山东大学 SHANDONG UNIVERSITY

Department of Physics, Shandong University

Compressed EWK study(ISRC1N2)

Chengxin Liao
liaocx@ihep.ac.cn

Mar, Wed 5, 2025

Outline

1. Hyperparameters optimization
2. Performance of Model
3. Backup

Hyperparameters optimization(HH)

Input(HH-Channel):

Sample:

Sig: ISRC1N2(mass_C1 = 100GeV, mass_N2 = 70GeV)->12180 entries
Bkg: 513850 entries

All input data(C1N2_100_70 and Bkg) already passed pre-selection

Strategy:

method: BDTG
Separate sig(bkg) into five folders, one for test, the other three for train, and last one for validation set,then traverse all possibilities.

```
Number of training and testing events
-------------------------------------------------------
Signal     -- training events            : 7311
Signal     -- testing events             : 2436
Signal     -- training and testing events: 9747
Background -- training events            : 308329
Background -- testing events             : 102770
Background -- training and testing events: 411099
```

Pre-Selection

$$had\text{-}had\ \ channel: nTaus \geq 2, nLeps = 0$$
$$pass\ MET\ trigger;\ MET \geq 200$$
$$1 \leq nBaseJet \leq 8$$
$$b - Veto$$
$$OS$$

Hyperparameters optimization(HH)

Variables(26):

Obj kinematics

Pt_tt

Angular correlations

dPhit1x
dEtatt
dPhiMax_xt
dPhiztt
dPhitt
dPhizxe
dPhiMin_xt
dPhit2x
dPhiMin_tj1
dRt2x
dRMax_xt
dRMin_tj
dRtt
sum_cos_dphi

Event kinematics

Mll(Invariant Mass of tau1 and tau2)
MIA
MT2_150
MET_Tau
Proj_tt
MstauA
MCT
frac_MET_tt
frac_MET_tau1
frac_MET_MeffInc_40
frac_MET_Meff

These vars are selected based on the importance

Hyperparameters optimization(HH)

Grid Search:

Ntrees: 200, 300, 400, 500
Max Depth: 6, 8, 10, 12
MinNodeSize: 1%, 2%, 3%
Learning Rate: 0.01, 0.05, 0.1

Binned significance: $Z = \sqrt{2((s_i + b_i)\log\left(1 + \frac{s_i}{b_i}\right) - s_i)}$

Show top Zn

	Model Name	Binned Significance	Max Zn	Max Zn Bin
500_12_1_005		14.2770	3.83857	199
300_10_1_01		13.9648	3.76965	198
200_6_1_01		13.9250	3.74940	198
500_6_3_01		14.2740	3.72616	199
400_10_1_01		13.9553	3.70167	199
300_6_2_01		13.9366	3.69620	199
300_10_2_01		14.0094	3.67743	199
300_8_1_01		14.0434	3.67624	198
200_8_1_01		14.1925	3.67005	198
400_12_1_005		14.1384	3.66529	199
200_6_2_01		14.2209	3.65978	199
200_6_3_01		13.7197	3.64427	199
500_10_1_01		13.8227	3.63722	198
500_8_1_01		13.8369	3.61405	198
400_10_2_01		14.2001	3.60950	199
500_6_1_005		14.0399	3.60132	197

Rebin result

```
Model Name,Binned Significance,Max Zn,Max Zn Bin,bin num
500_12_1_005,16.0862,3.8635,198,200
500_12_1_005,15.9967,3.62563,99,100
500_12_1_005,15.9318,3.62563,50,50
500_12_1_005,15.6612,3.07372,40,40
500_12_1_005,15.3086,2.45396,25,25
500_12_1_005,15.0825,2.20391,20,20
```

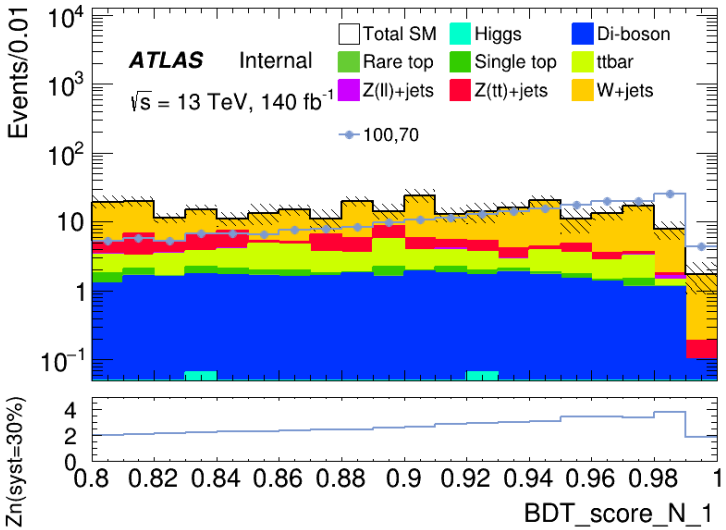
Compared with form result, there has a significant improvement in Zn

	Model Name	Binned Significance	Max Zn	Max Zn Bin
137	100_8_3_005	12.1380	3.27179	48
128	100_6_1_01	12.9663	3.24539	49
15	100_12_3_005	12.1254	3.22901	48
85	100_10_3_005	12.1150	3.20722	48
105	200_10_1_01	13.1608	3.19603	50
104	100_8_1_01	12.8853	3.18398	50
63	200_6_1_005	12.7673	3.17520	49
2	200_6_1_01	12.9052	3.17138	50
38	100_6_2_01	12.8248	3.16297	49
131	300_8_2_01	13.1256	3.16255	50
93	300_6_1_005	12.9703	3.14200	50
73	100_6_1_005	12.4457	3.14142	48
69	400_6_1_01	12.9285	3.14074	50
54	200_8_3_01	12.8685	3.13397	50
12	200_6_2_005	12.7035	3.12582	49
33	100_6_2_005	12.2453	3.11746	48
66	400_6_1_005	12.9369	3.10400	50
45	100_10_3_01	12.7388	3.10074	49
7	100_12_3_01	12.6318	3.10071	49
48	400_12_2_01	12.9393	3.09236	50
72	400_12_2_005	12.9179	3.06882	50
62	300_6_1_01	12.8501	3.06869	50
133	400_10_1_01	12.9846	3.06413	50
91	100_6_3_01	12.6291	3.06320	49
9	400_8_3_01	12.9337	3.06226	50
58	300_10_3_005	12.8854	3.05992	49
36	300_10_1_01	12.9906	3.05755	50
89	100_10_2_01	12.8406	3.05210	49
43	200_8_2_01	13.2400	3.04952	50
19	400_6_2_005	12.8263	3.04206	50

Performance of Model(HH)

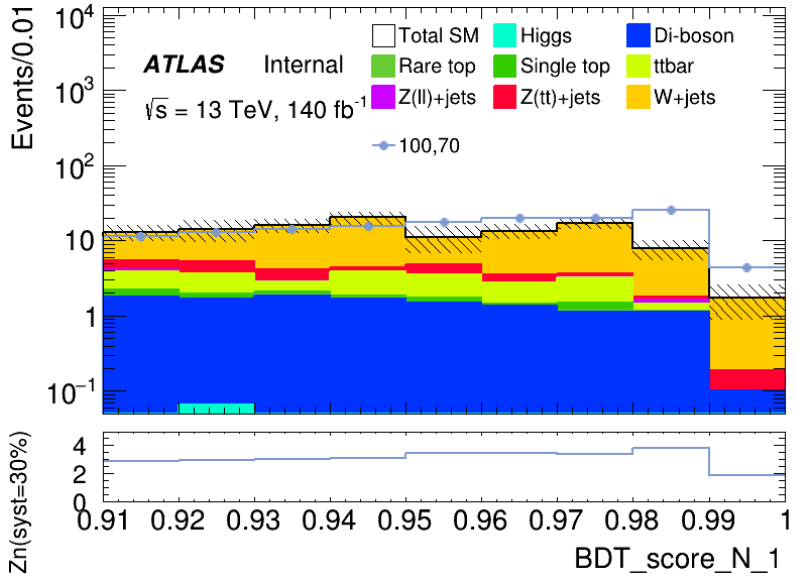
hyper parameter: NTrees=500, learning rate=0.05, max depth=12, MinNodeSize=1%(default)

Apply a rough cut at 0.80 to check the distribution



Precise cut at 0.91 to define signal region

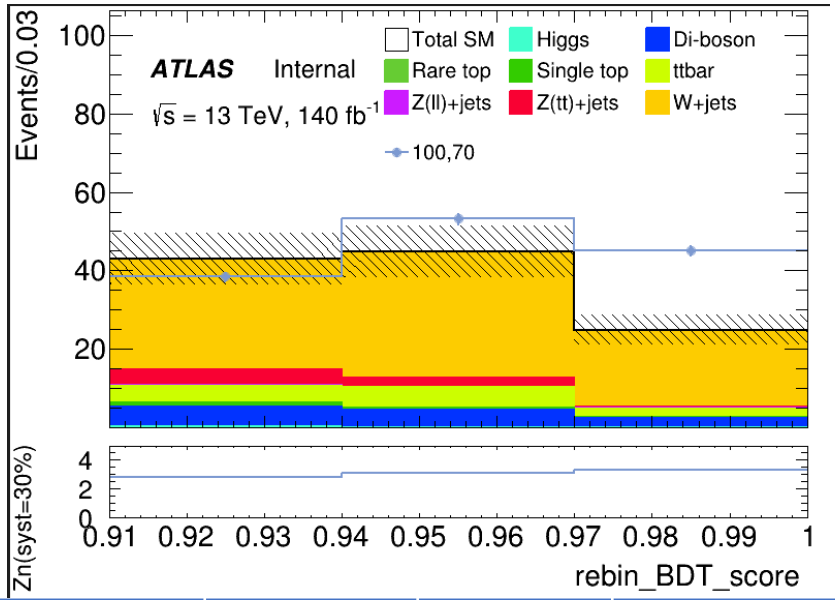
It has a wider peak than LH signal region



Rebin to: [0.91, 0.94, 0.97, 1.00]

Performance of Model(HH)

hyper parameter: NTrees=500, learning rate=0.05, max depth=12, MinNodeSize=1%(default)



Root of quadratic sum of Zn = 5.3163

bin	max Zn	C1N2ISR (100,70)	bkg	Higgs	OtherTop	SingleTop	TopPair	VV	Wjets	Zlljets	Zttjets
(0.91-0.95)	2.8271	**38.472+-1.192**	42.987+-6.489(15.09%)	0.150+-0.041	0.026+-0.020	0.958+-0.303	4.156+-0.802	5.144+-0.289	**28.198+-6.388**	0.265+-0.185	4.090+-0.675
(0.95-0.98)	3.0639	**53.206+-1.399**	44.788+-6.645(14.83%)	0.019+-0.014	0.026+-0.020	0.473+-0.239	5.158+-0.888	4.512+-0.303	**32.118+-6.547**	0.040+-0.026	2.435+-0.612
(0.98-1.00)	3.29906	**45.161+-1.287**	24.891+-3.088(12.40%)	0.007+-0.007	0.034+-0.017	0.348+-0.184	2.050+-0.570	2.263+-0.187	**19.479+-3.753**	0.196+-0.123	2.435+-0.612

TODO

1. BDT distribution of Validation Set & Binned BDT distribution of Data and Test Set

2. Finish Rebin Code

3. Summary of HH&LH channel ML results and the definition of SR

4. arrange the old code and optimize them(add README.md)

BackUp

An Interesting Method to Rebin

It can be proved that there are 2^{n-1} ways to rebin if histgram have n bins except 2 bins

2 bins

1. [0] [1] (separate)
2. [0+1] (merged)

Oredered the method and trun into binary number

3 bins

1. [0] [1] [2]
2. [0+1] [2]
3. [0] [1+2]
4. [0+1+2]

Based on Mathmatical Induction

2^{n-1} ways to rebin

4 bins

1. [0] [1] [2] [3]
2. [0+1] [2] [3]
3. [0] [1+2] [3]
4. [0] [1] [2+3]
5. [0+1+2] [3]
6. [0] [1+2+3]
7. [0+1] [2+3]
8. [0+1+2+3]

Example in 4 bins

000 (no walls): [0+1+2+3]

001 (wall at 2–3): [0+1+2] [3]

010 (wall at 1–2): [0+1] [2+3]

011 (walls at 1–2, 2–3): [0+1] [2] [3]

100 (wall at 0–1): [0] [1+2+3]

101 (walls at 0–1, 2–3): [0] [1+2] [3]

110 (walls at 0–1, 1–2): [0] [1] [2+3]

111 (all walls): [0] [1] [2] [3]

That's $2^3 = 8$ options (3 gaps).