Note

Feng-Yang Hsieh

1 CWoLa

The Classification Without Labels (CWoLa) is a weakly supervised learning method. The CWoLa approach trains a model to discriminate the mixed samples, which are mixtures of the original signal and background samples. The optimal classifier in the CWoLa approach is also the optimal classifier in the traditional fully supervised case where all label information is available. This section utilizes the CWoLa approach to train classifiers on di-Higgs samples.

1.1 Sample

This exercise's signal corresponds to the resonant Higgs boson pairs production in the four-b quarks channel. These Higgs boson pairs are produced via gluon-gluon fusion in the two Higgs doublet model (2HDM). The Higgs boson h ($m_h = 125$ GeV) pair is produced by the heavy CP-even scalar H with mass m_H ranging from 300 GeV to 1200 GeV. The background consists of QCD multi-jet events.

The CWoLa training samples M_1 and M_2 are the mixtures of the signal and background samples. The probability distribution of the mixed sample is a combination of the signal $p_s(x)$ and background $p_B(x)$ distributions:

$$p_{M_1}(x) = f_1 p_S(x) + (1 - f_1) p_B(x)$$
$$p_{M_2}(x) = f_2 p_S(x) + (1 - f_2) p_B(x)$$
(1)

where f_1, f_2 are the signal fractions, and x represents the observables used for the classification task.

DNN and SPANet network architectures are considered in this exercise. For DNN, the input features are summarised in Table 1, consisting of 16 variables. For SPANet, the input features are a list of final jets, each represented by their 4-momentum $(p_\mathrm{T}, \eta, \phi, M)$ and a boolean b-tag.

Table 1: Input variables used to train the dense neural network.

Reconstructed objects	Variables used for training	#
Higgs candidate	(p_T, η, ϕ, m)	8
Subjets	$\Delta R(j_1, j_2)$	2
b-tagging	Boolean for $j_i \in h_{1,2}^{\text{cand}}$	4
Di-Higgs system	p_T^{hh}, m_{hh}	2

1.2 Result

The CWoLa training utilizes samples with different signal fractions f_1, f_2 to train the classifiers. The results of CWoLa training are shown in Figure 1 with different signal fractions. When f_1 is far from 0.5, the results tend to approach those of the fully supervised case.

(a) DNN

(b) SPANet

Figure 1: The AUC of CWoLa training as a function of the signal fraction f_1. For simplicity, we set signal fraction f_2 equal to $1 - f_1$. The horizontal dashed line indicates the fully-supervised AUC.

When $f_1 = 0.5$ the mixed sample M_1 and M_2 have identical distributions, so the classifier can not learn anything in this case. In the case of DNN, the AUC is 0.5, as expected. However, for SPANet, the AUC is more than 0.7.

This is because SPANet is trained on both pairing and classification tasks simultaneously. The pairing part introduces asymmetries between signal and background samples, leading to the AUC that deviates from 0.5.

To investigate the effect of the pairing task on SPANet's performance, the weight of the

pairing component is set to zero, meaning that SPANet focuses solely on the classification task. Figure 2 shows the SPANet training results without pairing task. As expected, the AUC is close to 0.5 when $f_1 = 0.5$.
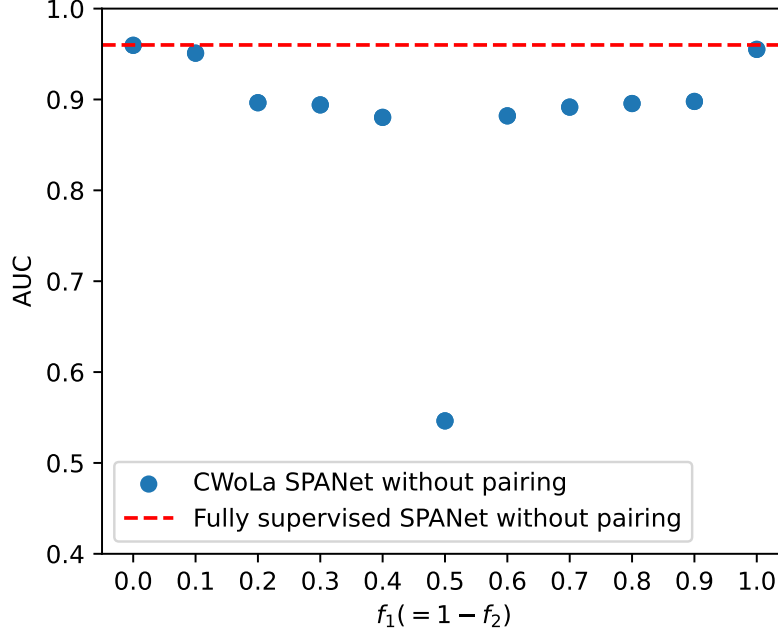

Figure 2: The AUC of CWoLa SPANet training as a function of the signal fraction f_1. For simplicity, we set signal fraction f_2 equal to $1 - f_1$. Here, SPANet is trained on the classification task only.

2 CWoLa hunting

The CWoLa hunting approach considers a variable m_{res}. For background, the m_{res} distribution is smooth while signal m_{res} distribution is expected to be localized near some m_0. Consequently, this variable could be used to create two mixed samples. Additional features that are uncorrelated with m_{res} can be used for training a classifier. This technic is first introduced by Reference [1].

2.1 Sample

The signal is the resonant Higgs boson pairs production in the four-b quarks channel. In this section, the Higgs boson pair is produced by the heavy CP-even scalar H with mass $m_H = 500$ GeV or $m_H = 1000$ GeV. The background consists of QCD multi-jet events. The

basic requirement is the "four-tag cut," which requires at least four b-tagged $R = 0.4$ anti-k_t jets with $p_T > 40$ GeV and $|\eta| < 2.5$. Only the events passing the four-tag cut are used in the following analysis.

The CWoLa hunting approach utilizes the signal and sideband regions to create the mixed training sample. The di-Higgs system's total invariant mass m_{hh} is utilized to determine the signal and sideband region. This quantity is computed from the four b-jets with the highest transverse momentum. Figure 3 presents the m_{hh} distribution of signal and background samples. Table 2 summarizes the signal and sideband regions. These signal and sideband regions are chosen such that the corresponding cross-sections are closed.

(a) $m_H = 500$ GeV (b) $m_H = 1000$ GeV

Figure 3: The total invariant mass m_{hh} distribution of signal and background samples. The signal region is between the red dashed lines. The sideband region is between the green dashed lines and excludes the signal region.

Table 2: The signal and sideband regions with different resonant samples. The unit is GeV.

m_H	Signal	Sideband
500	$[350, 550]$	$[250, 350] \cup [550, 700]$
1000	$[800, 1050]$	$[700, 800] \cup [1050, 1100]$

Table 3 is the cutflow table of the selection cuts. The number of events used in mixed training samples could be computed from these cross-sections. The training sample size is presented in Table 4.

Consider the DNN CWoLa classifier. The Higgs candidates are reconstructed by the min-ΔR pairing method. In the min-ΔR method, the four b-tagged jets with the highest p_T

Table 3: The cross sections for the di-Higgs signal and background processes at different selection cuts.

| m_H (GeV) | | Cross section (fb) | | S/B | $\mathcal{L} = 139$ fb^{-1} |
		Signal	Background		$S/\sqrt{B}$
500	Four tag	3.64	6.03e+03	6.03e-04	0.553
	Signal region	3.13	2.57e+03	1.22e-03	0.727
	Sideband region	0.35	2.36e+03	1.50e-04	0.086
1000	Four tag	0.081	6.03e+03	1.34e-05	0.0123
	Signal region	0.063	3.32e+02	1.90e-04	0.0408
	Sideband region	0.010	3.19e+02	3.03e-05	0.0064

Table 4: The training sample size for the mixed sample. The luminosity is $\mathcal{L} = 78$ fb^{-1} because the generated samples are not enough for now.

| m_H (GeV) | Mixed sample | True label | |
		Signal	Background
500	M_1	244	200k
	M_2	28	184k
1000	M_1	5	26k
	M_2	1	25k

are used to form the two Higgs boson candidates. The min-ΔR method selects the pairing configuration in which the higher-p_{T} jet pair has the smallest ΔR separation. The input features are similar to the previous case (Table 1), but the b-tagging information and the di-Higgs system's total invariant mass are excluded. For min-ΔR pairing, it only uses the b-tagged jets. Total invariant mass is already used to determine the signal and sideband region.

2.2 Training results

Table 5 presents the DNN classification training results. These numbers are evaluated from the pure samples, which consist of 5k signal events and 5k background events. The training datasets with and without signal events have similar results. This suggests that the DNN fails to distinguish the signal and background samples but learns the difference between the signal and sideband region. Moreover, the results also imply the input features may correlate to the total invariant mass of the di-Higgs system.

Table 5: The CWoLa DNN training results. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

m_H (GeV)		ACC	AUC
500	With signal	0.708 ± 0.002	0.770 ± 0.007
	No signal	0.705 ± 0.003	0.769 ± 0.009
1000	With signal	0.868 ± 0.024	0.925 ± 0.023
	No signal	0.850 ± 0.033	0.909 ± 0.026

Figure 4 shows the signal score distributions. Even though the signal scores are very different for signal and background distributions, the difference probably stems from the m_{hh} distribution.

There are two issues:

- The input features might correlated to the observables used to determine the signal and sideband region. We need to construct other independent input variables.

- The signal fraction is too low. It is hard to learn something about signal events.

2.3 Correlation matrix

The results in Section 2.2 imply that the di-Higgs system's total invariant mass is not independent of other input features. To find the variables that are highly dependent on

(a) $m_H = 500$ GeV

(b) $m_H = 1000$ GeV

Figure 4: The signal score distributions. We apply the CWoLa DNN on pure samples to obtain the signal score distributions.

the total invariant mass, the correlation coefficients are computed among these variables. Figure 5 and 6 are correlation coefficients on the 500 GeV and 1000 GeV cases, respectively.

Figure 5: The correlation coefficients among different variables, which are computed from 500 GeV testing sample, which consists of 5k signal and 5k background.

The results show that the transverse momentum p_T and the invariant mass m of Higgs candidates are highly correlated to the total invariant mass. Figure 7 shows the scatter plots of the transverse momentum of the leading Higgs candidate and the total invariant mass m_{hh}. These plots also explain why the DNN only trained on background samples can distinguish the signal and background events, because the background distribution in the signal and sideband regions are different.

Figure 6: The correlation coefficients among different variables, which are computed from 1000 GeV testing sample, which consists of 5k signal and 5k background.

(a) $m_H = 500$ GeV (b) $m_H = 1000$ GeV

Figure 7: The scatter plots of the transverse momentum of leading Higgs candidate $p_{\text{T}1}$ and total invariant mass m_{hh} distribution. The signal region is between the red dashed lines. The sideband region is between the green dashed lines and excludes the signal region.

2.4 Remove highly correlated features

Figure 5 and 6 show that the transverse momentum p_{T} and the invariant mass m of Higgs candidates are highly related to the total invariant mass m_{hh}. To investigate the impact of these highly correlated features on the discrimination power of CWoLa DNN models, we remove these input features and train the DNN model again.

Table 6: The CWoLa DNN training results. The transverse momentum and invariant mass of Higgs candidates are removed from samples. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

m_H (GeV)		ACC	AUC
500	With signal	0.526 ± 0.020	0.536 ± 0.053
	No signal	0.532 ± 0.015	0.543 ± 0.029
1000	With signal	0.586 ± 0.030	0.625 ± 0.046
	No signal	0.564 ± 0.024	0.583 ± 0.042

Table 6 summarizes the results of the CWoLa DNN training without p_{T} and m features. The training datasets with and without signal events still have similar results. Compared to the previous one (Table 5) the accuracy values are closer to 0.5. These results suggest that

the removed features have a significant contribution to the model's discrimination power, and the remaining parameters are hard to utilize to distinguish the signal and background events.

2.5 Transverse momentum cut testing

In Figure 3, the distribution of the background sample exhibits a gradual termination around 150 GeV. To investigate whether this termination is a result of the "four-tag cut", which requires $p_T > 40$ GeV, total invariant mass distributions with different p_T cuts are plotted in Figure 8. As the transverse momentum requirement increases from 40 GeV to 70 GeV, the termination point also shifts to larger values. Moreover, the termination remains gradual rather than an abrupt cut-off, suggesting that the gradual termination indeed results from the transverse momentum cut.

Figure 8: The total invariant mass m_{4j} distribution of background samples. The transverse momentum requirement is varied from 40 GeV to 70 GeV.

2.6 Enlarge the signal sample size

Another issue arises from the low signal fraction (Table 4), making DNN difficult to extract meaningful information about signal events. To investigate the impact of signal sample size, we increase the signal size manually and retrain the DNN model. The training sample sizes are summarized in Table 7.

Table 8 provides the results of the CWoLa DNN training without p_T and m features. For the 500 GeV case, the "0 times,", "1 times," and "10 times" samples yield similar results,

11

Table 7: The training sample size for the mixed sample. Various signal sizes are considered, and the background sizes are fixed for all cases. "1 times" represents the previous "With signal" case and "0 times" represents the previous "No signal" case.

| m_H (GeV) | Mixed sample | Signal | | | | Background |
		1 times	0 times	10 times	100 times	All
500	M_1	244	0	2438	24380	200k
	M_2	28	0	276	2760	184k
1000	M_1	5	0	49	492	26k
	M_2	1	0	8	75	25k

Table 8: The CWoLa DNN training results. The transverse momentum and invariant mass of Higgs candidates are removed from samples. 1 time and 0 times are the with signal and no signal case in Table 6. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

m_H (GeV)	times	ACC	AUC
500	1	0.526 ± 0.020	0.536 ± 0.053
	10	0.531 ± 0.027	0.533 ± 0.045
	100	0.634 ± 0.014	0.751 ± 0.030
	0	0.532 ± 0.015	0.543 ± 0.029
1000	1	0.586 ± 0.030	0.625 ± 0.046
	10	0.626 ± 0.027	0.678 ± 0.040
	100	0.621 ± 0.012	0.670 ± 0.023
	0	0.564 ± 0.024	0.583 ± 0.042

while "100 times" sample exhibits better performance. This suggests that the CWoLa DNN can extract meaningful information from the "100 times" sample. In the case of 1000 GeV, we can obtain better results when the signal sample size increases. The performance of 10 times and 100 times is similar. It seems that the training performance is saturated.

To further understand the behavior between 10 times and 100 times samples for the 500 GeV case, additional samples within this size range are generated, and the DNN is trained on these samples.

Figure 9: The accuracy of CWoLa DNN training as a function of the signal size. The unit of sample size is the size of the "1 times" case. The error bar is the standard deviation of 10 training. The grey band is the error bar of the "without signal" case.

Figure 9 is the training performance against the signal sample size. In this region, the performance increases when the signal size is increased. 120 times and 150 times samples are also generated and used in training. The accuracy is saturated at around 63%.

Similarly, for the 1000 GeV case, the DNN is trained on samples with sizes ranging from 1 to 10 times. Figure 10 is the training performance against the signal sample size. The performance is similar for all cases. The training accuracy is saturated at around 62%.

2.7 Training with deeper model

In Figure 10, the performance of CWoLa DNN is quickly saturated. To investigate the impact of the model structure, the deeper DNN model is trained. In Section 2.6, DNN consists of 2 hidden layers, while in this section we train the DNN with 4 hidden layers.

The DNN is trained on signal sample size ranging from 1 to 500 times. Table 9 and
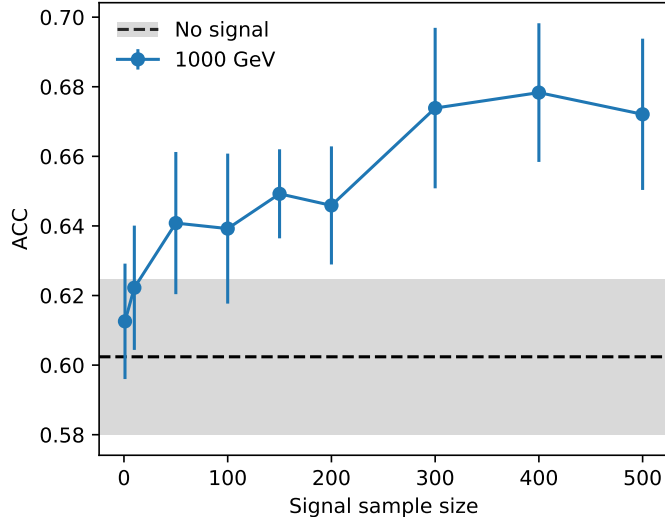
Figure 10: The performance of CWoLa DNN training as a function of the signal size. The unit of sample size is the size of the "1 times" case. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

Figure 11 are the training results. The performance is generally better than the previous results (Table 8), even for the without signal case. It seems that the previous model structure is too simple and it limits the training performance. For the signal size within 1 times to 300 times, the performance increases when the signal size is increased. After this region, the accuracy does not significantly improve. The accuracy is saturated at around 67%, but this value is still better than the previous ones.

Table 9: The CWoLa DNN training results. The transverse momentum and invariant mass of Higgs candidates are removed from samples. 1 time and 0 times are the with signal and no signal case in Table 6. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

m_H (GeV)	times	ACC	AUC
1000	1	0.613 ± 0.017	0.649 ± 0.021
	10	0.622 ± 0.018	0.673 ± 0.033
	100	0.639 ± 0.022	0.695 ± 0.034
	0	0.602 ± 0.022	0.643 ± 0.042

14

Figure 11: The performance of CWoLa DNN training as a function of the signal size. The unit of sample size is the size of the "1 times" case. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

3 Physical data augmentation

The physical augmentations are inspired by Reference [2], which considers the rotation and smearing augmentations. These augmentations reflect both the symmetries in the physical event and the experimental resolution of the detector.

3.1 Original training data

The signal is the resonant Higgs boson pairs production in the four-b quarks channel. In this section, the Higgs boson pair is produced by the heavy CP-even scalar H with mass $m_H = 500$ GeV. The background consists of QCD multi-jet events. The basic requirement is the "four-tag cut," which requires at least four b-tagged $R = 0.4$ anti-k_t jets with $p_T > 40$ GeV and $|\eta| < 2.5$. Only the events passing the four-tag cut are used in the following analysis.

The training samples consist of 50k signal events and 50k background events and the testing samples consist of 5k signal events and 5k background events.

The Higgs candidates are reconstructed by the min-ΔR pairing method. The input features are similar to the previous case (Table 1), but the b-tagging information is excluded.

3.2   Physical augmentation

We consider three different physical augmentations.

1. Azimuthal rotation: The final state is rotated by an angle ϕ randomly sampled from $[0, 2\pi]$.

2. $\eta - \phi$ smearing: The (η, ϕ) coordinate of Higgs candidates are resampled according to a Normal distribution centered on the original coordinate and with a standard deviation inversely proportional to the p_{T}

$$\eta' \sim \mathcal{N}\left(\eta, \frac{\Lambda}{p_{\mathrm{T}}}\right), \quad \phi' \sim \mathcal{N}\left(\phi, \frac{\Lambda}{p_{\mathrm{T}}}\right) \tag{2}$$

where η', ϕ' are the augmented coordinate, p_{T} is the transverse momentum of the Higgs candidate, and the smearing scale is set to be $\Lambda = 10$ GeV.

3. p_{T} smearing: The p_{T} of Higgs candidates are resampled according to

$$p'_{\mathrm{T}} \sim \mathcal{N}\left(p_{\mathrm{T}}, f(p_{\mathrm{T}})\right), \quad f(p_{\mathrm{T}}) = \sqrt{0.052 p_{\mathrm{T}}^2 + 1.502 p_{\mathrm{T}}} \tag{3}$$

where p'_{T} is the augmented transverse momentum, $f(p_{\mathrm{T}})$ is the energy smearing applied by `Delphes` (the p_{T}'s are normalised by 1 GeV).

Figure 12, 13 and 14 are the distributions before and after the augmentation. The distributions for the $\eta - \phi$ smearing are similar for both cases. For p_{T} smearing, the peak broadens and the transverse momentum distribution looks smoother.

For each type of augmentation, we test "n times augmentation" with different n. The n times augmentation means for one original sample, we generate n augmented samples. Additionally, we test another case that applies all augmentations at the same time.

3.3   Training results

Table 10 presents the DNN classification training results of the original sample. Table 11 are the training results of the augmented samples. For each type of augmentation, they all can improve the ACC by about 4%. The differences among the various augmentation are not significant. The 10-times augmentation has the best results, but the difference between the 5-times and 10-times augmentation is tiny. It seems that the performance of this classifier is saturated.
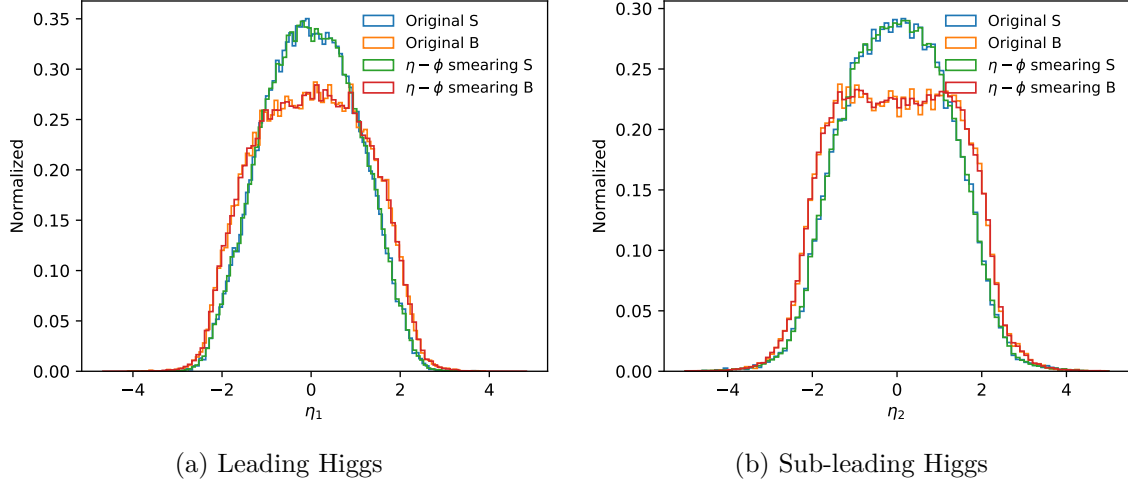
(a) Leading Higgs   (b) Sub-leading Higgs

Figure 12: The pseudorapidity distribution before and after the $\eta - \phi$ smearing augmentation. η_1 and η_2 are the pseudorapidities of the leading and the sub-leading Higgs candidate, respectively.

(a) Leading Higgs   (b) Sub-leading Higgs

Figure 13: The azimuthal angle distribution before and after the $\eta - \phi$ smearing augmentation. ϕ_1 and ϕ_2 are the azimuthal angles of the leading and the sub-leading Higgs candidate, respectively.

Table 10: The training results of original samples. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

	Original
ACC	0.845 ± 0.015
AUC	0.917 ± 0.005

17

(a) Leading Higgs   (b) Sub-leading Higgs

Figure 14: The transverse momentum distribution before and after the p_{T} smearing augmentation. $p_{\mathrm{T}1}$ and $p_{\mathrm{T}2}$ are the transverse momentum of the leading and the sub-leading Higgs candidates, respectively.

Table 11: The training results of augmented samples. "+ 3 times" means the training sample consists of the original sample and 3 times the augmented sample. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

Sample		Rotation	$\eta - \phi$ smear	p_{T} smear	All
+ 3 times	ACC	0.880 ± 0.007	0.879 ± 0.010	0.882 ± 0.003	0.875 ± 0.011
	AUC	0.950 ± 0.007	0.949 ± 0.008	0.951 ± 0.003	0.942 ± 0.012
+ 5 times	ACC	0.887 ± 0.002	0.887 ± 0.001	0.890 ± 0.002	0.889 ± 0.003
	AUC	0.955 ± 0.001	0.955 ± 0.001	0.957 ± 0.001	0.956 ± 0.001
+ 10 times	ACC	0.889 ± 0.001	0.889 ± 0.002	0.892 ± 0.002	0.892 ± 0.002
	AUC	0.956 ± 0.001	0.956 ± 0.001	0.958 ± 0.001	0.958 ± 0.000

18

3.4 Deeper model

In Section 3.3, the DNN model consists of 2 hidden layers, each containing 64 hidden nodes. To explore the impact of the model structure, the deeper DNN model is trained. We investigate the performance of the DNN model with 5 hidden layers.

Table 12: The training results of deeper DNN model. "+ 3 times" means the training sample consists of the original sample and 3 times the augmented sample. ACC is the best accuracy and AUC is the area under the ROC curve. The average and standard deviation of 10 training are presented.

Sample		Original	+ 3 times	+ 5 times	+ 10 times
All augmentation	ACC	0.864 ± 0.005	0.890 ± 0.002	0.890 ± 0.002	0.884 ± 0.005
	AUC	0.928 ± 0.005	0.957 ± 0.001	0.957 ± 0.001	0.949 ± 0.005

Table 12 are the training results with a deeper DNN model. Models are only trained on the "All augmentation" sample because from Table 11 we found that four augmentation methods yielded similar results. The results show that the augmented sample can improve ACC to 89%, even from the "+ 3 times" case and this accuracy value is similar to the previous test. These findings suggest that the classifier may have reached a saturation point and point out the difficulty of further improving accuracy on this test sample.

4 Hidden Valley model

4.1 Sample generation

The signal process is $f\bar{f} \to Z_V$, where Z_V is the massive gauge boson linking SM and the dark sector. The hidden Z_V boson would decay to a pair of dark quark $q_V \bar{q}_V$, which would lead to two jets in the detector. The signal sample is generated by `Pythia` and the detector simulation is done by `Delphes`. The anti-k_t algorithm is utilized for jet reconstruction with parameter $R = 0.8$. Some parameters are listed in Table 13.

The background sample is the SM QCD di-jet. This process is generated at $\sqrt{s} = 13$ TeV. Following are the `MadGraph` scripts for generating background samples:

```
generate p p > j j
output ppjj
launch ppjj
```

Table 13: The parameter setting for the Hidden Valley model. "490010x:m0 = 10.3306; x=1,2,3" means x should be replaced by 1,2,3 in Pythia card.

Parameter	Value	Pythia card
M_{Z_V}	5.5 TeV	4900023:m0 = 5500
$\sqrt{s}$	13 TeV	
Λ_{D}	10 GeV	HiddenValley:Lambda = 10.0
$m_{\pi_{\mathrm{D}}}$	10 GeV	4900x1:m0 = 10.0; x=11,21,31,22,32,33
$m_{\rho_{\mathrm{D}}}$	26.944 GeV	4900x3:m0 = 26.944; x=11,21,31,22,32,33
$m_{q,\mathrm{constituent}}$	10.3306 GeV	490010x:m0 = 10.3306; x=1,2,3

```
shower=Pythia8
detector=Delphes
analysis=OFF
madspin=OFF
done


Cards/delphes_card_CMS.dat


set run_card nevents 10000
set run_card ebeam1 6500.0
set run_card ebeam2 6500.0

set run_card ptj 700
set run_card etaj 2.2
set run_card mmjj 3000


done
```

4.2 Problem for generating signal sample

Error messages:

```
PYTHIA Error: input string not found in settings databases::
  HiddenValley:separateFlav    = on! Consider different flavours


PYTHIA Error: input particle not found in Particle Data Table:
```

```
4900102:m0                          = 10.3306
```

...

Solution: This problem arises from the `Pythia` version. At first, `Pythia 8.306` is used to generate signal samples. Some features are not included in this version. We should use `Pythia 8.307` at least. More details between 8.306 and 8.307 can be found in this page.

4.3 Sample selection

The selection cuts after the `Delphes` simulation:

- n_j cut: The number of jets should be greater than or equal to 2.

- p_T cut: The transverse momentum of two highest p_T jets should greater 750 GeV.

- η cut: The η range of two highest p_T jets are require $|\eta| < 2$.

- Signal region: Total invariant mass of two leading jets m_{jj} belonging to $[4700, 5500]$.

- Sideband region: Total invariant mass of two leading jets m_{jj} belonging to $[4300, 4700] \cup [5500, 5900]$.

Table 14 is the cutflow number at different selection cuts. Figure 15 is transverse momentum distribution after n_j cut. Figure 16 is the m_{jj} distribution after the η cut.

Table 14: The number of passing events and passing rates for signal and background processes at different selection cuts.

Cut	Signal	pass rate	Background	pass rate
Total	100000	1	100000	1
n_j cut	99996	1.00	99963	1.00
p_T cut	90901	0.91	57832	0.58
η cut	89800	0.90	55523	0.56
SR region	55844	0.56	1991	0.02
SB region	16079	0.16	3090	0.03

(a) Signal

(b) Background

Figure 15: The transverse momentum distribution of leading and sub-leading jets. The red dashed lines are the p_T cut.

(a) My plot

(b) Zong-En's plot

Figure 16: The total invriant mass m_{jj} distribution of signal and background samples. In my plot, the signal region is between the red dashed lines. The sideband region is between the green dashed lines and excludes the signal region.

4.4 Jet image

We can construct the jet image from the event passing the selection cuts described in Section 4.3. The jet image is constructed for each jet separately, so for each event, we would obtain two jet images. They are treated as two channels of a picture. The following steps construct the jet image:

1. Centralization: Compute the p_{T} weighted center in (η, ϕ) coordinates, then shift this point to origin.

2. Rotation: Rotate the highest intensity axis to the η axis.

3. Flipping: Flip the highest intensity quadrant to the first quadrant.

4. Pixelating in $\eta \in [-1, 1]$, $\phi \in [-1, 1]$ box, with 75×75 pixels.

Figure 17 is the jet images of a single event. Figure 18 is the average jet image.

(a) Signal   (b) Background

Figure 17: The jet images of the leading jet in the signal region. The η' and ϕ' are the coordinates after the preprocessing (centralization, rotation, flipping).

4.5 Datasets

The total cross-section of the background events is 6837 fb. From Table 14, we can compute the corresponding cross-sections of signal and sideband regions are 136.1 fb and 211.2 fb, respectively. Thus, the numbers of events in signal and sideband region are 19k and 29k at luminosity $\mathcal{L} = 139$ fb^{-1}.

The training sample sizes are summarized in Table 15.

(a) Signal (b) Background

Figure 18: The average jet images of the leading jet in the signal region. The η' and ϕ' are the coordinates after the preprocessing (centralization, rotation, flipping). The number of the signal events is 56k. The number of the background events is 20k.

Table 15: The training sample size for the mixed sample. We set sensitivity $S/\sqrt{B} = 1$ in the signal region and evaluate the number of events in the signal region. Then, the number of events in the sideband region can be obtained from Table 14.

	True label	
Mixed sample	Signal	Background
Signal region	138	19k
Sideband region	40	29k

The pure testing sample consists of 10k signal events and 10k background events, which are selected from the signal region.

4.6   Training CNN

The CNN model structure is summarized in Figure 19. The internal node uses the rectified leaner unit (ReLU) as the activation function. The loss function is Categorical cross-entropy, and we use the Adam optimizer to optimize the loss function.

4.7   Hidden Valley model training results

The CNN is trained on samples with sensitivity $S/\sqrt{B}$ ranging from 1 to 10. Figure 20 presents the training results. These numbers are evaluated from the pure samples. The CNN cannot learn useful information for the case with a sensitivity $S/\sqrt{B} \leq 5$, so the ACC is 0.5. After this region, the accuracy demonstrates a significant improvement. It seems that the CNN model surpasses a certain threshold.

4.8   Data process procedure

1. Generate the sample file in `.root` format. Following Section 4.1.

2. Apply the selection cuts described in Section 4.3 and save the event passing the cuts in `HDF5` format. The file contains the information listed below

   - The $(p_\mathrm{T}, \eta, \phi)$ of leading and sub-leading jet constituents.
   - Total invariant mass m_{jj}.
   - Type of event: 1 for signal, 0 for background.

3. Make mixed sample in `HDF5` format. Following Section 4.5, we can compute the size of datasets.

4. Apply data augmentation in `HDF5` format. Following Section 4.9.

5. Generate the jet image from `HDF5` data.

Note the preprocessing is done in step 5. If we do the preprocessing many times, then the rounding errors would break the jet image.

| InputLayer | input: | [(None, 75, 75, 2)] |
| | output: | [(None, 75, 75, 2)] |

| BatchNormalization | input: | (None, 75, 75, 2) |
| | output: | (None, 75, 75, 2) |

| Conv2D | input: | (None, 75, 75, 2) |
| | output: | (None, 75, 75, 32) |

| MaxPooling2D | input: | (None, 75, 75, 32) |
| | output: | (None, 37, 37, 32) |

| Conv2D | input: | (None, 37, 37, 32) |
| | output: | (None, 37, 37, 32) |

| MaxPooling2D | input: | (None, 37, 37, 32) |
| | output: | (None, 18, 18, 32) |

| Conv2D | input: | (None, 18, 18, 32) |
| | output: | (None, 18, 18, 64) |

| MaxPooling2D | input: | (None, 18, 18, 64) |
| | output: | (None, 9, 9, 64) |

| Conv2D | input: | (None, 9, 9, 64) |
| | output: | (None, 9, 9, 64) |

| Flatten | input: | (None, 9, 9, 64) |
| | output: | (None, 5184) |

| Dense | input: | (None, 5184) |
| | output: | (None, 64) |

| Dense | input: | (None, 64) |
| | output: | (None, 64) |

| Dense | input: | (None, 64) |
| | output: | (None, 64) |

| Dense | input: | (None, 64) |
| | output: | (None, 2) |

Figure 19: For first and second Conv2D layers, the kernel sizes are 5×5. For third and fourth Conv2D layers, the kernel sizes are 3×3.
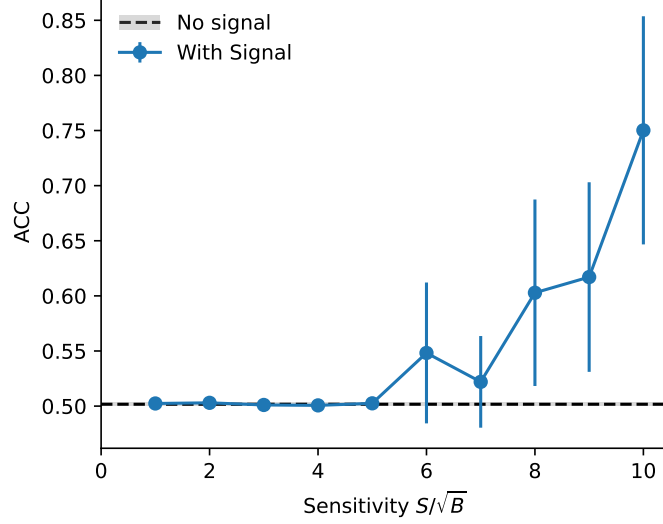
Figure 20: The performance of CWoLa CNN training as a function of the sensitivity. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. For sensitivity less than 5, the error bar is too small to see.

4.9   Data augmentation

To reduce the threshold, the data augmentation technique is tested. Similar to the Section 3.2, we apply the $\eta - \phi$ smearing on our training sample. Specifically, the (η, ϕ) coordinate of jet constituents are resampled according to a Normal distribution centered on the original coordinate and with a standard deviation inversely proportional to the p_T

$$\eta' \sim \mathcal{N}\left(\eta, \frac{\Lambda}{p_\mathrm{T}}\right), \quad \phi' \sim \mathcal{N}\left(\phi, \frac{\Lambda}{p_\mathrm{T}}\right) \tag{4}$$

where η', ϕ' are the augmented coordinate, p_T is the transverse momentum of the jet constituent, and the smearing scale is set to be $\Lambda = 100$ MeV.

Figure 21 and 22 are the jet image before and after the augmentation. The jet images look similar before and after the augmentation, but not the same.

We generate samples with sensitivity $S/\sqrt{B}$ ranging from 1 to 10. Then, apply the data augmentation to make larger samples. These samples are used for CNN training. Figure 23 presents the training results. These numbers are evaluated from the pure samples. The ACC is better than previous results (blue curve in Figure 20) for the "+1 times" curve for all sensitivities. The data augmentation technique seems to suppress the threshold and improve the training performance.

Figure 24 presents the training results with larger samples. For the "+3 times" curve, there is a strange result, the ACC is worse than the "+1 times" case for sensitivities $S/\sqrt{B} \geq$

27
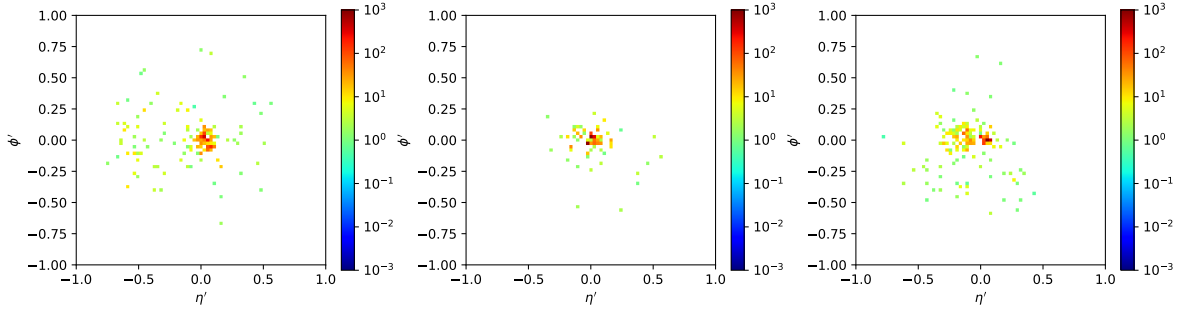
(a) Before augmentation
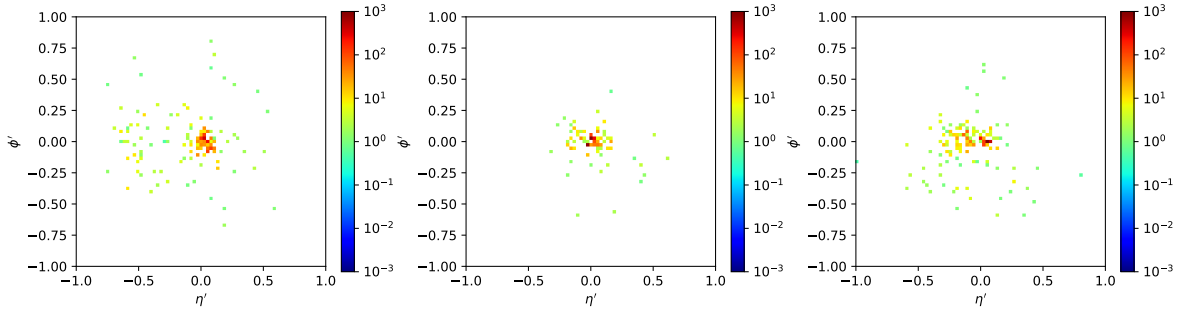
(b) After augmentation

Figure 21: The jet images before and after the $\eta - \phi$ smearing augmentation.
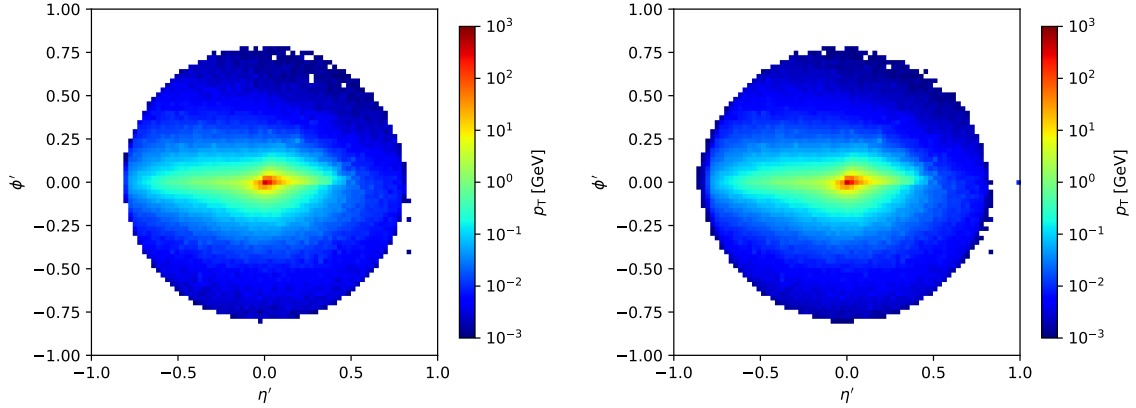
(a) Before augmentation

(b) After augmentation

Figure 22: The average jet images before and after the $\eta - \phi$ smearing augmentation.

Figure 23: The performance of CWoLa CNN training as a function of the sensitivity. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

3. There are some problems in the training procedure, e.g. overfitting.

Figure 25 shows the average jet image for different augmentation times.

4.10   Enlarge the training sample size

To validate the correctness of the data augmentation results, we generated larger samples by scaling the luminosity. To compare the results with "+1 times" augmentation, we scale the luminosity to $\mathcal{L} = 139 \times 2$ fb^{-1}. It is important to note that the number of signal events is not equal for "+1 times" and "luminosity $\times 2$" at the same sensitivity because the number of signal events is computed from $S/\sqrt{B}$ for a given B, resulting in the number of "+1 times" signal sample is greater by a factor of $\sqrt{2}$. Table 16 is an example. For a fair comparison, where the sizes of signal and background events are the same, we should compare the "+1 times" results with the point on the "luminosity $\times 2$" curve corresponding to $\sqrt{2}$ times the sensitivity.

Figure 26 presents the training results with larger samples. The results of "luminosity $\times 2$" is better than the original sample, but still worse than the "+1 times" case. Even considering the $\sqrt{2}$ scale factor, the corresponding points' performance is still worse. There may be something wrong with the data processing or training codes.

Figure 24: The performance of CWoLa CNN training as a function of the sensitivity. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

(a) +1 times

(b) +3 times

Figure 25: The average jet images of the $\eta - \phi$ smearing augmentation with different times.

Table 16: The training sample size for the original, "+1 times" and "luminosity $\times 2$" mixed sample. We set sensitivity $S/\sqrt{B} = 1$ in the signal region and evaluate the number of events in the signal region.

Mixed sample	Original		+1 times		luminosity $\times 2$	
	Signal	Background	Signal	Background	Signal	Background
Signal region	138	19k	276	38k	194	38k
Sideband region	40	29k	80	58k	56	58k

30

Figure 26: The performance of CWoLa CNN training as a function of the sensitivity. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

4.11 Check the codes and make more plots

The data process steps from 2 to 5 (Section 4.8) have been checked and samples are processed again.

Figure 27 is the p_T distribution of the jet constituents. The distributions of leading and sub-leading jet constituents are similar.

Figure 28 and 29 are the η, ϕ distributions before and after the augmentation. Because the smearing scale is small, the distributions look almost the same for both cases.

Figure 30 shows the average jet image for various cases.

Figure 31 is training results with re-processed samples.

4.12 Sensitivity difference

The CWoLa CNN is applied for event selection. The threshold is set for a given background efficiency ε_b (false positive rate). For a given ε_b, the sensitivity would be scale by a factor $\varepsilon_s/\sqrt{\varepsilon_b}$.

Figure 32 is the sensitivity improvement of orginal, "+1 times" and "luminosity ×2" samples. The 10% curves give more stable results. The lower background efficiencies could achieve better results at the range with higher sensitivities, but the standard deviation would be much larger. Figure 33 compares the sensitivity improvement of various training samples with the same background efficiency. The "+1 times" samples provide the best threshold

31

Figure 27: The transverse momentum distribution of jet constituents.

(a) Leading Higgs

(b) Sub-leading Higgs
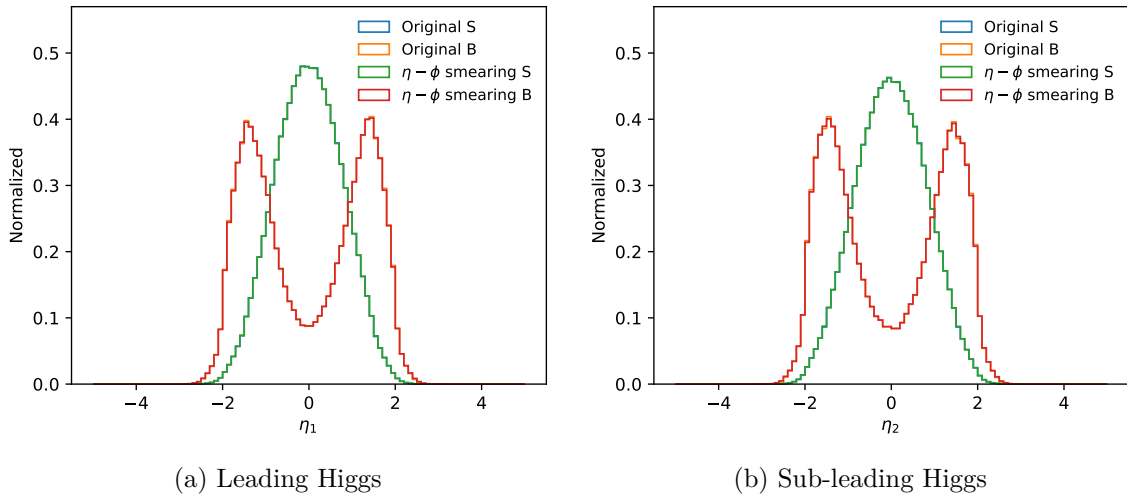
Figure 28: The pseudorapidity distribution before and after the $\eta-\phi$ smearing augmentation. η_1 and η_2 are the pseudorapidities of the leading and sub-leading jet constituents.

(a) Leading Higgs (b) Sub-leading Higgs

Figure 29: The azimuthal angle distribution before and after the $\eta - \phi$ smearing augmentation. ϕ_1 and ϕ_2 are the azimuthal angles of the leading and sub-leading jet constituents.

(a) Original (b) +1 times augmentation (c) Luminosity $\times 2$
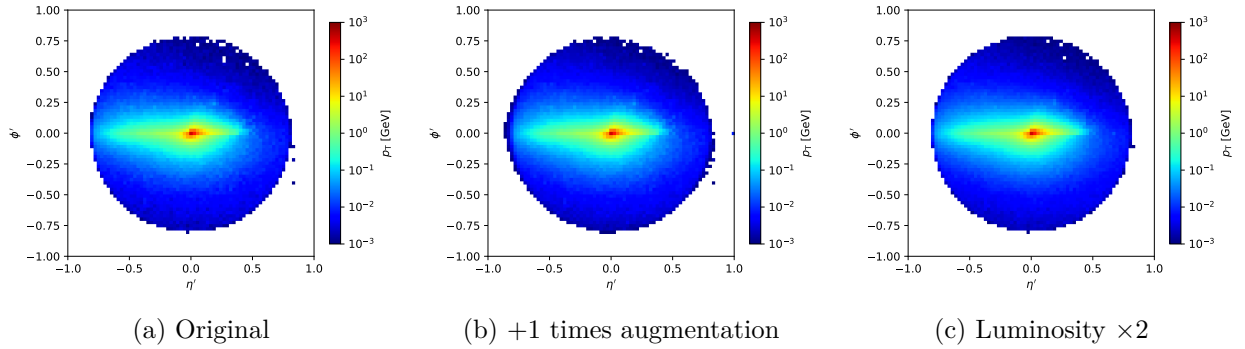
Figure 30: The average jet images of different cases. These plots are made from mixed samples.
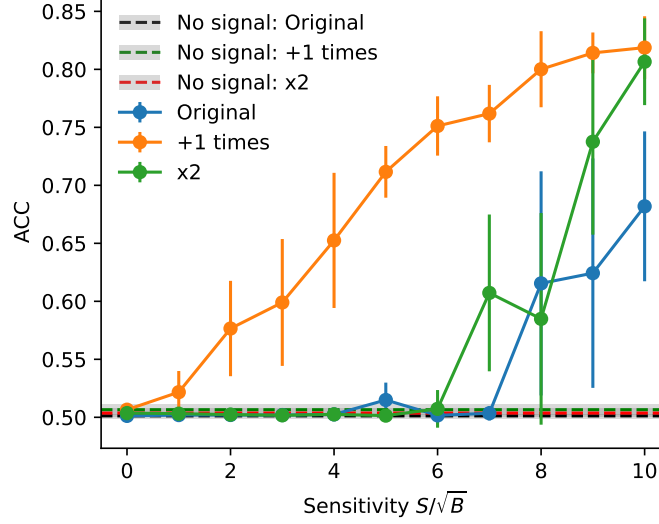
Figure 31: The performance of CWoLa CNN training with re-processed samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case.

among all.

4.13   Training and true performance

In our intuition, we would expect that"+1 times" and "luminosity ×2" samples should exhibit similar performance, or the "luminosity ×2" sample might provide better results since the data augmentation generates more samples from the known data. However, the results in Figure 33 contradict our intuition. Since we use the CWoLa approach, our comparison is not based on the training performance on mixed samples; instead, we evaluate the testing performance on pure samples, referred to as "true performance." The relationship between training and true performance may not be a monotonic function and can be quite complicated.

Figure 34 is the scatter plot for training and true accuracy. Training ACC values are close for all cases, but true ACC values can be much different. The true ACC has a range that is ten times larger than the training ACC range. For training ACC, both "+1 times" and "luminosity ×2" samples show similar performance, slightly better than the original results. For the true ACC, the "+1 times" provides much better results. It seems that "+1 times" samples are more effective in improving true ACC.

Figure 35 includes results with"+3 times" training . The "+3 times" samples provide much better training ACC, but they do not achieve higher performance for true ACC.

(a) Original (b) +1 times augmentation (c) Luminosity ×2
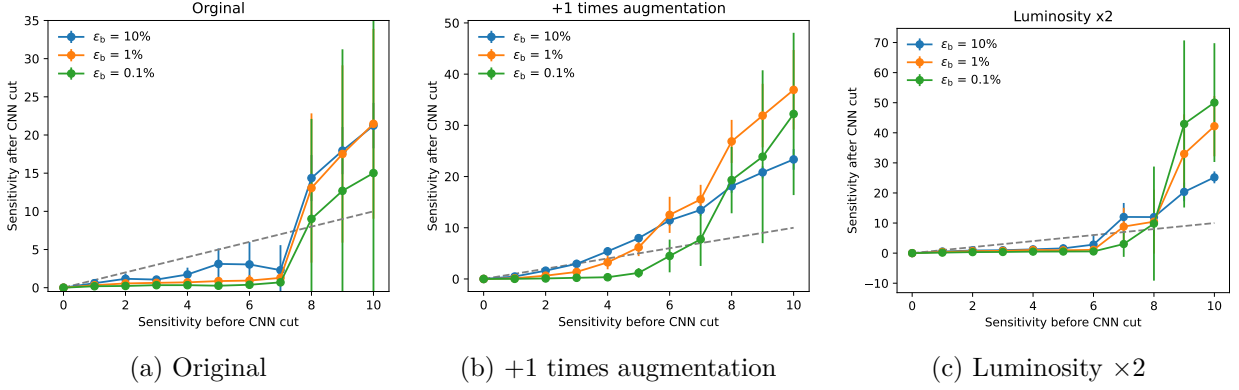
Figure 32: The sensitivities before and after the CWoLa CNN selection. The thresholds are chosen from $\varepsilon_{\mathrm{b}} = 10\%, 1\%, 0.1\%$. The slope of the dashed line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.
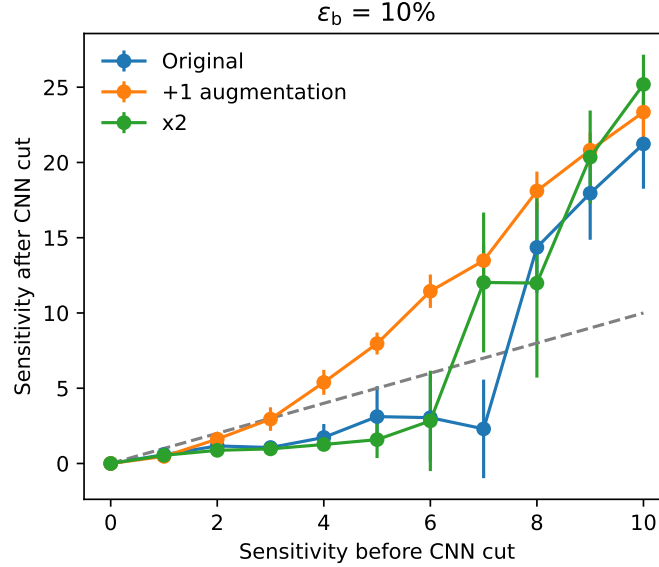

Figure 33: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_{\mathrm{b}} = 10\%$. The slope of the dashed line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The "+1 times" samples provide the best threshold.
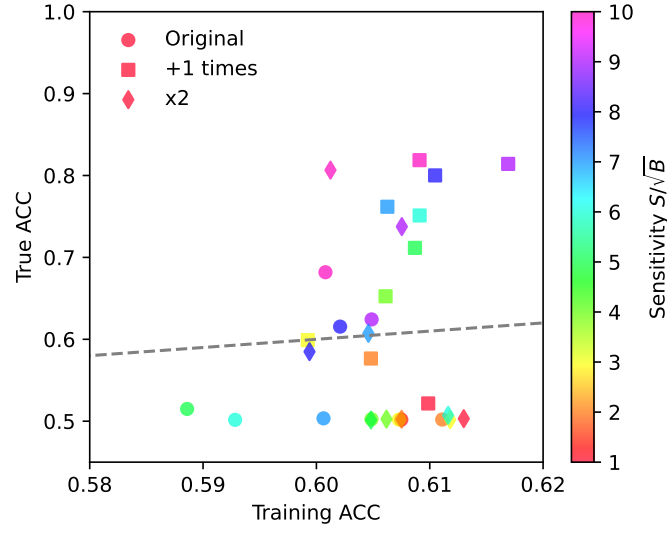
Figure 34: Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC. The range of training ACC is very small, roughly 0.03, while the range of true ACC is about 0.3.
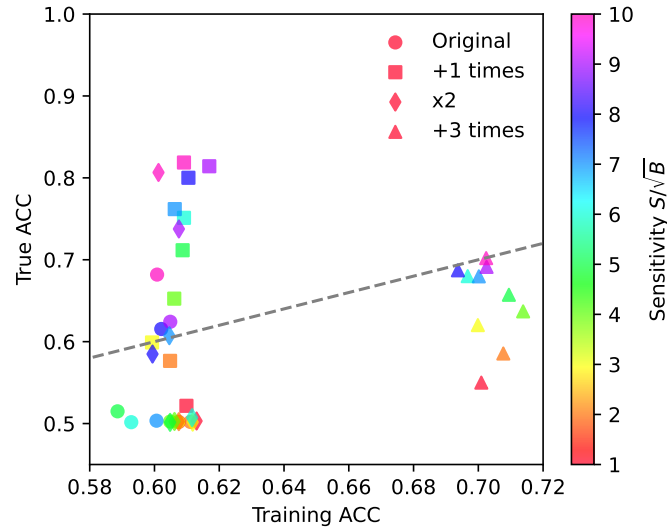
Figure 35: Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

4.14 Miscellaneous test

4.14.1 Sensitivity 11 to 20

To find out the training performance when sensitivity is greater than 10, we generate the samples with sensitivity ranging from 11 to 20. The training performance is summarized in Figure 36. The ACC is increased when the sensitivity is increased. When the $S/\sqrt{B} > 13$, the ACC improvement is slowed down, where the ACC is 85%. The training ACC is still similar for all cases, but the true ACC exhibits better value when the sensitivity is higher.
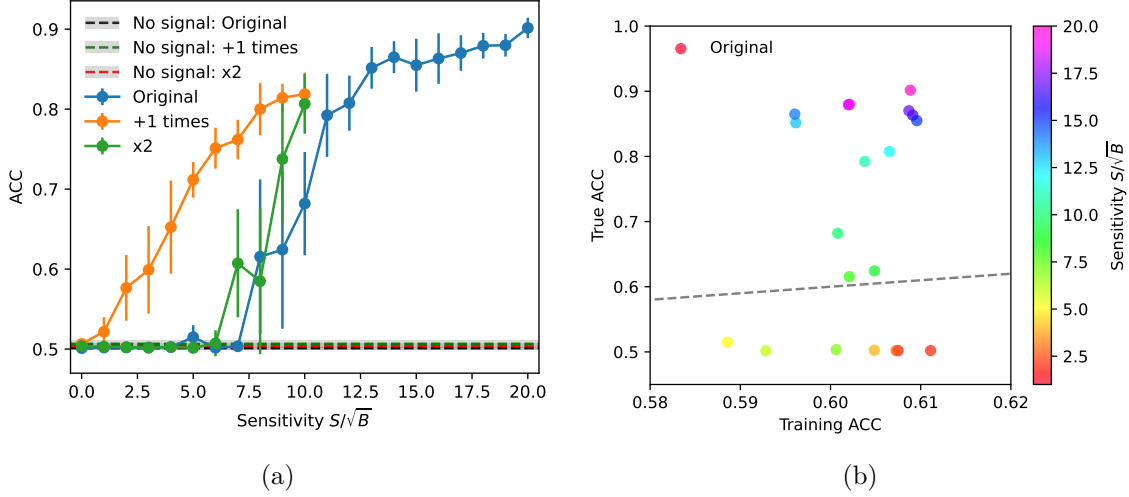
Figure 36: (a) The performance of CWoLa CNN training with higher sensitivity samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

4.14.2 +1 copy sample

To investigate the impact of the data augmentation technique, we generate a sample by duplicating the original sample without any augmentation. This sample is called "+1 copy." Figure 37 is the ACC curve and ACC distribution. The ACC is close "+1 augmentation" case when $S/\sqrt{B} \leq 4$, while the ACC is worse when $S/\sqrt{B} \geq 5$. The training ACC is much better than the original and "+1 augmentation" samples, but the true ACC can not achieve higher values.

Figure 38 is the sensitivity improvement. The performance of the "+1 copy" sample is worse than "+1 augmentation". It exhibits the worst results in high-sensitivity regions. The data augmentation is effective in improving the true ACC.

(a)   (b)
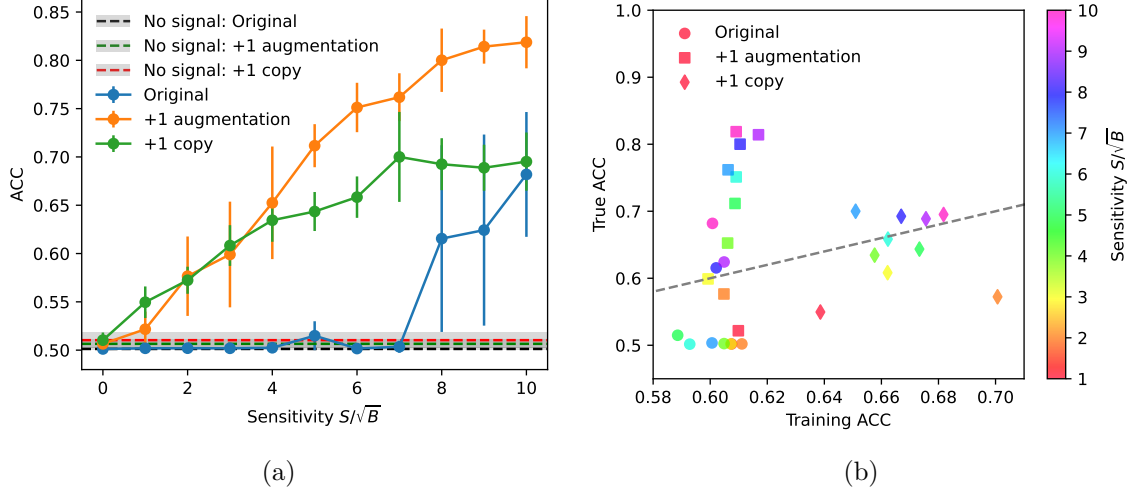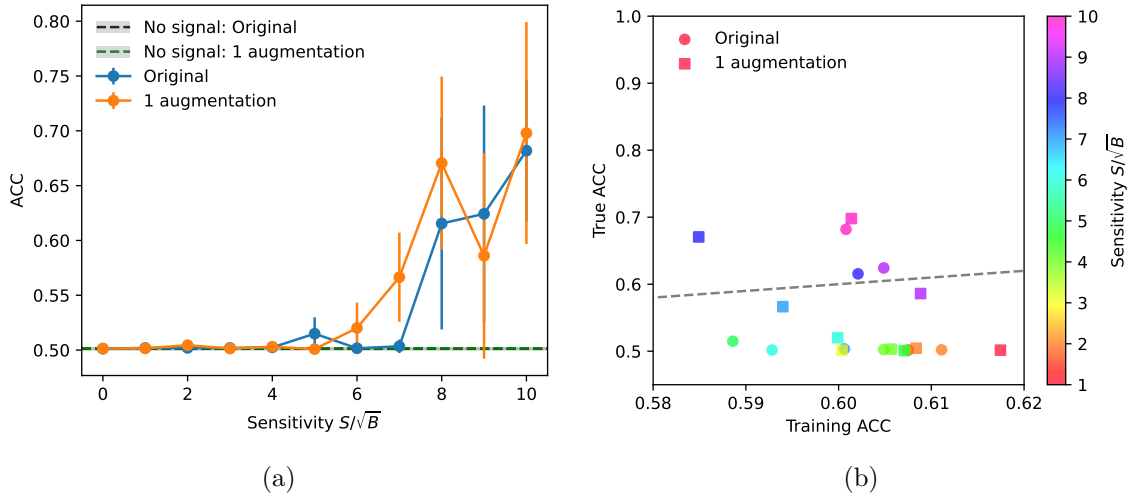
Figure 37: (a) The performance of CWoLa CNN training with "+1 copy" samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.
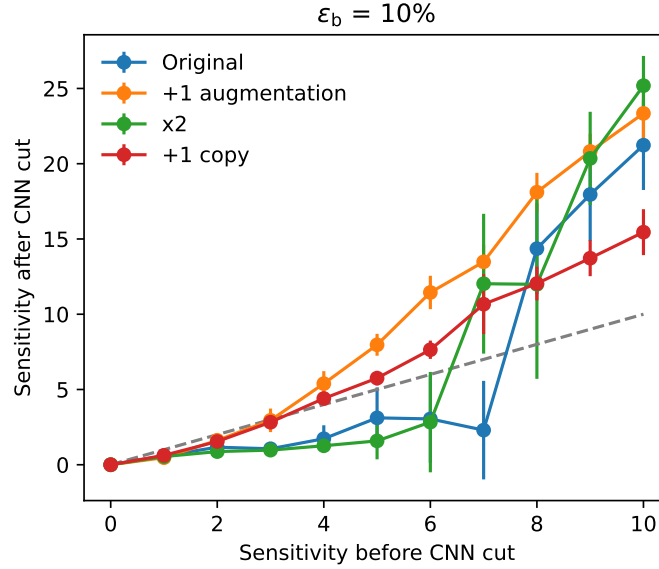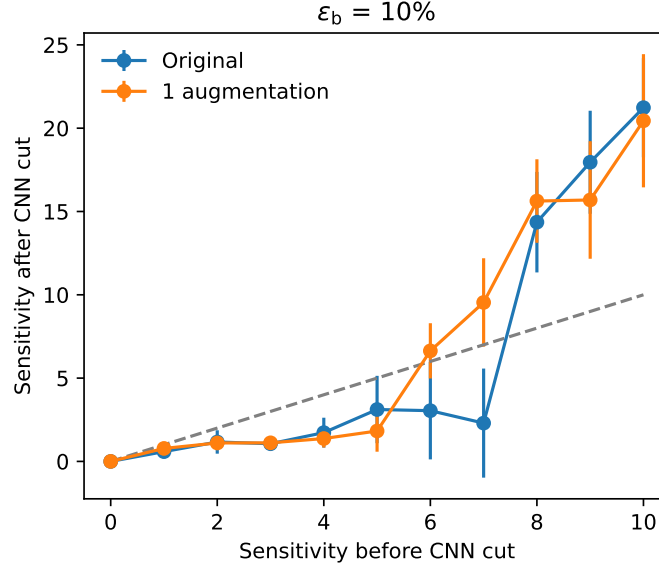

Figure 38: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The "+1 times" samples provide the best threshold.

4.15 Train on only augmented sample

To diagnose augmented samples, we train the CWoLa CNN with only augmented datasets. The training samples are generated from the original dataset and only the augmented part is kept for training. We test "1 augmentation" and "2 augmentation" cases. They consist of pure augmented samples. Their sizes are 1 time and 2 times the original dataset.

Figure 39 shows the ACC curve and ACC distribution with the "1 augmentation" case. The ACC of the "1 augmentation" sample is better than the "original" one at $S/\sqrt{B} = 7$, but they have similar performance on other sensitivities. Figure 40 is the sensitivity improvement. The threshold of the "1 augmentation" sample is lower than the "original" one, but they exhibit similar performance on other sensitivities. These results are consistent with the ACC curve. The difference at $S/\sqrt{B} = 7$ may come from the training fluctuation.

(a)   (b)

Figure 39: (a) The performance of CWoLa CNN training with "1 augmentation" samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

Figure 41 is the training results with "2 augmentation" samples. The results are similar to the "+1 augmentation" case. Figure 42 is the sensitivity improvement. The performance of the "2 augmentation" samples is almost the same as "+1 augmentation".

Figure 40: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_{\mathrm{b}} = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The "1 augmentation" samples provide a lower threshold.
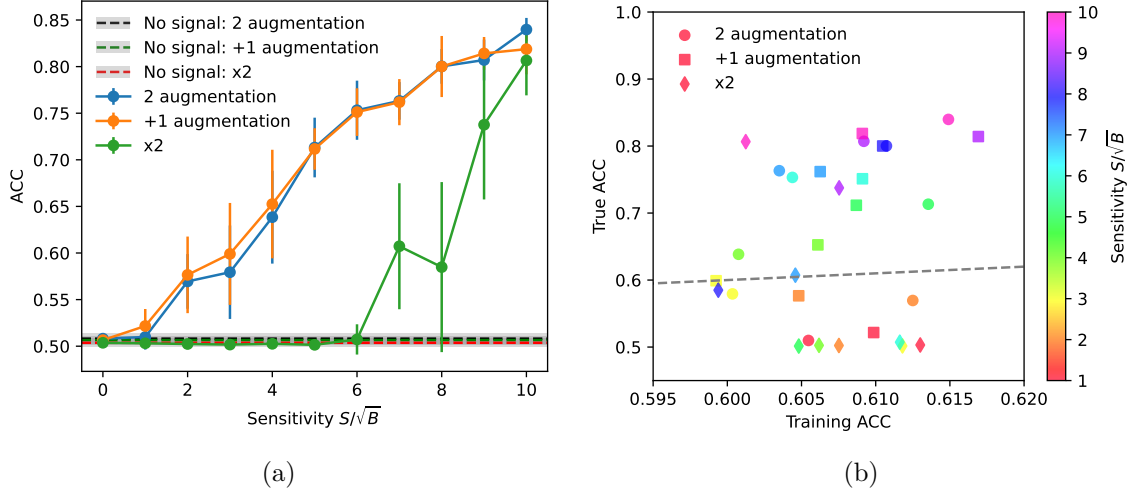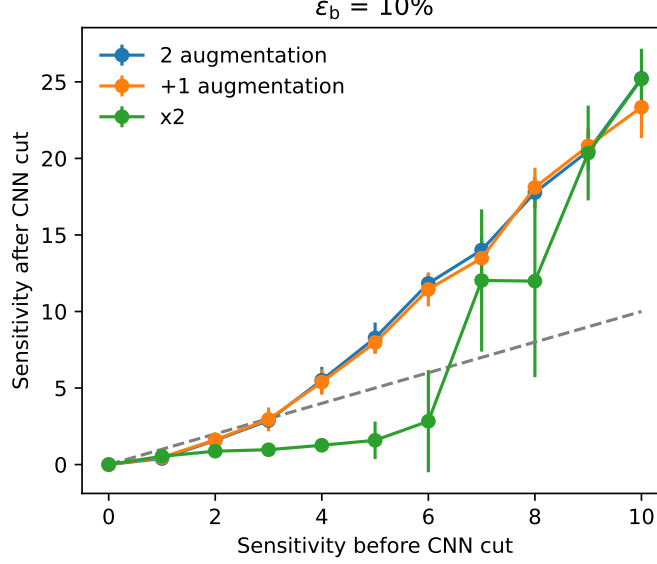

(a)   (b)

Figure 41: (a) The performance of CWoLa CNN training with "2 augmentation" samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

40

Figure 42: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_{\rm b} = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. The "2 augmentation" and "+1 augmentation" samples perform similarly.

4.16 Change the smearing scale

To investigate the impact of the smearing scale Λ of data augmentation, we generate samples with various Λ for CWoLa CNN training. We test $\Lambda = 200, 500$ MeV samples.

Figure 43 and 44 are the jet image with different smearing scales. When the smearing scale increases the jet image is more different.

Figure 45 is the training results with various smearing scales. In the low-sensitivity region, the performance is better for the lower smearing scale, while in the high-sensitivity region, the performance is better for the higher smearing scale. Figure 46 is the sensitivity improvement. The results are consistent with the ACC curve. $\Lambda = 100$ MeV exhibits the lower threshold.
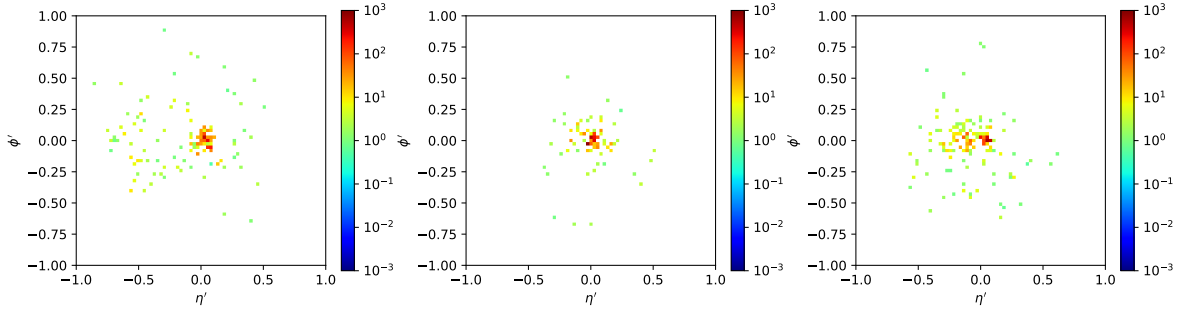
4.17 Duplicate training sample

To compare the results with Section 4.14.2, we generate a sample by duplicating the original sample without further shuffle. Therefore, the first n-th sample and $n + 1$-th to the final sample should be the same. This dataset is called "+1 copy without shuffle." Figure 47 is the average jet image of first n-th sample and $n + 1$-th to the final sample. These two images are the same.

(a) Original

(b) Λ = 100 MeV

(c) Λ = 200 MeV

(d) Λ = 500 MeV

Figure 43: The jet images with different smearing scale Λ.

(a) Original

(b) $\Lambda = 100$ MeV
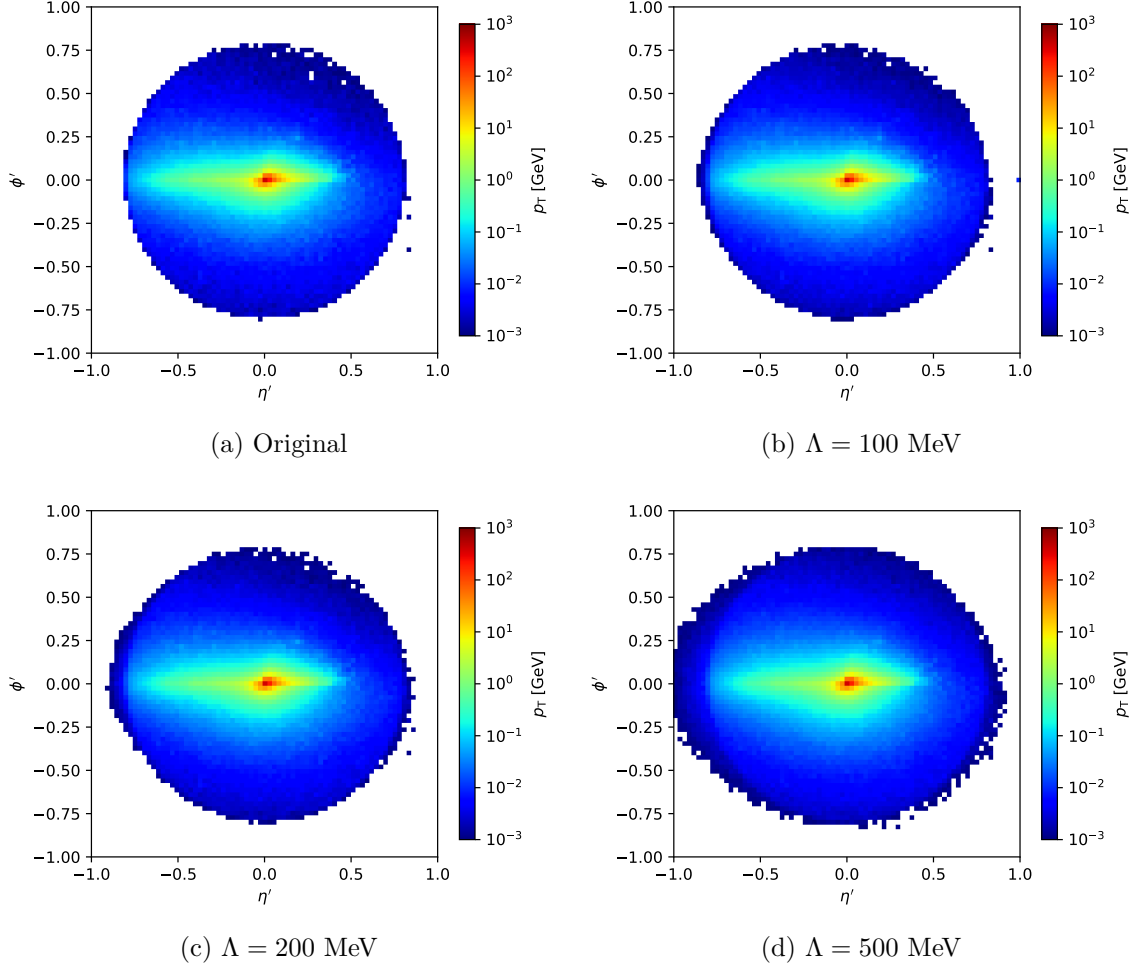
(c) $\Lambda = 200$ MeV

(d) $\Lambda = 500$ MeV

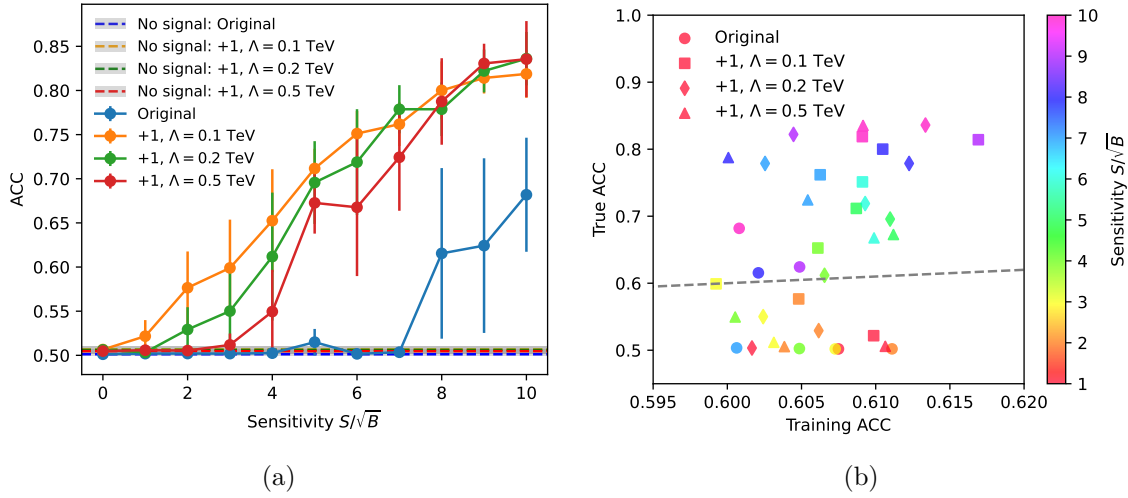Figure 44: The average jet images with different smearing scale Λ.

Figure 45: (a) The performance of CWoLa CNN training with various smearing scales. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. "+1" means the +1 time augmentation sample. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

Figure 46: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training. "+1" means the +1 time augmentation sample.
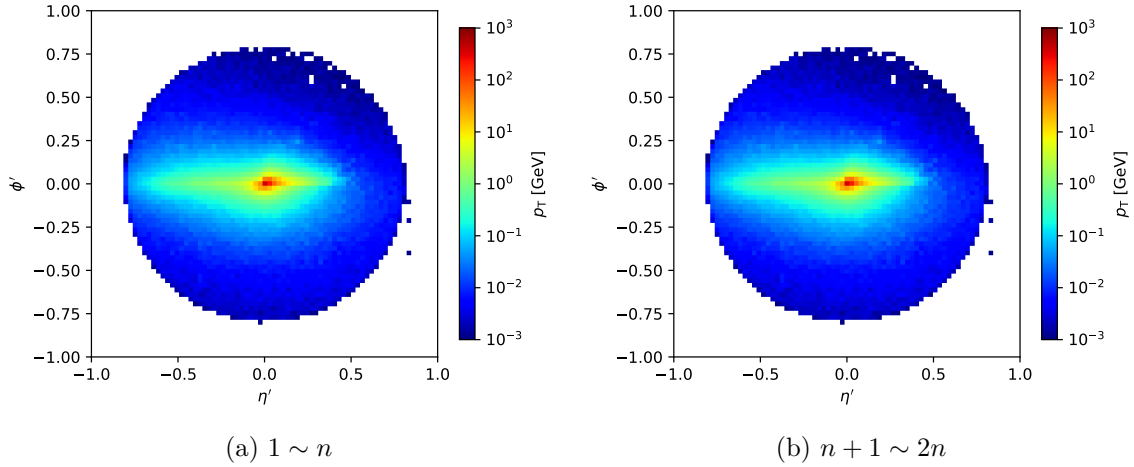
(a) $1 \sim n$ (b) $n + 1 \sim 2n$

Figure 47: The average jet image of first n-th sample and $n + 1$-th to the final sample.

Figure 48 is the ACC curve and ACC distribution. The training results for "+1 copy" and "+1 copy without shuffle" do not have significant differences and they all have better performance than the original training sample.

Figure 49 is the sensitivity improvement. The performance of the "+1 copy" and "+1 copy without shuffle" samples are similar.

4.17.1   Remove shuffle in training code

I remove default shuffles in the training codes. In the default setting, the training sample would be shuffled before training and testing sample splitting and each epoch training. Figure 50 is the ACC curve and ACC distribution. The training results of removing shuffling codes are still similar to the previous one and all "+1 copy" samples have better performance than the original training sample.

Figure 51 is the sensitivity improvement. The performance of the "+1 copy" sample and "+1 copy" with removing shuffling codes are similar.

References

[1] Jack H. Collins, Kiel Howe, and Benjamin Nachman. Anomaly Detection for Resonant New Physics with Machine Learning. *Phys. Rev. Lett.*, 121(24):241803, 2018.

[2] Barry M. Dillon, Luigi Favaro, Friedrich Feiden, Tanmoy Modak, and Tilman Plehn. Anomalies, Representations, and Self-Supervision. 1 2023.
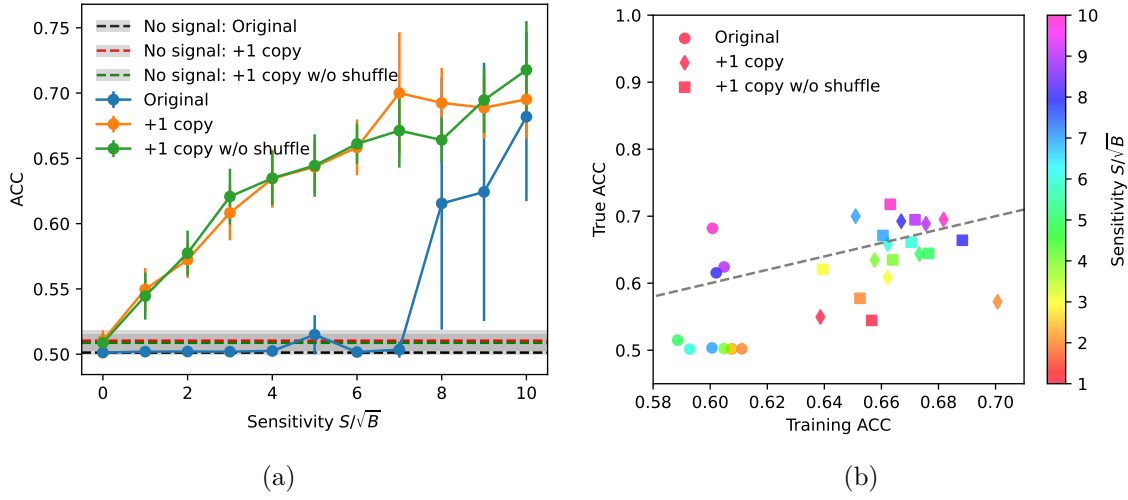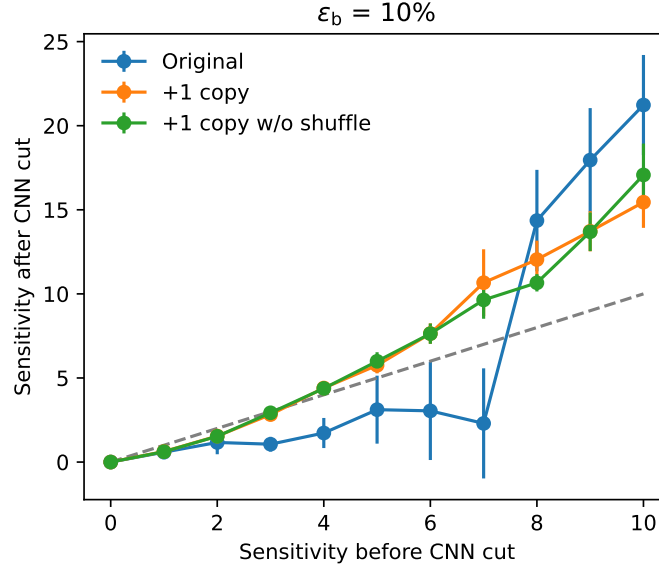
Figure 48: (a) The performance of CWoLa CNN training with "+1 copy" and "+1 copy without shuffle" samples. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.

Figure 49: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_{\rm b} = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.
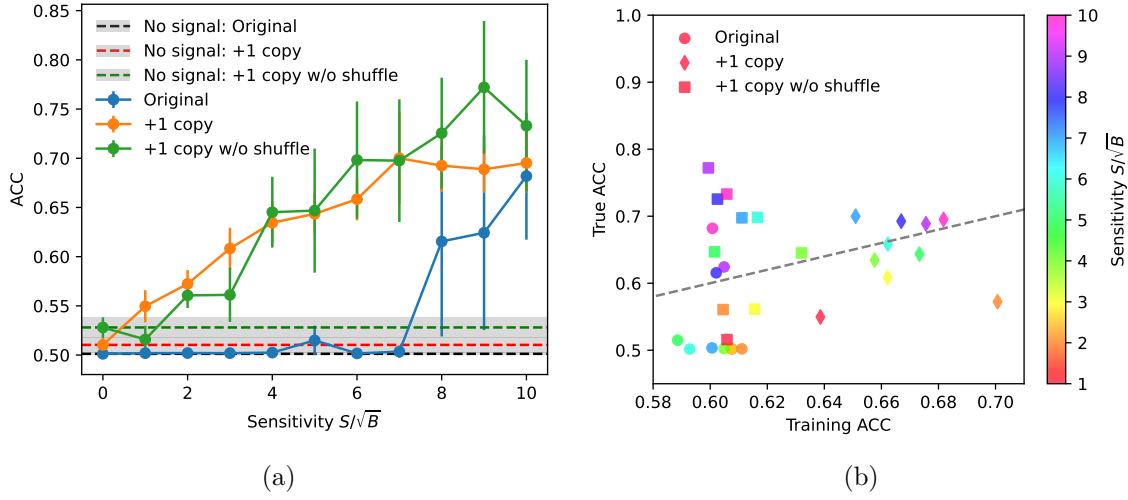
Figure 50: (a) The performance of CWoLa CNN training with "+1 copy" sample and "+1 copy" with removing shuffling codes. The error bar is the standard deviation of 10 times training. The grey band is the error bar of the "without signal" case. (b) Scatter plot for training ACC and true ACC. The slope of the grey dashed line is 1, representing the same training and true ACC.
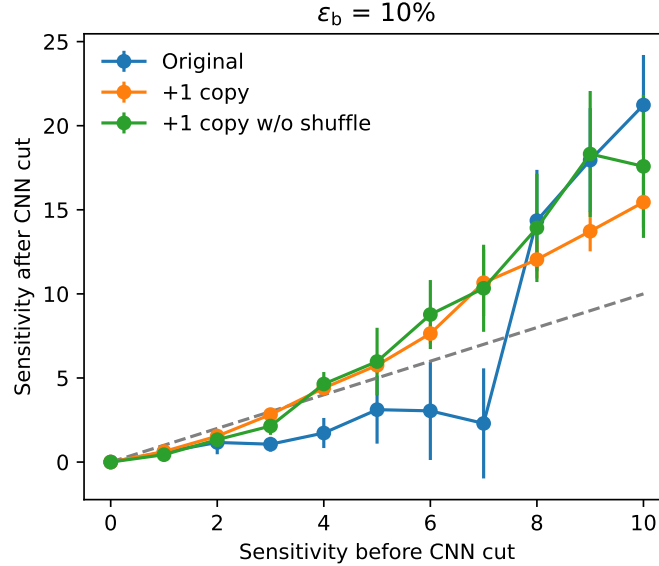

Figure 51: The sensitivities before and after the CWoLa CNN selection. The threshold is chosen from $\varepsilon_b = 10\%$. The slope of the dashed grey line is 1, representing the same performance before and after the selection. The error bar is the standard deviation of 10 times training.