*R*esearch in
*A*stronomy and
*A*strophysics

A machine learning method to separate cosmic ray electrons from protons from 10 to 100 GeV using DAMPE data

Hao Zhao[1,2], Wen-Xi Peng[1], Huan-Yu Wang[1,2], Rui Qiao[1], Dong-Ya Guo[1], Hong Xiao[3] and Zhao-Min Wang[4,5]

[1] Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China; *zhaohao@ihep.ac.cn, pengwx@ihep.ac.cn*

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] Changzhou Institute of Technology, Changzhou 213032, China

[4] Gran Sasso Science Institute (GSSI), Via Iacobucci 2, I-67100 L'Aquila, Italy

[5] Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Lecce, Via per Arnesano, I-73100 Lecce, Italy

Abstract DArk Matter Particle Explorer (DAMPE) is a general purpose high energy cosmic ray and gamma ray observatory, aiming to detect high energy electrons and gammas in the energy range 5 GeV to 10 TeV and hundreds of TeV for nuclei. This paper provides a method using machine learning to identify electrons and separate them from gammas, protons, helium and heavy nuclei with the DAMPE data acquired from 2016 January 1 to 2017 June 30, in the energy range from 10 to 100 GeV.

Key words: astroparticle physics — methods: data analysis — cosmic rays

1 INTRODUCTION

The high energy spectrum of cosmic ray electrons (CREs) has been investigated both on the ground and in space in recent years (Aguilar et al. 2014; Picozza et al. 2007; Aharonian et al. 2008), in order to gain knowledge on CREs' origin and propagation, and to try to search for signals of dark matter annihilation (Pospelov & Ritz 2009). The direct detection of CREs in space is more accurate than ground detections since there is no atmosphere between the primary CREs and the detectors. Due to the small effective area of satellite experiments, the electron flux could be measured up to a few TeV. A direct detection of electrons up to 10 TeV may reveal some unknown physics related to dark matter or nearby sources (Kobayashi et al. 2004; Yin et al. 2013).

The DArk Matter Particle Explorer (DAMPE) was successfully launched into a Sun-synchronous orbit at an altitude of 500 km on 2015 December 17 from the Jiuquan Satellite Launch Center (Chang et al. 2017). DAMPE mainly focuses on the detection of Galactic cosmic rays (GCRs), the potential signal of dark matter annihilation and gamma-ray astronomy. It consists of four sub-detectors, a Plastic Scintillator strip Detector (PSD), a Silicon-Tungsten tracKer-converter (STK), a BGO (Bi4Ge3O12) imaging calorimeter and a NeUtron Detector (NUD). The PSD consists of two layers of scintillators to measure the charge of passing nuclei and as veto detector for gamma rays; the STK consists of 12 layers of single-sided silicon detectors. It can measure the trajectory and charge of charged particles, identify nuclei up to oxygen and the tungsten plates inside can convert photons into e^+e^- pairs. The BGO calorimeter is made of 14 layers where each layer contains 22 BGO bars. It measures the energy profile of an electromagnetic shower induced by electrons, positrons or photons, or a hadronic shower induced by protons and nuclei. The NUD records the secondary neutrons produced in a BGO shower and contributes to the identification of electrons/protons.

It is a big challenge to get clean samples of electrons from GCR data since the number of background protons is approximately 10^3 times larger than that of elec-

trons above hundreds of GeV. The traditional cut-based method used in DAMPE Collaboration et al. (2017) relies on a full understanding of the detector but does not take into account all the relations between the variables used and in this way could miss some hidden piece of information. The machine learning method has been widely used in data analysis of various physics experiments in the last decades and usually gives better results than the traditional cut-based method (Roe et al. 2005; Abdollahi et al. 2017). In this paper, we present an analysis using the machine learning method to separate electrons from background, mainly protons, in energy range from 10 to 100 GeV.

The whole procedure used in the analysis is introduced in Section 2. The data analysis is presented in Section 3 and Section 4 in detail. A discussion about the method is provided in Section 5 and the conclusion given is in Section 6.

2 ANALYSIS PROCEDURE

The analysis starts with DAMPE 2A ROOT files (Monte Carlo (MC) data from version 5.3.0), which have already been properly calibrated and reconstructed (Zhang et al. 2016). A flow chart illustrating the machine learning method is shown in Figure 1.

The whole process starts with an energy correction process to correct the energy of electron events (Sect. 3.1). Since there are lots of events with large incident angles, events in which the shower is partially recorded in the BGO or events from the bottom of the detector, the energies may not be precisely reconstructed. A pre-selection is applied to eliminate those events in both MC data and on-orbit data (Sect. 3.2). After the pre-selection, we extract some characteristic variables from BGO (which represents features of the shower) of all the surviving electrons and protons in the MC data (Sect. 4.1), then feed the variables into the machine learning training algorithm to yield a "discriminator" to distinguish electrons from protons (Sect. 4.3).

3 DATA PRE-SELECTION

3.1 Energy Correction

As the first step, the energy of an electron event needs to be corrected. The total energy of the incident electron may not be fully collected because some energies of the shower might be deposited in the supportive structure or leak from the side or bottom of the detector. So, it is necessary to correct the deposited energy from electrons to compute their original energy. We use the MC data to evaluate the deposited energy distribution for electrons. An example is shown in Figure 2. The distribution of deposited energy/MC truth energy ratio approximates a Gaussian distribution.

We fit data displayed in Figure 2 with a Gaussian distribution, and calculate the mean and sigma values. The mean value is taken as the ratio between the deposited energy and the MC truth energy of incident electrons for all the events in this energy bin. The electron energy resolution is based on the MC data from 10 to 100 GeV, which is the sigma from the Gaussian fit, as shown in Figure 3. This resolution is well-matched with the beam test result, indicating a good configuration for the simulation (Zhang et al. 2016).

The energy correction process needs to be applied to all on-orbit data since we cannot tell if a given event results from an electron or not. Through this process, the energies of electrons are corrected but the energies of protons are also changed. Since our final objective is to get a clean electron sample and reconstruct the electron spectrum later, this is not a concern as long as the electron energy is precise.

3.2 Pre-selection Cut

After the energy correction, we need to select events whose energies and trajectories can be reconstructed accurately. We call this procedure 'pre-selection'. Several cuts are implemented to remove events that have poor energy reconstruction, both on MC and on-orbit data. Here are those cuts in order:

(1) There is at least one *Globtrack* (Chang et al. 2017) reconstructed successfully. The *Globtracks* of an event are its STK tracks which match the BGO reconstructed track. Then we select the one with the smallest chi-square value from the track reconstruction for further analysis.
(2) Use PSD reconstructed energy to eliminate He and heavier nuclei.
(3) Use PSD reconstructed energy to eliminate photons.
(4) *Globtrack* goes through both the top and bottom layer of the BGO calorimeter.
(5) The bar with the maximum deposited energy is not on the edge of the calorimeter.
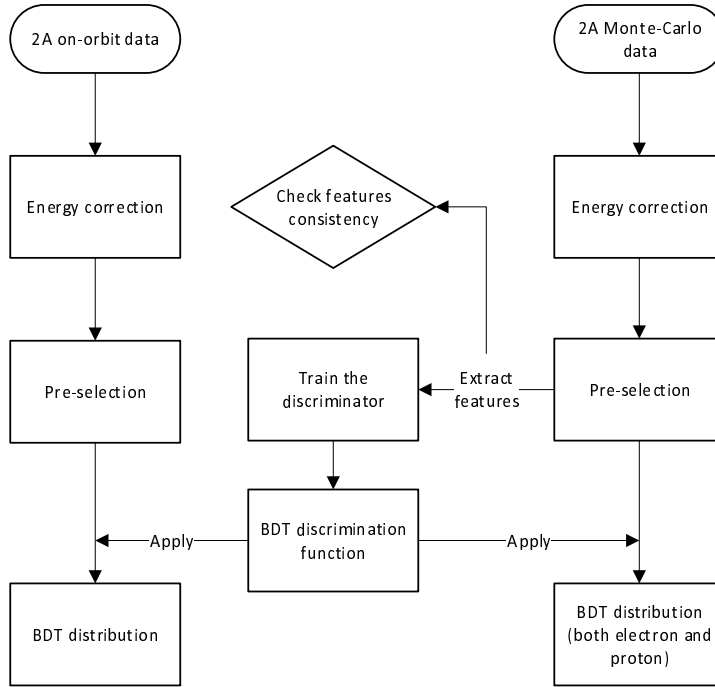
Fig. 1 A flow chart illustrating the machine learning analysis.
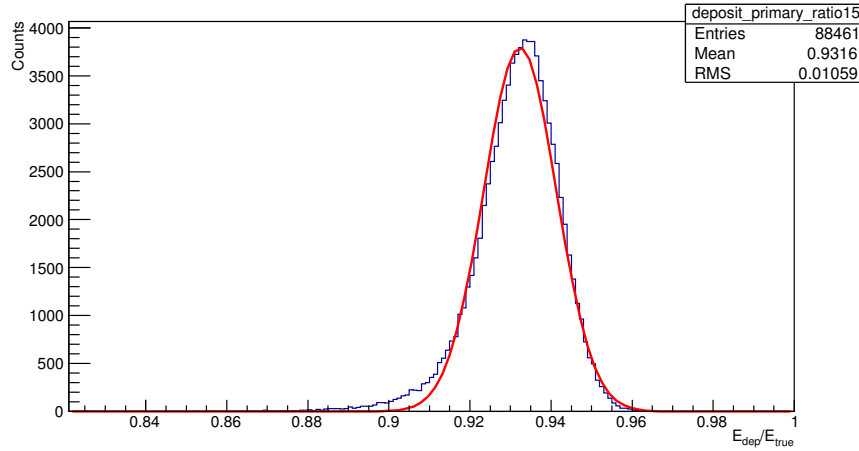
Fig. 2 A Gaussian fit to the distribution of energy ratio $E_{\mathrm{dep}}/E_{\mathrm{true}}$ for MC electrons deposited in the 86.6–100 GeV energy bin, where E_{dep} is the deposited energy of electrons and E_{true} is the true energy. Fit result: $\mathrm{mean} = 0.929$; $\mathrm{sigma} = 0.011$.

(6) The ratio between the total *RMS* and *HorizontalRMS* (defined in Sect. 4.1) is larger than 15 to exclude particles incident from the flank of DAMPE.

(7) The energy deposition ratio of the first BGO layer is less than 0.2. This cut tries to eliminate those events with large incident angles, but which cannot be removed by cut 4 since the reconstructed *Globtrack* of these events is inaccurate.

Figure 4 shows the distribution of deposited energy after all cuts. Those events with y-axis value close to 0 are fully deposited events which need to be preserved. It is obvious that most events with small deposited energy have been excluded by these cuts. After all cuts are applied, about 0.1% of partially deposited events remain in the whole sample, which is acceptable considering the error from this effect is about 0.1%.
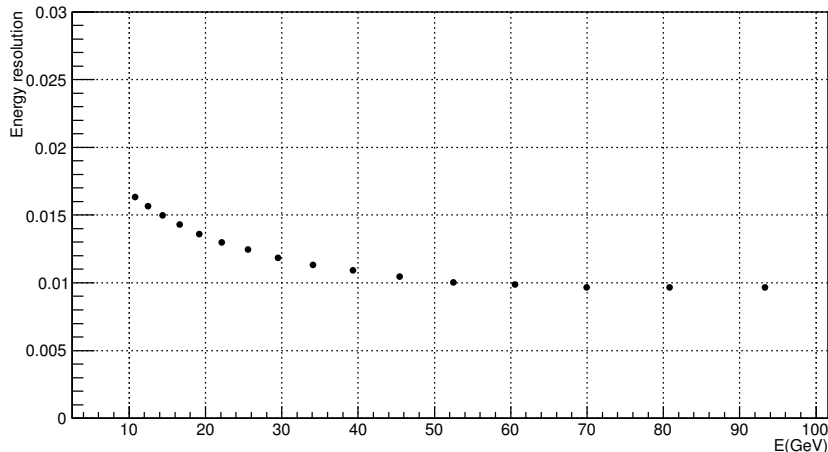
Fig. 3 Energy resolution from the Gaussian fit of MC electrons versus the deposited energy.

Table 1 Response of the MC Electrons and Protons to the Cuts

	Electrons		Protons	
Cut 0	2.57×10^7	100%	2.46×10^7	100%
Cut 1	1.39×10^7	54.1%	1.36×10^7	55.3%
Cut 2	6 426 500	46.2%	4 488 086	33.0%
Cut 3	3 613 769	56.2%	1 526 301	34.1%
Cut 4	1 488 657	41.2%	690 336	45.2%
Cut 5	1 419 133	95.9%	600 799	87.0%
Cut 6	1 305 428	92.1%	488 117	81.2%
Cut 7	1 249 552	96.1%(4.9%[1])	464 585	99.7%(1.9%)

Notes: The percentage shown in the table is the ratio of surviving particles after the corresponding cut compared to number of particles before this cut. [1] The value in brackets is the percentage of surviving events for all cuts.

Moreover, the pre-selection can depress the proton/electron ratio due to differing profiles of electromagnetic and hadronic showers. Table 1 shows the statistics of electrons and protons with these cuts based on the MC data. The bottom row shows the percentage of surviving particles compared to all particles. Consequently, the percentage of electrons is 3 times more than the percentage of protons, so the pre-selection has a background rejection power of 3 in the energy range 10–100 GeV.

4 DATA ANALYSIS

We now should consider how to increase electron rejection power and separate them from the severe background events. As the BGO calorimeter has a much larger radiation length than the nuclear interaction length, an electromagnetic shower is typically thinner and smaller than a hadronic shower with the same deposited energy.

In order to quantify the shape of a shower, we define some characteristic variables to describe its profile.

4.1 Define Characteristic Variables

The characteristic variables are mainly based on the differences between electromagnetic showers and hadronic showers in the BGO calorimeter. These variables can be studied using the MC data and are listed in the following. The whole detector is defined in a three-dimensional coordinate system. The orientation of positive z is vertical to the layer of BGO crystals pointing at the STK detector while x and y are defined in the plane that is parallel with the layer.

4.1.1 Energy deposition fraction in each layer

The variable $Efrac_i$ is simply defined as the energy deposited in the i-th layer divided by the total deposited

(a)

(b)
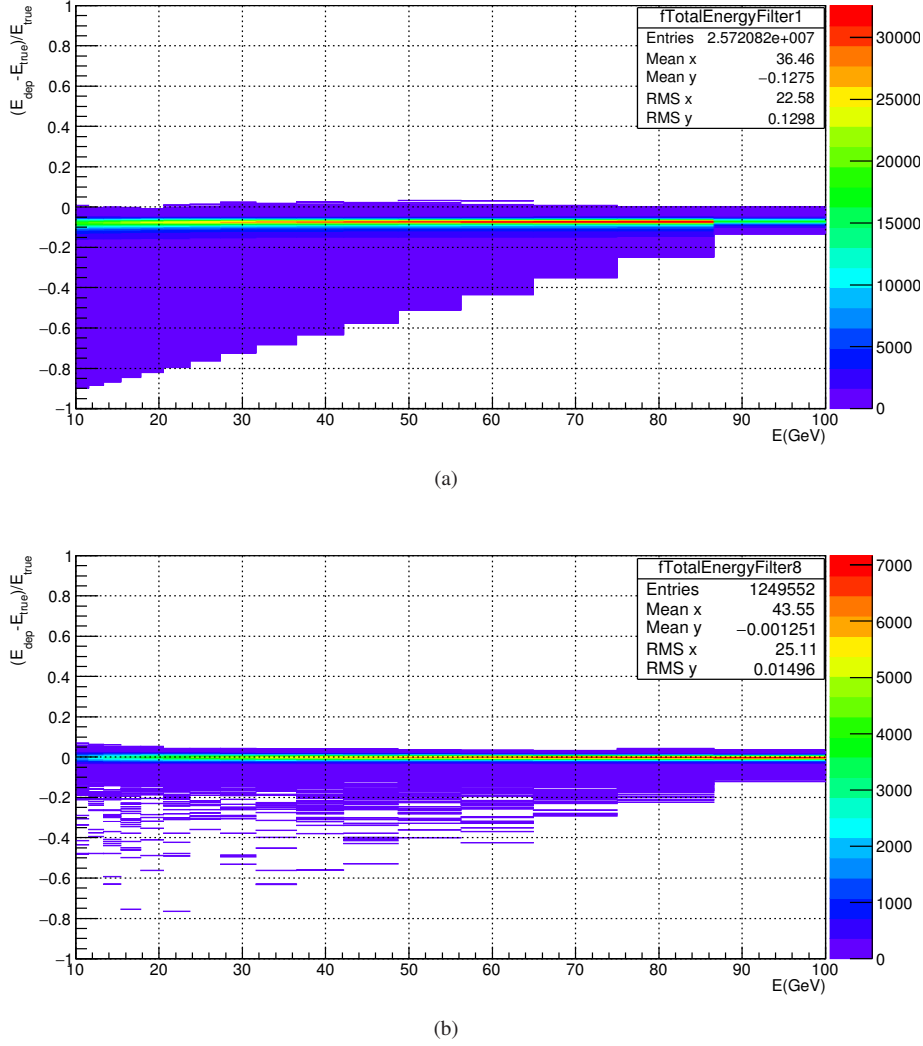
Fig. 4 Deposited energy of events. (a) Before all cuts. (b) After all cuts. The x axis is the deposited energy. The y axis is $(E_{\mathrm{dep}} - E_{\mathrm{true}})/E_{\mathrm{true}}$, where E_{dep} is the deposited energy. E_{true} is the true energy.

energy, where i is the layer number from 0 to 13. Since a hadronic shower tends to extend more deeply than an electromagnetic shower in the BGO calorimeter, the energy deposition fraction of the last few layers has strong discriminating power.

4.1.2 Root mean square

The *Root Mean Square* (*RMS*) of the deposited energies in the i-th layer is defined as

$$RMS_i = \sum_{j=0}^{22} E_{ij} \times (d_{ij} - d_i^{\mathrm{cog}})^2, \qquad (1)$$

where E_{ij} is the deposited energy of the j-th bar in the i-th layer, d_{ij} is the one-dimensional coordinate of the

j-th bar in i-th layer and d_i^{cog} is the one-dimensional coordinate of the i-th layer's center of gravity, defined as

$$d_i^{\mathrm{cog}} = \sum_{j=0}^{22} E_{ij} \times d_{ij}/E_i. \qquad (2)$$

Here E_i is the deposited energy of the i-th layer, so we can get a total of 14 *RMS* values for BGO. Total *RMS* used in Section 3.2 is simply the summation of all 14 *RMS* values. Based on the definition, the *RMS* indicates the lateral development of the shower. The electromagnetic shower is expected to be thinner than the hadronic shower when the incident particle has the same energy, resulting in a smaller *RMS* value.

4.1.3 RMS_r and RMS_l

RMS_r (Li et al. 2016) is the root mean square where the center of gravity of energy deposited is along the *Globtrack*. RMS_r is defined as

$$RMS_r = \sqrt{\sum_{i=1}^{N} E_i \times D_i^2 / E_{\text{total}}}, \qquad (3)$$

where N represents all the triggered BGO bars, E_{total} is the total deposited energy in BGO and D_i is the distance between the i-th BGO bar and the *Globtrack*.

RMS_l is similar to RMS_r except that D_i is the distance between the i-th BGO bar and the center of gravity, which is defined as

$$d_{\text{cog}} = \sum_{i=0}^{N} E_i \times d_i / E_{\text{total}}, \qquad (4)$$

where d_i is the three-dimensional coordinates of the i-th BGO bar. These two variables are evaluated for the whole BGO detector. RMS_r and RMS_l also indicate the lateral development of the shower. These two values are less dependent on the structure of the detector since they are only calculated using the actual track of the incident particle.

4.1.4 FValue

$FValue_i$ of the i-th layer is defined as

$$FValue_i = \sum_{j=0}^{22} E_{ij} \times (d_{ij} - d_i^{\text{cog}})^2 \times Efraction_i, \quad (5)$$

where $Efraction_i$ is the fraction representing the deposited energy of the i-th layer to the total deposited energy. Other parameters are the same as defined in Equation (1). $FValue_i$ is a combination of *RMS* and $Efrac_i$, which shows stronger discrimination power than these two variables alone (Chang et al. 2008).

4.1.5 L factor

L factor (Li et al. 2016) is a combination of RMS_l and *FValue*, defined as

$$L = \left(\frac{RMS_l}{60} \right)^{1-\alpha} \left(\frac{F_{13} + F_{14}}{0.1} \right)^{\alpha}, \qquad (6)$$

where F_{13} and F_{14} are $FValue_{13}$ and $FValue_{14}$, respectively; α is defined as

$$\alpha = 0.5 + \frac{1}{\pi} \arctan \left(5 \log_{10} \left(\frac{E}{50 \text{GeV}} \right) \right). \qquad (7)$$

L factor also considers the energy factor, which allows the variable to have steady performance over the whole energy range.

4.1.6 Chi-square

Longo & Sestili (1975) present an equation to describe the electromagnetic shower as

$$dE/dx = k_{\text{norm}} t^a e^{-bt}, \qquad (8)$$

where a and b are coefficients, and k_{norm} is defined as

$$k_{\text{norm}} = E b^{a+1} / \Gamma(a+1), \qquad (9)$$

where Γ is Euler's Gamma Function. We use this equation to fit the shower and compute the associated *Chi-square*. The *Chi-square* from the fit is a good estimator to distinguish electromagnetic showers from hadronic showers. Obviously, the *Chi-square* of an electromagnetic shower is smaller compared to the *Chi-square* of a hadronic shower with the same deposited energy.

4.1.7 tmax

The variable *tmax* is derived from Equation (8) after the fit to the shower. It is defined as

$$tmax = a/b. \qquad (10)$$

The variable *tmax* also represents the radiation length where the energy loss $\frac{dE}{dx}$ become the largest.

4.1.8 LongitudinalRMS

LongitudinalRMS is defined as

$$LongitudinalRMS = \sum_{i=0}^{14} E_i \times (dz_i - dAll_{\text{cog}}), \quad (11)$$

where dz_i is the z coordinate of the i-th layer and $dAll_{\text{cog}}$ is the center of gravity for all layers, which is calculated as

$$dAll_{\text{cog}} = \sum_{i=0}^{14} E_i \times d_i / E, \qquad (12)$$

where E is the total deposited energy. *LongitudinalRMS* indicates the longitudinal development of the shower, which turns out to have strong discrimination capability as shown in Figure 6.

HorizontalRMS used in Section 3.2 is defined in the same way as in Equation (11) except that d_i is the x or y coordinate of the i-th bar.

4.2 Variable Comparison

Since the discriminator will be trained using the MC data, it is crucial to check the consistency between the MC data and the on-orbit data. First of all, we use three variables, *FValue*, *L_factor* and *RMS*, to select electrons and protons preliminarily from on-orbit data for variable comparison. Then all variables in every energy bin of the MC and the on-orbit data are compared. All the variables defined in Section 4.1, which show good consistency in every energy bin, will be used in further analysis. An example of the comparison for variable (*RMS*) is displayed in Figure 5. The match between MC and on-orbit data is good based on a visual inspection in this energy bin. After passing the comparison test in all energy bins, this variable (*RMS*) will be used for training.

4.3 Machine Learning

We use the boosted decision trees (BDT) method from the toolkit TMVA (Hoecker et al. 2007) to train the network. A decision tree is a binary tree structured classifier, in which each internal node is labeled with an input feature. The boosting of a decision tree extends this concept from one tree to several trees which form a forest. Boosting is a way of enhancing the classification and regression performance (and increasing the stability with respect to statistical fluctuations in the training sample) of typically weak multi-variant analysis (MVA) methods by sequentially applying an MVA algorithm to reweighted (boosted) versions of the training data and then taking a weighted majority vote of the sequence of MVA algorithms thus produced. It was introduced into classification techniques in the early 1990s and in many cases this strategy has resulted in dramatic performance increases (Freund et al. 1999; Schapire 2003).

MC electrons and protons passing pre-selection are used for the machine learning. Half of the sample is used as training sample and the other half as test sample. The values of the variables for every MC event, forming a vector, will be input into the training algorithm. The algorithm will generate a discriminator which can distinguish electrons from protons. Every energy bin in range 10–100 GeV is trained separately. We show the training result for the 65.0–75.0 GeV energy bin as an example. The distributions of some well distinguished variables are shown in Figure 6.

The Receiver Operating Curves (ROCs) (Bradley 1997) in the 65.0–75.0 GeV energy bin that result from four BDT methods, namely BDT with adaptive boosting (BDT), BDT with gradient boosting (BDTG), BDT with bagging (BDTB), and BDT with decorrelation and adaptive boosting (BDTD) (Hoecker et al. 2007), are shown in Figure 7. The background rejection power of the BDT, the selected method, is about 0.9997 at signal selection efficiency of 0.9. Also, in a certain energy bin, the statistic corresponding to protons is mainly contributed by protons with higher energy (Zhang et al. 2015). Based on MC data, this so-called "hard suppression" has a suppression power of 9. With hard suppression of 9, the pre-selection cut's suppression power of 3 we mentioned before and the suppression power provided by BDT, this analysis produces a total of $3 \times 9 \div (1 - 0.9997) = 9.0 \times 10^4$ for background rejection power at 0.9 signal selection efficiency after pre-selection in the 65.0–75.0 GeV energy bin. Similar background rejection powers are obtained in other energy bins of 10–100 GeV.

Applying the discriminator to both training and testing samples, the BDT response distributions of both electrons and protons in the 65.0–75.0 GeV energy bin are obtained as shown in Figure 8. These distributions are considered as templates in the template fit to calculate the counts of electrons and protons, a topic which is not the content of this paper. Also, an overtraining check is performed to check if the distributions of training sample and test sample are consistent. The signal value of 0.194 and the background value of 0.295 indicate no overtraining is present.

5 DISCUSSION

There are also many other machine learning methods such as Support Vector Machine (SVM), Artificial Neutral Network (ANN) and K-Nearest Neighbor (KNN) available in the TMVA package. It is crucial to choose appropriate and strong distinguishing characteristic variables in order to have higher background rejection power. It is worth pointing out some of our BDT training settings: number of trees is 1000; max depth of a tree is 3; number of grid points in variable range used in finding optimal cut in node splitting is 20. We have altered some of these parameters in order to get better training results. Since the BDT result we obtained is almost perfect, the tuning did not yield much difference in this case.

In addition, the field of deep learning has developed very fast over the past several years years and has been used in many areas, including data analysis in physics (Baldi et al. 2014; Adam-Bourdarios et al. 2015;
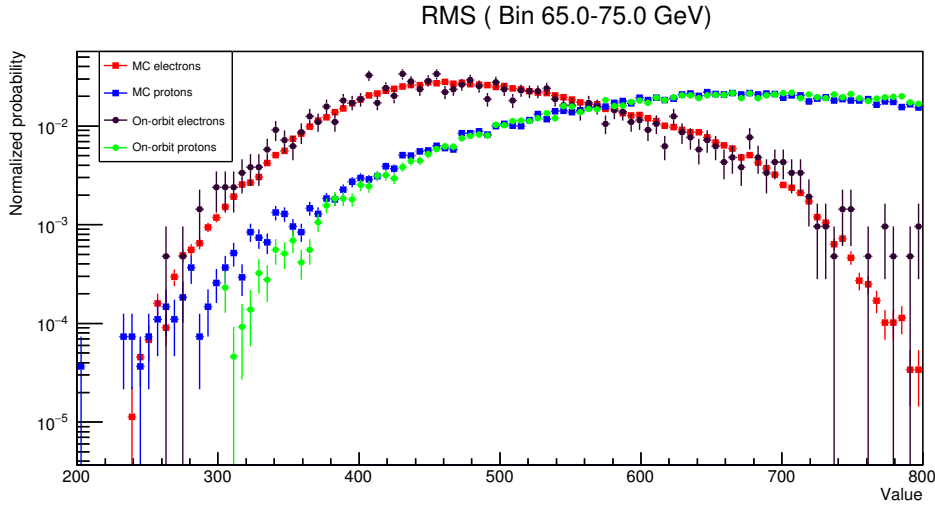
Fig. 5 The MC and on-orbit data comparison of variable *RMS*, only in the 65.0–75.0 GeV bin. *Green*: on-orbit proton; *Black*: on-orbit electron; *Blue*: MC proton; *Red*: MC electron.
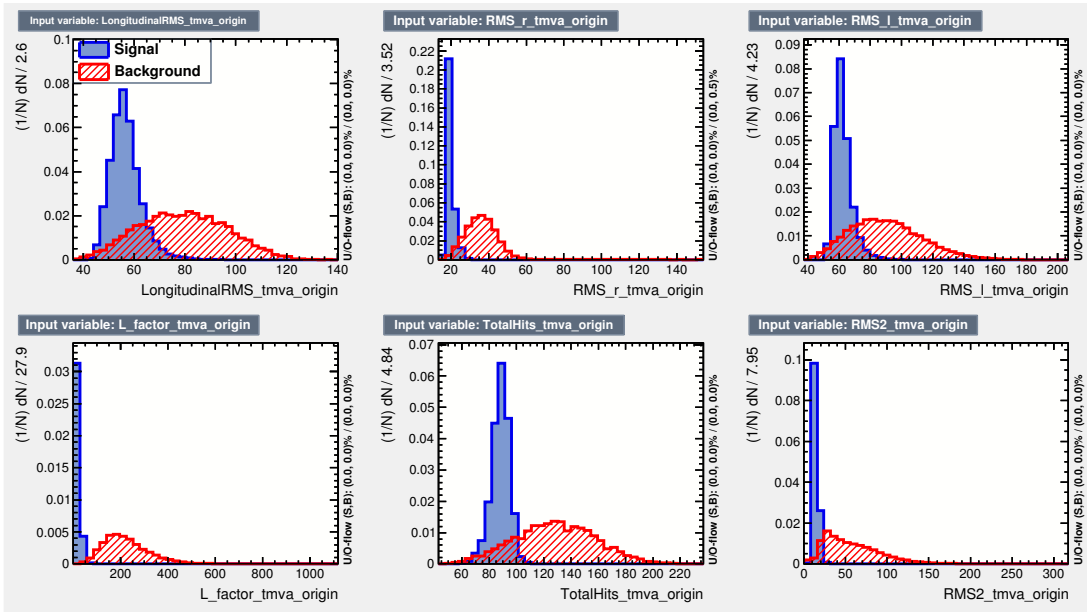

Fig. 6 The variable distributions of *LongitudinalRMS*, RMS_r, RMS_l, $Lfactor$, *TotalHits* and *RMS* for the MC electrons and protons in the 65.0–75.0 GeV energy bin. *Blue*: electrons; *red*: protons.

Sadowski et al. 2015). Deep learning and conventional machine learning both offer ways to train models and classify data. Unlike the conventional machine learning method, which demands the user extract relevant features from a classified object in order to train a model, the deep learning method skips the manual step of extracting features. Instead, one can feed an object directly into the deep learning algorithm which then predicts it. Deep learning is very powerful when dealing with complex objects, which is often the case in physics researches. We also expect an equally good, if not better, performance than conventional machine learning when deep learning is applied to DAMPE data for electron/proton separation.

We cannot use other detectors except for the BGO calorimeter to select pure electrons as a control sample or training sample from the on-orbit data for discrimina-
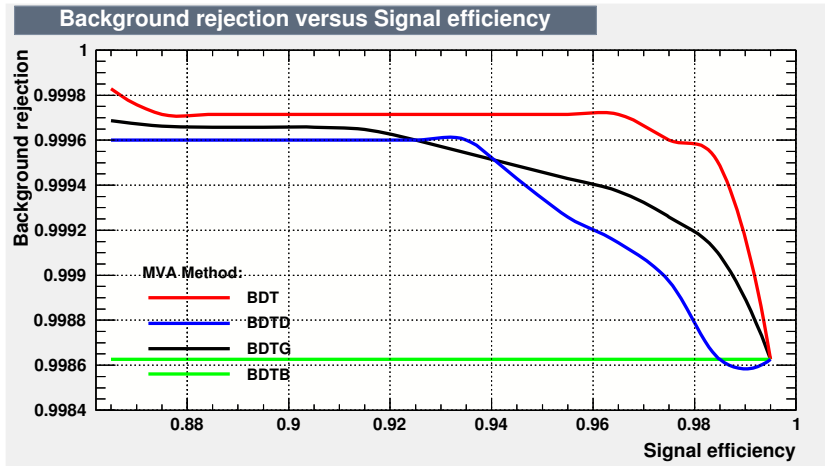
Fig. 7 ROC in the 65.0–75.0 GeV energy bin, where the x axis is the signal selection efficiency and the y axis is the background rejection power. The results of the four different BDT methods are shown for comparison.
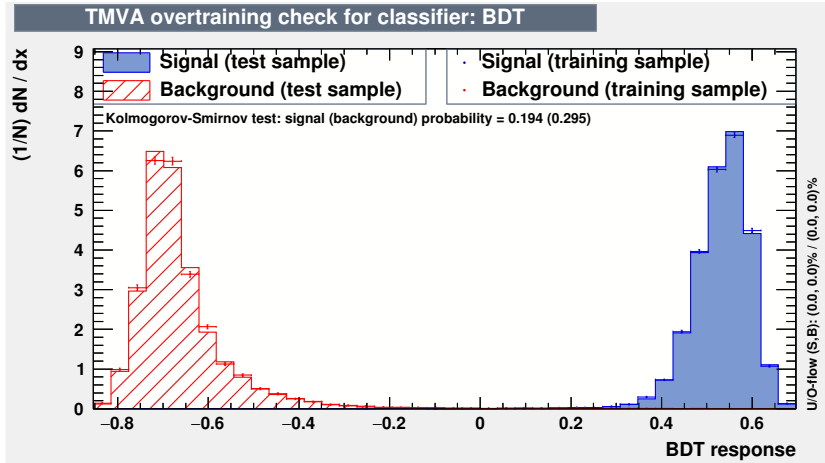
Fig. 8 The output distributions of the BDT method from both training and test samples in the 65.0–75.0 GeV energy bin. The Kolmogorov-Smirnov test is performed to check overtraining.

tor training like AMS-02. Our analysis relies strongly on consistency between data generated with the MC technique and real measurements.

6 CONCLUSIONS

Here the machine learning method is utilized to separate electrons from protons in DAMPE data which were acquired from 2016 January 1 to 2017 June 30 in the energy range 10–100 GeV. This machine learning method makes good use of all information available from the BGO calorimeter. It turns out that DAMPE can separate electrons from protons in the energy range 10–100 GeV with background rejection power of 9.0×10^4 using the machine learning method, which is higher compared to the background power rejection of $>10^4$ at a signal ef-

ficiency of 90% provided by the traditional cut-based method used in DAMPE Collaboration et al. (2017).

References

Abdollahi, S., Ackermann, M., Ajello, M., et al. 2017, Phys. Rev. D, 95, 082007

Adam-Bourdarios, C., Cowan, G., Germain, C., et al. 2015, in NIPS 2014 Workshop on High-energy Physics and Machine Learning, 19

Aguilar, M., Aisa, D., Alvino, A., et al. 2014, Physical Review Letters, 113, 121102

Aharonian, F., Akhperjanian, A. G., Barres de Almeida, U., et al. 2008, Physical Review Letters, 101, 261104

Baldi, P., Sadowski, P., & Whiteson, D. 2014, Nature Communications, 5, 4308

Bradley, A. P. 1997, Pattern recognition, 30, 1145

Chang, J., Adams, J. H., Ahn, H. S., et al. 2008, Nature, 456, 362

Chang, J., Ambrosi, G., An, Q., et al. 2017, Astroparticle Physics, 95, 6

DAMPE Collaboration, Ambrosi, G., An, Q., et al. 2017, Nature, 552, 63

Freund, Y., Schapire, R., & Abe, N. 1999, Journal-Japanese Society For Artificial Intelligence, 14, 1612

Hoecker, A., Speckmayer, P., Stelzer, J., et al. 2007, physics/0703039

Kobayashi, T., Komori, Y., Yoshida, K., & Nishimura, J. 2004, ApJ, 601, 340

Li, X., et al. 2016, Software Development of Dark Matter Particle Explorer and Researches on Dark Matter Indirect Detection, PhD thesis, Purple Mountain Observatory

Longo, E., & Sestili, I. 1975, Nuclear Instruments and Methods, 128, 283

Picozza, P., Galper, A. M., Castellini, G., et al. 2007, Astroparticle Physics, 27, 296

Pospelov, M., & Ritz, A. 2009, Physics Letters B, 671, 391

Roe, B. P., Yang, H.-J., Zhu, J., et al. 2005, Nuclear Instruments and Methods in Physics Research A, 543, 577

Sadowski, P., Collado, J., Whiteson, D., & Baldi, P. 2015, in NIPS 2014 Workshop on High-energy Physics and Machine Learning, 81

Schapire, R. E. 2003, in Nonlinear Estimation and Classification (Springer), 149

Yin, P.-F., Yu, Z.-H., Yuan, Q., & Bi, X.-J. 2013, Phys. Rev. D, 88, 023001

Zhang, Z., Zhang, Y., Dong, J., et al. 2015, Nuclear Instruments and Methods in Physics Research A, 780, 21

Zhang, Z., Wang, C., Dong, J., et al. 2016, Nuclear Instruments and Methods in Physics Research A, 836, 98