Vector Boson Fusion versus Gluon Fusion

Chen-Hsun Chen\textsuperscript{1}, Kingman Cheung\textsuperscript{1,2,3}, Yi-Lun Chung\textsuperscript{1}, and Pai-Hsien Hsu\textsuperscript{1}

\textsuperscript{1} Department of Physics, 
National Tsing Hua University, 
Hsinchu 300, Taiwan

\textsuperscript{2} Physics Division, National Center for Theoretical Sciences, Hsinchu, Taiwan

\textsuperscript{3} Division of Quantum Phases and Devices, 
School of Physics, Konkuk University, 
Seoul 143-701, Republic of Korea

(Dated: June 12, 2017)

Abstract

Vector-boson fusion (VBF) is a clean probe of the electroweak-symmetry breaking (EWSB), which inevitably suffers from some level of contamination due to the gluon fusion (ggF). In addition to the jet variables used in the current experimental analysis, we analyze a few more jet-shape variables defined by the girth and integrated jet-shape. Taking $H \rightarrow WW^{*} \rightarrow \nu \nu \mu \mu$ and $H \rightarrow \gamma \gamma$ as examples, we perform the analysis with a new technique of 2-step boosted-decision-tree method, which significantly reduces the contamination of the ggF in the VBF sample, thus, providing a clean environment in probing the EWSB sector.
I. INTRODUCTION

The origin of mass is one of the most fundamental questions for our existence. Particle physics explains the origin of mass by the electroweak symmetry breaking (EWSB). Before the electroweak symmetry is broken the whole Universe is filled up with a Higgs field and every particle is massless. When this Higgs field develops a vacuum expectation value (VEV), a particular direction in the field space is chosen and the symmetry is broken. Particles then acquire masses proportional to the VEV of the Higgs field.

The discovery of the Higgs boson at the Large Hadron Collider (LHC) in 2012 [1] was a remarkable evidence of the EWSB and its properties help us to fully understand the nature of the EWSB. The long-sought standard model (SM) Higgs boson was proposed more than 50 years ago, which breaks the electroweak symmetry in order to give masses to gauge bosons and fermions. If the discovered boson is really the SM Higgs boson or something similar, the investigation of its properties would give a lot of information about the EWSB.

The measurements of the properties of the Higgs boson, including mass, total width, production cross sections, and branching ratios will give us a lot of information on its gauge and Yukawa couplings, thus indirectly the details inside the EWSB sector, which could be as complicated as one can imagine. The current dominant production mechanism of the Higgs boson is the gluon fusion (ggF), followed by a small fraction by vector-boson fusion (VBF). Although the ggF could provide useful information on the top-Yukawa coupling, the VBF is the ultimate testing ground for probing the EWSB section, because the longitudinal component of the $W$ and $Z$ bosons originate from the EWSB sector itself.

The approach of isolating the VBF from ggF relies on the properties of the jets involved in the process and a few techniques were developed two decades ago, namely, forward-jet tagging [2] and central-jet vetoing [3]. The two accompanying jets carry most of the jet energy of the incoming quark partons, and thus they are very energetic and very forward. One can also make use of the wide rapidity gap between those two jets [9]. On the other hand, the jets involved in the ggF come directly from the QCD radiation. Ideally, with all the sophisticated jet selection cuts the event samples selected from experimental data should have almost entirely come from the VBF. Nevertheless, with much improved accuracy in the N$^3$LO calculation of ggF [10] the level of ggF in such selection is indeed not negligible but a substantial fraction of the VBF+ggF sample. We shall use the word “contamination”
of the VBF sample to denote the fraction of ggF in the VBF+ggF sample. Thus, the “contamination” of the VBF sample due to ggF is defined by

\[
\frac{\text{ggF}}{\text{VBF} + \text{ggF}}.
\]

It stands at a level about 25% in the current experimental studies [4, 5]. The pure the VBF sample, the better one can probe the EWSB sector. The current experimental status of discriminating the VBF from ggF was based on a set of jet kinematical variables ($M_{jj}$, $\Delta\eta_{jj}$, ...), a set of jet-shape variables, and those kinematic variables depending on the decay channel of the Higgs boson. A standard boosted-decision-tree (BDT) approach was employed to achieve the current purity of the VBF sample and to reduce the contamination of the ggF. In this study, we employ a 2-step BDT analysis to further reduce the contamination by ggF, thus a purer VBF sample is achieved without significant loss in event rates. This is the main result of this work. We illustrate our analysis for the decay channels of $H \rightarrow WW^* \rightarrow e\nu\mu\nu$ and $H \rightarrow \gamma\gamma$.

The organization is as follows. In the next section, we describe the Monte-Carlo simulations, and in Sec. III procedures in the BDT analysis. We present the results in Sec. IV and conclude in Sec. V.

II. EVENT SAMPLES PREPARATION

In order to compare directly with the current status on purity of VBF samples of ATLAS [4, 5], we follow their preparation of event samples as closely as possible. We simulate the event samples for Higgs boson production including those via VBF and ggF using the POWHEG [11-13] generator at next-to-leading-order (NLO), with input parton distribution functions (PDFs) CT10 [14], and the mass and width of the Higgs taken at $m_H = 125$GeV and $\Gamma_H = 4.07$MeV. The Higgs boson samples are normalized to the cross sections given in the ATLAS analysis for 13TeV. Note that for $H \rightarrow \gamma\gamma$ a parton-level cut $105 \leq m_{\gamma\gamma} \leq 160$GeV (Higgs window) is applied.

All Higgs boson events are then showered and decayed into either $WW + \text{jets}$ or $\gamma\gamma + \text{jets}$ by PYTHIA 8 [7] and passed to DELPHES [8] for detector-level simulation. Note that for the channel

1 In this study, although we generate the VBF Monte-Carlo sample and ggF sample separately, we shall keep using “contamination” to denote the fraction of ggF in the sum VBF+ggF events.
\( H \rightarrow WW^* \) each of the \( W \) bosons further decays into a charged lepton and a neutrino. Note that the charged-lepton flavors from the \( W \) boson pair are required to be different, i.e, \( e^+\mu^- \) or \( e^-\mu^+ \).

Table I summarizes the event generators and the cross sections for each process.

| Process | MC generator                  | \( \sigma \cdot B \) (pb) | Number of Events |
|---------|-------------------------------|---------------------------|------------------|
| VBF     | Powheg +Pythia 8              | 0.0232                    | 553240           |
| ggF     | Powheg +Pythia 8              | 0.297                     | 1936340          |
| \( t\bar{t} \) | MadGraph5_AMC@NLO +Pythia 8 | 22.6                      | 3319440          |
| \( WW \) | Powheg +Pythia 8              | 3.10                      | 3319440          |

TABLE I. Monte Carlo generators, cross sections and the generated number of events (non-normalized) used to model each signal and background process in \( WW \) decay channel at \( \sqrt{s} = 13 \) TeV.

In the \( WW \) decay channel, we consider two main backgrounds: the SM \( t\bar{t} \) and \( WW \) production. The \( t\bar{t} \) events are generated at NLO using the MadGraph5_AMC@NLO [6], while the \( WW \) events are generated with Powheg at NLO [15]. After that, the \( t\bar{t} \) and \( WW \) events are showered and each top quark decays into \( b + W \) with Pythia 8 [7]. The \( W \) bosons further decay into \( \ell + \nu \), and the flavors of two charged leptons in each event are required to be different. Events are then passed into Delphes for detector simulations. The event generators, cross sections, and the generated number of events for these backgrounds are also tabulated in Table I.

In the diphoton channel, we only consider one source of background: \( \gamma\gamma + jj \), which are generated at leading-order (LO) using the MadGraph5_AMC@NLO [6]. Each of the jet in \( \gamma\gamma + jj \) events is then showered into multi-jets with Pythia 6 [16]. Finally, events are passed into Delphes for detector simulations. The event generators, cross sections, and the generated number of events for the backgrounds in the diphoton decay channel are also listed in Table II.

**III. METHODS IN BOOSTED DECISION TREES (BDT)**

The dedicated event samples will undergo a series of analysis tools or methods, including pre-selection cuts and boosted decision tree (BDT) [17], in order to enhance the purity of the VBF among the Higgs signals and backgrounds. In general, each signal and background event has to first pass a set of kinematic preselection cuts, and then is further selected according to the BDT.
TABLE II. Monte Carlo generators, cross sections, and the generated number of events (non-normalized) used to model each signal and background process in diphoton decay channel at $\sqrt{s} = 13$ TeV output. In each decay channel, we present four different methods of BDT, including the standard BDT, which mainly follows the method in ATLAS so that we can make directly comparison to the other three new methods of BDT. Tables [III] and [IV] summarize the procedures for $H \rightarrow WW^*$ and $H \rightarrow \gamma\gamma$, respectively. The details are described in the following two subsections.

A. $H \rightarrow WW^* \rightarrow e\nu\mu\nu$

The event samples for the VBF $H \rightarrow WW^*$ signal, ggF, and the SM backgrounds have to pass the preselection cuts which were given in the current ATLAS analysis for the SM Higgs boson decaying into $WW^*$ in the different lepton-flavor category, which are described as follows:

1. $N_j \geq 2$;
2. $p_T^j > 25$ GeV ($|\eta^j| < 2.4$) and $p_T^j > 30$ GeV ($2.4 < |\eta^j| < 4.4$);
3. $p_T^{\ell_1} > 25$ GeV and $p_T^{\ell_2} > 15$ GeV;
4. $m_{\ell\ell} > 10$ GeV, where $m_{\ell\ell}$ is the invariant mass of two leading leptons;
5. $N_b \leq 0$;
6. Outside-lepton veto (OLV), and central-jet veto (CJV) [II]

Standard BDT  Following the current procedures of the ATLAS analysis, the signal sample of VBF and the background samples of simulated ggF, simulated $t\bar{t}$, and simulated $WW$ events are used to train the BDT. We call this one the standard BDT, with which we shall compare. The following 8 variables are fed into the BDT:

| Process     | Generator                     | $\sigma \cdot B$ (pb) | Number of Events |
|-------------|-------------------------------|------------------------|------------------|
| VBF         | POWHEG +Pythia 8              | 0.862                  | 200000           |
| ggF         | POWHEG +Pythia 8              | 11.1                   | 800000           |
| $\gamma\gamma$+jj | MadGraph5_AMC@NLO +Pythia 6  | 4.12                   | 2000000          |
1. $m_{jj}$: invariant mass of two leading jets;
2. $\Delta \eta_{jj} \equiv |\eta_{j1} - \eta_{j2}|$;
3. $p_T^{\text{sum}} \equiv p_T^{\ell\ell} + p_T^{\text{miss}} + \sum p_T^j$;
4. $\sum m_\ell_j \equiv m_{\ell_1,j_1} + m_{\ell_1,j_2} + m_{\ell_2,j_1} + m_{\ell_2,j_2}$;
5. $\sum C_\ell \equiv \sum |\eta_\ell - \sum \eta_j/2| / \Delta \eta_{jj}$;
6. $m_{\ell\ell}$;
7. $\Delta \phi_{\ell\ell}$;

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
Objective & Standard BDT & 11-Var BDT & 7-Var BDT & 2-step BDT \\
\hline
Preselection & $N_j \geq 2, N_b \leq 0,$ & $p_T^j > 25 \text{ GeV} \ (|\eta^j| < 2.4) \ & \ &  \\
 & & $\& p_T^j > 30 \text{ GeV} \ (2.4 < |\eta^j| < 4.4),$ & &  \\
 & & $p_T^{\ell 1} > 25 \text{ GeV}, p_T^{\ell 2} > 15 \text{ GeV},$ & &  \\
 & & $m_{\ell\ell} > 10 \text{ GeV},$ & & OLV, CJV \\
\hline
1st step & Signal sample & VBF & VBF & VBF \ & Bkg. sample & ggF & & ggF & & $t\bar{t}$ & & $t\bar{t}$ & & ggF \ & BDT inputs & $m_{jj}, \Delta \eta_{jj}, p_T^{\text{sum}}$, & $m_{jj}, \Delta \eta_{jj}, p_T^{\text{sum}}$, & $m_{jj}, \Delta \eta_{jj}, p_T^{\text{sum}}$, & $m_{jj}, \Delta \eta_{jj}, p_T^{\text{sum}}$, \ & & $\sum m_\ell_j, \sum C_\ell$, & $\sum m_\ell_j, \sum C_\ell$, & $\sum m_\ell_j, \sum C_\ell$, & $\sum m_\ell_j, \sum C_\ell$, \ & & $m_{\ell\ell}, \Delta \phi_{\ell\ell}, m_T$ & $m_{\ell\ell}, \Delta \phi_{\ell\ell}, m_T$ & $m_{\ell\ell}, \Delta \phi_{\ell\ell}, m_T$ & $m_{\ell\ell}, \Delta \phi_{\ell\ell}, m_T$ \ & & $\sum g_j, \Psi_c, \Psi_s$ & & & & & \\
\hline
2nd step & Signal sample & - & - & - & VBF \ & Bkg. sample & - & - & - & ggF \ & BDT inputs & & & $m_{jj}, \Delta \eta_{jj}$, & & $p_T^{\text{sum}}, \sum m_\ell_j$, & & $\sum g_j, \Psi_c, \Psi_s$, \ & & & & & & \\
\hline
\end{tabular}
\caption{Summary of each analytic method for $H \rightarrow WW^*$}
\end{table}
8. transverse mass: 

\[ m_T \equiv \sqrt{(E_T^{\ell\ell} + p_T^{\nu\nu})^2 - |p_T^{\ell\ell} + p_T^{\nu\nu}|^2}, \]

where \( E_T^{\ell\ell} = \sqrt{(p_{\nu\nu}^T)^2 + (m_{\nu\nu})^2} \), \( p_T^{\nu\nu}(p_T^{\ell\ell}) \) is the vector sum of the neutrino (lepton) transverse momenta, and \( p_T^{\nu\nu}(p_T^{\ell\ell}) \) is its modulus.

The distributions of these variables for signal and backgrounds are shown in Fig. [1] in which we can clearly see the capability of each of the variables in discriminating between the signal and backgrounds.

**11-variable BDT** The signal and background training samples are the same as the standard BDT. In addition to the 8 variables in standard BDT, 3 more jet-shape variables [18] are employed in this 11-variable BDT analysis:

1. girth summed over two leading jets: 
\[
\sum g_j \equiv \sum_{j,i \in j} \frac{p_T^{\ell\ell, r_{ij}}}{p_T^{\ell\ell}}
\]

2. the central integrated jet shape: 
\[
\Psi_c \equiv \frac{1}{N} \sum_{j=1}^{N} \frac{1}{2} \sum_{i \in j} \frac{p_T^{\ell\ell, (0<r_{ij}<0.1)}}{p_T^{\ell\ell}}
\]

3. the side integrated jet shape: 
\[
\Psi_s \equiv \frac{1}{N} \sum_{j=1}^{N} \frac{1}{2} \sum_{i \in j} \frac{p_T^{\ell\ell, (0.1<r_{ij}<0.2)}}{p_T^{\ell\ell}}
\]

The distributions of these jet-shape variables for the signal and backgrounds are shown in Fig. [2]

**7-variable BDT** Analyzing the distributions shown in Figs. [1] and Fig. [2] we find that 7 of the variables are sufficient in distinguishing between the VBF events and the others: \( m_{jj} \), \( \Delta \eta_{jj} \), \( p_T^{\text{sum}} \), \( \sum m_{ij} \), \( \sum g_j \), \( \Psi_c \), and \( \Psi_s \). Thus, in this method only these 7 variables are used in discriminating VBF from the ggF and backgrounds. The signal and background training samples are the same as the standard BDT.

**2-step BDT** This is the new approach that we adopt in this study. We separate the training of the BDT in two steps, in which the BDT is trained for VBF against the SM backgrounds and against the ggF, respectively.

- The first step: the VBF signal sample is trained against the SM background samples of \( t\bar{t} \) and \( WW \) events. In this step, the variables used are the same as the standard BDT.

- The second step: after imposing the selection cuts obtained in the first-step-BDT output \( O_{\text{BDT}}^1 \), the event samples will further undergo the second-step BDT, in which the VBF signal sample is trained against the ggF sample only. In this step, the variables used are the same as 7-Var BDT.
B. \( H \to \gamma\gamma \)

Similar to the procedures in \( H \to WW^* \), the events samples for the VBF \( H \to \gamma\gamma \) signal, ggF, and the SM background have to pass the preselection cuts, which were given in the current ATLAS analysis for the SM \( H \to \gamma\gamma \) in the VBF enriched category. The requirements are described as follows:

1. \( N_j \geq 2; \)
2. \( p_T^j > 25 \text{ GeV} \) (\( |\eta^j| < 2.4 \)) and \( p_T^j > 30 \text{ GeV} \) (\( 2.4 < |\eta^j| < 4.4 \));
3. \( \Delta \eta_{jj} > 2; \)
4. \( 105 \leq m_{\gamma\gamma} \leq 160 \text{ GeV} ; \)
5. \( p_T^{\gamma_1} \geq 0.35 m_{\gamma\gamma} \) and \( p_T^{\gamma_2} \geq 0.25 m_{\gamma\gamma} ; \)
6. \( |\eta^*| < 5, \) where \( |\eta^*| \equiv |\eta_{\gamma\gamma} - (\eta_{j1} + \eta_{j2})|/2. \)

**Standard BDT** Following the current procedures in the ATLAS analysis, the signal sample of VBF and the background samples of ggF events and simulated \( \gamma\gamma + jj \) events are used to train the BDT. Again, this is the standard BDT. The following 6 variables are inputs to the BDT:

1. \( m_{jj}; \)
2. \( \Delta \eta_{jj}; \)
3. \( p_{T_{\text{Et}}} \equiv |(p_T^{\gamma_1} + p_T^{\gamma_2}) \times \hat{t}|, \) where \( \hat{t} = (p_T^{\gamma_1} - p_T^{\gamma_2}) / |p_T^{\gamma_1} - p_T^{\gamma_2}|; \)
4. \( \Delta R_{\gamma,j}^{\text{min}} \equiv \) the minimum separation between the leading/subleading photon and the leading/subleading jet;
5. \( |\eta^*|; \)
6. \( \phi^* \equiv \) the azimuthal angle between the diphoton and the dijet system.

The distributions of these variables for the signal and backgrounds are shown in Fig. 3.
| Objective | Standard BDT | 9-Var BDT | 5-Var BDT | 2-step BDT |
|-----------|--------------|-----------|-----------|------------|
| Preselection | $N_j \geq 2,$ | $p_T^j > 25 \text{GeV (} |\eta^j| < 2.4 \text{)} \& p_T^j > 30 \text{GeV (} 2.4 < |\eta^j| < 4.4 \text{)},$ | $\Delta \eta_{jj} > 2,$ | $105 \leq m_{\gamma\gamma} \leq 160 \text{GeV},$ |
|           |              | $p_T^{j1} \geq 0.35m_{\gamma\gamma}$ and $p_T^{j2} \geq 0.25m_{\gamma\gamma}$ |              | $|\eta^*| < 5$ |

1st step

| Signal sample | VBF | VBF | VBF | VBF |
|---------------|-----|-----|-----|-----|
| Bkg. sample   | ggF & $\gamma + jj$ | ggF & $\gamma + jj$ | ggF & $\gamma + jj$ | $\gamma + jj$ |
| BDT inputs    | $m_{jj}, \Delta \eta_{jj}, p_T$, | $m_{jj}, \Delta \eta_{jj}, p_T$, | $m_{jj}, \Delta \eta_{jj}$, | $m_{jj}, \Delta \eta_{jj}, p_T$, |
|               | $\Delta R_{\gamma,j}$, $|\eta^*|$, $\phi^*$ | $\Delta R_{\gamma,j}$, $|\eta^*|$, $\phi^*$ | $\sum g_j, \Psi_c, \Psi_s, \Delta R_{\gamma,j}$, $|\eta^*|$, $\phi^*$ |
|               | $\sum g_j, \Psi_c, \Psi_s$ | $\sum g_j, \Psi_c, \Psi_s$ | $\sum g_j, \Psi_c, \Psi_s$ | $\sum g_j, \Psi_c, \Psi_s$ |

2nd step

| Signal sample | - | - | - | VBF |
|---------------|---|---|---|-----|
| Bkg. sample   | - | - | - | ggF |
| BDT inputs    | - | - | - | $m_{jj}, \Delta \eta_{jj}$, |
|               | | | | $\sum g_j, \Psi_c, \Psi_s$ |

TABLE IV. Summary of each analytic method for $H \rightarrow \gamma\gamma$

9-variable BDT The signal and background training samples are the same as the standard BDT. In addition to the 6 variables in the standard BDT, 3 more jet-shape variables are used in this 9-variable BDT: $\sum g_j, \Psi_c, \Psi_s$, whose distributions are shown in Fig. 4

5-variable BDT Analyzing the distributions of the above 9 variables we find five most powerful variables in discriminating between VBF and ggF. They are $m_{jj}, \Delta \eta_{jj}, \sum g_j, \Psi_c, \Psi_s$, as shown in Fig. 3 and Fig. 4

2-step BDT Again, this is the new approach that we are adopting in this study. We separate the training of the BDT in two steps:

- The first step: the VBF signal sample is trained against the background sample of $\gamma + jj$ events. In this step, the variables used are the same as the standard BDT.
• The second step: after imposing the selection cuts obtained in the first-step-BDT output $O_{1\text{BDT}}$, the event samples will further undergo the second-step BDT, in which the VBF signal samples is trained against the ggF sample. In this step, the variables used are the same as 5-Var BDT.

IV. RESULTS

A. $H \rightarrow WW^* \rightarrow e\nu\mu\nu$

Figure 5 shows the linear correlations between any two of the variables used in the 11-Var BDT for the channel $H \rightarrow WW^*$. From the figure we can see very strong correlations appear among the 3 jet-shape variables, and among $\sum m_{\ell j}$, $\Delta\eta_{jj}$, and $m_{jj}$ in both the signal and backgrounds. A sizeable correlation also appears between $m_{\ell\ell}$ and $\Delta\phi_{\ell\ell}$ in both the signal and backgrounds. In addition, in order to avoid overtraining in BDT analyses, we show the BDT output distributions for both the training and testing samples in Fig. 6.

The results of our analyses for the channel $H \rightarrow WW^*$ are summarized in Table V. The final numbers of the remained VBF events for all methods are all around 5.1. Comparing between the standard BDT and the 11-Var BDT, the latter which used 3 jet-shape variables, can enhance the VBF purity and at the same time reduce the ggF contamination by about 2%. When we focus on distinguishing just between the VBF and ggF event samples, the 7-Var BDT using the most powerful 7 variables is introduced and can further decrease the ggF contamination by about 1%. However, this method sacrifices the discrimination between the VBF sample and the other SM backgrounds, and thus lowers the VBF purity to only 50.5%.

To overcome the problem in the 7-Var BDT, we perform the analysis with a new 2-step BDT method. In the first step, we use the 9 variables as in the standard BDT to discriminate between the VBF and the SM backgrounds including $t\bar{t}$ and $WW$. Whereas in the second step, we focus on discriminating the VBF and ggF using the most powerful discriminators as those used in 7-Var BDT. Figure 7 shows the 2-step BDT output distributions after both steps. Figure 8 shows the VBF purity and ggF contamination versus the cut values of $O_{1\text{BDT}}$ (each event has a larger value than the cut value). It is clear and evident that we shall have purer VBF signal sample when we impose a more stringent cut. Also, the ggF contamination increases slightly as the cut gets more
TABLE V. Summary of the results for the event numbers of each process, VBF purity, and ggF contamination in WW decay channel, after applying cuts on various methods of BDT. Here the ggF contamination is defined as $\frac{N(\text{ggF})}{N(\text{ggF}) + N(\text{VBF})}$. The event numbers are normalized to 5.8 fb$^{-1}$.

| BDT method       | Event number | VBF purity of all processes | ggF contamination |
|------------------|--------------|-----------------------------|------------------|
|                  |              | VBF | ggF | $t\bar{t}$ | WW |
| Standard BDT     | 5.13         | 0.73 | 0.40 | 0.45              | 76.42% | 12.38% |
| 11-Var BDT       | 5.11         | 0.61 | 0.32 | 0.43              | 79.05% | 10.66% |
| 7-Var BDT        | 5.11         | 0.55 | 2.89 | 1.58              | 50.49% | 9.70%  |
| 2-step BDT ($O_{\text{bdt}}^1 > 0.9$) | 5.10 | 0.44 | 0.51 | 0.56              | 77.09% | 7.93%  |

severe. The first-step-BDT output cut value is optimized at 0.9 to obtain the highest purity of VBF and the lowest ggF contamination. As shown in Table V with this new method of 2-step BDT we can highly reduce the ggF contamination down from 12.38 to 7.93%, and at the same time maintain the VBF purity of 77%.

B. $H \rightarrow \gamma\gamma$

In Fig. 9, we show the linear correlations between any two variables that we have used in the channel $H \rightarrow \gamma\gamma$ analyses. We can see that strong correlations among the 3 jet-shape variables, and between $\Delta n_{jj}$ and $m_{jj}$ in both the signal and background samples. In addition, in order to avoid overtraining in the BDT analyses, we show the BDT output distributions for both the training and testing samples as shown in Fig. [10].

The results of our analyses in the channel $H \rightarrow \gamma\gamma$ are summarized in Table VI. We control the VBF efficiency at 5.4% for comparison. The 9-Var BDT, which adds 3 new jet-shape variables compared to the standard BDT, can enhance the VBF purity and at the same time reduce the ggF contamination by about 2%. In order to focus on distinguishing between the VBF and ggF event samples, the 5-Var BDT, which uses the most powerful 5 variables, is introduced and can further decrease the ggF contamination by about 2%. However, this method sacrifices the discrimination from the other SM backgrounds and lowers the VBF purity to only 24.6%.

Similar to the previous channel, we attempt the 2-step BDT method to this case. We use the
TABLE VI. Summary of the results for the event numbers of each process, VBF purity, ggF contamination in diphoton decay channel, after applying cuts on various methods of BDT. The event numbers are normalized to 13.3 fb$^{-1}$.

| BDT method       | VBF efficiency | Event number VBF \(\gamma\gamma + jj\) | VBF purity of all processes | ggF contamination |
|------------------|----------------|------------------------------------------|----------------------------|-------------------|
| Standard BDT     | 5.4%           | 6.19 1.44 10.41                         | 34.3%                      | 18.89%            |
| 9-Var BDT        | 5.4%           | 6.20 1.28 9.59                          | 36.3%                      | 17.08%            |
| 5-Var BDT        | 5.4%           | 6.19 1.12 17.86                         | 24.6%                      | 15.33%            |
| 2-step BDT \((O_{\text{bdt}}^1 > 0.9)\) | 5.4%       | 6.20 0.85 9.48                          | 37.5%                      | 12.00%            |

standard 6 variables in the first step to discriminate between the VBF and \(\gamma\gamma + jj\) background. In the second step, we separate between the VBF and ggF using the most powerful 5 discriminators as those used in 5-Var BDT. Figure[11] shows the 2-step BDT output distribution in both steps. Figure[12] shows the VBF purity and ggF contamination versus the cut value of \(O_{\text{BDT}}^1\). Again, we will have a purer VBF signal sample but with a slightly larger ggF contamination when we apply a more stringent cut value. The cut value of \(O_{\text{BDT}}^1\) is optimized at 0.9 for the highest purity of VBF and the lowest ggF contamination. As shown in Table[VI], the ggF contamination is substantially reduced from 18.89% to 12.00%, and at the same time achieve a small increase in VBF purity (37.5%).

V. CONCLUSIONS

We have studied the performance of the approach of 2-step boosted decision trees. We have followed as closely as the way that the ATLAS generated the event samples of VBF, ggF, and the corresponding SM backgrounds in the channels of \(H \rightarrow WW^*\) and \(H \rightarrow \gamma\gamma\). In the first step, we trained the VBF signal against the SM backgrounds without the ggF sample, while in the second step we trained the VBF signal against the ggF sample.

We have demonstrated with our new approach of 2-step BDT, we can achieve a significant reduction of the ggF contamination from 12% (19%) down to 8% (12%) for \(H \rightarrow WW^* (H \rightarrow \gamma\gamma)\). The ggF contamination that we obtained by the standard BDT in the channel \(H \rightarrow \gamma\gamma\) is somewhat smaller (about 6%) than that obtained in ATLAS[5]. We presume the discrepancy is due to the uncertainty in detector simulations as we use DELPHES while ATLAS uses GEANT4.
At the same time, we can maintain or slightly improve the overall purity of the VBF sample among all the backgrounds.

The approach of this study can be applied to other decay channels, such as $H \rightarrow ZZ^*$, $\tau \tau$, and $b \bar{b}$. Further investigations can include optimization of the number of variables used in each step in the 2-step BDT. Actually, one can use various ways to rank the importance of each variable.

ACKNOWLEDGMENTS

This research was supported in parts by the MoST of Taiwan under Grant Nos. MOST-105-2112-M-007-028-MY3 and MOST-103-2112-M-007-??

[1] G. Aad et al. [ATLAS Collaboration], Phys. Lett. B 716, 1 (2012) [arXiv:1207.7214 [hep-ex]]; S. Chatrchyan et al. [CMS Collaboration], Phys. Lett. B 716, 30 (2012) [arXiv:1207.7235 [hep-ex]].

[2] V. D. Barger, K. m. Cheung, T. Han, J. Ohnemus and D. Zeppenfeld, Phys. Rev. D 44, 1426 (1991). doi:10.1103/PhysRevD.44.1426

[3] V. D. Barger, K. m. Cheung, T. Han and D. Zeppenfeld, Phys. Rev. D 44, 2701 (1991) Erratum: [Phys. Rev. D 48, 5444 (1993)]. doi:10.1103/PhysRevD.48.5444, 10.1103/PhysRevD.44.2701

[4] The ATLAS collaboration [ATLAS Collaboration], ATLAS-CONF-2016-112; G. Aad et al. [ATLAS Collaboration], Phys. Rev. D 92, no. 1, 012006 (2015) doi:10.1103/PhysRevD.92.012006 [arXiv:1412.2641 [hep-ex]].

[5] The ATLAS collaboration [ATLAS Collaboration], ATLAS-CONF-2016-067; G. Aad et al. [ATLAS Collaboration], Phys. Rev. D 90, no. 11, 112015 (2014) doi:10.1103/PhysRevD.90.112015 [arXiv:1408.7084 [hep-ex]].

[6] J. Alwall et al., JHEP 1407, 079 (2014) doi:10.1007/JHEP07(2014)079 [arXiv:1405.0301 [hep-ph]].

[7] T. Sjostrand, S. Mrenna and P. Z. Skands, Comput. Phys. Commun. 178 (2008) 852 doi:10.1016/j.cpc.2008.01.036 [arXiv:0710.3820 [hep-ph]].
[8] J. de Favereau et al. [DELPHES 3 Collaboration], JHEP 1402, 057 (2014) [arXiv:1307.6346 [hep-ex]].

[9] N. Kauer, T. Plehn, D. L. Rainwater and D. Zeppenfeld, Phys. Lett. B 503, 113 (2001) doi:10.1016/S0370-2693(01)00211-8 [hep-ph/0012351].

[10] C. Anastasiou, C. Duhr, F. Dulat, E. Furlan, T. Gehrmann, F. Herzog, A. Lazopoulos and B. Mistlberger, JHEP 1605, 058 (2016) doi:10.1007/JHEP05(2016)058 [arXiv:1602.00695 [hep-ph]]; C. Anastasiou, C. Duhr, F. Dulat, F. Herzog and B. Mistlberger, Phys. Rev. Lett. 114, 212001 (2015) doi:10.1103/PhysRevLett.114.212001 [arXiv:1503.06056 [hep-ph]].

[11] S. Alioli, P. Nason, C. Oleari and E. Re, JHEP 1006, 043 (2010) doi:10.1007/JHEP06(2010)043 [arXiv:1002.2581 [hep-ph]].

[12] P. Nason and C. Oleari, JHEP 1002, 037 (2010) doi:10.1007/JHEP02(2010)037 [arXiv:0911.5299 [hep-ph]].

[13] S. Alioli, P. Nason, C. Oleari and E. Re, JHEP 0904, 002 (2009) doi:10.1088/1126-6708/2009/04/002 [arXiv:0812.0578 [hep-ph]].

[14] H. L. Lai, M. Guzzi, J. Huston, Z. Li, P. M. Nadolsky, J. Pumplin and C.-P. Yuan, Phys. Rev. D 82, 074024 (2010) doi:10.1103/PhysRevD.82.074024 [arXiv:1007.2241 [hep-ph]].

[15] T. Melia, P. Nason, R. Rontsch and G. Zanderighi, JHEP 1111, 078 (2011) doi:10.1007/JHEP11(2011)078 [arXiv:1107.5051 [hep-ph]].

[16] T. Sjostrand, S. Mrenna and P. Z. Skands, JHEP 0605, 026 (2006) doi:10.1088/1126-6708/2006/05/026 [hep-ph/0603175].

[17] A. Hocker et al., “TMVA - Toolkit for Multivariate Data Analysis,” PoS ACAT , 040 (2007) [physics/0703039 [PHYSICS]].

[18] J. Shelton, doi:10.1142/9789814525220_0007 [arXiv:1302.0260 [hep-ph]].
FIG. 1. Distributions of various variables used in the standard BDT in the $H \to WW^*$ channel for the VBF signal, ggF, and the SM backgrounds.
FIG. 2. Distributions of the 3 jet-shape variables used in the $H \to WW^*$ channel.
FIG. 3. Distributions of the 6 variables used in the channel $H \rightarrow \gamma\gamma$ for the VBF signal, ggF, and the SM background $\gamma\gamma jj$. 
FIG. 4. Distributions of the 3 jet-shape variables used in the channel $H \rightarrow \gamma \gamma$.

FIG. 5. Linear correlations between any two variables in the channel $H \rightarrow WW^* \rightarrow e\nu\mu\nu$ used in the 11-Var BDT. The left panel shows the VBF while the right panel includes ggH, $tt\bar{t}$, and WW.
FIG. 6. BDT output distributions of training and test samples in $H \rightarrow WW^*$. (Top-left) Method of standard BDT. (Top-right) Method of 11-Var BDT. (Middle-left) Method of 7-Var BDT. (Middle-right) The first step in the 2-step BDT. (Bottom) The second step in the 2-step BDT. Note that half of sample is used as the training sample and the other half as the testing sample.
FIG. 7. 2-step BDT output distributions in first step(left) and second step(right) for each process in $H \rightarrow WW^*$. 

FIG. 8. Combined plot of VBF purity and ggF contamination in $H \rightarrow WW^*$ versus the cut value on $O_{\text{BDT}}$. 
FIG. 9. Linear correlation of each variable used in BDT in $H \to \gamma\gamma$. In this figure, the signal (left) denotes VBF, and the background (right) includes ggH as well as $\gamma\gamma + jj$. 
FIG. 10. BDT output distribution of training and testing samples in $H \rightarrow \gamma\gamma$. (Top-left) Method of standard BDT. (Top-right) Method of 9-Var BDT. (Middle-left) Method of 5-Var BDT. (Middle-right) The first step in the 2-step BDT. (Bottom) The second step in the 2-step BDT. Note that half of sample is used as the training sample and the other half as the testing sample.
FIG. 11. 2-step BDT output distributions in first step (left) and second step (right) for each process in $H \rightarrow \gamma \gamma$.

FIG. 12. Combined plot of VBF purity and ggF contamination in $H \rightarrow \gamma \gamma$ versus the cut value on $O_{\text{BDT}}^1$. 
