

# Discrimination of converted photons and neutral pions at high energies<sup>\*</sup>

XIAO Hong(肖虹)<sup>1,2;1)</sup> CHEN Guo-Ming(陈国明)<sup>1</sup> TAO Jun-Quan(陶军全)<sup>1</sup>  
 BIAN Jian-Guo(卞建国)<sup>1</sup> FAN Jia-Wei(范嘉伟)<sup>1,2</sup> LIANG Song(梁松)<sup>1,2</sup>  
 MENG Xiang-Wei(孟祥伟)<sup>1</sup> SHEN Yu-Qiao(沈玉乔)<sup>1,2</sup> TANG Zhi-Cheng(唐志成)<sup>1</sup>  
 WANG Jian(王健)<sup>1</sup> WANG Jian(王健)<sup>1,2</sup> WANG Xian-You(汪先友)<sup>1,3</sup> WANG Zheng(王征)<sup>1</sup>  
 XU Ming(徐明)<sup>1,2</sup> XU Wei-Wei(许伟伟)<sup>1,2</sup> YANG Min(杨民)<sup>1</sup> YAN Qi(严琪)<sup>1,2</sup>

<sup>1</sup> Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China

<sup>2</sup> Graduate University of Chinese Academy of Sciences, Beijing 100049, China

<sup>3</sup> Theoretical Physics Institute, Chongqing University, Chongqing 400044, China

**Abstract:** In the LHC experiment, the  $H \rightarrow \gamma\gamma$  channel provides a clean final state with an effective mass peak that is reconstructed with great precision, despite the small branching ratio. As a consequence, the  $H \rightarrow \gamma\gamma$  channel is one of the most promising channels for the Higgs discovery in the very low mass region. In order to increase the sensitivity of the Higgs search, background rejection rate is very important, so  $\gamma/\pi^0$  discrimination is one of the key points in the analysis. At least 40% of photons will convert with the experience of ATLAS and CMS. We constructed electromagnetic calorimeter (ECAL) in GEANT4 simulation, using 6 variables which have different shapes between converted  $\gamma$  and  $\pi^0$ , with the TMVA (Toolkit for Multivariate Data Analysis) to do the separation. With this method we can get 30% to 60%  $\pi^0$  rejection efficiency when keeping 90% converted  $\gamma$  efficiency, in the region of transverse momentum 15 GeV to 75 GeV, not only in MC simulation but also in real data.

**Key words:** converted  $\gamma$ ,  $\pi^0$ , discrimination, TMVA

**PACS:** 11.30.Rd, 14.70.Bh, 07.05.Mh      **DOI:** 10.1088/1674-1137/36/10/011

## 1 Introduction

In the LHC experiment, jet production rate is very large. These jets contain energetic  $\pi^0$ s which decay into two photons, and leave an energy deposit in the electromagnetic calorimeter (ECAL). Because the two photons are too close to each other, they will be misinterpreted as one photon of the same energy. For this reason, it is hard to separate  $\pi^0$  and  $\gamma$ . Neutral pion is the main reducible background for high energy photon analysis like the  $H \rightarrow \gamma\gamma$  channel, so it is very important to improve the rejection power of the background in order to increase signal significance.

Since the material budget of the tracker is large,

at least 40% of photons will convert into two electrons with the experience of ATLAS [1] and CMS [2]. In the diphoton final state like  $H \rightarrow \gamma\gamma$ , the percentage of events which contain at least one converted photon will be even higher. In this paper, we develop a method to discriminate  $\pi^0$  and converted  $\gamma$  taking their different shower shapes into consideration using the TMVA (Toolkit for Multivariate Data Analysis).

## 2 The detector description

One of the primary goals of LHC is to search for Higgs boson.  $H \rightarrow \gamma\gamma$  is one of the most promising channels for low mass Higgs discovery, because of the clean final state topology. In order to get a better sen-

Received 18 January 2012, Revised 15 March 2012

<sup>\*</sup> Supported by Natural Science Foundation of China (10435070, 10773011, 10721140381, 10099630) and China Ministry of Science and Technology (2007CB16101, 2010CB833000)

1) E-mail: xiaoh@ihep.ac.cn

©2012 Chinese Physical Society and the Institute of High Energy Physics of the Chinese Academy of Sciences and the Institute of Modern Physics of the Chinese Academy of Sciences and IOP Publishing Ltd

sensitivity on the Higgs search, converted  $\gamma$  and  $\pi^0$  discrimination is extremely important. We constructed an ECAL (Electromagnetic Calorimeter) with geometry like the barrel region of the CMS detector at the LHC [3] with GEANT4 package.

The homogeneous calorimeter is made of 61200 lead tungstate ( $\text{PbWO}_4$ ) crystals. There are 36 super-modules in the whole part, 18 in each half barrel, each covering  $20^\circ$  in  $\phi$ , and each super-module has 85 crystals in  $\eta$  direction (see Fig. 1). They are labelled 1 to 85 from  $z = 0$  which will be used as  $i\eta$  in Section 4.1. Each super-module has 4 modules. Module 1 contains 500 crystals, while modules 2, 3 and 4 each contain 400 crystals. Modules are assembled in units of  $2 \times 5$  crystals called sub-modules. The long side of the sub-modules is mounted parallel to the  $z$  axis. The crystals have 17 possible shapes and two orientations of the internal angles, so that crystals are classified as left and right crystals and there are 34 types in total. The crystals are 3 degrees off-pointing with respect to the vector from the nominal interaction vertex both in  $\eta$  and  $\phi$  direction in order to avoid cracks aligned with particle trajectories. The crystal's cross section is roughly  $0.0174 \times 0.0174$  in  $\eta$  and  $\phi$ , or  $22 \text{ mm} \times 22 \text{ mm}$  at the front face and  $26 \text{ mm} \times 26 \text{ mm}$  at the rear face. The crystal length is 230 mm according to  $25.8X_0$ , and the center of the front faces of the crystals is at a radius 1.29 m from the  $z$  axis [4].

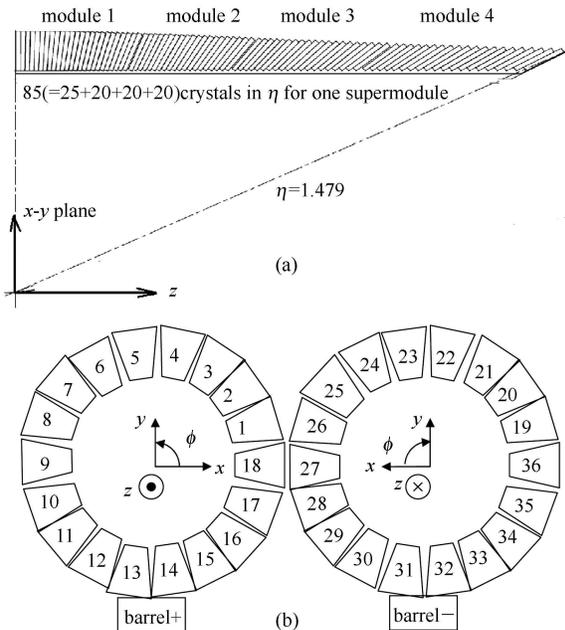


Fig. 1. ECAL super-module numbering scheme in the geometry description used in this paper (a) The longitudinal section of the ECAL (one quarter) (b) The transverse view of the ECAL.

We also constructed the same tracker geometry in front of the ECAL and the magnetic field which is 3.8 T along the  $z$  direction as the CMS experiment in the GEANT4 simulation. Details of the geometry were described in Ref. [4].

### 3 The photon reconstruction

Photon reconstruction in the ECAL depends on the shower deposits of photon energy in the crystals. The ECAL is made of  $\text{PbWO}_4$ , whose Moliere radius ( $R_M$ ) is about 2 cm. Within radius  $R_M$  there is about 90% of the energy of the electromagnetic cascades, while 99% is contained in  $3.5R_M$  [2]. According to our experience, the energy of a single photon in  $3 \times 3$  crystals is about 94%, while there is approximately 97% in  $5 \times 5$  crystals. Unconverted photons measured in such a fixed area will give the best performance. However, converted photons produce Bremsstrahlung and cause an extended spread in the  $\phi$  direction in the ECAL, influenced by the strong magnetic field and the material in front of the calorimeter. So we use the following method which is called a clustering algorithm to get the shower shape which is called a super-cluster.

First, we look for a single crystal maxima which appears as a bump in the array of crystal energy deposits. In order to get rid of the effect of noise, we set the crystal energy threshold to 80 MeV, so there will be a hole when the crystal energy deposit is lower than the crystal energy threshold. Then the left bumps are selected as seed crystals. The clustering algorithm is based on these seed crystals. Second, starting from the most energetic seed, we collect the crystals in a  $5 \times 5$  array whose energy deposition center is our seed crystal as shown in Fig. 2(a). Then one seed clustering is completed. All the crystals we include in the seed clustering will form a cluster shape, which we call a basic-cluster. After all these steps, we can start another seed clustering with the same procedures as the previous one. We should notice that one crystal should not be collected when the crystal is already clustered by another seed, therefore no double counting will happen; more details can be found in Fig. 2(b). Because of the spread of converted photons in the ECAL, one basic-cluster is not enough to collect all the energy. In order to get all the photon energy, we collect all of the basic-clusters in the super-cluster range and use them as a whole. In terms of the experience, the super-cluster window is 0.06 in  $\phi$  direction and 0.80 in  $\eta$  direction. The window corresponds to both positive and negative direction.

The photon position is calculated by the energy-weighted mean position of the crystals in the super-cluster

$$x = \frac{\sum x_i \cdot W_i}{\sum W_i}, \quad (1)$$

where  $x_i$  is the position of the crystal  $i$  in the super-cluster, and  $W_i$  is the weight of the crystal in the whole super-cluster which is defined as

$$W_i = W_0 + \lg \frac{E_i}{\sum E_i}, \quad (2)$$

where  $W_0 = 4.7$  [5].

In the following analysis the  $5 \times 5$  array whose center is the seed will be frequently used [6].

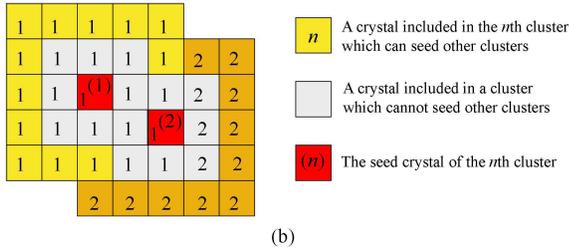
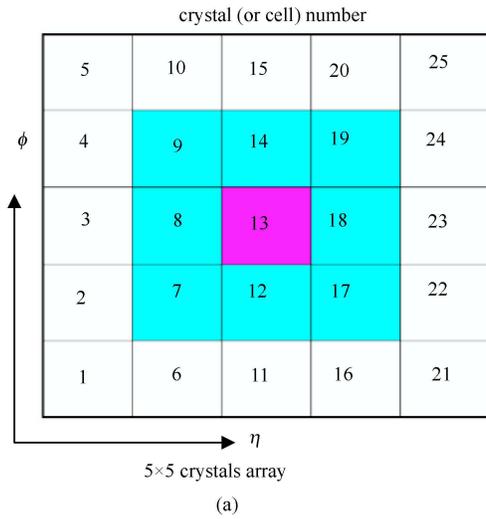


Fig. 2. (a) Illustration of the  $5 \times 5$  clustering algorithm and the numbering of the  $5 \times 5$  array; (b) Illustration of two overlapping multi  $5 \times 5$  clusters.

## 4 The converted $\gamma/\pi^0$ discrimination

In this section we use several variables, most of which describe the shower shapes of converted  $\gamma/\pi^0$  in the electromagnetic calorimeter. These variables are selected from lots of shower shape variables according to the separation power in converted  $\gamma/\pi^0$  discrimination. In the analysis, we use the TMVA package in ROOT to get the discrimination result [7].

### 4.1 Variable description

The shower shape variables are used as criteria to distinguish converted  $\gamma/\pi^0$  because of the sensitivity to differences between single and overlapping converted photon showers [8].

In this analysis we use 6 variables,  $\sigma_{i\eta i\eta}$ , CEP,  $\sigma_\eta$ ,  $\sigma_\phi$ ,  $E_{\text{over}P}$ ,  $r_9$ . The detailed descriptions are as follows.

$\sigma_{i\eta i\eta}$  is the eigenvalue of the local covariance between  $\eta$  and  $\eta$ ,

$$\sigma_{i\eta i\eta} = \sqrt{\frac{\sum_{i=1}^n ((i\eta_i - i\eta_{\text{seed}}) - \langle i\eta \rangle)^2 \times w_i}{cs \times \sum_{c=1}^{25} w_i}}, \quad (3)$$

$$\langle i\eta \rangle = \frac{\sum_{i=1}^{25} E_i \times (i\eta_i - i\eta_{\text{seed}})}{\sum_{i=1}^{25} E_i},$$

where

$$w_i = \text{MAX} \left( 0, w_0 + \lg \frac{E_i}{S_{25}} \right),$$

$w_0 = 4.7$ ,  $cs = \text{crystalsize} = 0.01745$ ,  $i\eta_i$ ,  $i\eta_{\text{seed}}$  are crystal labels in  $\eta$  direction of the crystal and the seed which are described in Section 2.

CEP is defined as the covariance between  $\eta$  and  $\phi$ ,

$$\text{CEP} = \frac{\sum_{i=1}^n (\eta_i - \langle \eta \rangle)(\phi_i - \langle \phi \rangle) \times w_i}{\sum_{c=1}^{25} w_i}, \quad (4)$$

$$\langle \eta \rangle = \frac{\sum_{i=1}^{25} E_i \times \eta_i}{\sum_{i=1}^{25} E_i}, \quad \langle \phi \rangle = \frac{\sum_{i=1}^{25} E_i \times \phi_i}{\sum_{i=1}^{25} E_i},$$

$\sigma_\eta$  is defined as the width of the shower in eta direction,

$$\sigma_\eta = \sum_{i=1}^n \sqrt{\frac{E_i^{5 \times 5}}{E_{\text{seed}}^{5 \times 5}} \times (\eta_{\text{seed}} - \eta_i)^2}, \quad (5)$$

$\sigma_\phi$  is defined as the width of the shower in phi direction,

$$\sigma_\phi = \sum_{i=1}^n \sqrt{\frac{E_i^{5 \times 5}}{E_{\text{seed}}^{5 \times 5}} \times (\phi_{\text{seed}} - \phi_i)^2}. \quad (6)$$

The above four variables sum up all the crystals in single photon reconstruction which pass the selection in Section 3. Seed crystal is the local crystal energy maxima.  $E_{\text{seed}}^{5 \times 5}$  is the energy sum of the crystals in

the  $5 \times 5$  region whose center is the seed crystal.  $E_i^{5 \times 5}$  is the energy sum of the crystals in the  $5 \times 5$  region whose center is the crystal  $i$ .  $\eta_{seed}$ ,  $\phi_{seed}$ ,  $\eta_i$  and  $\phi_i$  are the position of the crystal center of the seed crystal and crystal  $i$  respectively.

$E$  over  $P$  is defined as the photon super-cluster energy divided by converted photon momentum,

$$E_{overP} = \frac{E_{super-cluster}}{P_{converted\gamma}}, \quad (7)$$

$r_9$  is the ratio between the energy of center  $3 \times 3$  region (crystal 7, 8, 9, 12, 13, 14, 17, 18, 19 in Fig. 2(a)) and the whole super-cluster energy,

$$r_9 = \frac{E_{3 \times 3}}{E_{super-cluster}}. \quad (8)$$

Super-cluster is defined in Section 3.

## 4.2 The sample discription

Although we simulate 200000 events in every  $E_T$  bin for both single  $\gamma$  and  $\pi^0$  samples,  $\gamma$  and  $\pi^0$  samples are produced uniformly in both  $\eta$  and  $\varphi$  directions. In order to get a more typical shape to separate converted  $\gamma$  and  $\pi^0$ , we only choose the events which contain at least one converted  $\gamma$ . We choose  $E_T$  bins based on the sensitive region of standard model  $H \rightarrow \gamma\gamma$ . The detailed events are listed in Table 1.

From Fig. 3 which shows variable distributions, we can see the differences of shower shape between converted  $\gamma$  and  $\pi^0$ , while converted  $\gamma$  is signal,  $\pi^0$  is background.

Table 1. The number of selected converted  $\gamma/\pi^0$  events in each  $E_T$  bin.

$E_T/\text{GeV}$	converted $\gamma$ events	$\pi^0$ events
15–25	48138	54250
25–35	53901	66003
35–45	56274	72556
45–55	57419	77248
55–65	59193	81088
65–75	59436	82694

## 4.3 Discrimination result with the TMVA

The 6 variables in Section 4.1 are used as the input variables of the TMVA classifiers to perform the converted  $\gamma/\pi^0$  discrimination with the samples described in the above subsection, while half of the samples are used for training and the other half for testing.

The Boosted Decision Trees (BDT) and Multi-layer Perceptron (MLP) methods are used in the TMVA analysis [7].

MLP which is a recommended neural network method, is optimized by the improved back propagation algorithm based on an artificial neural network, thus it has a clear speed advantage compared with a traditional artificial neural network. It has 3 kinds of layers, the input layer (our input variables), the hidden layer and the output layer (output result). Every layer has several neurons while there is no connection

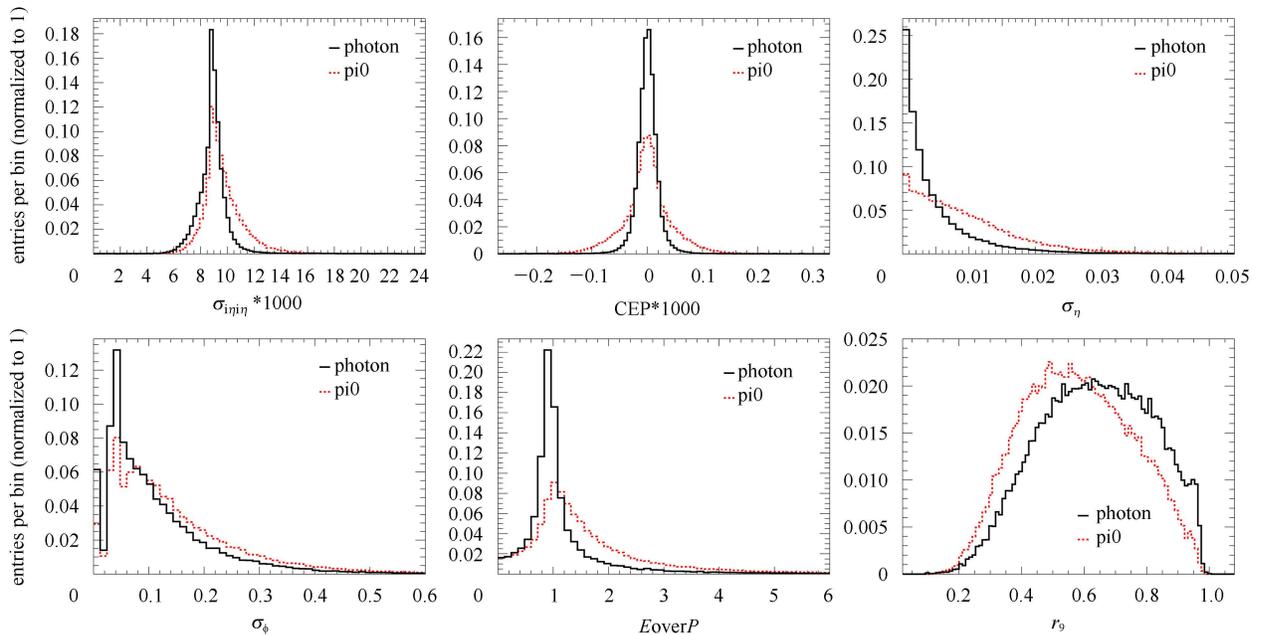


Fig. 3. Different distributions between converted  $\gamma/\pi^0$  for 6 variable distributions in  $E_T$  region 35 GeV to 45 GeV.

within a layer and full connection between layers. We can have several hidden layers. Using the improved back propagation algorithm on each layer, we can get the MLP result.

BDT represents an extension to a single decision tree. A single decision tree has a two dimensional tree structure which can be separated as signal-like or background-like leaf node, and then the leaf node can be separated further. The separation will be terminated when the nodes are dominated by signal or background. However, there is fluctuation in one decision tree, so other reweighted trees are added into the training. The result of BDT is the majority vote

of the trees.

Figure 4 shows the analysis results with the TMVA package. The background rejection changes with the signal efficiency from the testing samples are shown in Figs. 4(a) and 4(b) for the 6  $E_T$  bins with BDT and MLP methods respectively. The plots show the higher the  $E_T$  is, the lower the rejection power we can get. The numbers in the legend show the background rejection power for different  $E_T$  bins when 90% signal efficiency is kept. We can get more detailed information on what we can get in Table 2. Figs. 4(c) and 4(d) show the TMVA result of signal training, testing samples, background training and

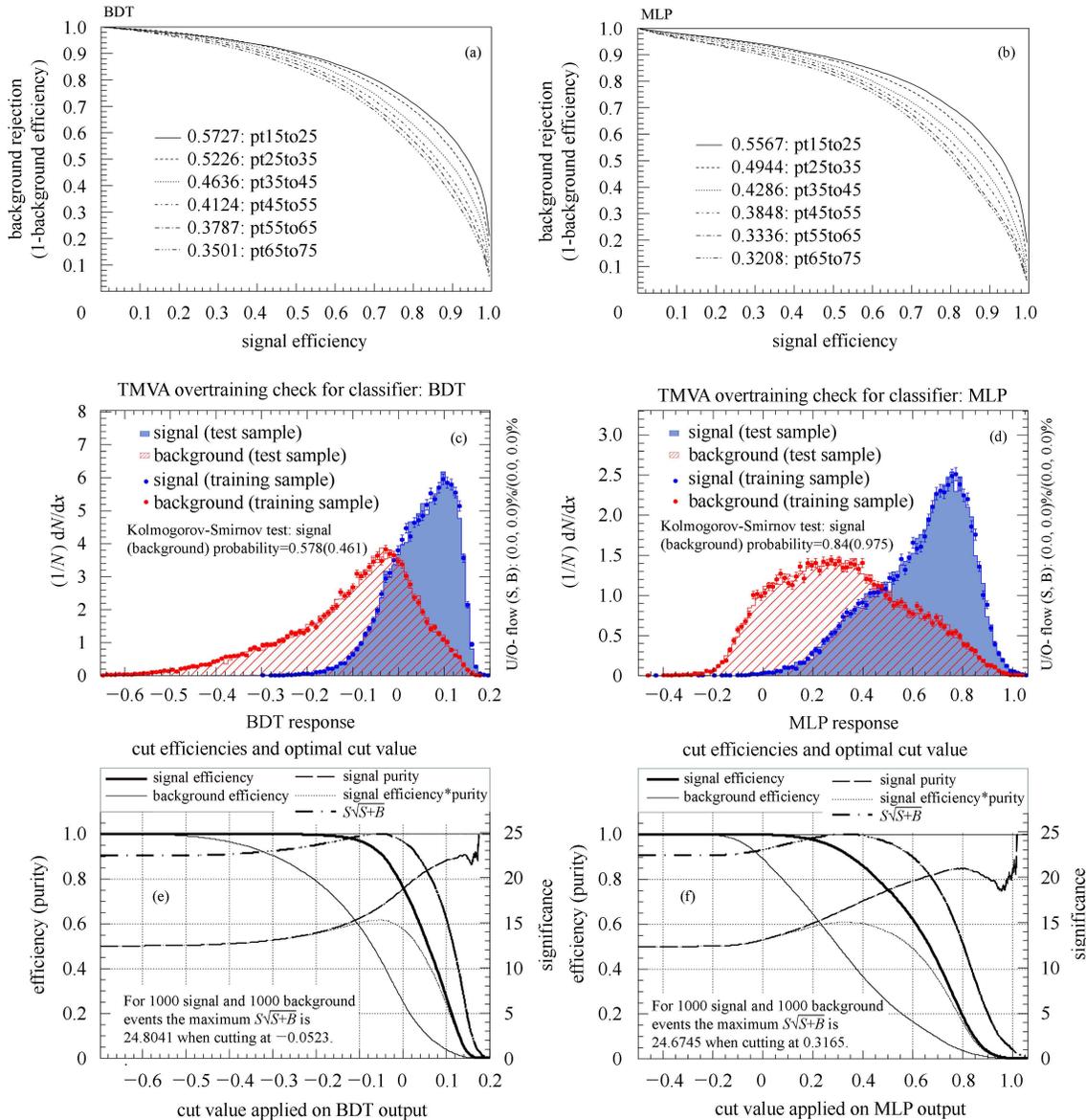


Fig. 4. Results from the TMVA analysis. Background rejection versus signal efficiency of (a) BDT method and (b) MLP method; Responses and overtraining check for (c) BDT method and (d) MLP method with 35–45 GeV samples; Efficiency and purity changes with the (e) BDT output and (f) MLP output with 35–45 GeV samples.

testing samples. From the distributions, the responses of training and testing samples agree well for both of the signal and background, which means no over-training exists during the analysis. Figs. 4(e) and 4(f) show the signal efficiency, background efficiency, signal purity and significance  $\left(\frac{S}{\sqrt{S+B}}\right)$  for both methods.

The background ( $\pi^0$ ) rejections for different  $E_T$  bins are listed in Table 2, when 90% signal (converted  $\gamma$ ) efficiency is kept. It can be seen that the BDT and MLP have almost the same  $\pi^0$  rejection efficiency for the same  $E_T$  bin while the BDT method can give more robust results.

Table 2. The  $\pi^0$  rejection efficiency if 90% converted  $\gamma$  efficiency is kept in each  $E_T$  bin with BDT and MLP method.

$E_T/\text{GeV}$	BDT	MLP
15–25	57.27%±0.4%	55.87%±0.4%
25–35	52.26%±0.4%	49.44%±0.4%
35–45	46.36%±0.4%	42.86%±0.4%
45–55	41.24%±0.4%	38.48%±0.4%
55–65	37.87%±0.4%	33.36%±0.4%
65–75	35.01%±0.4%	32.08%±0.4%

LHC has taken data for two years. We found the variables we use in this paper agree very well between data and MC which is simulated with GEANT4 package. Only  $\sigma_{i\eta i\eta}$  in data has a 0.9% shift with respect to the average of the simulated photon  $\sigma_{i\eta i\eta}$  values [9]. Because we can not find pure photon signal and background samples with enough statistics for the TMVA, in order to make data and MC agree with each other,

we corrected our MC simulation for the observed shift and train the samples with the TMVA again. We get almost the same result (Table 3) with Table 2.

Table 3. The  $\pi^0$  rejection efficiency if 90% converted  $\gamma$  efficiency is kept in each  $E_T$  bin with BDT and MLP method after data correction.

$E_T/\text{GeV}$	BDT	MLP
15–25	56.23%±0.4%	54.11%±0.4%
25–35	51.45%±0.4%	48.01%±0.4%
35–45	46.15%±0.4%	41.66%±0.4%
45–55	41.74%±0.4%	38.32%±0.4%
55–65	38.47%±0.4%	30.70%±0.4%
65–75	36.22%±0.4%	30.03%±0.4%

## 5 Discussion and conclusion

In this paper, we construct an ECAL detector and use GEANT4 to simulate data. 6 variables which have discrimination power between converted  $\gamma$  and  $\pi^0$  are selected as inputs of the TMVA. BDT and MLP classifiers are used as the analysis methods to get the discrimination result in 6  $E_T$  bins. From the result we find the higher the  $E_T$  is, the lower the rejection power we can get. With  $E_T$  between 15GeV to 75 GeV, about 30%–60%  $\pi^0$  rejections can be obtained when 90% converted  $\gamma$  efficiency is kept. From the analysis results, this method is helpful in converted  $\gamma/\pi^0$  discrimination related analysis not only in MC but also in real data, such as light Higgs (with mass around 120 GeV) searches with the  $H \rightarrow \gamma\gamma$  channel in the LHC experiments, by suppressing the jet backgrounds which contain a lot of neutral pions.

## References

- Aad G et al. (ATLAS collaboration). Expected Photon Performance in the ATLAS Experiment, Tech. Rep. ATLAS-PHYS-PUB-2011-007, CERN, Geneva
- Nakamura K et al. (Particle Data Group). Journal of Physics G: Nuclear and Particle Physics, 2010, **37**: 075021
- Adair A et al. (CMS collaboration). J. Instrum., 2008, **3**: S08004. 361
- Bayatian G L et al. CMS Physics Technical Design Report Volume I: Detector Performance and Software, Technical Design Report CMS. Geneva: CERN, 2006
- Awes T C, Obenshain F E, Plasil F et al. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 1992, **311**: 130
- TAO, Jun-Quan et al. Chin. Phys. C, 2011, **35**: 269
- Hoecker A, Speckmayer P, Stelzer J et al. 2007, ArXiv Physics e-prints
- Grunhaus J, Kananov S. Nucl. Instrum. Methods A, 1997, **391**: 351
- Khachatryan V, Sirunyan A M, Tumasyan A et al. Phys. Rev. Lett., 2011, **106**: 082001