PROCEEDINGS OF SCIENCE

The components discrimination of CRs by LHAASO-KM2A from 1 PeV to 10 PeV with deep neural network

Yangxuan Diao,[a,*] **Hu Liu,**[a] **Fengrong Zhu,**[a] **Feng Zhang**[a] **and Xi Liu**[a]

[a] *Southwest Jiaotong University, School of Physical Science and Technology, No. 999, Xi'an Road, Chengdu, China*

E-mail: zhufr@home.swjtu.edu.cn

The origin of the cosmic rays(CRs) in the knee region has been a controversial issue, and how to distinguish their composition for ground-based detectors is still very challenging. The most common method currently used to identify the components of CRs is based on calculation and extraction of composition sensitive variables from the reconstructed information. This method can be constrained by the limited computing power so that deeper relationships between the features of the CR components cannot be extracted. In this research, an updated deep neural network (DNN) model is developed to distinguish between primary CR components from extensive air showers detected by KM2A arrays. Our preliminary result shows that the DNN model with AUC value 0.900 in Proton-Helium identification is more effective for identification of CR in the knee region than the traditional method with AUC value 0.750, and similar result is obtained for Fe-MgAlSi identification.

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

ICRC2023
38th International Cosmic Ray Conference
The Astroparticle Physics Conference

*Speaker

PoS(ICRC2023)546

1. Introduction

CRs are energetic particles from outer space that are emitted through an outburst or evolution at some source[1]. In order to understand the mechanisms associated with the propagation of CRs and their ultimate source, many CR observatories have been built in recent decades, mainly in space, on the ground, and underground (or underwater)[5]. Large High Altitude Air Shower Observatory (LHAASO) is one of the many ground-based detectors built in Haizishan, Daocheng County, Sichuan Province,China. There are three sub-arrays in LHAASO, One Square Kilometer Array (KM2A), Water Cherenkov Detector Array (WCDA), and Wide Field Cherenkov Telescope Array (WFCTA)[5], respectively. One of the scientific purpose of LHAASO is to find the origin of these CRs, as well as to measure and map the energy spectrum of each component of the knee region to investigate the physical origin mechanism of the knee[2, 3], the search for dark matter is also one of its goals[6].

In the previous work[6] of component discrimination, a graph neural network (GNN) and a simple convolutional neural network (CNN) was implemented to identify lighter component of CRs, but it was limited by the redundancy of the network and the information limitation of the feature map[6]. In this paper, we not only extract the ED, MD information from Chao Jin's feature map[6], but also take into account the information of lateral distribution of secondary particles. Then we use the most component-sensitive features to distinguish the light and heavy components of the CRs,which are the most contaminated part of CRs.

For the identification of the components of CRs, it is necessary to accurately select the target signal from the CR background[4]. To achieve this, it is necessary to clarify the correlation characteristics of each particle or the correlation between particles. At present, the most commonly used method is based on filtering multiple variables of the original CR. This method requires a lot of labor costs, and is also an identification method based on the experience obtained from various variable analysis[6]. However, the deeper relationship between variables can not be extracted, which will lose the corresponding feature information and the underlying information of its features. The deep learning method can be used to analyze the data volume that is 2 to 3 orders of magnitude higher than that of manual analysis[7]. This order of magnitude reading enables the underlying relationship between CR features to be extracted.

2. Simulation Data

2.1 Data selection

We uses EAS simulations generated by the CORSIKA software package, a software derived from the Monte Carlo-based simulation method. The hadronic interaction models used are QGSjet-II-04, EPOS-LHC at the high energy end, and EGS4 at the electromagnetic interaction.

In order to ensure the accuracy of the identification and the quality of the data, it is necessary to perform another filtering on each root file to ensure that the noise is reduced and the simulated data can be used for physical analysis[5]. In order to reduce the impact of punch-through effect, only the secondary particles triggered by the detector 40m away from the shower core are considered.

Table 1: KM2A full array simulates data filtering conditions.

Parameter	Filter criteria
NtrigE	>50
NpE2	>=32
NhitM	>10
NpW	>78
NfiltE	>40
Age	0.6-2.4
E	1PeV-10PeV

Usually, each energy segment of the primary component is divided into five bins, and the energy is distributed equally after taking the logarithm, for example, 6.0-6.2, 6.2-6.4, 6.4-6.6, 6.6-6.8, 6.2-6.4, 6.8-7.0.

To verify the quality of the filtered primary particles, their core resolution and angular resolution are calculated. The core resolution is defined as the difference between the reconstructed core position and the true core position at 68% of the events included, and the core resolution contains 1 sigma precision. As shown in Eqs. (1), where (x_T, y_T) is the real core position and (x_R, y_R) is the reconstructed core position.

The core resolution will be different at different zenith angles and energies, as the left image of Fig. 2 shows the performance of the primary particles at different zenith angles with different energies after two different filter criteria[5].

$$R_{68} = \sqrt{(x_T - x_R)^2 + (y_T - y_R)^2} \tag{1}$$

Similarly, the angular resolution is defined as the difference between the reconstructed zenith

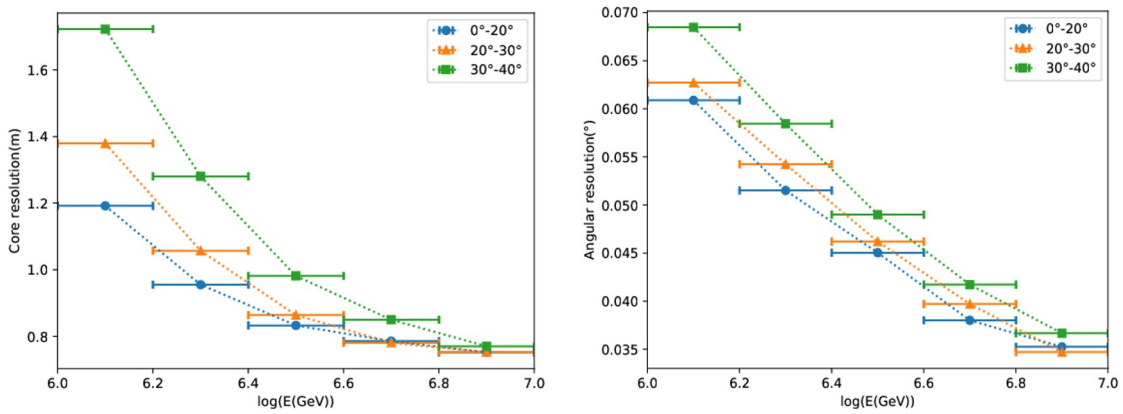

Figure 1: The left panel is the variations of the core resolution of the five primary components at different energies after averaging. The green, orange and blue dots are the results of full-array core resolution under the our filter criteria conditions. The right panel is the angular resolution of the five primary components was averaged at different energies. The green, orange and blue dots are the results of the full-array angular resolution under our filter criteria conditions.

angle and the original true zenith angle at the included 68% events. As shown in Eqs. (2), where A_T is the original zenith angle and A_R is the reconstructed zenith angle.

$$A_{68} = |A_T - A_R| \tag{2}$$

The angular resolution can be different at different zenith angles and energies, as the left image of Fig. 2 shows the performance of the primary particles at different zenith angles with different energies after two different filter criteria[5].In the left image of Fig. 2, the angular resolution and core resolution also show the full improvement of data quality under the new filter criteria conditions.

2.2 Component-sensitive feature selection

After filtering the data quality, in order to input the data into our neural network model, it is also necessary to perform component sensitivity selecting for each feature related to the primary particle shower, and then select the feature that exhibits the most component sensitivity for each primary particle.

For the frontal slope, which is the angle between the frontal plane of the primary component shower and the horizontal plane, the distribution of the angle varies on each primary component. The features about total number of triggered electromagnetic particles and total number of triggered muons associated with the baseline model are also taken into account in this statistic, and their analyses are taken as logarithms with a base of 10, and it can be found that both have distinct differences between the different primary components[2]. Finally it is found that feature of the slope of the secondary gamma produced by each shower is more sensitive to the primary component than feature the slope of the secondary muons produced by each shower. On the data from 1 PeV to 10 PeV, the features that are relatively sensitive to the components were selected as the frontal slope, total number of triggered electromagnetic particles, total number of triggered muons, and the slope of the secondary gamma produced by each shower.

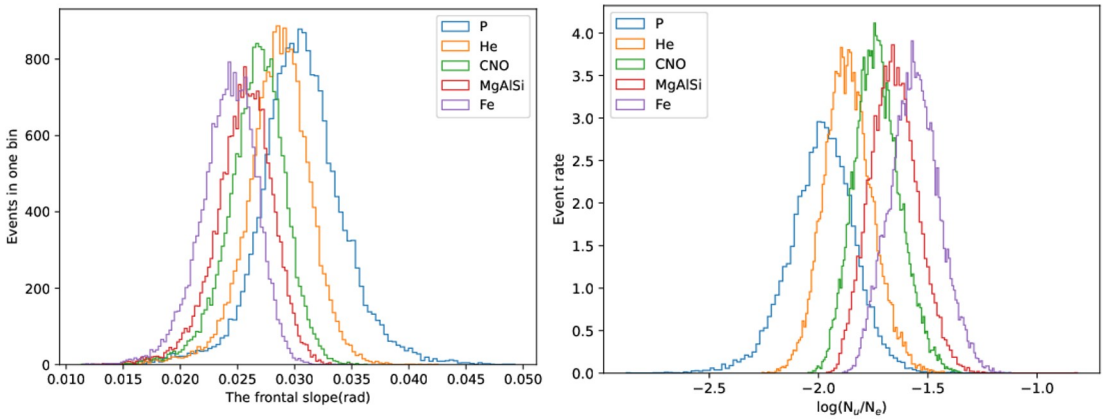

Figure 2: The left image is the features to be selected with certain compositional sensitivity to the five primary components, the horizontal coordinate is the feature and the vertical coordinate is the number of events. The feature in the figure is the distribution of the frontal slope for each shower. The right image is the distribution of B-values at different original components.

3. Model Identification and Evaluation

$$\mathrm{B} = \log_{10}\left(\frac{N_u}{N_e}\right) \tag{3}$$

3.1 Model setup

Before investigating our DNN method, the traditional methods in component discrimination need to be analyzed and used as a baseline for the comparison of our deep learning method. The traditional analysis method relies on empirically set sensitive parameters for composition identification, which mainly involve the number of electromagnetic particles and the number of muons triggered and recorded from the ED and MD detectors in KM2A. As shown in the right image of Fig. 2. The baseline model is calculated as follows Eqs. (3), N_u is the number of muons, N_e is the number of electromagnetic particles.

In DNN model, various characteristic variables enter the input layer and perform regression approximation between characteristic variables in the hidden layer, and finally obtain the corresponding regression or classification results in the output layer[10].

In the multi-layer perceptron applied on the LHAASO experiment, as in Fig. 3, the hidden layer uses relu continuous activation functions(see Eqs. (4)) to stimulate the input data on the neuron[10]. After reaching the output end, the corresponding loss function used to reduce the gap between the real value and the predicted value, and then back propagation(BP) carried out[8].

Table 2: Data input to the model at different energy bins.

Rec_E	Input datasets
$(10^6 \sim 10^{6.2})$GeV	14000
$(10^{6.2} \sim 10^{6.5})$GeV	12800
$(10^{6.5} \sim 10^{7})$GeV	9000

Table 3: Division of datasets under different tasks.

	P		I	
	Signal	Background	Signal	Background
Train	17900	17900	17900	17900
Evaluation	17900	17900	17900	17900
Predict	1000	990	1000	980

Table 4: Comparison of AUC values of different models in different energy bins

	DNN(task P)	DNN(task I)	Baseline(task P)	Baseline(task I)
$(10^6 \sim 10^{6.2})$GeV	0.841	0.842	0.681	0.516
$(10^{6.2} \sim 10^{6.5})$GeV	0.836	0.858	0.541	0.928
$(10^{6.5} \sim 10^{7})$GeV	0.900	0.889	0.750	0.542

$$y_{\text{relu}} = \begin{cases} 1 & (a \geq 0) \\ 0 & (a < \mathbf{0}) \end{cases} \tag{4}$$

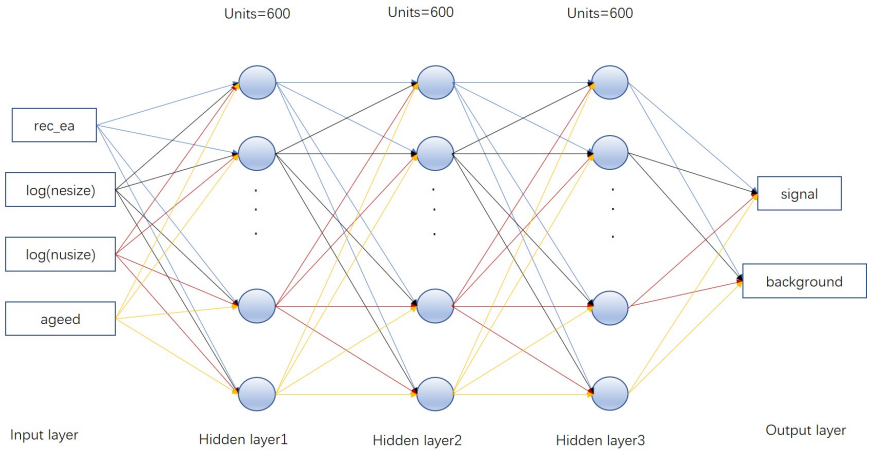

Figure 3: The structure of the DNN model applied to the LHAASO-KM2A simulation data. Three hidden layers with 600 neurons are used, a relu activation function is used from the input layer to the middle hidden layer, BN (batch normalization) is performed before each input layer, and an Adam adaptive optimizer is used for training. A sigmoid activation function is used in the input layer to output the range of binary classification. The input is the filtered 4 features, and the output is the identified signal with background particles.

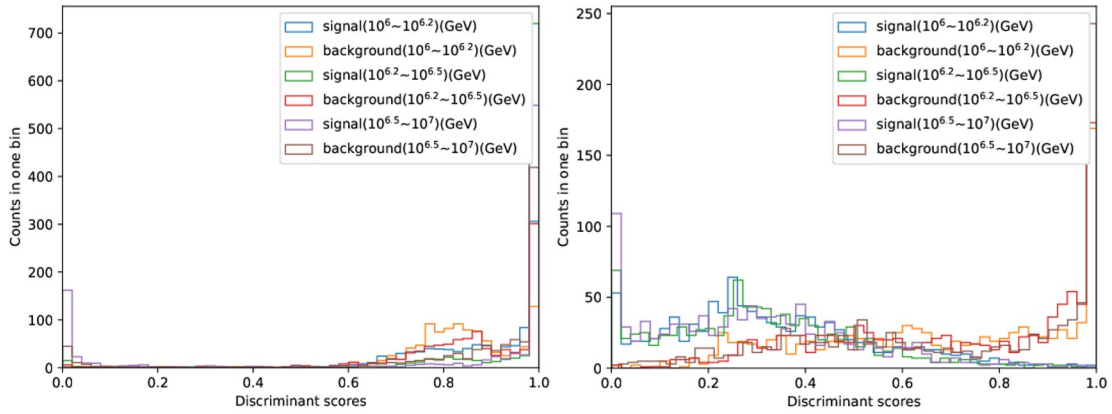

Figure 4: Distribution of discriminant scores of DNN model output on task P and task I, left for task P and right for task I.

3.2 Test on KM2A

After filtering the data, we have generated 94,936 events of the five CR components. The dataset can then be partitioned to substitute into the DNN model. The data set is divided into a training set (Train), a validation set (Val) and a prediction set (Predict).

The training set is used for continuous nonlinear fitting of the model; the validation set (test set) is the analysis and evaluation the error of the output predicted values and the true values after one training iterations, so as to prevent the model from overfitting[11]. the prediction set is finally the data set for comparing the results of the output after model convergence, which is used for the physical analysis and the evaluation of the effects produced by the model classification.

This paper focuses on the identification of lighter and heavier components of CRs, because it is difficult to distinguish protons from light components, and similar to iron nuclei, these particles with similar mass numbers in the CRs, which are highly contaminated with each other[12], as can be found qualitatively in the right image of the left image of Fig. 2. we divided the work into task P and task I before substituting the model. In the task P, protons are used as signal and helium as background; in the task I, iron are used as signal and magnesium aluminum silicon as background. In the evaluation, we used the H3a model[2] to normalize the prediction set to achieve a discriminate effect under the proportion of CR real current flux.

In order to maximize the model effect, the data are divided as evenly as possible on three energy bin, and the division results are shown in Table 2.Then the best DNN model is trained on each energy bin, and finally the DNN with three hidden layers is found to be the best model after debugging.

4. results

The corresponding discrimination score is the output of the model. Fig. 4 shows the distribution of the discriminant scores of the DNN output, which shows that the background is mostly distinguished in the task I, which also shows that the task I is slightly better than the task P in the ability to distinguish the background.

To reflect the contamination from the background, the true positive rate (TPR) needs to be compared with the false positive rate (FPR). Where the TPR is the fraction of correctly identified signals

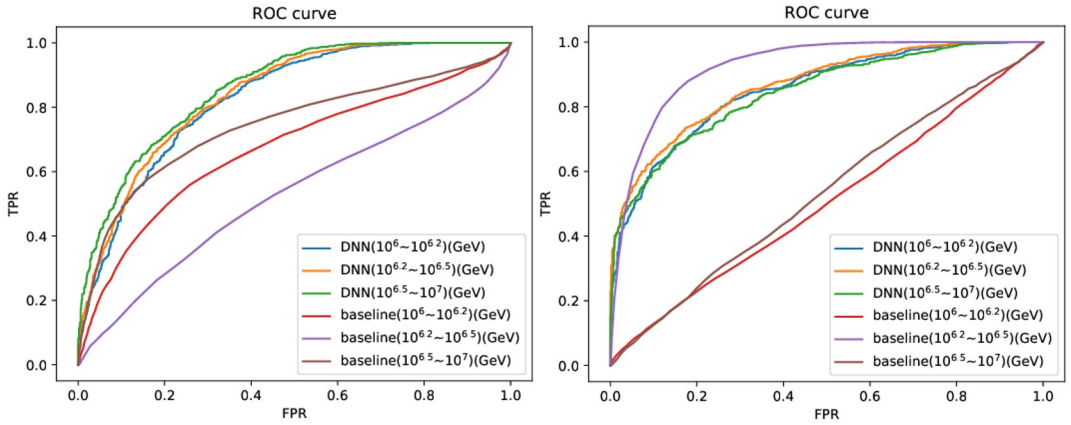

Figure 5: The left figure shows the ROC comparison of the task P at different energy intervals for different models. The blue, orange and green lines in the figure are the results made by the DNN model; the red, purple and soil lines are the results made by the baseline model. The right figure is the ROC curve of task I.

over the total number of signals identified by the model, the FPR is the fraction of misidentified signals over the total number of backgrounds identified by the model.

Generally, ROC curve and AUC value are used to analyze the ability of model identification for binary classification [6]. As shown in Fig. 6, respectively, in the task P, the baseline model generally performs worse than the DNN model, see the right image of Fig. 2, which is related to the serious contamination of protons by the helium nucleus. In the task I, the baseline model performs well at $(10^{6.2}\sim10^{6.5})$ GeV, but the iron nuclei in the rest of the energy range are basically indistinguishable from the background, which also suggests that the identification of iron nuclei in the knee region may be mainly contributed by the identification ability of the intermediate energy range. We also uses AUC(area under roc curve) to quantify the specific identification ability of the model[11], details can be obtained from Table 4.

5. conclusion

At the time when deep learning is widely used in various fields, we created a DNN neural network model based on the LHAASO-KM2A full-array experiment to test its classification ability. By extracting the shower information of primary particles, we found four features that are relatively sensitive to the components, that is, on the basis of integrating ED and MD information, two kinds of secondary particle information are added: frontal slope and gamma slope. The information is integrated into the model, and finally we find that the DNN model shows better identification in the knee region and outperforms the traditional physical baseline model. In addition, we also analyzed the causes of contamination in two groups: proton and helium nuclei, iron nuclei and magnesium-aluminum-silica, and found that the former were more contaminated than the latter, but this gap was greatly reduced by the DNN model.

References

[1] B. Commercon, A. Marcowith, Cosmic-ray propagation in the bi-stable interstellar medium: I.conditions for cosmic-ray trapping, Astronomy and Astrophysics 622 (2019).

[2] Gaisser, T.K., Stanev, T. Tilav, S. Cosmic ray energy spectrum from measurements of air showers. Front. Phys. 8, 748–758 (2013).

[3] H. Sato, Cosmic-Ray Energy Spectrum and High-Energy Particles in Supernova, Progress of Theoretical Physics 30, 804 (1963).

[4] B. D. Piazzoli, Chapter 4 cosmic-ray physics, Chinese Physics C 46, 85 (2022).

[5] X.-H. Ma, Chapter1 LHAASO instruments and detector technology, Chinese Physics C 46, 1 (2022).

[6] C. Jin, S. zhan Chen, Classifying cosmic- ray proton and light groups in LHAASO-KM2A experiment with graph neural network, Chinese Physics C 44, 133 (2020).

[7] X. Zuo, Design and performances of prototype muon detectors of LHAASO-KM2A, Nuclear Instruments and Methods in Physics Research, Section A. Accelerators, Spectrometers, Detectors and Associated Equipment 789, 143 (2015).

[8] G. Hinton, Teh, Unsupervised discovery of nonlinear structure using contrastive backpropagation, Cognitive science 30, 725 (2006).

[9] F. EG, Extended tanh-function method and its applications to nonlinear equations, Physics Letters, A 277, 212 (2000).

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86, 2278 (1998).

[11] H. Kwon, Flexion: A quantitative metric for flexibility in dnn accelerators, IEEE Computer Architecture Letters 20 (2021).

[12] S. Mollerach and E. Roulet, Progress in high-energy CR physics, Progress in Particle and Nuclear Physics 98, 85 (2018).