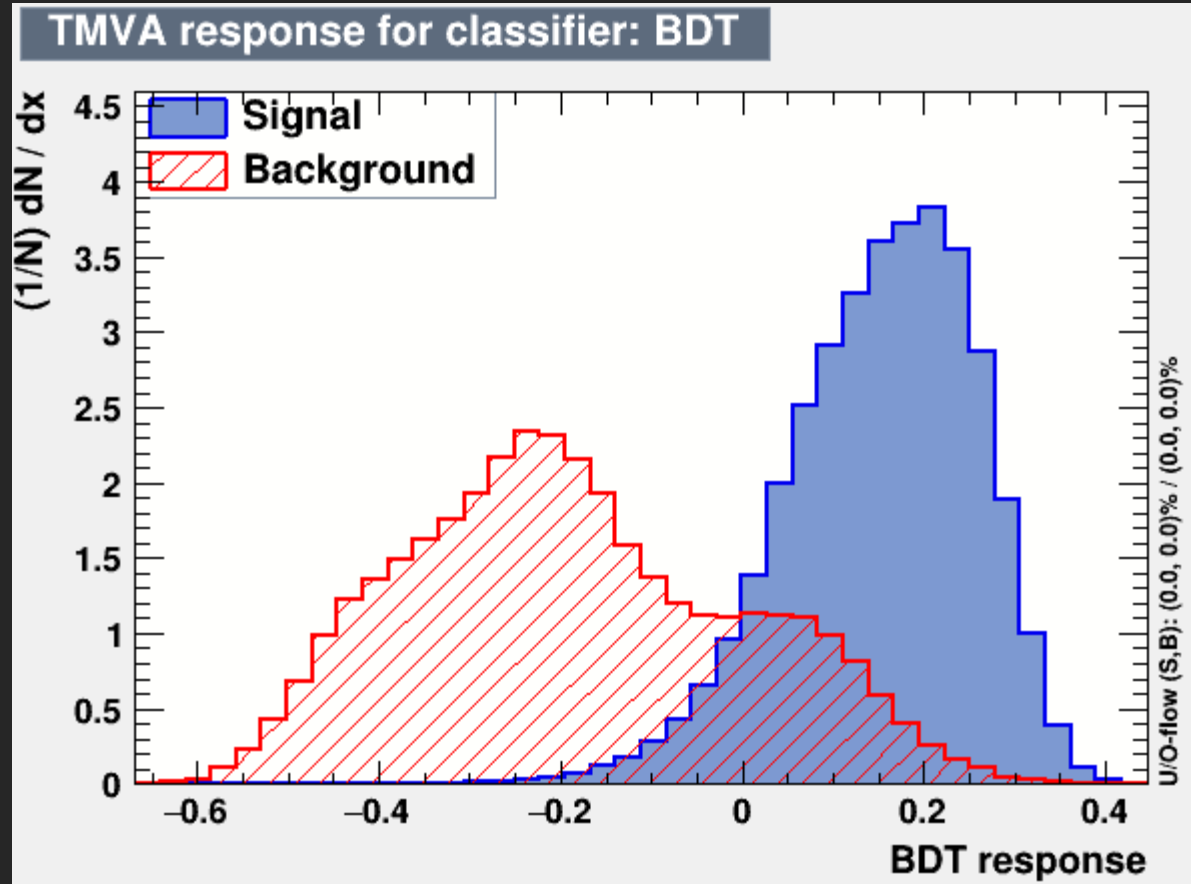
Training

```
TMVAClassification(TString myMethodList = "BDT",
                   TString bkg_file_name = "evtgen-all-gp-to-etap-skim-4g-training-jef.root",
                   TString sig_file_name = "evtgen-ggpi0-gp-to-etap-skim-4g-training-jef.root",
                   TString output_file_name = "tmva-test.root")
```

Training Outputs .xml weight files for testing as well as a root file we can use with TMVAGui to evaluate our training methods

Method: Testing

Testing

```
TMVAClassificationApplication(TString myMethodList = "BDT",
                    TString inputName = "evtgen-all-gp-to-etap-skim-4g-testing-jef.root",
                    TString weightfile = "dataset/mva_weight/TMVAClassification_BDT.weights.xml",
                    TString outfileName = "tmva-test-app.root")
```

Takes our .xml weight file and uses it to test on our data sets

Outputs a root file which we can use to see our tested data and make manual cuts

Method Comparison: TMVA response for classifiers

BDT

Fisher
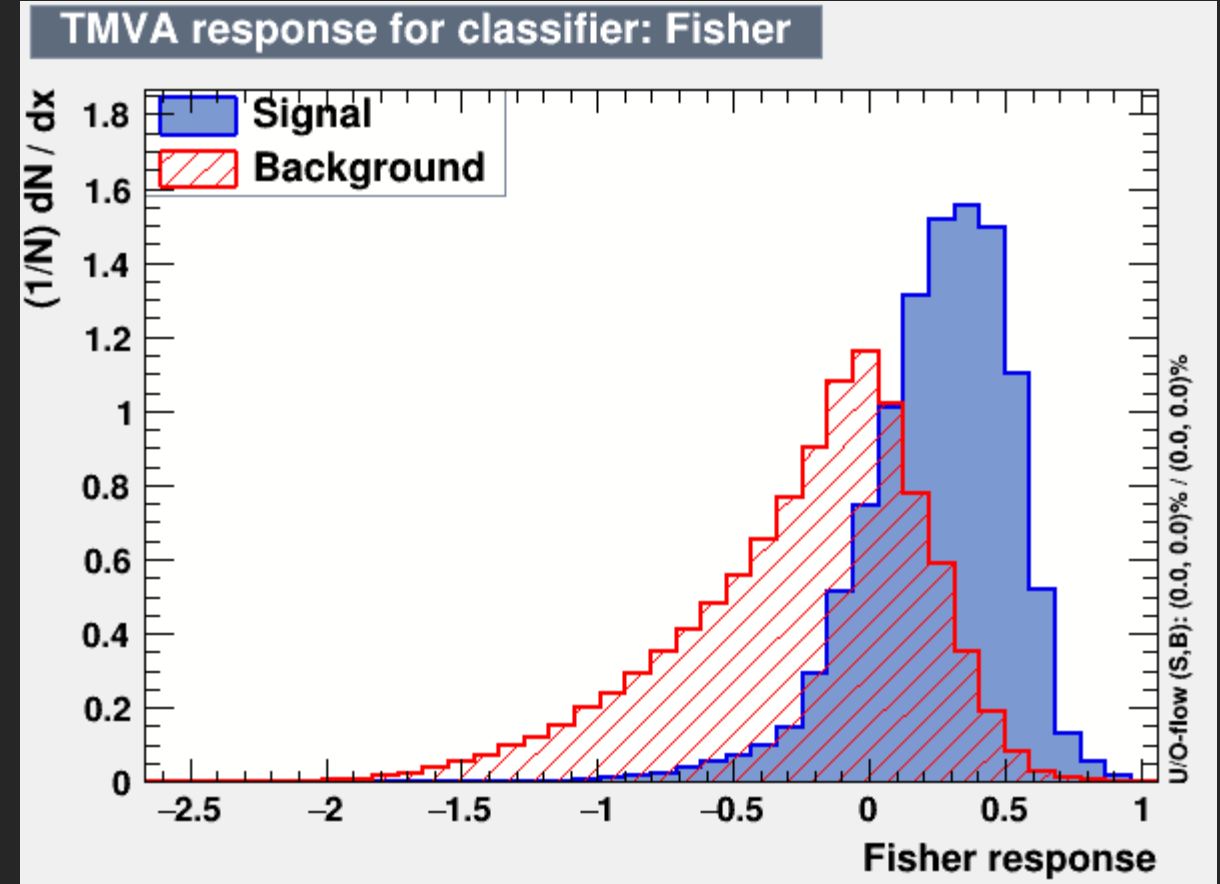
BDT Separates signal from Background more efficiently

Method Comparison: Background Rejection vs. Signal Efficiency

BDT
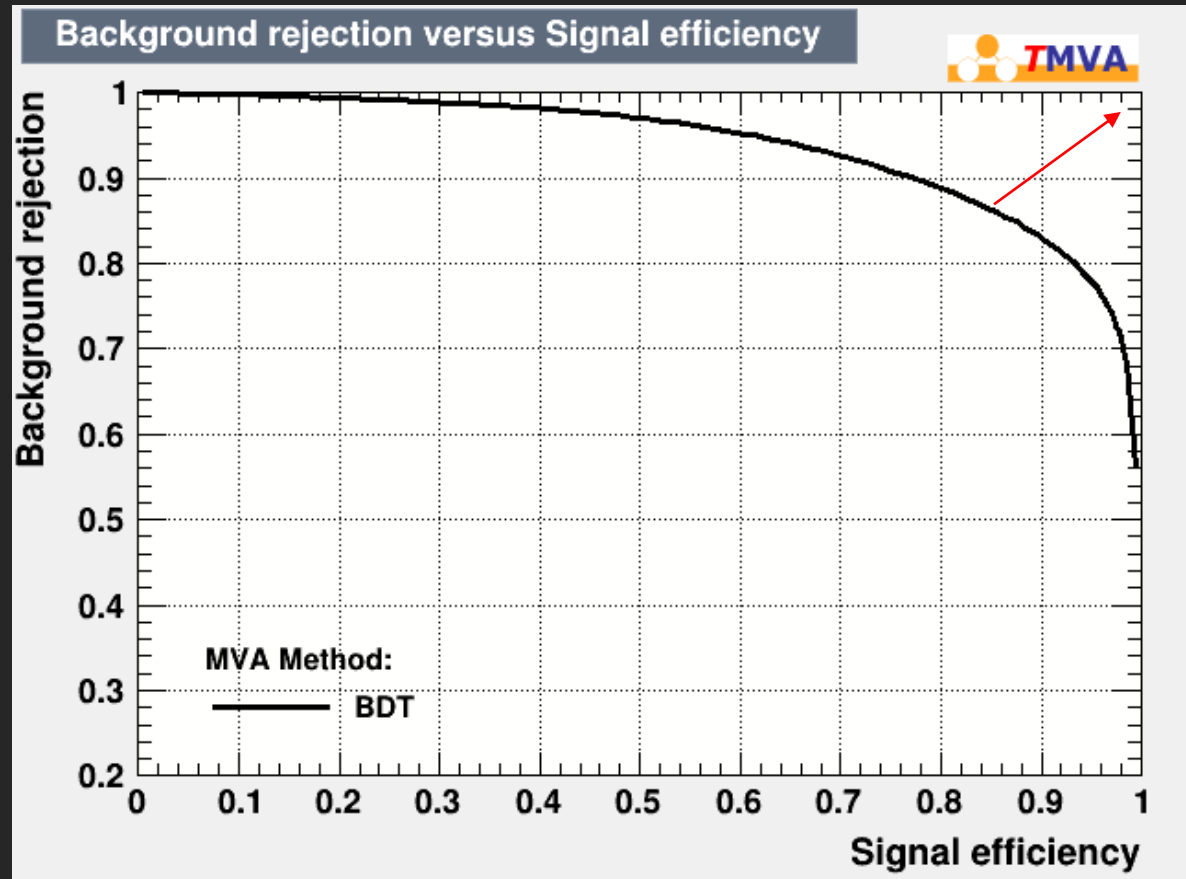
Fisher
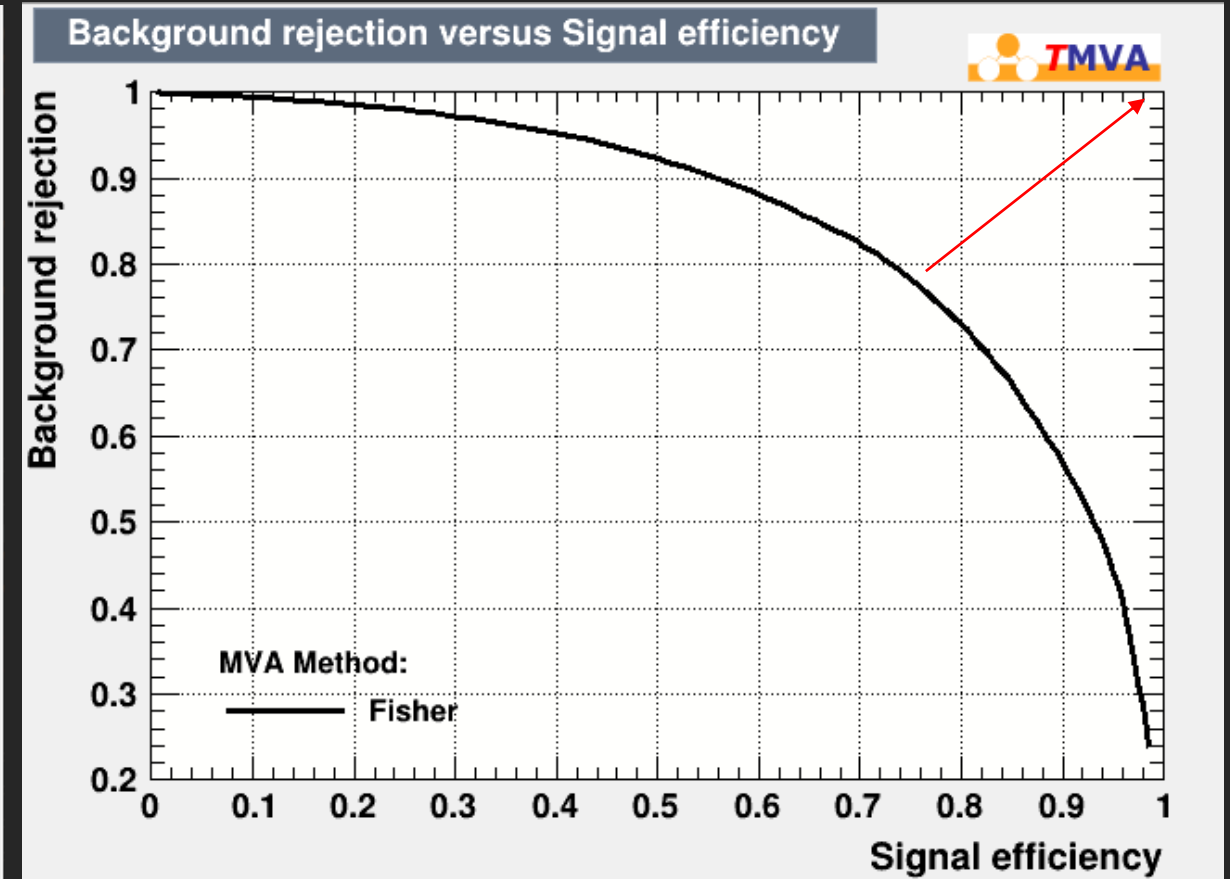
BDT Background Rejection vs Signal Efficiency more ideal than Fisher

Method: Overtraining Modifications

BDT changes for overtraining:

Ntrees 400 -> 850 (The number of trees, or iterations. More trees generally leads to overtraining.)

MinNodeSize 2.5% -> 10% (Minimum percentage of events in a terminal tree node(leaf))

nCuts 10 -> 30 (Number of cuts in a range for determining optimal node splitting)

Background/Signal training ratio from 90/10 to 50/50 (What weight of our background and signal is used for training)

Test for Overtraining
Kolmogorov-Smirnov Test: Ideally has a 0.5 (identically distributed data).

Overtraining modifications to the default BDT

Method: Overtraining Comparison
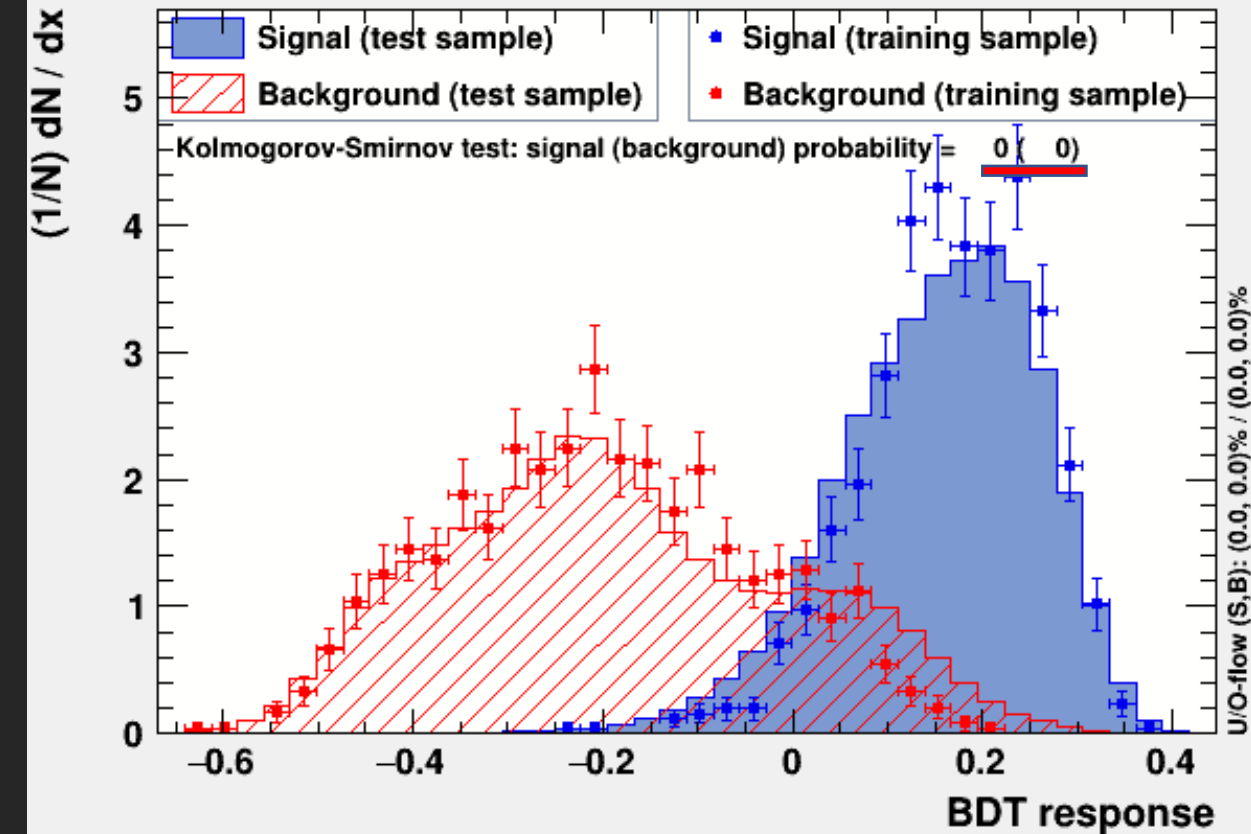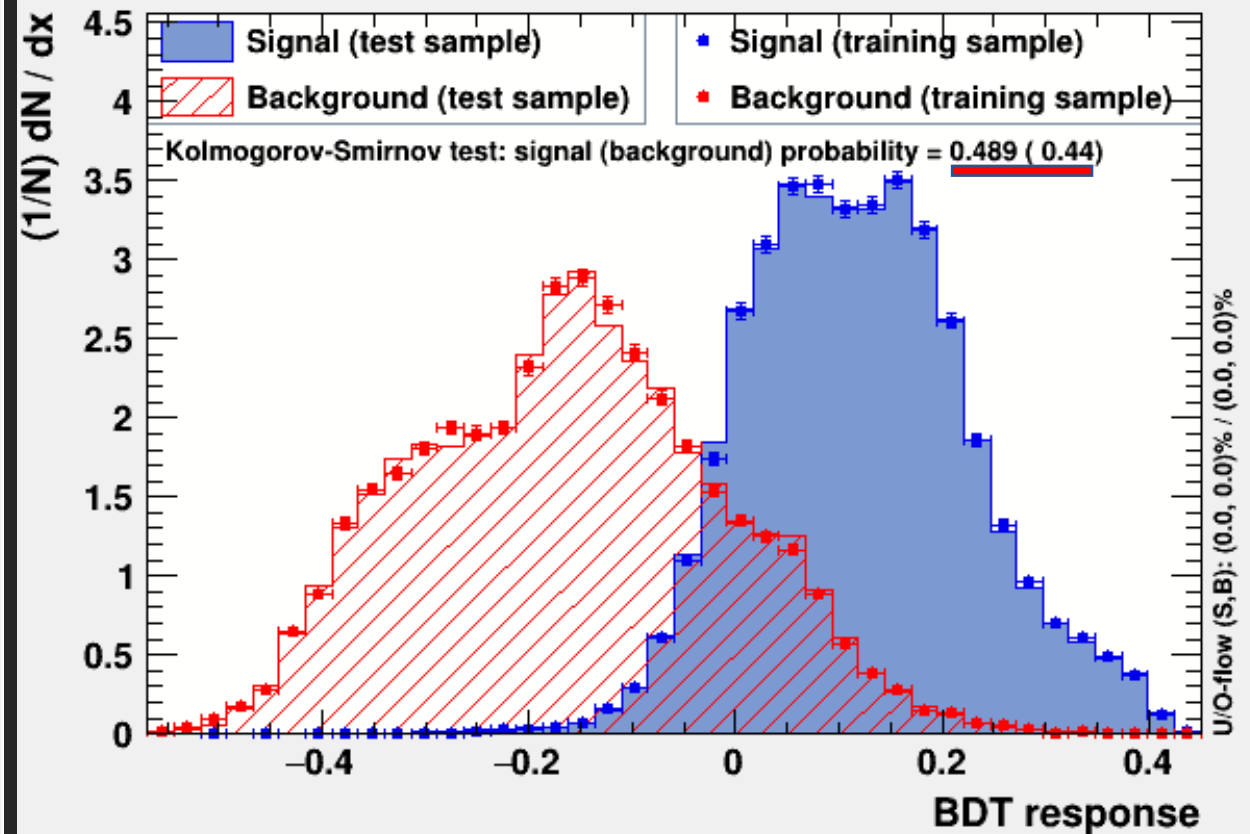
The BDT modifications benefit the KS Test coefficient greatly.

Input Variable Sets

Input
extraE
Elasticity_cons
Helicity
Mandelstam_t
Transversemomentum_eta
Transversemomentum_pi0
Transversemomentum_2g
m_shower_x0
m_shower_x1
m_shower_x2
m_shower_x3
m_shower_y0
m_shower_y1
m_shower_y2
m_shower_y3

Spectator
Chi_2
Gammabeam_energy
Missingmasssquare
Coplanarity
Insertnumber
Fcalnumber
Signal
Weight
Invariantmass2g
Invariantmass4g
Rawinvariantmass4g
Rawinvariantmass2gpi0
m_shower_z0
m_shower_z1
m_shower_z2
m_shower_z3
m_shower_e0
m_shower_e1
m_shower_e2
m_shower_e3

We tested various combinations of input variables to see which achieved the best result. Some spectator variables were originally negligible input variables.

Keeping extraE and showers enabled in each, there are 64 combinations tested.

Spectators are of no consequence to the training and testing, however we can interchange input and spectator variables, if desired.

Variable sets were compared according to their total number of signal at a control BDT_Prob value.

Variables: Entries vs. Trials
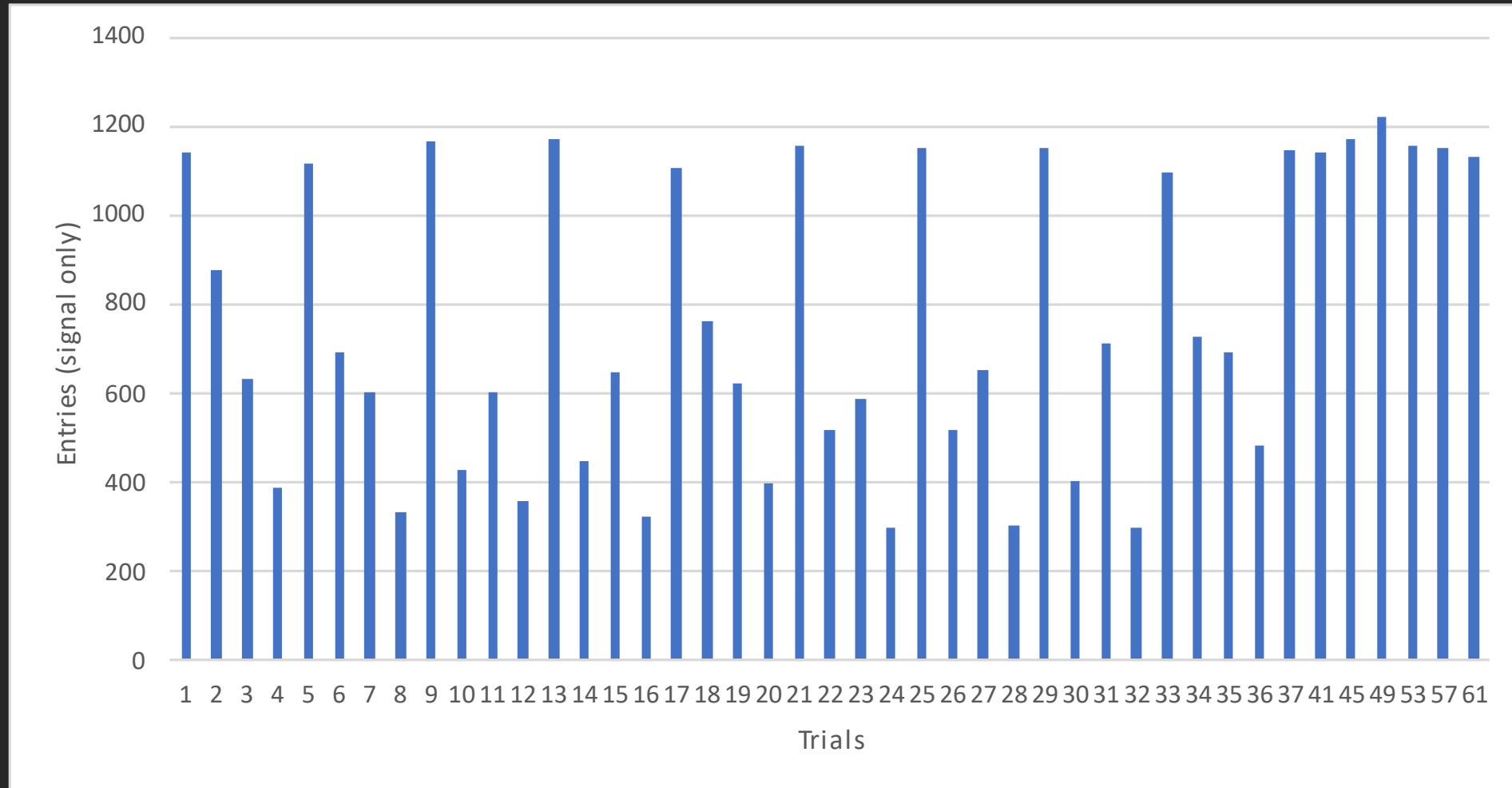
Trials 1,5,9,13,17… (1 +4n) have the highest amount of entries

Input Variable Sets

A checkmark represents a disabled variable

The best results are visually ones every 1,5,9,13..61 All of which have elast_cons and helicity enabled

Test Num (post)	elast_cons	helicity	mandelstam	trans.eta	trans.pi0	trans.2g
1						
2	✓					
3		✓				
4	✓	✓				
5			✓			
6	✓		✓			
7		✓	✓			
8	✓	✓	✓			
9				✓		
10	✓			✓		
11		✓		✓		
12	✓	✓		✓		
13			✓	✓		
14	✓		✓	✓		
15		✓	✓	✓		
16	✓	✓	✓	✓		
17					✓	
18	✓				✓	
19		✓			✓	
20	✓	✓			✓	
21			✓		✓	
22	✓		✓		✓	
23		✓	✓		✓	
24	✓	✓	✓		✓	
25				✓	✓	
26	✓			✓	✓	
27		✓		✓	✓	
28	✓	✓		✓	✓	
29			✓	✓	✓	
30	✓		✓	✓	✓	
31		✓	✓	✓	✓	
32	✓	✓	✓	✓	✓	
33						✓
34	✓					✓
35		✓				✓
36	✓	✓				✓
41				✓		✓
45			✓	✓		✓
49					✓	✓
53			✓		✓	✓
57				✓	✓	✓
61			✓	✓	✓	✓

Variables: Entries vs. Trials narrowed
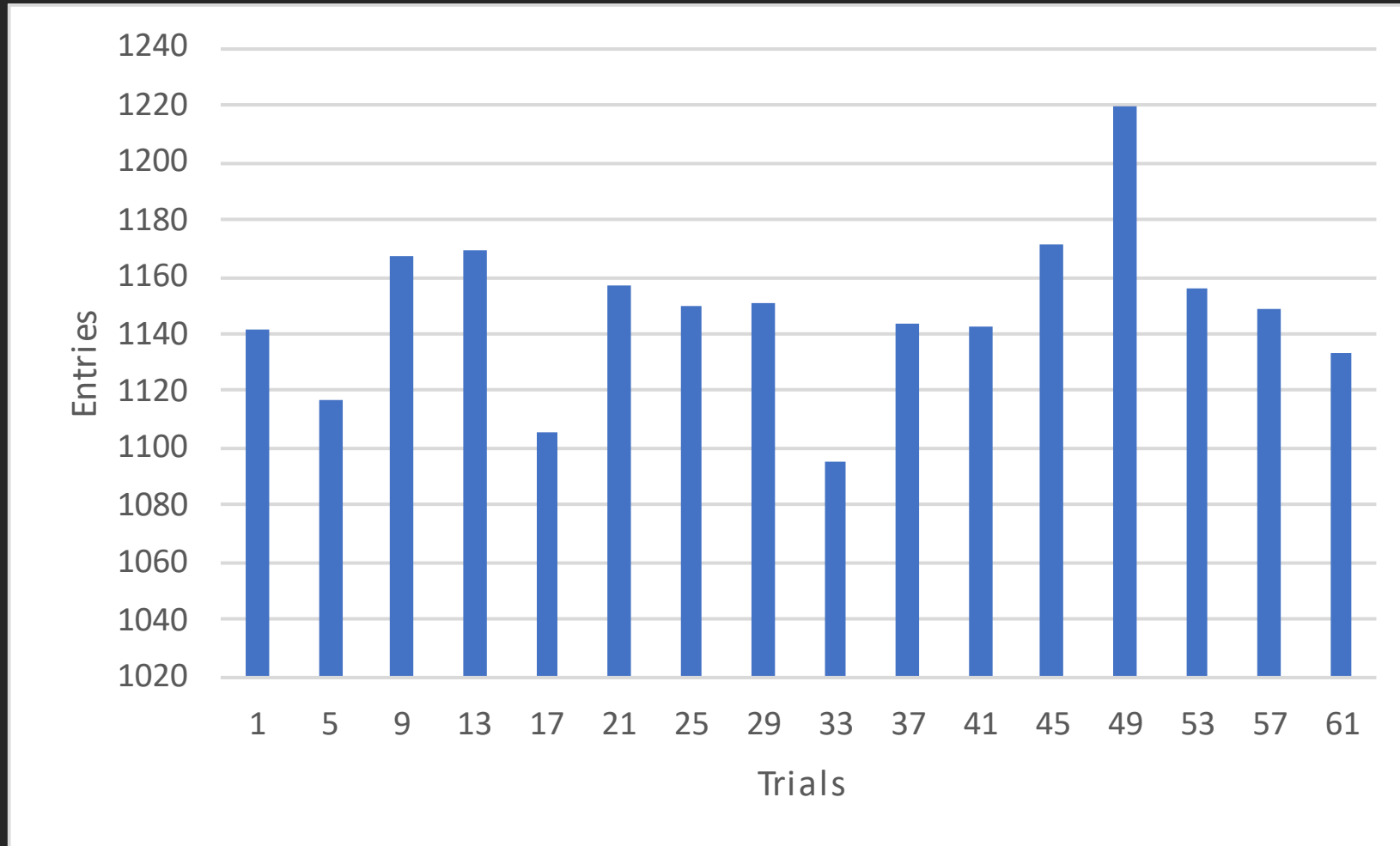
Trial 49 has the highest entries: elasticity_cons, helicity, extraE, mandlestam_t, trans.mom._eta, m_shower x,y
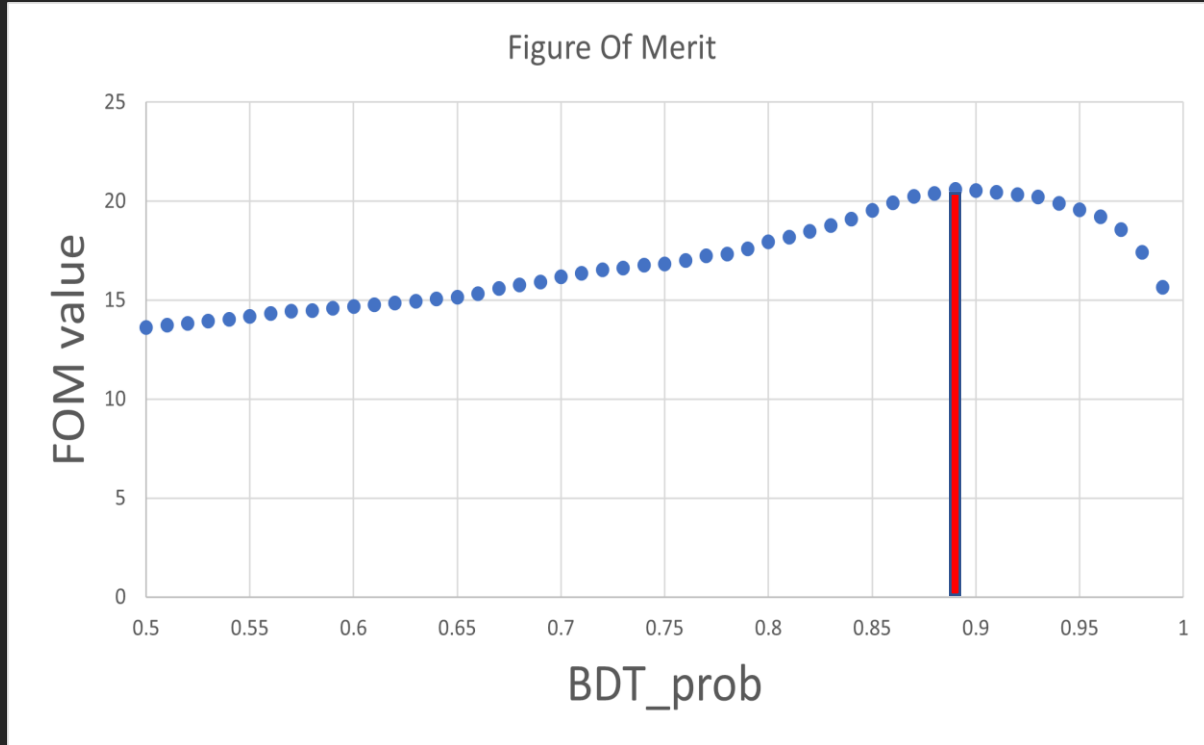
BDT_Prob

BDT_Prob is the ROOT variable we use to manually cut our trained and tested data

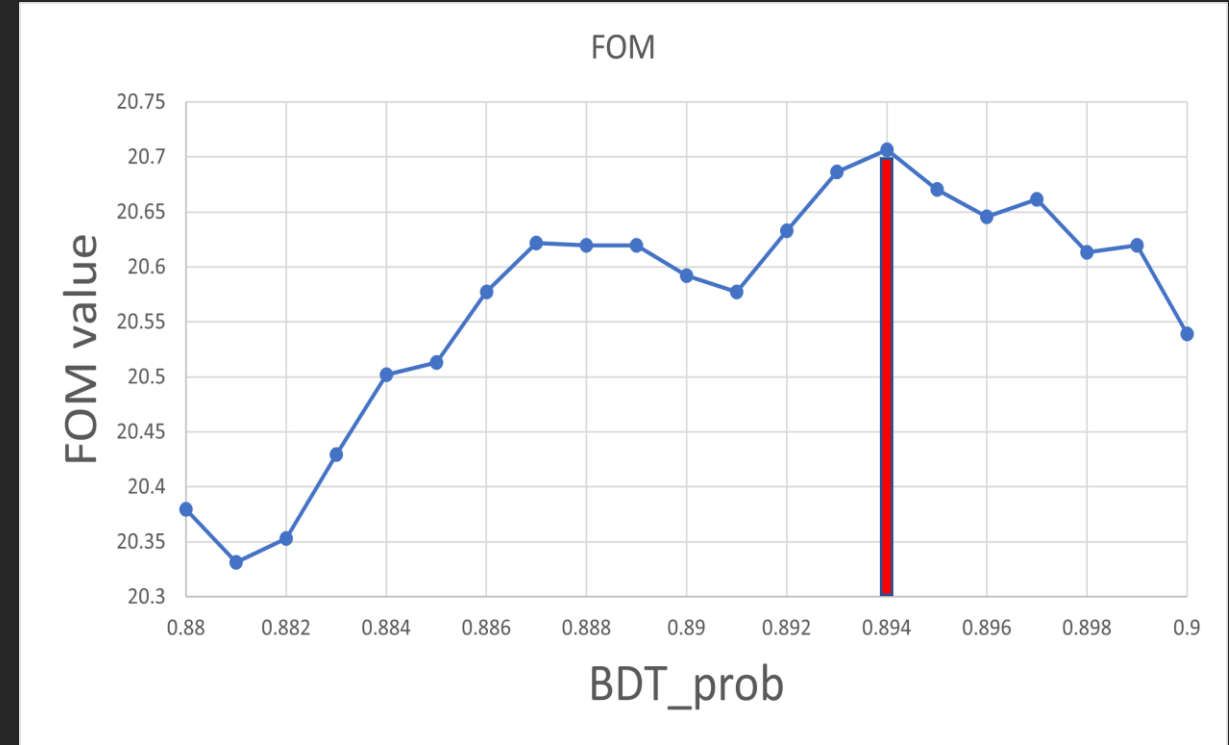What value of BDT_Prob will give us the best background suppression?

$$\text{Figure of Merit:} \quad \frac{signal}{\sqrt{signal + background}}$$

We want the largest value for our Figure-of-Merit
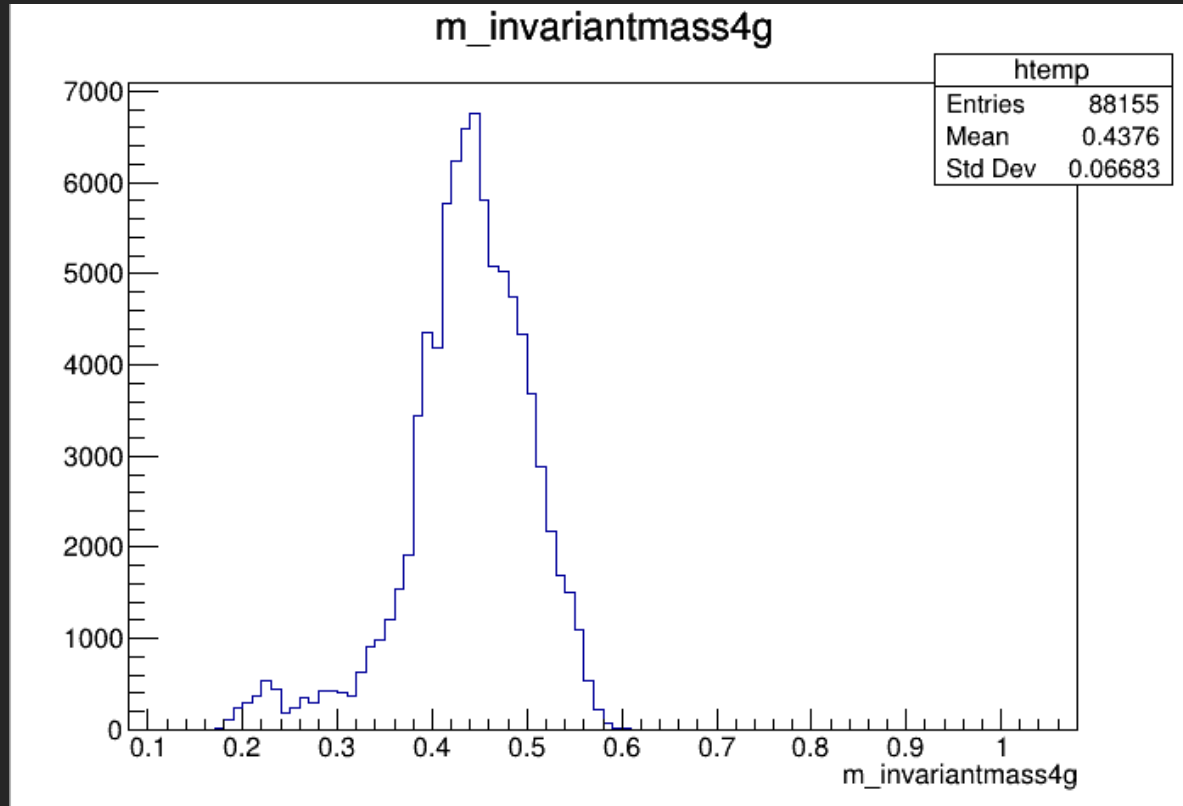
BDT_Prob
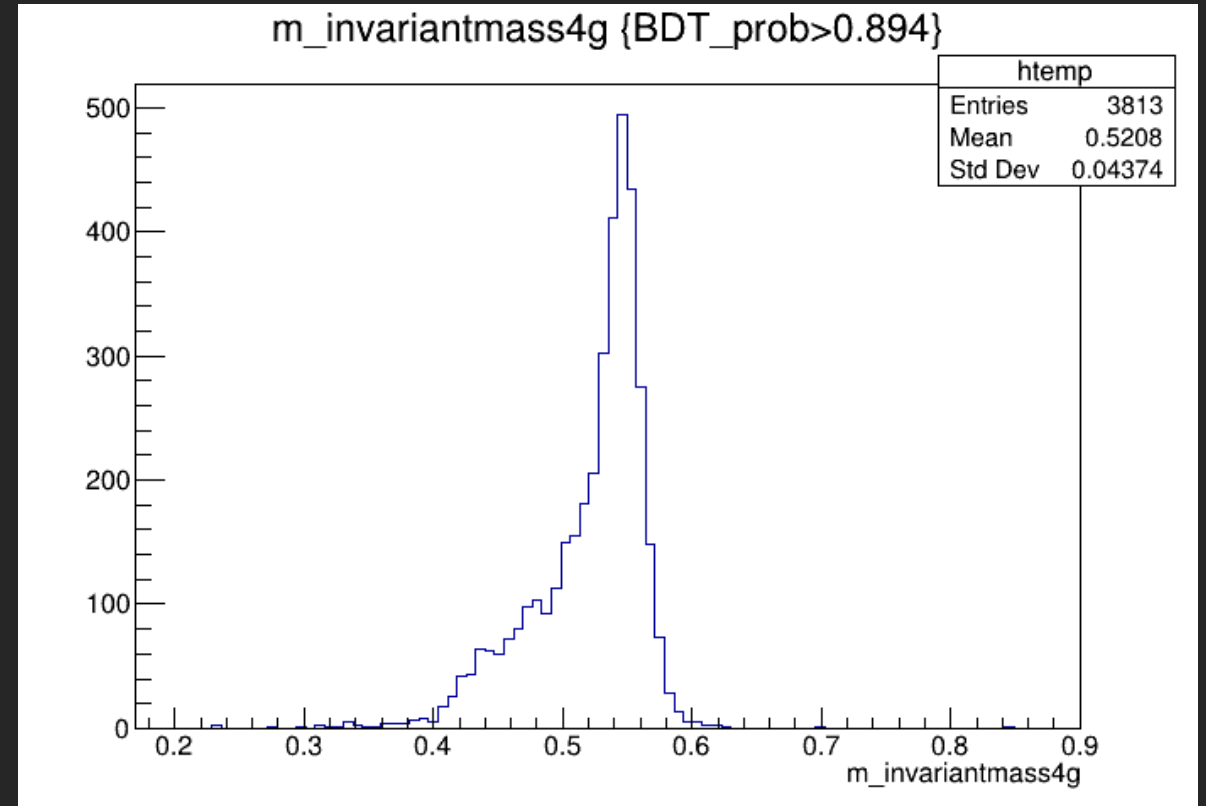

Overall (0.5 – 1)

Zoom (0.88-0.9)

BDT_Prob has best results at 0.894

Conclusion

Raw

BDT, Variable Set, and Cuts

Conclusion

- BDT with overtraining modifications and 50/50 training ratio
- Variables: elasticity_cons

    extraE

    helicity

    mandlestam_t

    transversemomentum_eta

    m_shower x,y

- BDT_Prob value: 0.894