Identification of proton and gamma in LHAASO-KM2A simulation data with deep learning algorithms

F. Zhang[a], F. R. Zhu[a], S. M. Liu[a], Y. C. Hao[b], C. He[c], J. Hou[c], Z. Li[d] . for the LHAASO Collaboration
a School of Physical Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
b Graduate School of Tangshan, Southwest Jiaotong University, Tangshan 063000, China
c School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
d Key Laboratory of Particle Astrophysics, Institute of High Energy Physics, CAS, Beijing, China

ONLINE ICRC 2021 THE ASTROPARTICLE PHYSICS CONFERENCE Berlin | Germany 37th International Cosmic Ray Conference 12–23 July 2021

Abstract:

Identification of proton and gamma plays an essential role in ultra-high energy gamma-ray astronomy with LHAASO-KM2A. In this work, two neural networks (deep neural networks (DNN) and graph neural networks (GNN)) are applied to distinguish proton and gamma in the LHAASO-KM2A simulation data. The receiver operating characteristic (ROC) curves are used to evaluate the quality of the model. Both KM2A-DNN and KM2A-GNN models give higher Area Under Curve (AUC) scores than the traditional baseline model.

1. Introduction

The Large High Altitude Air Shower Observatory (LHAASO) is a multi-component ground detector array. It is located at a high altitude (4410 m a.s.l.) in Daocheng, Sichuan Province, China. LHAASO consists of a kilometer array with an area of 1.3 km² (KM2A), 78,000 m² Water Cherenkov detector array (WCDA) and 18 Wide Field air Cherenkov Telescopes array (WFCTA). The LHAASO-KM2A occupies the major area and is composed of two sub-arrays, 5195 electromagnetic particle detectors (ED) and 1188 underground water Cherenkov tanks for muon detectors (MD). In order to discriminate showers with the core located within the central area from the outside ones, the ED detectors are divided into two parts, the central part with 4901 detectors and an out-skirt ring with 294 detectors .Fig 1 shows the layout of these datectors.

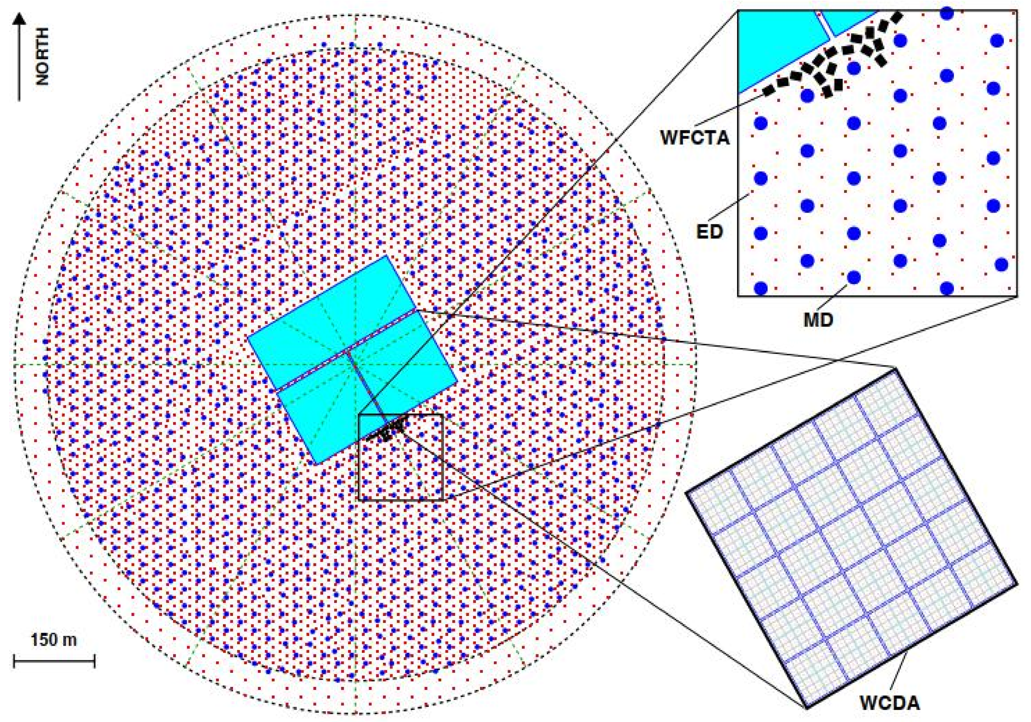

Fig. 1: Layout of the LHAASO experiment. The insets show the details of one pond of the WCDA, and the EDs (red points) and MDs (blue points) of the KM2A. The WFCTA located at the edge of the WCDA is also shown.

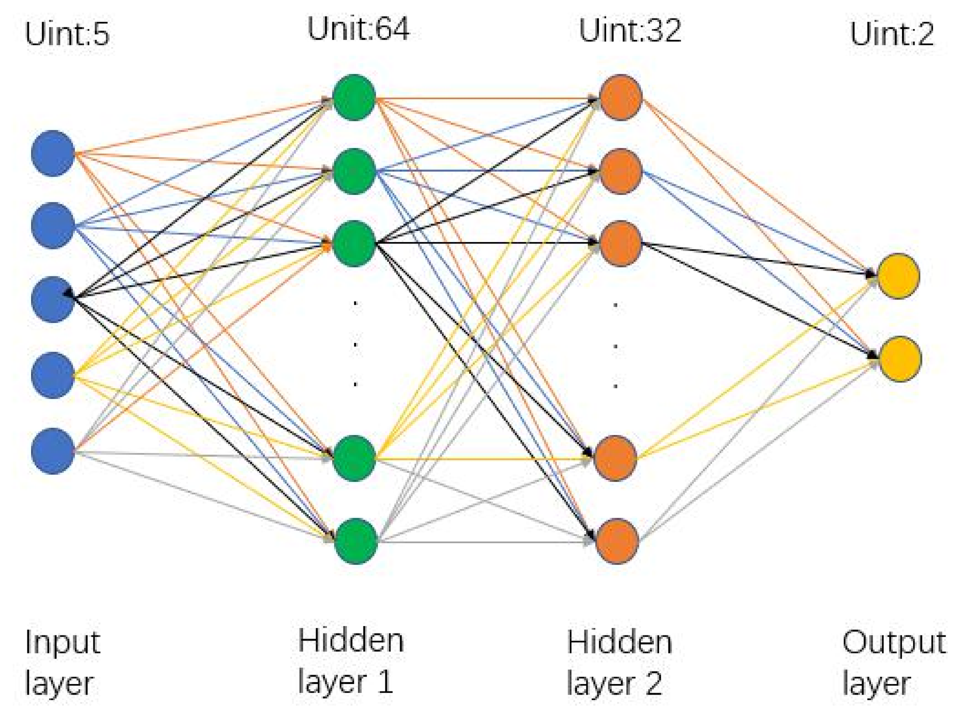

Fig2. KM2A-DNN model structure diagram

DNN model refers to fully connected neuronal structures that do not contain convolution units or temporal associations. Fig 2 shows the structure of KM2A-DNN model.

GNN is a connectionist model for learning graphs that contain a large number of connections. As information propagates between nodes of the graph, GNN can capture graph independence. Unlike a standard neural network, GNN maintains a state that represents information from an artificially specified depth.

2. Datasets

In this work, the Monte Carlo simulation is used to generate the event data for training[1, 2]. The KM2A detector simulation is performed based on framework of the Geant4 package (v4.10.00)[3–5] for KM2A half an array.

Filtering out noise, reconstruction and events selection is done accrding to this paper[6]. After the screening, 1996894 proton events and 1025774 gamma events survived.

The partition of KM2A-DNN model and KM2A-GNN model data sets is shown in Table 1 and Table 2, respectively.

Table 1: The number of the proton and gamma for each data sets of KM2A-DNN model

Data set	10^{12}-10^{13}eV		10^{13}-10^{14}eV		10^{14}-10^{15}eV		10^{15}-10^{16}eV	
	Proton	Gamma	Proton	Gamma	Proton	Gamma	Proton	Gamma
Train	60000	20000	300000	100000	50000	30000	6000	6000
Test	116871	17512	953683	79937	157755	68515	15123	10335

Table 2: The number of the proton and gamma for each data sets of KM2A-GNN model

Data set	10^{12}-10^{13}eV		10^{13}-10^{14}eV		10^{14}-10^{15}eV		10^{15}-10^{16}eV	
	Proton	Gamma	Proton	Gamma	Proton	Gamma	Proton	Gamma
Train	688659	183061	572017	431521	122202	92187	14946	11274
Validation	196759	52303	163434	123291	34915	26339	4270	3221
Test	98381	26151	81717	61646	17458	13169	2136	1611

3. Baseline , DNN and GNN model

There is less muon in the shower induced by gamma than by proton[7]. The ratio of collected signals from the MD and ED is a component sensitive estimator and is adopted widely on the CR experiments, Nu/Ne was formulated as the physics-base method named baseline model. The distribution of Nu/Ne with respect to gamma and proton components at different energy levels is shown in Fig 3.

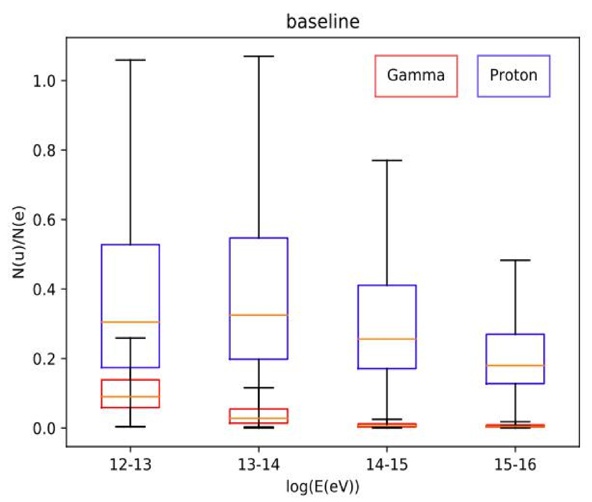

Fig 3.The distribution of Nu/Ne with respect to gamma and proton components at different energy levels

In the training process of KM2A-DNN and KM2A-GNN models, Learning rate of the model is set to 0.001. Focal loss of cross entropy loss function is used as loss function. AdamOptimizer is used as the optimizer.

4. Results

ROC curve is used to evaluate the model performance. In order to obtain the physical significance of ROC curve in this paper, Fig 4 is taken as an example to give a detailed introduction. In this figure, when accuracy of gamma equal to 0.85, the contamination of proton is 0.27 and 0.17 for baseline model and KM2A-DNN model respectively.

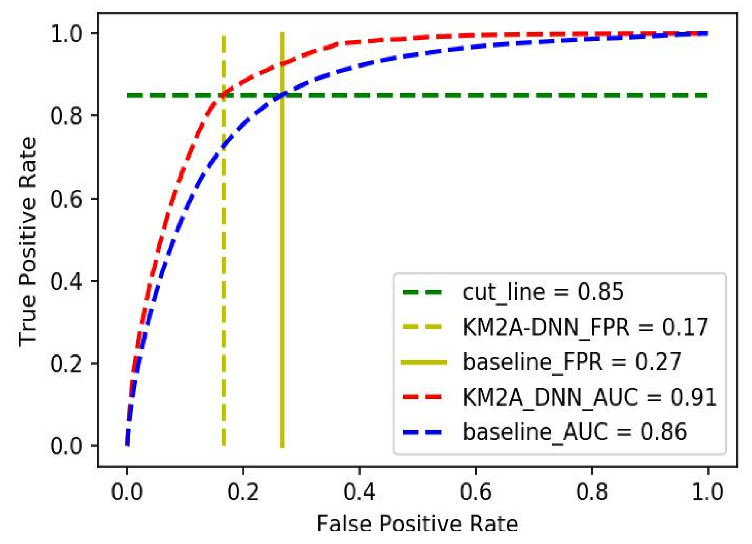

Fig 4. Physical significance of ROC curve

The conclusion that KM2A-DNN model is better than baseline model in discriminating ability can be obtained. Area Under Curve (AUC) score is a parameter used to quantitatively compare good models with bad ones. The AUC score of each figure is indicated in the lower right corner. The comparison of ROC curve between baseline model and KM2A-DNN model and KM2A-GNN model are shown in in Fig 5 and Fig 6, respectively.

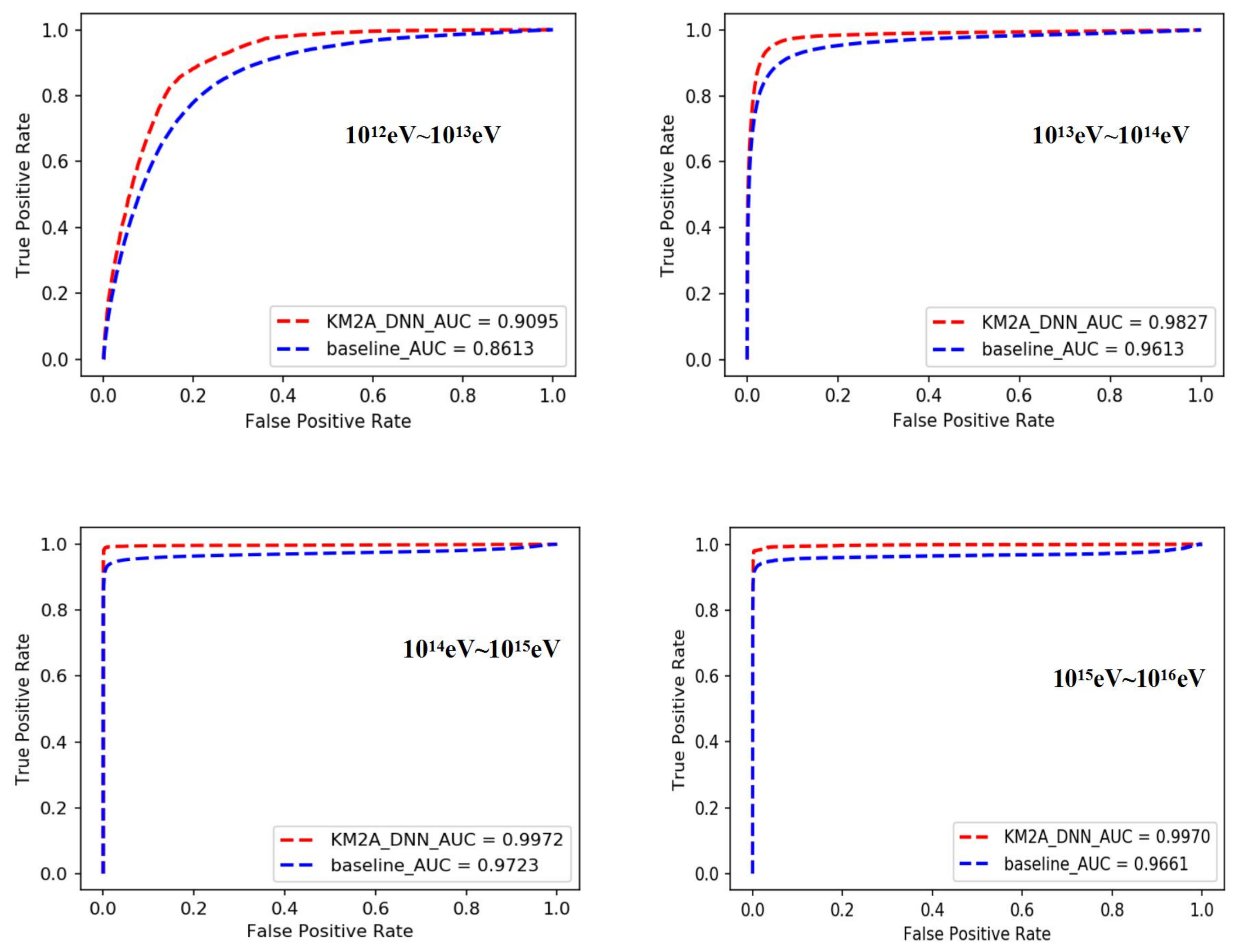

Fig 5. Comparison of ROC curve between baseline model and KM2A-DNN model in different energy ranges, the AUC score is indicated in the lower right corner.

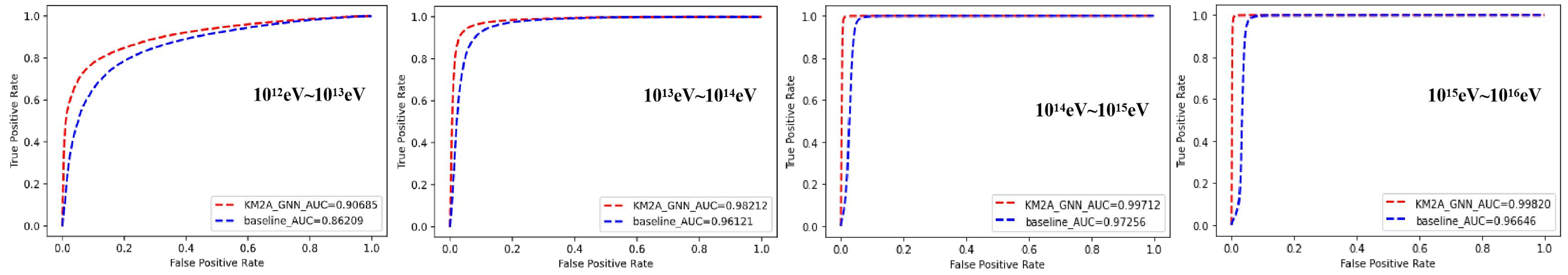

Fig 6. Comparison of ROC curve between baseline model and KM2A-GNN model in different energy ranges, the AUC score is indicated in the lower right corner.

Acknowledgement

This work is supported in China by the Fundamental Research Funds for the Central Universities (grant numbers 2682020CX77, 2682020CX73, 2682020CX74). It is also supported by the Science and Technology Department of Sichuan Province (grant numbers 2021YFSY0031, 2020YFSY0016), and by NSFC (grant number 11947404), and by National Key R&D program of China (grant number 2018YFA0404201).

Reference

[1] CORSIKA: a Monte Carlocode to simulate extensive air showers, 1998
[2] Heck, D.,et al., 1998, Forschungszentrum Karlsruhe Report No. FZKA 6019
[3] S.Z. Chen, et al., Nuclear Electronics and Detection Technology,37: 1101-1105 (2017)
[4] S.Z. Chen, et al., proceeding of 36th ICRC, (2019)
[5] Agostinelli S,et al., Nuclear Instruments and Methods in Physics Research Section A, 506:250 (2003)
[6] F.Aharonian, et al.,Chinese Physics C,2021,45(02):522-534
[7] Apel W D,et al., Physical Review Letters, 107(17): 171104 (2011)