PROCEEDINGS OF SCIENCE

Neural Networks for Gamma Ray/Cosmic Ray Separation in Air Shower Observation with a Large Area Surface Scintillation Detector Array

S.Okukawa,[a,*] **K. Hara,**[b] **K. Hibino,**[c] **Y. Katayose,**[a] **K. Kawata,**[d] **M. Ohnishi,**[d] **T. Sako,**[d] **T.K. Sako,**[d] **A. Shiomi,**[b] **M. Shibata**[a] **and M. Takita**[d]

[a] *Faculty of Engineering, Yokohama National University, Yokohama 240-8501, Japan*
[b] *College of Industrial Technology, Nihon University, Narashino 275-8576, Japan*
[c] *Faculty of Engineering, Kanagawa University, Yokohama 221-8686, Japan*
[d] *Institute for Cosmic Ray Research, University of Tokyo, Kashiwa 277-8582, Japan*

E-mail: okukawa-sousuke-dt@ynu.jp

The Tibet ASγ experiment has been observing cosmic rays in the energy range from TeV to several tens of PeV using the Tibet-III air shower array since 1998. In 2014, they added the underground water Cherenkov muon detector (MD) to separate cosmic gamma rays from the background cosmic rays, and started hybrid observation using these two detectors. This study developed methods to separate gamma-ray-induced air showers and hadronic cosmic-ray-induced ones using the measured particle number density distribution to improve the sensitivity of cosmic gamma-ray measurement using the Tibet-III array data alone before the installation of the MD. We tested two kinds of approaches based on Neural Networks. The first method used feature values representing the shower particle spread from the measured particle number density distribution and the second method used image data. In order to compare the separation performance of the each method, we analyzed Monte Carlo air shower events of vertically incident direction with mono initial energy gamma rays and protons. A separation method with Multi-Layer Perceptron (MLP) based on multiple feature values has the AUC (Area Under the Curve) values of 0.748 for gamma-ray energy of 10 TeV and 0.845 for 100 TeV. A separation method with Convolutional Neural Network (CNN) using the image data has the AUC values of 0.781 for gamma-ray energy of 10 TeV and 0.901 for 100 TeV, which are about 5 % higher than those of MLP.

38th International Cosmic Ray Conference (ICRC2023)
26 July - 3 August, 2023
Nagoya, Japan

*Speaker

PoS(ICRC2023)786

1. Introduction

Tibet ASγ has been conducting the hybrid experiment since 2014. However, for approximately 15 years from 1998, gamma-ray observations were performed using the Tibet-III air shower array alone. It would be meaningful to reanalyze data accumulated over a long period without MD for sudden phenomena such as flares [1, 2] using a new method that increases the sensitivity of gamma-ray measurements without MD. In recent years, machine learning algorithms, particularly a technique known as deep learning, have made significant advancements and demonstrated high effectiveness in various tasks such as image recognition and voice recognition. Many of those deep learning techniques are primarily implemented using neural networks.
In this study, to improve the gamma-ray sensitivity of the Tibet-III measurement data before 2014, we investigated separation methods of gamma-ray-induced air showers (hereafter called gamma-showers) and proton-induced air showers (hereafter called proton-showers) using neural networks. In section 2, we briefly describe the configuration of the Tibet-III air shower array. In section 3, we applied the two type separation methods to particle density data measured with the Tibet-III. To compare their separation performance, we analyzed Monte Carlo data for single-energy, vertically incident gamma-showers and proton-showers(γ/p-separation). One method used specific feature values defined with the particle number density data. First, we tested a Multi-Layer perceptron neural network with multiple feature values. The other method involved image data converted from the air shower density and analyzing it using a CNN.

2. Tibet-III air shower array

The Tibet ASγ experiment is located at the Yangbajing Cosmic Ray Observatory in China (longitude 90°.522 E, latitude 30°.102 N, altitude 4300 m above sea level).

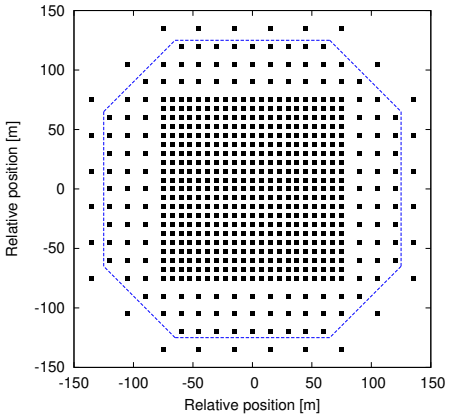


Figure 1: Top view of the Tibet air shower array (Tibet-III). The black squares represent surface 0.5 m^2 scintillation counters composing Tibet-III. The air shower array, consists of 597 scintillation counters. The enclosed area by the dashed blue line indicates the inner area used in the selection condition (2) of subsubsection 3.1.1.

Each detector comprises a 3 cm thick plastic scintillator under a 5 mm thick lead plate with a detection area of 0.5 m^2, and photomultiplier tubes (PMTs). Each detector is arranged at grid points

with a spacing of 7.5 m, covering a detection area of approximately 65,700 m^2, as shown in Figure 1. Until 2014 for approximately 15 years, observations were conducted only with Tibet-III. In 2014, a water-Cherenkov-type muon detector (MD) was installed 2 m below the Tibet-III array, allowing for hybrid observations with two kinds of detectors since 2015. In this study, we analyzed Monte Carlo data based on the assumption of the Tibet-III AS array consisting of 597 plastic scintillator detectors, as shown in Figure 1.

3. Separation methods for gamma-ray-induced air shower and cosmic-ray-induced air shower and those performances

We applied two neural network methods to particle number densities data measured by the Tibet-III and then compared the performance of two methods. Firstly, we defined multiple feature values to represent the characteristics of the lateral spread of shower particles from the shower core and their azimuthal spread of them and tried a Multi-Layer Perceptron (MLP) type neural network with multiple feature values. Next, we applied a Convolutional Neural Network (CNN) method to image data of particle number density distribution.
To simplify evaluation and comparison of the separation performances of two methods, we used vertically incident gamma-showers and proton-showers with mono initial energies generated by a Monte Carlo simulation.

3.1 Monte Carlo data of air showers measured with Tibet-III

First, we generated vertical gamma-ray air showers at 10 TeV and 100 TeV using the simulation code CORSIKA (version 7.6400) [3]. In this study, we used the EPOS LHC model [4] for the high-energy range and the FLUKA model [5] for the low-energy range. In Tibet-III, the total sum of detected particle densities on the ground is used to determine the primary energy. The total sum of particle densities differs between gamma-ray air showers and proton air showers, so for 10 TeV gamma rays, to achieve a similar number of detected particles, 21 TeV proton air showers are generated and separated (10 TeV-γ/p-separation). Similarly, for 100 TeV gamma-ray air showers, 165 TeV proton air showers are generated and separated (100 TeV-γ/p-separation).
Next, the response of each scintillation detector to the charged particles and gamma rays was calculated using Geant4.10.02. Finally, we applied the same analysis methods as those used in the observation experiment to the recorded particle densities and hit timings in each detector and reconstructed the energy and arrival direction of the primary particles. This procedure allowed the creation of simulated data which have the same format as the experimental data.

3.1.1 Air shower reconstruction

The measured energy loss in each scintillation detector was converted into the particle number density (ρ [/m^2]), and the energy of primary gamma rays and primary cosmic rays was calculated using the total value of ρ for all detectors ($\Sigma\rho$). The shower reconstruction used the hit timing to estimate the arrival direction [6]. In this study, we imposed the following selection condition to use well-reconstructed air shower events from the detected data: (1)At least four detectors hit within the time width of coincidence of 600 ns; (2) Detectors with $\rho \geq 0.6$ are at least four in the inner detectors of the array.Here, the inner detectors indicate the detectors inside the outermost detectors

of the AS array as shown Fig. 1; (3)Five or more of the top six detectors with the highest number of detected particles are contained in the inner detectors; (4)The location of the air-shower core determined by the analysis must be within a 50-mradius from the array center; (5)The residual error from the air shower front, which is defined as a reverse-conic type [7–9], must be less than 1.0 m. After analyzing MC data on the above selection condition, we obtained the $\Sigma\rho$ distributions of 10 TeV and 100 TeV gamma-ray-induced air showers. To separate gamma-ray- and hadronic cosmic-ray-induced air showers that have the same valuer of $\Sigma\rho$, we randomly thinned proton events. After matching the shape of proton distribution to gamma-ray distribution, we used these thinned proton events in the following analysis.

3.2 Separation with a MLP with multi-feature values

We defined the following six feature values representing the characteristics of the spread of air shower particles as inputs of MLP.

3.2.1 Definition of feature values

[Feature value (R)]
For the aim to represent the average lateral spread of the secondary particles, we tested the following equation,

$$R = \frac{\sum_i \rho_i \times r_i}{\sum_i \rho_i} \quad , \tag{1}$$

where ρ_i is the particle number density detected by the i-th detector, and r_i is the distance of the i-th detector from the shower core. Rs of gamma-ray-induced air showers tended to be smaller than proton-induced air showers for both energies.
[Feature value (N_j)]
We intended to extract multiple high-density particle regions in proton-induced air showers and tested the following equation,

$$N_j = \rho_j \times r_j \quad , \tag{2}$$

where ρ_j represents the j-th largest value of particle density within a shower, and r_j is the distance from the shower axis recorded by the detector corresponding to ρ_j. In the case of the 100 TeV gamma-ray-induced and 165 TeV proton-induced air showers, we used only detectors located at distances greater than 50 m from the shower core since the particle density near the shower core was high for both, and the difference between the two was not very significant. The N_j value for the gamma-ray-induced air shower tended to be smaller than that for the proton-induced air shower.
[Feature value($E_{\Sigma\rho}$,E_n)]
The fluctuations in the production point of π^0 mesons in the early stages of proton shower development may affect the density distribution of the shower at the ground, potentially resulting in an asymmetric distribution in the circumferential direction. So, we divided the region into eight circular segments with the shower core at the center, and defined the feature values using particle number density and the number of hit detectors in each region. First, we determined the reference direction of division for each air shower event using a weighted average $\bar{\varphi}$ of the angles. The weighted average $(\bar{x}, \bar{y})$ of the hit detector coordinates were calculated using the following equations,

$$\left(\bar{x} = \frac{\sum_i \rho_i \times (x_i - x_{\rm core})}{\sum_i \rho_i}, \bar{y} = \frac{\sum_i \rho_i \times (y_i - y_{\rm core})}{\sum_i \rho_i}\right) \quad .$$

The weighted average $\bar{\varphi}$ was then calculated as $\bar{\varphi} = \arctan(\frac{\bar{y}}{\bar{x}})$. Here, (x_i, y_i) represents the position of the i-th detector, and $(x_{\rm core}, y_{\rm core})$ represents the position of the reconstructed shower core obtained from the analysis. We then defined $E_{\Sigma\rho}$ and E_n as feature values in the following equations,

$$E_{\Sigma\rho} = \sqrt{\frac{1}{8}\sum_{j=1}^{8}\left\{(\sum\rho)_j - \frac{(\sum\rho)_{\rm all}}{8}\right\}^2} \quad , \tag{3}$$

$$E_n = \sqrt{\frac{1}{8}\sum_{j=1}^{8}\left(n_j - \frac{n_{\rm all}}{8}\right)^2} \quad , \tag{4}$$

where $(\sum\rho)_j$ represents the sum of particle number density in the j-th region out of the eight regions, and $(\sum\rho)_{\rm all}$ represents the sum of particle number density in all regions. n_j represents the number of hit detectors in the j-th region, and $n_{\rm all}$ represents the total number of hit detectors in all regions. To avoid the influence of core position fluctuations, especially near the shower core, we used only detectors located at distances greater than 30 m from the shower core for the 10 TeV-γ/p-separation. And, for the 100 TeV-γ/p-separation, we used only detectors located at a distance of at least 50 m from the shower core.

3.2.2 MLP model and Analysis procedure

This study tested an MLP using the Keras neural network library [10]. The MLP has an input dimension of 6, a hidden layer with 16 nodes, and an output dimension of 1. For the input, we used the six features defined in Section 3.2.1: R, N_1, N_2, N_3, $E_{\Sigma\rho}$, and E_n. The activation function used in the hidden layer is the sigmoid function. The sigmoid function is also used as the output function, the output value (P_γ value) is expected to be close to 0 for proton events and close to 1 for gamma-ray events if the training is performed correctly.

3.2.3 Separation performance with a MLP

The above method yielded the P_γ distributions shown in Figure 2a and 2b. To quantitatively evaluate the γ/p-separation performance of the MLP, we used the ROC curves and AUC values from the P_γ distributions, as shown in Figure 3. The AUC values calculated from these ROC curves are AUC = $0.748^{+0.010}_{-0.010}$ for 10 TeV-γ/p-separation and AUC = $0.845^{+0.008}_{-0.008}$ for 100 TeV-γ/p-separation. The separation performance improves as the energy increases.

3.3 Separation with a CNN with image-like data of particle density distribution

MLP method needs to extract feature values from particle density distribution in advance Therefore, finding the optimal feature values is essential. CNN has an advantage that automatically optimizes the parameters for extracting features from the image data in its computational process. We created 2D image-like data from the positions of each surface detector and the detected particle density data. Then we inputted those data into a Convolutional Neural Network (CNN) known for its image recognition capabilities and attempted γ/p-separation.

(a) 10 TeV gamma-rays, 21 TeV protons **(b)** 100 TeV gamma-rays, 165 TeV protons

Figure 2: The normalized distributions of P_γ of MLP. Figure (a) shows the result of the separation of 10 TeV gamma-ray-incuded air shower and 21 TeV proton-incuded air shower. Figure (b) shows 100 TeV gamma rays, and 165 TeV protons result. The red histograms show gamma-ray-induced air showers, and the green histograms show proton-induced air showers.

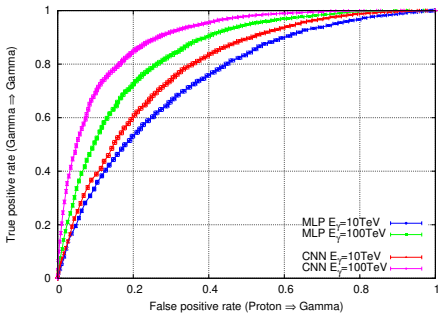


Figure 3: ROC curves for separation with MLP and CNN. The blue curve shows the ROC with MLP for 10 TeV-γ/p-separation. The green curve shows the ROC with MLP for 100 TeV-γ/p-separationton. The Red and magenta curves show ROCs with CNN for 10 TeV-γ/p-separation and 100 TeV-γ/p-separation.

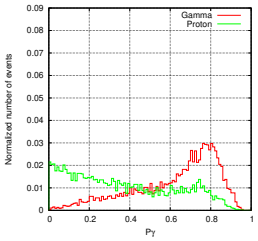
3.3.1 Design of image-like data

To convert the particle density data from 597 detectors arranged in a grid pattern into numerical array data in image format, the range of 307.5 m by 307.5 m was divided into a 41×41 grid. We stored the value of detected particle density in each element of the numerical array. However, we set the value to zero for array elements corresponding to positions without detectors. In addition, to utilize a CNN designed for RGB color image input, where each pixel consists of three components, we stored only one component of the particle density and zeros for the other two components. This procedure created numerical array data in image format with dimensions of $41 \times 41 \times 3$, representing the spread of particle density.

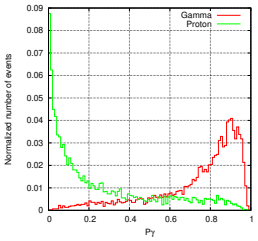
3.3.2 CNN model and Analysis procedure

In this study, we constructed a CNN architecture consisting of consecutive processes, referencing the Inception-v3 [11], and implemented it using Keras. The input data size changes after each process, and the final output is scalar value P_γ. To match the input data size for the Inception-v3 ($299 \times 299 \times 3$), we enlarged the image-like data size to ($246 \times 246 \times 3$), which is six times larger.

The same set of air shower events used for training and testing in the MLP approach were also used in the CNN approach. Similar to MLP, 10 % of the training data was set aside as validation



(a) 10 TeV gamma-rays, 21 TeV protons



(b) 100 TeV gamma-rays, 165 TeV protons

Figure 4: The normalized distributions of P_γ of CNN. Figure (a) shows the result of the separation of 10 TeV gamma-ray-incuded air shower and 21 TeV proton-incuded air shower. Figure (b) shows 100 TeV gamma rays, and 165 TeV protons result. The red histograms show gamma-ray-induced air showers, and the green histograms show proton-induced air showers.

data, and the validation loss was calculated at each epoch. The training results from the epoch with the lowest validation loss were used for testing.

3.3.3 Separation performance with CNN

As a result of analyzing the image-like data created in Section 3.3.1 using the CNN in Section 3.3.2, the P_γ distributions shown in Figure 4a and Figure 4b were obtained. To quantitatively evaluate the performance of γ/p-separation, similar to the case of MLP, evaluation was conducted using ROC curves and AUC. The AUC values for 10 TeV-γ/p-separation are $0.781^{+0.009}_{-0.009}$, and for 100 TeV-γ/p-separation, they are $0.901^{+0.006}_{-0.006}$. It can be observed that, similar to MLP, the selection performance improves with increasing energy. Additionally, the ROC curves of this CNN, as shown in Figure 3, have larger values compared to the MLP case in the previous section. As a result, the AUC values are also slightly higher, indicating a superior separation performance by a few percentage points.

4. Summary

This work developed two types of separation methods of gamma-ray showers and proton showers, using the detected particle density measured by the Tibet-III air shower array, and investigated the separation performance of the two methods. To compare the performance of each separation method, we generated Monte Carlo simulations of vertically incident gamma-ray showers with single energies of 10 TeV and 100 TeV, as well as proton showers with energies of 21 TeV and 165 TeV, which have similar detected particle densities to the gamma-ray showers. We applied the separation methods to these showers. First, we employed a Multi-Layer Perceptron (MLP) where we input the six feature values and performed the analysis. The separation performance of MLP was evaluated with AUC values of 0.748 for 10 TeV gamma-ray shower separation and 0.845 for 100 TeV gamma-ray shower separation. Next, we created image-like data representing the spread of particle density in air showers from the measured data and performed separation using Convolutional Neural

Network (CNN). The AUC values for CNN separation were 0.781 and 0.901, respectively, which were approximately 5 % higher than those of MLP. CNN has the advantage of automatically optimizing filter parameters for feature extraction, eliminating the need for manual data categorization or feature value caluculation. It was also found to have higher separation performance compared to MLP

References

[1] M. Amenomori et al., Astrophys. J., **598**, 242 (2003)
DOI: 10.1086/378350

[2] A.A.Abdo et al., Science, **331**, 739 (2010)
DOI: 10.1126/science.1199705

[3] Heck D. et al., Forschungszentrum Karlsruhe Report FZKA 6019 (1998)

[4] T. Pierog et al., Phys. Rev. C , **92**, 034906 (2015)
DOI: https://doi.org/10.1103/PhysRevC.92.034906

[5] Battistoni G.et al. , Annals of Nuclear Energy, **82**, 10 (2015)
DOI: https://doi.org/10.1016/j.anucene.2014.11.007

[6] Amenomori, M. et al., Phys. Rev. Lett., **123**, 051101 (2019)
DOI: https://doi.org/10.1103/PhysRevLett.123.051101

[7] M. Amenomori et al, Astrophys. J., **678**, 1165 (2008)
DOI: 10.1086/529514

[8] S. Kato et al., Exp. Astron., **52**, 85 (2021)
DOI: ttps://doi.org/10.1007/s10686-021-09796-8

[9] M.Amenomori, et al., Nucl. Instr. Meth. in Phys. Res. A, **288**, 619 (1990)
DOI: https://doi.org/10.1016/0168-9002(90)90161-X

[10] F. Chollet, others. (2015)
Keras. GitHub. Retrieved from https://github.com/fchollet/keras

[11] C.Szegedy et al., Proc. CVPR, pp.2818 (2016)
DOI: 10.1109/CVPR.2016.308