# Note

Feng-Yang Hsieh

# 1 Higgs Production

We want to apply deep learning methods to distinguish vector boson fusion (VBF) from gluon-gluon fusion (GGF) and Higgs production at the LHC.

We want to apply the CWoLa method, so we can use the real data without knowing the true label.

# 2 Sample Preparation

## 2.1 Monte Carlo samples

We consider Standard Model (SM) di-photon Higgs events produced via GGF and VBF channels at a center-of-mass energy of $\sqrt{s} = 14$ TeV. The Higgs boson events are generated using `MadGraph 3.3.1` [1] for both GGF and VBF production. The Higgs decays into the di-photon final state, and the parton showering and hadronization are simulated using `Pythia 8.306` [2]. The detector simulation is conducted by `Delphes 3.4.2` [3]. Jet reconstruction is performed using `FastJet 3.3.2` [4] with the anti-$k_t$ algorithm [5] and a jet radius of $R = 0.4$. These jets are required to have transverse momentum $p_{\mathrm{T}} > 25$ GeV.

The following `MadGraph` scripts generate Monte Carlo samples for each production channel.

**GGF Higgs Sample Generation**

```
generate p p > h QCD<=99 [QCD]
output GGF_Higgs
launch GGF_Higgs

shower=Pythia8
detector=Delphes
```

```
analysis=OFF
madspin=OFF
done

set run_card nevents 100000
set run_card ebeam1 7000.0
set run_card ebeam2 7000.0

set run_card use_syst False

set pythia8_card 25:onMode = off
set pythia8_card 25:onIfMatch = 22 22
done
```

### VBF Higgs Sample Generation

```
define v = w+ w- z
generate p p > h j j $$v
output VBF_Higgs
launch VBF_Higgs

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

set run_card nevents 100000
set run_card ebeam1 7000.0
set run_card ebeam2 7000.0

set run_card use_syst False

set pythia8_card 25:onMode = off
set pythia8_card 25:onIfMatch = 22 22
done
```

## 2.2 Event selection

The selection cuts after the `Delphes` simulation:

- $n_\gamma$ cut: The number of photons should be at least 2.
- $n_j$ cut: The number of jets should be at least 2.
- $m_{\gamma\gamma}$ cut: The invariant mass of two leading photons $m_{\gamma\gamma}$ are required 120 GeV $\leq m_{\gamma\gamma} \leq$ 130 GeV.

Table 1 summarizes the cutflow number at different selection cuts.

Table 1: Number of passing events and passing rates for GGF and VBF Higgs production at different selection cuts.

| Cut | GGF | pass rate | VBF | pass rate |
|---|---|---|---|---|
| Total | 100000 | 1 | 100000 | 1 |
| $n_\gamma$ cut | 48286 | 0.48 | 53087 | 0.53 |
| $n_j$ cut | 9302 | 0.09 | 42860 | 0.43 |
| $m_{\gamma\gamma}$ cut | 8864 | 0.09 | 40694 | 0.41 |

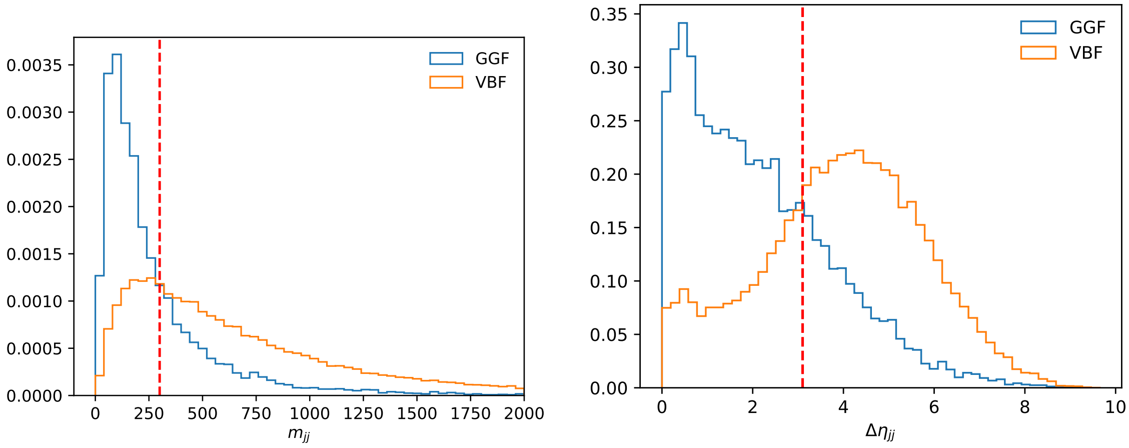
Figure 1 shows the distributions of $m_{jj}$ (the invariant mass of the two leading jets) and $\Delta\eta_{jj}$ (the pseudorapidity difference between the two leading jets). The scatter plot of $m_{jj}$ versus $\Delta\eta_{jj}$ is presented in Figure 2.

## 2.3 Event image

The inputs for the neural networks are event images [6, 7, 8]. These images are constructed from events that pass the kinematic selection criteria described in section 2.2. Each event image has three channels corresponding to calorimeter towers, tracks, and photons. The following preprocessing steps are applied to all event constituents:

1. Translation: Compute the $p_\mathrm{T}$-weighted center in the $\phi$ coordinates, then shift this point to the origin.
2. Flipping: Flip the highest $p_\mathrm{T}$ quadrant to the first quadrant.
3. Pixelation: Pixelate in a $\eta \in [-5, 5]$, $\phi \in [-\pi, \pi]$ box, with $40 \times 40$ pixels

Figure 3 shows the event images for GGF and VBF production modes.



(a) $m_{jj}$ distribution

(b) $\Delta\eta_{jj}$ distribution

Figure 1: Distributions of the invariant mass $m_{jj}$ and pseudorapidity difference $\Delta\eta_{jj}$ of the two leading jets. Red dashed lines are selection cuts used to construct mixed datasets.

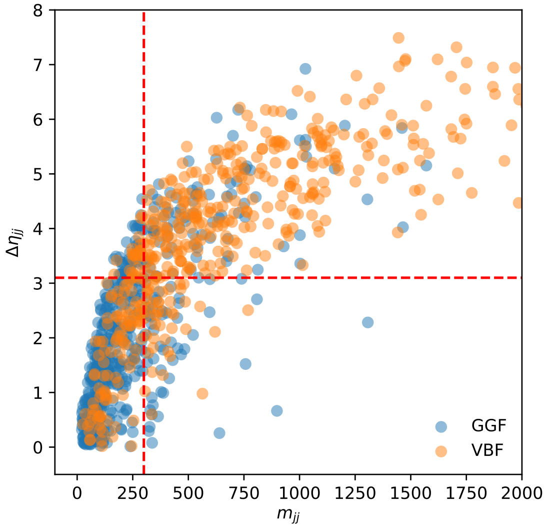
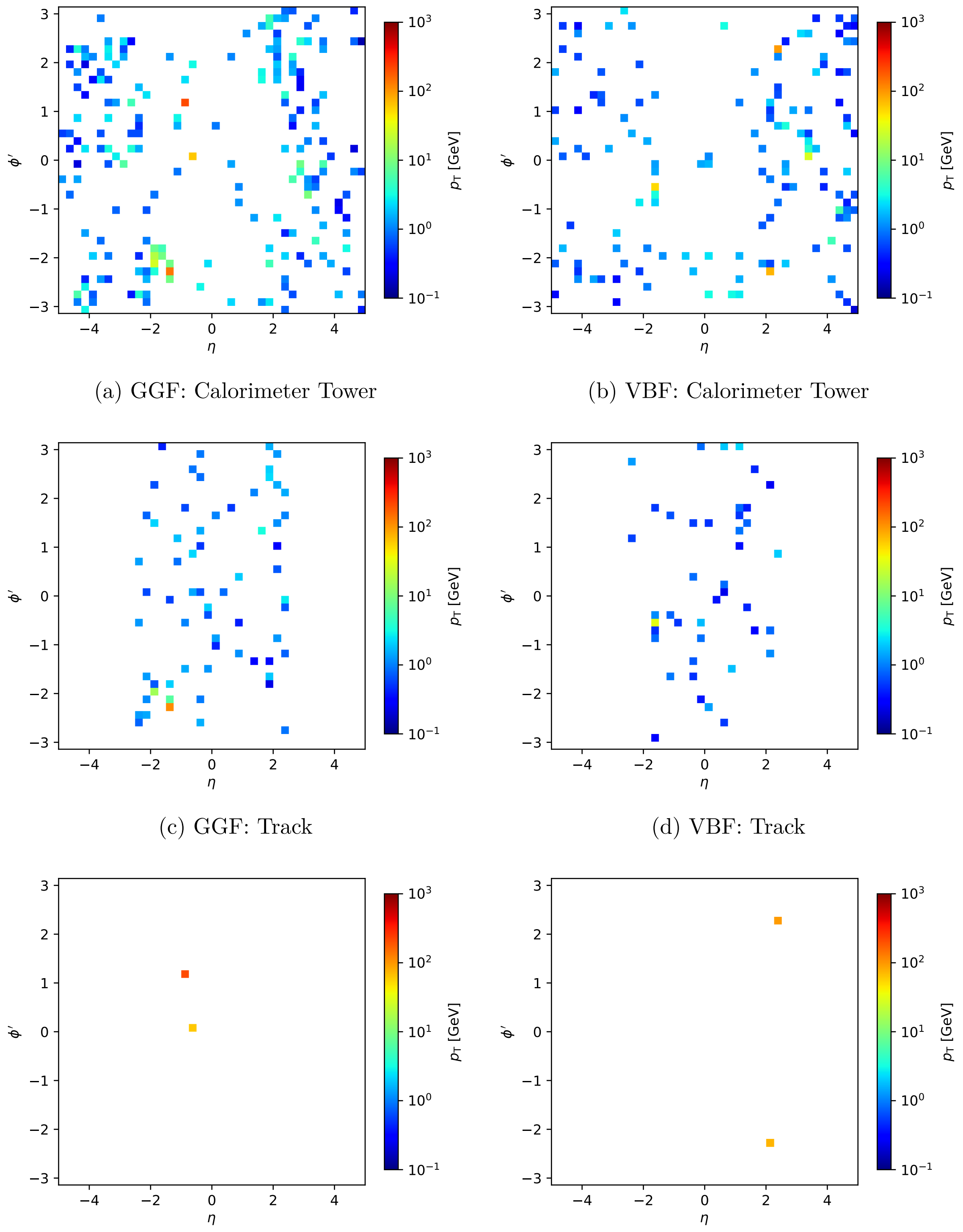


Figure 2: Scatter plot of $m_{jj}$ versus $\Delta\eta_{jj}$. Red dashed lines are selection cuts used to construct mixed datasets.



(a) GGF: Calorimeter Tower

(b) VBF: Calorimeter Tower

(c) GGF: Track

(d) VBF: Track

(e) GGF: Photon

(f) VBF: Photon

Figure 3: Event images for GGF and VBF production, separately shown for calorimeter towers, tracks, and photons.

## 2.4 Mixed datasets

Based on figure 1, we set selection cuts of $m_{jj} > 300$ GeV and $\Delta\eta_{jj} > 3.1$. We consider three cases: applying each cut individually and simultaneously. These cuts define the signal region (SR), which is VBF-like, and the background region (BR), which is GGF-like. Table 2 summarizes the cutflow results for different selection criteria.

Table 2: Number of passing events and passing rates for GGF and VBF Higgs production under different selection cuts.

| Cut | GGF | pass rate | VBF | pass rate |
|---|---|---|---|---|
| Total | 100000 | 1.00 | 100000 | 1.00 |
| $n_\gamma$ cut | 9302 | 0.09 | 42860 | 0.43 |
| $n_j$ cut | 9302 | 0.09 | 42860 | 0.43 |
| $m_{\gamma\gamma}$ cut | 8864 | 0.09 | 40694 | 0.41 |
| $m_{jj}$ cut: SR | 2695 | 0.03 | 29496 | 0.29 |
| $m_{jj}$ cut: BR | 6169 | 0.06 | 11198 | 0.11 |
| $\Delta\eta_{jj}$ cut: SR | 2317 | 0.02 | 28160 | 0.28 |
| $\Delta\eta_{jj}$ cut: BR | 6547 | 0.07 | 12534 | 0.13 |
| $m_{jj}, \Delta\eta_{jj}$ cuts: SR | 1832 | 0.02 | 26446 | 0.26 |
| $m_{jj}, \Delta\eta_{jj}$ cuts: BR | 5684 | 0.06 | 9484 | 0.09 |

The total cross-section for VBF production is $\sigma_{\text{VBF}} = 4.278\ \text{pb}^{-1}$ at NNLO and for GGF production is $\sigma_{\text{GGF}} = 54.67\ \text{pb}^{-1}$ at N3LO, as referenced in this link. The branching ratio for the di-photon decay channel is $\Gamma(h \to \gamma\gamma) = 2.270 \times 10^{-3}$, as given in this link.

Assuming the luminosity of $\mathcal{L} = 300\ \text{fb}^{-1}$, we can estimate the number of events belonging to the SR and BR. These results are summarized in table 3.

# 3 Training CNN

The total sample sizes are mentioned in section 2.4. We allocate 80% of the data for training and 20% for validation. The testing set consists of the SR's 10,000 VBF and 10,000 GGF events.

The convolutional neural network (CNN) model structure is summarized in figure 4. The internal node uses the rectified linear unit (ReLU) as the activation function. The loss function is the binary cross-entropy. The `Adam` optimizer minimizes the loss value. The learning rate is $10^{-4}$, and the batch size is 512. We employ the early stopping technique to

Table 3: The number of events of mixed datasets under different selection cuts.

(a) $m_{jj} > 300$ GeV

| | GGF | VBF |
|---|---|---|
| BR | 2297 | 326 |
| SR | 1003 | 859 |

(b) $\Delta\eta_{jj} > 3.1$

| | GGF | VBF |
|---|---|---|
| BR | 2437 | 365 |
| SR | 863 | 820 |

(c) $m_{jj} > 300$ GeV, $\Delta\eta_{jj} > 3.1$

| | GGF | VBF |
|---|---|---|
| BR | 2116 | 276 |
| SR | 682 | 770 |

prevent over-training issues with patience of 10.

The training results are summarized in table 4. The performance of the $\Delta\eta_{jj}$ cuts is better than the $m_{jj}$ cut. Moreover, when both cuts are applied together, the performance is slightly worse than when applying either cut individually.

Table 4: The CNN training results. The ACC and AUC are evaluated based on 10 training. The selection cuts of $m_{jj} > 300$ GeV and $\Delta\eta_{jj} > 3.1$ are applied.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Cut | ACC | AUC | ACC | AUC |
| $m_{jj}$ | $0.712 \pm 0.023$ | $0.741 \pm 0.041$ | $0.576 \pm 0.010$ | $0.596 \pm 0.014$ |
| $\Delta\eta_{jj}$ | $0.828 \pm 0.043$ | $0.889 \pm 0.050$ | $0.604 \pm 0.014$ | $0.630 \pm 0.015$ |
| $m_{jj}, \Delta\eta_{jj}$ | $0.753 \pm 0.022$ | $0.792 \pm 0.035$ | $0.573 \pm 0.007$ | $0.596 \pm 0.008$ |

## 3.1 More events

This section assumes the luminosity of $\mathcal{L} = 3000$ fb$^{-1}$. The number of events belonging to the SR and BR are summarized in table 5.

The training results are summarized in table 6. All datasets' performance is better than the results in table 4. The $\Delta\eta_{jj}$ cut performs better than the $m_{jj}$ cut. Moreover, when both cuts are applied together, the performance is slightly worse than the $\Delta\eta_{jj}$ cut but better than $m_{jj}$. These results are similar to the previous one.

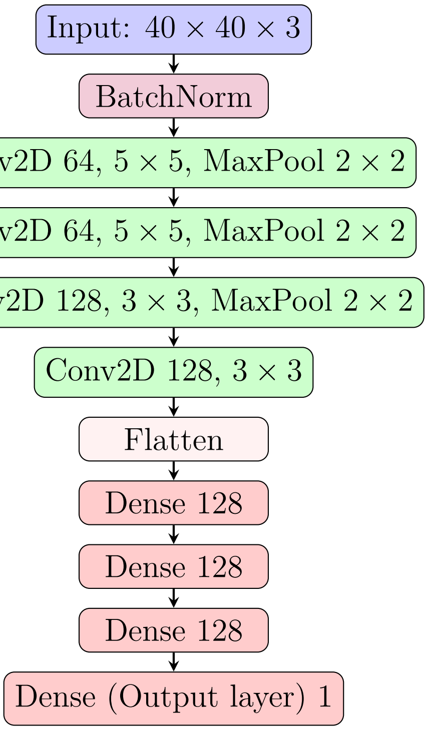

Input: $40 \times 40 \times 3$
BatchNorm
Conv2D 64, $5 \times 5$, MaxPool $2 \times 2$
Conv2D 64, $5 \times 5$, MaxPool $2 \times 2$
Conv2D 128, $3 \times 3$, MaxPool $2 \times 2$
Conv2D 128, $3 \times 3$
Flatten
Dense 128
Dense 128
Dense 128
Dense (Output layer) 1

Figure 4: The architecture of the CNN model with key hyperparameters.

Table 5: The number of events of mixed datasets under different selection cuts.

(a) $m_{jj} > 300$ GeV

| | GGF | VBF |
|---|---|---|
| BR | 22967 | 3262 |
| SR | 10034 | 8593 |

(b) $\Delta\eta_{jj} > 3.1$

| | GGF | VBF |
|---|---|---|
| BR | 24375 | 3652 |
| SR | 8626 | 8204 |

(c) $m_{jj} > 300$ GeV,
$\Delta\eta_{jj} > 3.1$

| | GGF | VBF |
|---|---|---|
| BR | 21162 | 2763 |
| SR | 6821 | 7705 |

Table 6: The CNN training results. The ACC and AUC are evaluated based on 10 training. The selection cuts of $m_{jj} > 300$ GeV and $\Delta\eta_{jj} > 3.1$ are applied.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Cut | ACC | AUC | ACC | AUC |
| $m_{jj}$ | $0.907 \pm 0.002$ | $0.969 \pm 0.002$ | $0.598 \pm 0.008$ | $0.625 \pm 0.009$ |
| $\Delta\eta_{jj}$ | $0.931 \pm 0.004$ | $0.979 \pm 0.002$ | $0.615 \pm 0.005$ | $0.648 \pm 0.006$ |
| $m_{jj}, \Delta\eta_{jj}$ | $0.929 \pm 0.003$ | $0.978 \pm 0.002$ | $0.608 \pm 0.004$ | $0.638 \pm 0.005$ |

# 4 $p_{\mathrm{T}}$ normalization

To remove the potential dependence of the input samples on $m_{jj}$, we standardize the event images to remove the difference in input data distributions between the SR and BR. We calculate the mean and standard deviation of the event image transverse momentum and use these values to standardize each event image. We standardize each channel separately.

The number of events in the SR and BR are the same as previously in table 5.

The training results are summarized in table 7. The $m_{jj}$ cut performs better than the previous one (table 6).

Table 7: The CNN training results with $p_{\mathrm{T}}$ normalization technique. The ACC and AUC are evaluated based on 10 training. The selection cuts of $m_{jj} > 300$ GeV and $\Delta\eta_{jj} > 3.1$ are applied.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Cut | ACC | AUC | ACC | AUC |
| $m_{jj}$ | $0.874 \pm 0.004$ | $0.946 \pm 0.003$ | $0.624 \pm 0.005$ | $0.663 \pm 0.006$ |
| $\Delta\eta_{jj}$ | $0.928 \pm 0.005$ | $0.979 \pm 0.002$ | $0.597 \pm 0.005$ | $0.630 \pm 0.006$ |
| $m_{jj}, \Delta\eta_{jj}$ | $0.917 \pm 0.003$ | $0.973 \pm 0.002$ | $0.603 \pm 0.004$ | $0.636 \pm 0.006$ |

# 5 Different cut setting

We set selection cuts of $m_{jj} > 225$ GeV and $\Delta\eta_{jj} > 2.3$ to ensure the SR and BR datasets have similar sizes. Table 8 summarizes the cutflow results for different selection criteria.

Assuming the luminosity of $\mathcal{L} = 3000$ fb$^{-1}$, we can estimate the number of events belonging to the SR and BR. These results are summarized in table 9

Table 8: Number of passing events and passing rates for GGF and VBF Higgs production under different selection cuts.

| Cut | GGF | pass rate | VBF | pass rate |
|---|---|---|---|---|
| Total | 100000 | 1.00 | 100000 | 1.00 |
| $n_\gamma$ cut | 9302 | 0.09 | 42860 | 0.43 |
| $n_j$ cut | 9302 | 0.09 | 42860 | 0.43 |
| $m_{\gamma\gamma}$ cut | 8864 | 0.09 | 40694 | 0.41 |
| $m_{jj}$ cut: SR | 3638 | 0.04 | 32993 | 0.33 |
| $m_{jj}$ cut: BR | 5226 | 0.05 | 7701 | 0.08 |
| $\Delta\eta_{jj}$ cut: SR | 3611 | 0.04 | 32914 | 0.33 |
| $\Delta\eta_{jj}$ cut: BR | 5253 | 0.05 | 7780 | 0.08 |
| $m_{jj}, \Delta\eta_{jj}$ cuts: SR | 2842 | 0.03 | 31113 | 0.31 |
| $m_{jj}, \Delta\eta_{jj}$ cuts: BR | 4457 | 0.04 | 5900 | 0.06 |

Table 9: The number of events of mixed datasets under different selection cuts.

(a) $m_{jj} > 225$ GeV

| | GGF | VBF |
|---|---|---|
| BR | 19457 | 2244 |
| SR | 13544 | 9612 |

(b) $\Delta\eta_{jj} > 2.3$

| | GGF | VBF |
|---|---|---|
| BR | 19557 | 2267 |
| SR | 13444 | 9589 |

(c) $m_{jj} > 225$ GeV, $\Delta\eta_{jj} > 2.3$

| | GGF | VBF |
|---|---|---|
| BR | 16594 | 1719 |
| SR | 10581 | 9064 |

The training results are summarized in table 10. The results are better than the table 7 by 1%. Similarly, the $m_{jj}$ cut performs best.

Table 10: The CNN training results with $p_{\mathrm{T}}$ normalization technique. The ACC and AUC are evaluated based on 10 training. The selection cuts of $m_{jj} > 225$ GeV and $\Delta\eta_{jj} > 2.3$ are applied.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Cut | ACC | AUC | ACC | AUC |
| $m_{jj}$ | $0.864 \pm 0.004$ | $0.940 \pm 0.004$ | $0.632 \pm 0.006$ | $0.673 \pm 0.007$ |
| $\Delta\eta_{jj}$ | $0.913 \pm 0.006$ | $0.972 \pm 0.003$ | $0.605 \pm 0.007$ | $0.640 \pm 0.009$ |
| $m_{jj}, \Delta\eta_{jj}$ | $0.896 \pm 0.007$ | $0.961 \pm 0.004$ | $0.616 \pm 0.005$ | $0.653 \pm 0.006$ |

# 6 Supervised training

This section tests the supervised training on CNN. The training, validation, and testing sample sizes are summarized in table 11. The events passing all selection requirements (section 2.2) are considered.

Table 11: Sizes of various samples used for supervised training.

| | Training | Validation | Testing |
|---|---|---|---|
| GGF | 100k | 25k | 25k |
| VBF | 100k | 25k | 25k |

The training results are summarized in table 12. These results demonstrate the upper limit of CNN training.

Table 12: The CNN training results with $p_{\mathrm{T}}$ normalization technique. The ACC and AUC are evaluated based on 10 training.

| ACC | AUC |
|---|---|
| $0.784 \pm 0.001$ | $0.861 \pm 0.001$ |

## 6.1 Testing sample in SR and BR

The testing events used to evaluate the table 12 are all events passing the selection and not restricted to the particular SR. Thus, to make a fair comparison with previous results, we must evaluate the training performance on the events in SR and BR.

The new testing dataset consists of the 10,000 VBF and 10,000 GGF events from SR and BR. The numbers of SR and BR events are computed from table 8.

The training results of table 10 are re-evaluated on the new testing set and shown in table 13. The results are better than the table 10. It seems that the events in the BR can be distinguished better than those in the SR.

Table 13: The CNN training results with $p_{\mathrm{T}}$ normalization technique. The ACC and AUC are evaluated based on 10 training. The selection cuts of $m_{jj} > 225$ GeV and $\Delta\eta_{jj} > 2.3$ are applied.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Cut | ACC | AUC | ACC | AUC |
| $m_{jj}$ | $0.863 \pm 0.004$ | $0.940 \pm 0.002$ | $0.716 \pm 0.003$ | $0.780 \pm 0.004$ |
| $\Delta\eta_{jj}$ | $0.914 \pm 0.004$ | $0.972 \pm 0.003$ | $0.702 \pm 0.003$ | $0.754 \pm 0.003$ |
| $m_{jj}, \Delta\eta_{jj}$ | $0.896 \pm 0.006$ | $0.962 \pm 0.004$ | $0.723 \pm 0.003$ | $0.780 \pm 0.002$ |

## 7 Use jet tagging results to construct mixed datasets

This section uses the jet tagging results to construct the mixed datasets.

Assuming the luminosity of $\mathcal{L} = 3000$ fb$^{-1}$, we can estimate the number of events belonging to the SR and BR. The SR and BR are defined based on the number of gluon jets $n_g$ and quark jets $n_q$. The selection results are summarized in table 14.

Table 14: The number of events of mixed datasets under different selection cuts. Here, $agbq$ means that $n_g = a, n_q = b$.

(a) SR: $2q0g$; BR: $1q1g, 0q2g$

| | GGF | VBF |
|---|---|---|
| SR | 16828 | 10229 |
| BR | 16865 | 1596 |

(b) SR: $2q0g, 1q1g$; BR: $0q2g$

| | GGF | VBF |
|---|---|---|
| SR | 30752 | 11779 |
| BR | 2941 | 47 |

(c) SR: $2q0g$; BR: $0q2g$

| | GGF | VBF |
|---|---|---|
| SR | 16828 | 10229 |
| BR | 2941 | 47 |

For now, we use the true information from `Delphes` and do not consider the mis-tagging case.

The training results are summarized in table 15. All different jet-tagging conditions produced similar performance. However, the results are worse than those of kinematic cuts (table 13).

Table 15: The CNN training results with $p_{\mathrm{T}}$ normalization technique. The ACC and AUC are evaluated based on 10 training.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| SR: $2q0g$; BR: $1q1g, 0q2g$ | $0.623 \pm 0.005$ | $0.642 \pm 0.005$ | $0.653 \pm 0.008$ | $0.706 \pm 0.009$ |
| SR: $2q0g, 1q1g$; BR: $0q2g$ | $0.934 \pm 0.000$ | $0.689 \pm 0.012$ | $0.662 \pm 0.006$ | $0.719 \pm 0.008$ |
| SR: $2q0g$; BR: $0q2g$ | $0.900 \pm 0.000$ | $0.740 \pm 0.010$ | $0.655 \pm 0.008$ | $0.710 \pm 0.009$ |

The training results without $p_{\mathrm{T}}$ nomalization are summarized in table 16. All different jet-tagging conditions produced similar performance. However, the results are worse than the ones with $p_{\mathrm{T}}$ normalization (table 15) by 2%.

Table 16: The CNN training results without $p_{\mathrm{T}}$ normalization technique. The ACC and AUC are evaluated based on 10 training.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| SR: $2q0g$; BR: $1q1g, 0q2g$ | $0.614 \pm 0.007$ | $0.632 \pm 0.011$ | $0.646 \pm 0.008$ | $0.690 \pm 0.011$ |
| SR: $2q0g, 1q1g$; BR: $0q2g$ | $0.934 \pm 0.000$ | $0.695 \pm 0.015$ | $0.643 \pm 0.009$ | $0.689 \pm 0.011$ |
| SR: $2q0g$; BR: $0q2g$ | $0.900 \pm 0.000$ | $0.743 \pm 0.011$ | $0.632 \pm 0.007$ | $0.677 \pm 0.008$ |

## 7.1 Loss weighted

Since the sample sizes are unbalanced, we add the class weights. The weights are proportional to the reciprocal of the number of events.

The training results with class weights are summarized in table 17. All different jet-tagging conditions produced similar performance.

# 8 Total scaling of transverse momentum

The $p_{\mathrm{T}}$ normalization removes the magnitude information of the input datasets. Thus, we would expect the training performance of the $p_{\mathrm{T}}$ normalization datasets would be worse than the one without it. However, table 15 and 16 shows the opposite results.

Table 17: The CNN training results without $p_{\mathrm{T}}$ normalization technique. The ACC and AUC are evaluated based on 10 training.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| SR: $2q0g$; BR: $1q1g, 0q2g$ | $0.621 \pm 0.006$ | $0.635 \pm 0.007$ | $0.645 \pm 0.009$ | $0.688 \pm 0.013$ |
| SR: $2q0g, 1q1g$; BR: $0q2g$ | $0.934 \pm 0.000$ | $0.679 \pm 0.016$ | $0.624 \pm 0.005$ | $0.662 \pm 0.008$ |
| SR: $2q0g$; BR: $0q2g$ | $0.900 \pm 0.000$ | $0.730 \pm 0.013$ | $0.621 \pm 0.005$ | $0.658 \pm 0.008$ |

To explore the reason why the $p_{\mathrm{T}}$ normalization could improve the training performance, we try the total $p_{\mathrm{T}}$ scaling, which computes the mean and standard deviation of all input samples. Then, use these values to standardize the input datasets.

### 8.1 Results

The training results with $p_{\mathrm{T}}$ scaling are summarized in table 18. All different jet-tagging conditions produced similar performance. However, the results are worse than the ones with $p_{\mathrm{T}}$ normalization (table 15).

Table 18: The CNN training results with $p_{\mathrm{T}}$ scaling technique. The ACC and AUC are evaluated based on 10 training. The selection cuts on the number of gluon jets are applied.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| SR: $2q0g$; BR: $1q1g, 0q2g$ | $0.622 \pm 0.004$ | $0.637 \pm 0.008$ | $0.638 \pm 0.009$ | $0.678 \pm 0.011$ |
| SR: $2q0g, 1q1g$; BR: $0q2g$ | $0.934 \pm 0.000$ | $0.673 \pm 0.032$ | $0.619 \pm 0.019$ | $0.652 \pm 0.029$ |
| SR: $2q0g$; BR: $0q2g$ | $0.900 \pm 0.000$ | $0.733 \pm 0.011$ | $0.621 \pm 0.006$ | $0.657 \pm 0.009$ |

The training results with $p_{\mathrm{T}}$ normalization are summarized in table 19.

The training results without $p_{\mathrm{T}}$ normalization are summarized in table 20.

## 9 Data augmentation

To improve the training performance, we will consider various data augmentation methods.

Table 19: The CNN training results with $p_{\mathrm{T}}$ normalization technique. The ACC and AUC are evaluated based on 10 training. The selection cuts on the number of gluon jets are applied.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| SR: $2q0g$; BR: $1q1g, 0q2g$ | $0.615 \pm 0.005$ | $0.632 \pm 0.007$ | $0.650 \pm 0.011$ | $0.703 \pm 0.015$ |
| SR: $2q0g, 1q1g$; BR: $0q2g$ | $0.934 \pm 0.000$ | $0.662 \pm 0.014$ | $0.630 \pm 0.008$ | $0.675 \pm 0.011$ |
| SR: $2q0g$; BR: $0q2g$ | $0.900 \pm 0.000$ | $0.716 \pm 0.012$ | $0.640 \pm 0.007$ | $0.690 \pm 0.009$ |

Table 20: The CNN training results without $p_{\mathrm{T}}$ normalization technique. The ACC and AUC are evaluated based on 10 training. The selection cuts on the number of gluon jets are applied.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| SR: $2q0g$; BR: $1q1g, 0q2g$ | $0.620 \pm 0.004$ | $0.636 \pm 0.005$ | $0.643 \pm 0.006$ | $0.686 \pm 0.007$ |
| SR: $2q0g, 1q1g$; BR: $0q2g$ | $0.934 \pm 0.000$ | $0.680 \pm 0.014$ | $0.624 \pm 0.010$ | $0.660 \pm 0.016$ |
| SR: $2q0g$; BR: $0q2g$ | $0.900 \pm 0.000$ | $0.727 \pm 0.010$ | $0.628 \pm 0.008$ | $0.666 \pm 0.011$ |

## 9.1 $p_{\mathrm{T}}$ smearing

The $p_{\mathrm{T}}$ smearing method simulates detector resolution effects on the transverse momentum of event constituents. This method resamples the transverse momentum $p_{\mathrm{T}}$ of event constituents according to the normal distribution:

$$p'_{\mathrm{T}} \sim \mathcal{N}\left(p_{\mathrm{T}}, f(p_{\mathrm{T}})\right), \quad f(p_{\mathrm{T}}) = \sqrt{0.052 p_{\mathrm{T}}^2 + 1.502 p_{\mathrm{T}}}, \tag{1}$$

where $p'_{\mathrm{T}}$ is the augmented transverse momentum, and $f\left(p_{\mathrm{T}}\right)$ is the energy smearing function applied by `Delphes` (the $p_{\mathrm{T}}$'s are normalized in units of GeV). The preprocessing is applied after the $p_{\mathrm{T}}$ smearing augmentation.

The training results of the $2q0g$ datasets (Table 14 (a)) are summarized in table 21.

## 9.2 $\phi$ shifting

The $\phi$ shifting method shifts entire events by a random angle $\Delta\phi \in [-\pi, \pi]$ to enlarge the diversity of training datasets.

The training results of the $2q0g$ datasets are summarized in table 22.

Table 21: CNN training results with different augmentation sizes. The ACC and AUC are evaluated based on 10 training.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| Original | $0.615 \pm 0.005$ | $0.632 \pm 0.007$ | $0.650 \pm 0.011$ | $0.703 \pm 0.015$ |
| +5 | $0.625 \pm 0.006$ | $0.653 \pm 0.009$ | $0.661 \pm 0.010$ | $0.714 \pm 0.012$ |
| +10 | $0.629 \pm 0.005$ | $0.658 \pm 0.005$ | $0.666 \pm 0.008$ | $0.721 \pm 0.009$ |
| +15 | $0.629 \pm 0.003$ | $0.660 \pm 0.003$ | $0.661 \pm 0.015$ | $0.710 \pm 0.018$ |

Table 22: CNN training results with different augmentation sizes. The ACC and AUC are evaluated based on 10 training.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| Original | $0.615 \pm 0.005$ | $0.632 \pm 0.007$ | $0.650 \pm 0.011$ | $0.703 \pm 0.015$ |
| +5 | $0.641 \pm 0.003$ | $0.680 \pm 0.004$ | $0.683 \pm 0.010$ | $0.736 \pm 0.013$ |
| +10 | $0.642 \pm 0.006$ | $0.684 \pm 0.008$ | $0.686 \pm 0.008$ | $0.739 \pm 0.011$ |
| +15 | $0.643 \pm 0.005$ | $0.685 \pm 0.006$ | $0.687 \pm 0.009$ | $0.742 \pm 0.010$ |

## 9.3 $\eta - \phi$ smearing

We apply the $\eta - \phi$ smearing on the training samples. Specifically, the $(\eta, \phi)$ coordinates of constituents are resampled according to a normal distribution centered on the original coordinate and with a standard deviation inversely proportional to the $p_{\mathrm{T}}$

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

where $\eta', \phi'$ are the augmented coordinates, $p_{\mathrm{T}}$ is the transverse momentum of the constituent, and the smearing scale is set to be $\Lambda = 100$ MeV.

The training results on the $2q0g$ datasets are summarized in Table 23. The +5 and +10 augmentation cases show performance comparable to the original dataset. However, applying +15 augmentations degrades the performance, suggesting that introducing too many augmented samples may lead the training in the wrong direction.

## 9.4 Without pre-processing

The $\phi$ shifting seems to cancel the $\phi$ translation in the pre-processing. Thus, we expect the model trained on the $\phi$ shifting dataset could perform similarly to the no pre-processing

Table 23: CNN training results with different augmentation sizes. The ACC and AUC are evaluated based on 10 training.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| Original | $0.615 \pm 0.005$ | $0.632 \pm 0.007$ | $0.650 \pm 0.011$ | $0.703 \pm 0.015$ |
| +5 | $0.618 \pm 0.004$ | $0.640 \pm 0.006$ | $0.658 \pm 0.009$ | $0.711 \pm 0.013$ |
| +10 | $0.617 \pm 0.004$ | $0.641 \pm 0.006$ | $0.654 \pm 0.010$ | $0.705 \pm 0.012$ |
| +15 | $0.612 \pm 0.006$ | $0.628 \pm 0.008$ | $0.635 \pm 0.009$ | $0.679 \pm 0.013$ |

datasets.

The testing results of the $2q0g$ datasets are summarized in table 24. The performance of pre-processing datasets is generally better than that without pre-processing. The reason may be that the original datasets are applied pre-processed. Thus, the samples have higher density for the $\phi$ center at 0. The model would prefer to learn these events first.

We can train the model on only the augmented datasets to ensure the effect of the original samples.

Table 24: CNN training results with different augmentation sizes. The ACC and AUC are evaluated based on 10 training.

| | w/ pre-processing | | w/o pre-processing | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| Original | $0.637 \pm 0.007$ | $0.686 \pm 0.008$ | $0.625 \pm 0.006$ | $0.669 \pm 0.008$ |
| +5 | $0.682 \pm 0.011$ | $0.735 \pm 0.013$ | $0.669 \pm 0.011$ | $0.720 \pm 0.015$ |
| +10 | $0.685 \pm 0.008$ | $0.739 \pm 0.010$ | $0.673 \pm 0.008$ | $0.726 \pm 0.010$ |
| +15 | $0.688 \pm 0.007$ | $0.743 \pm 0.009$ | $0.674 \pm 0.008$ | $0.726 \pm 0.008$ |

## 9.5 Only augmentation datasets

In section 9.4, we found that the performance of pre-processing datasets is generally better than without pre-processing. We train the model on only the augmented datasets to ensure the effect of the original samples.

The testing results of the only augmented sample are summarized in table 25. The performance without original samples is similar to that with original samples. It seems that the impact of the original datasets is limited. For 10, 15 augmentation cases, with and without original samples perform almost the same.

Table 25: CNN training results with different augmentation sizes. The ACC and AUC are evaluated based on 10 training. Here, $+x$ contains original and augmented samples; $=x$ contains only augmented samples.

| | w/ pre-processing | | w/o pre-processing | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| +5 | $0.682 \pm 0.011$ | $0.735 \pm 0.013$ | $0.669 \pm 0.011$ | $0.720 \pm 0.015$ |
| =5 | $0.682 \pm 0.007$ | $0.736 \pm 0.009$ | $0.668 \pm 0.007$ | $0.718 \pm 0.010$ |
| +10 | $0.685 \pm 0.008$ | $0.739 \pm 0.010$ | $0.673 \pm 0.008$ | $0.726 \pm 0.010$ |
| =10 | $0.687 \pm 0.010$ | $0.740 \pm 0.012$ | $0.675 \pm 0.009$ | $0.726 \pm 0.011$ |
| +15 | $0.688 \pm 0.007$ | $0.743 \pm 0.009$ | $0.674 \pm 0.008$ | $0.726 \pm 0.008$ |
| =15 | $0.687 \pm 0.007$ | $0.741 \pm 0.010$ | $0.672 \pm 0.009$ | $0.725 \pm 0.012$ |

# 10 Removing photon information

To investigate the role of photon information in model training, we conduct two exercises:

- **Case 1:** Remove both the photon channel and photon features in the Tower channel.
- **Case 2:** Remove the photon channel.
- **Case 3:** Remove the photon features in the Tower channel.

We consider the $2q0g$ datasets. The training results are summarized in Tables 26, 27, and 28.

Table 26: CNN training results with both the photon channel and Tower photon features removed (Case 1). The ACC and AUC are evaluated based on 10 training.

| | $M_1/M_2$ | | $S/B$ | |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| Original | $0.633 \pm 0.005$ | $0.664 \pm 0.008$ | $0.690 \pm 0.005$ | $0.750 \pm 0.008$ |
| +5 | $0.644 \pm 0.004$ | $0.687 \pm 0.005$ | $0.693 \pm 0.007$ | $0.746 \pm 0.006$ |
| +10 | $0.645 \pm 0.004$ | $0.689 \pm 0.005$ | $0.697 \pm 0.009$ | $0.751 \pm 0.011$ |
| +15 | $0.645 \pm 0.004$ | $0.689 \pm 0.005$ | $0.698 \pm 0.009$ | $0.753 \pm 0.010$ |

Case 1 leads the performance improvement compared to the full-feature baseline (Table 22). Case 2 shows a little better performance of the all-feature input case. Case 3 demonstrates more improvement than case 2 but still worse than case 1.

Table 27: CNN training results with only the photon channel removed (Case 2). The ACC and AUC are evaluated based on 10 training.

|  | $M_1/M_2$ |  | $S/B$ |  |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| Original | $0.621 \pm 0.005$ | $0.640 \pm 0.008$ | $0.661 \pm 0.007$ | $0.715 \pm 0.010$ |
| +5 | $0.636 \pm 0.004$ | $0.673 \pm 0.006$ | $0.673 \pm 0.009$ | $0.727 \pm 0.011$ |
| +10 | $0.639 \pm 0.004$ | $0.677 \pm 0.005$ | $0.677 \pm 0.007$ | $0.728 \pm 0.010$ |
| +15 | $0.640 \pm 0.005$ | $0.679 \pm 0.007$ | $0.678 \pm 0.007$ | $0.731 \pm 0.010$ |

Table 28: CNN training results with the Tower photon features removed (Case 3). The ACC and AUC are evaluated based on 10 training.

|  | $M_1/M_2$ |  | $S/B$ |  |
|---|---|---|---|---|
| Datasets | ACC | AUC | ACC | AUC |
| Original | $0.629 \pm 0.004$ | $0.653 \pm 0.006$ | $0.670 \pm 0.008$ | $0.726 \pm 0.009$ |

# References

[1] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *JHEP*, vol. 07, p. 079, 2014.

[2] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, "An introduction to PYTHIA 8.2," *Comput. Phys. Commun.*, vol. 191, pp. 159–177, 2015.

[3] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, "DELPHES 3, A modular framework for fast simulation of a generic collider experiment," *JHEP*, vol. 02, p. 057, 2014.

[4] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual," *Eur. Phys. J. C*, vol. 72, p. 1896, 2012.

[5] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-$k_t$ jet clustering algorithm," *JHEP*, vol. 04, p. 063, 2008.

[6] A. Butter *et al.*, "The Machine Learning landscape of top taggers," *SciPost Phys.*, vol. 7, p. 014, 2019.

[7] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, “Jet-images — deep learning edition,” *JHEP*, vol. 07, p. 069, 2016.

[8] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, “Deep-learning Top Taggers or The End of QCD?,” *JHEP*, vol. 05, p. 006, 2017.