PROCEEDINGS OF SCIENCE

A machine learning approach to identify the air shower cores for the GRAPES-3 experiment

M. Chakraborty,[a,*] S. Ahmad,[c] A. Chandra,[c] S.R. Dugad,[a] U.D. Goswami,[l] S.K. Gupta,[a] B. Hariharan,[a] Y. Hayashi,[b] P. Jagadeesan,[a] A. Jain,[a] P. Jain,[d] S. Kawakami,[b] H. Kojima,[e] S. Mahapatra,[i] P.K. Mohanty,[a] R. Moharana,[j] Y. Muraki,[g] P.K. Nayak,[a] T. Nonaka,[h] A. Oshima,[e] S. Paul,[a] B.P. Pant,[j] D. Pattanaik,[a,i] G.S. Pradhan,[k] M. Rameez,[a] K. Ramesh,[a] L.V. Reddy,[a] R. Sahoo,[k] R. Scaria,[k] S. Shibata,[e] K. Tanaka,[f] F. Varsi[d] and M. Zuberi[a]

[a] *Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India*
[b] *Graduate School of Science, Osaka City University, Osaka 558-8585, Japan*
[c] *Aligarh Muslim University, Aligarh 202002, India*
[d] *Indian Institute of Technology Kanpur, Kanpur 208016, India*
[e] *College of Engineering, Chubu University, Kasugai, Aichi 487-8501, Japan*
[f] *Graduate School of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan*
[g] *Institute for Space-Earth Environmental Research, Nagoya University, Nagoya 464-8601, Japan*
[h] *Institute for Cosmic Ray Research, Tokyo University, Kashiwa, Chiba 277-8582, Japan*
[i] *Utkal University, Bhubaneshwar 751004, India*
[j] *Indian Institute of Technology Jodhpur, Jodhpur 342037, India*
[k] *Indian Institute of Technology Indore, Indore 453552, India*
[l] *Dibrugarh University, Dibrugarh 786004, India*

E-mail: pkm@tifr.res.in

The GRAPES-3 experiment located in Ooty consists of a dense array of 400 plastic scintillator detectors spread over an area of 25,000 m^2 and a large area (560 m^2) tracking muon telescope. Everyday, the array records about 3 million showers in the energy range of 1 TeV – 10 PeV induced by the interaction of primary cosmic rays in the atmosphere. These showers are reconstructed in order to find several shower parameters such as shower core, size, and age. High-energy showers landing far away from the array often trigger the array and are found to have their reconstructed cores within the array even though their true cores lie outside, due to reconstruction of partial information. These showers contaminate and lead to an inaccurate measurement of energy spectrum and composition. Such showers are removed by applying quality cuts on various shower parameters, manually as well as with machine learning approach. This work describes the improvements achieved in removal of such contaminated showers with the help of machine learning.

*Speaker

PoS(DLCP2022)001

1. Introduction

The long standing mystery of the sources, acceleration and propagation mechanism of cosmic rays (CRs) can be probed by various methods, like the study of energy spectrum and composition or the energy dependence of anisotropy in CR flux. These require a better precision in energy measurement of the showers. However, due to several factors, often the reconstructed energy is improper as studied using simulated data and can hamper the precise measurement of energy spectrum. Figure 1 shows the reconstructed energy spectrum deviating from the input energy spectrum generated from simulation after passing the simulated CRs through the GRAPES-3 detector response. This is caused due to the presence of high energy showers landing hundreds of metres away from the array which also trigger the array, and such showers often have their reconstructed cores inside the array due to mis-reconstruction. This work demonstrates methods to remove such contaminated showers by using simple cut based and machine learning strategies, and the subsequent improvements in energy spectrum measurements achieved by using these methods.

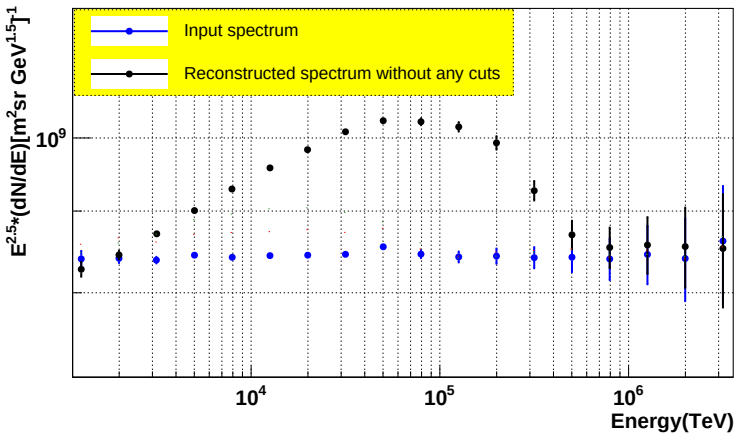

Figure 1: Deviation observed in the reconstructed energy spectrum from simulation due to the presence of contaminated showers

2. The GRAPES-3 experiment

The GRAPES-3 (Gamma Ray Astronomy at PeV EnergieS Phase-3) experiment is located at Ooty (11.4°N, 76.7°E, 2200 m a.s.l.), India. The extensive air shower (EAS) array consists of 400 plastic scintillator detectors. Each of these detectors records particle densities and relative arrival times of particles in an EAS [1]. The scintillator array covers an area of 25000 m^2. GRAPES-3 uses two level trigger, level-0 trigger is a simple three fold coincidence of three consecutive line of detectors in 100 ns time window and level-1 trigger requires at least 10 detectors hit in 1 µs time window. The scintillator detectors are arranged in hexagonal geometry, with an inter-detector separation of 8 m. GRAPES-3 also has a 560 m^2 tracking muon detector consists of 3712 proportional counters (PRCs) [2]. A schematic of GRAPES-3 with the fiducial area ($\sim$ 14560 m^2) is shown in Figure 2.

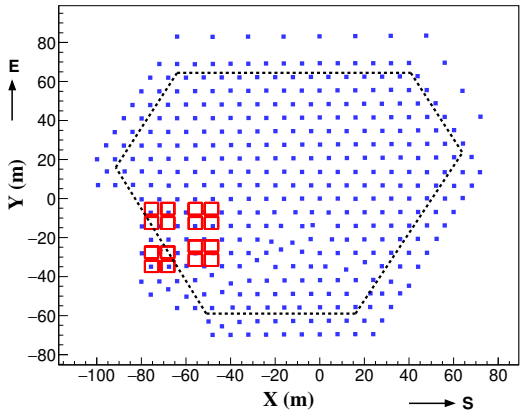

Figure 2: Schematic of GRAPES-3 air shower array, the area enclosed by the black dashed line marks the fiducial area

3. Simulation

The Monte Carlo simulation of EAS is carried out using CORSIKA (version 7.69) package. 5×10^8 EAS with proton primaries and spectral index -2.5, having energies within 1 TeV – 10 PeV were simulated using the hadronic interaction generators SIBYLL and FLUKA. The showers are thrown within a circular area of radius 'r' such that the trigger fraction is less than 1% outside this circle. The value of this radius 'r' is energy dependent. The typical value ranges within 100-300 m for showers with energies within 100 TeV, 300-500 m for showers having energies of a few hundred TeV, and within 500-800 m for showers with energies above a PeV. The simulated showers are then passed through the detector response of GRAPES-3 array calculated using GEANT4, followed by shower reconstruction similar to data in order to obtain the shower parameters. The particle densities recorded in the detectors are fitted by the well known Nishimura-Kamata-Greisen (NKG) function to obtain the shower parameters as described in [3], namely, the shower size (N_e), age (s) and the shower core (X_c, Y_c), as shown in Figure 3.

$$\rho_i = \frac{N_e}{2\pi r_m^2} \frac{\Gamma(4.5-s)}{\Gamma(s)\Gamma(4.5-2s)} \left(\frac{r_i}{r_m}\right) \left(1+\frac{r_i}{r_m}\right)^{s-4.5} \tag{1}$$

where ρ_i is the expected particle density at i^{th} detector, r_i is the distance of i^{th} detector from shower core. r_m is the Moliere radius which is 103 m at Ooty.

The NKG fit is performed by negative log-likelihood minimisation. For an expected density of ρ_i, the probability p_i of detecting n_i particles in the i^{th} detector can be expressed by a Poisson distribution,

$$p_i = \frac{(\rho_i A_i cos\theta)^{n_i}}{n_i!} e^{-\rho_i A_i cos\theta} \tag{2}$$

where A_i is the detector area which is 1 m^2. Thus, the likelihood (L) for all the detectors is calculated as,

$$L = \prod_i p_i \tag{3}$$

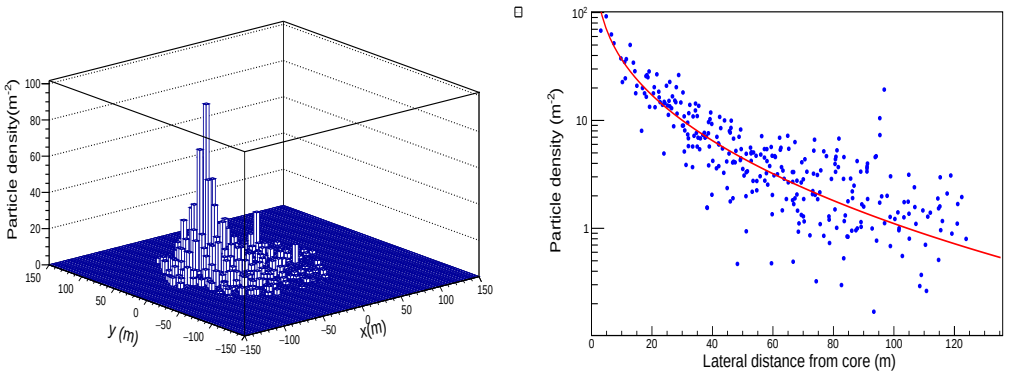

Figure 3: The shower profile (left) and its lateral distribution fitted by NKG function (right) of a shower recorded by GRAPES-3. The core is around the region where the highest particle densities are recorded.

Calculation of products is computationally intensive. Hence, the product is converted to a summation by taking its natural logarithm.

$$ln(L) = \sum_i (n_i \times ln(\rho_i cos\theta) - \rho_i cos\theta - ln(n_i!)) \tag{4}$$

The negative log likelihood, $-ln(L)$, is minimised in order to fit and obtain the shower parameters as described in [4].

4. Analysis

After reconstructing the simulated showers, the following shower selection criteria are applied:

- Showers with successful NKG fit
- Zenith angle within 25°
- Showers with reconstructed cores within the fiducial area.

Several showers having their true cores outside are observed to have their reconstructed cores inside. Such showers are seen to have a lower shower size due to partial information of the shower available with the array as shown in Figure 4. The contamination is seen to increase with increase in the true energy of the shower as shown in Figure 5. This is due to increase in the lateral spread of air showers with energy.

The energy reconstruction of showers is dependent on the shower size and zenith, hence these contaminated showers are viewed as low energy showers by the array and leads to contamination in lower shower size bins. The observable parameter for data is shower size, hence we select variables from logarithmic shower size bins of bin width $10^{0.2}$. Data quality cuts are devised on these variables to remove such contaminated showers. For the rest of this document, the well-reconstructed showers having both their true and reconstructed cores within the fiducial area will be referred to as "signal" and the mis-reconstructed showers described above are referred as "background". The variables selected are as follows:

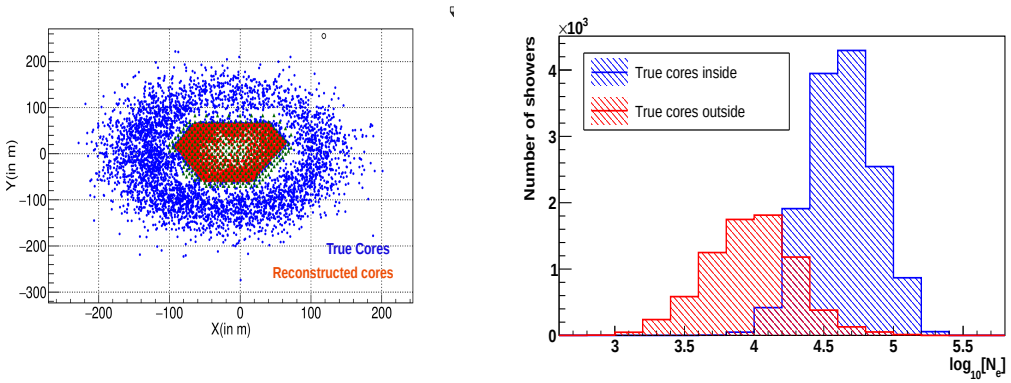

Figure 4: Showers with true cores outside but mis-reconstructed cores inside having energies within 100 TeV $\leq E <$ 158 TeV (left), and the shower size distributions for showers with the same energy are shown for well and mis-reconstructed showers (right). The mis-reconstructed showers have lower shower size.

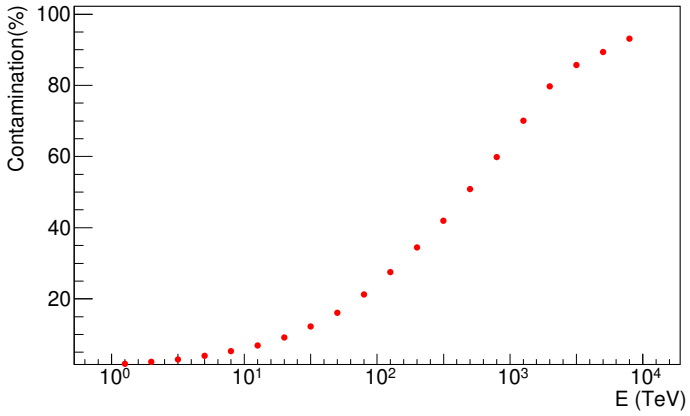

Figure 5: Contamination increases with increase in true energy

- **PSumRatio**: It is defined as the ratio between the total particle density collected outside the fiducial area and the total particle density collected within the fiducial area. This ratio is higher when the true cores lie outside the array
- **LnNKGP**: The mis-reconstructed showers have a poor NKG reconstruction. The best value of the log-likelihood function obtained from NKG fit as described in equation 4, denoted by LnNKGP, is shown in Figure 6.
- **Age**: Age is mostly very high in the case of improper core reconstruction as shown in Figure 6 as the NKG curves get flatter.
- **Age error**: Error in age parameter.
- **LnCErr**: Error in constant parameter of NKG fit.

Manual cuts are applied chronologically on the above variables. Cuts are chosen at the value where the signal significance, given by $S/\sqrt{S+B}$, attains its maximum. Here, S and B are signal and

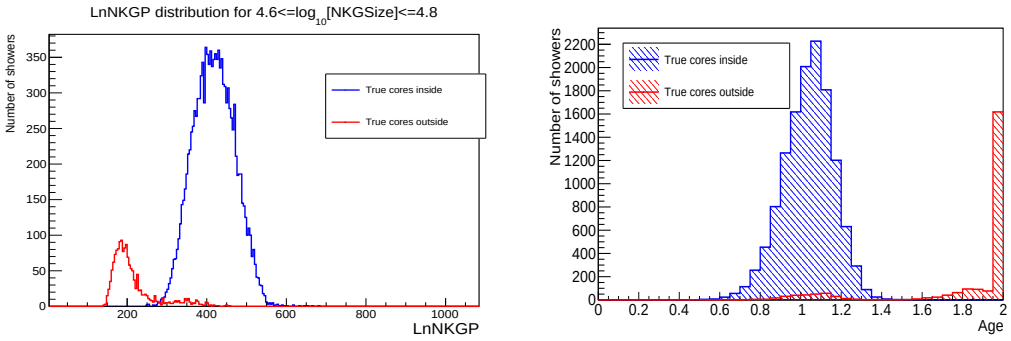

Figure 6: LnNKGP (left) and Age (right) distributions for shower size $4.6 \leq log_{10}[N_e] < 4.8$, both the variables show a clear separation for signal and background events. The degree of separation is lesser in lower shower size bins and increases with increase in shower size.

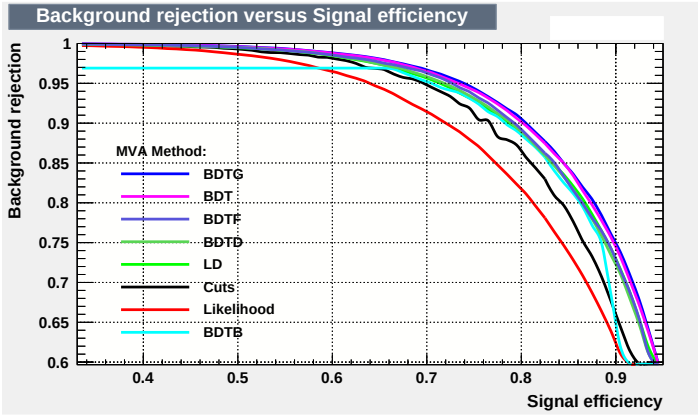

Figure 7: The performance of different ML methods to decide the ML method to be used.

background events respectively. Further conditions are also applied in selecting these cut values as described in [5].

Handling multiple variables and determining quality cuts on each of them for every shower size bin is tedious. Hence we use machine learning (ML) in order to device cuts. The simulated data set is divided into two equal parts for training and testing and using the default parameters, the various ML methods were tested as shown in Figure 7 using the TMVA package of ROOT [6]. Boosted decision tree with gradient boost is seen to have the highest area under ROC, and was used for further analysis. The hyperparameters and input variables were optimized to ensure that the integral of receiver-operator-characteristic curves (ROC) for train dataset is close to 1. In order to avoid possible overtraining, we ensure that the Kolmogorov-Smirnov (KS) probability between train and test distributions of the BDT output variable is above 0.05. Additionally, the area under ROC of the two statistically independent test and train samples were observed to be consistent with each other as shown in Figure 8. The BDT output variable also shows a clear separation and cuts were applied on the BDT output variable in order to remove contaminated showers.

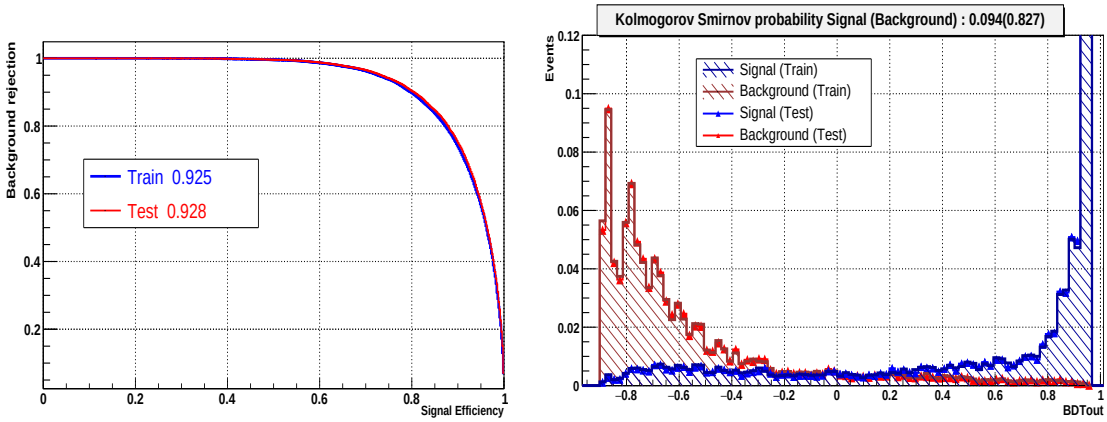

Figure 8: ROC (left) and KS test (right) results for $4.0 \leq log_{10}[N_e] < 4.2$. The BDT output variable is seen to show a clear separation for "signal" and "background".

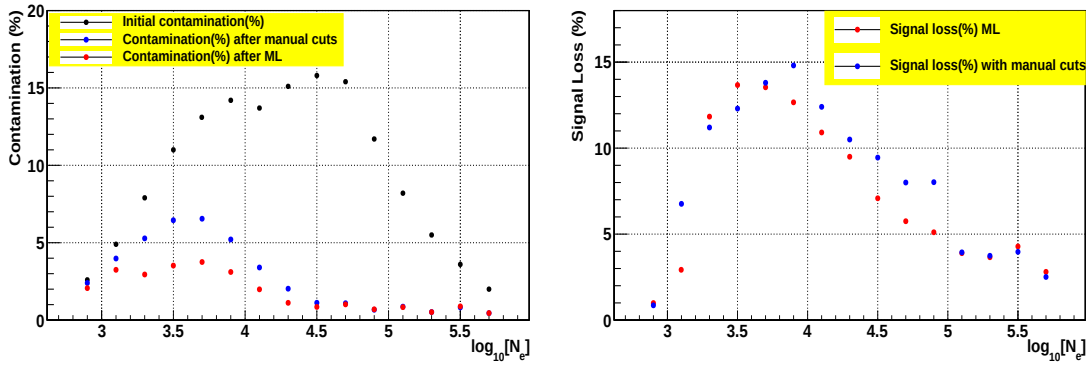

Figure 9: Reduction in contamination and the corresponding signal loss as a function of $log_{10}[N_e]$

5. Results and discussions

The resulting contamination and signal loss after applying manual cuts and ML as a function of shower size have been shown in Figure 9. It can be seen that ML achieves a better reduction in contamination than manual cut results with a comparable signal loss. A maximum contamination of 18% was brought down to 8% using manual cuts and within 5% by the use of ML without any significant increase in signal loss. This also improves energy spectrum measurements as shown in Figure 10. The deviation from expected spectrum is further reduced by the use of ML.

Thus, ML is an effective tool to remove contaminated showers and can help in improving energy dependent studies performed by GRAPES-3.

Acknowledgements We thank D.B. Arjunan, A.S. Bosco, V. Jeyakumar, S. Kingston, K. Manjunath, S. Murugapandian, S. Pandurangan, B. Rajesh, V. Santhoshkumar, M.S. Shareef, C. Shobana, and R. Sureshkumar for their efforts in maintaining the GRAPES-3 experiment. The first author would also like to thank Rahul Tiwary (TIFR, Mumbai) for his help with TMVA.

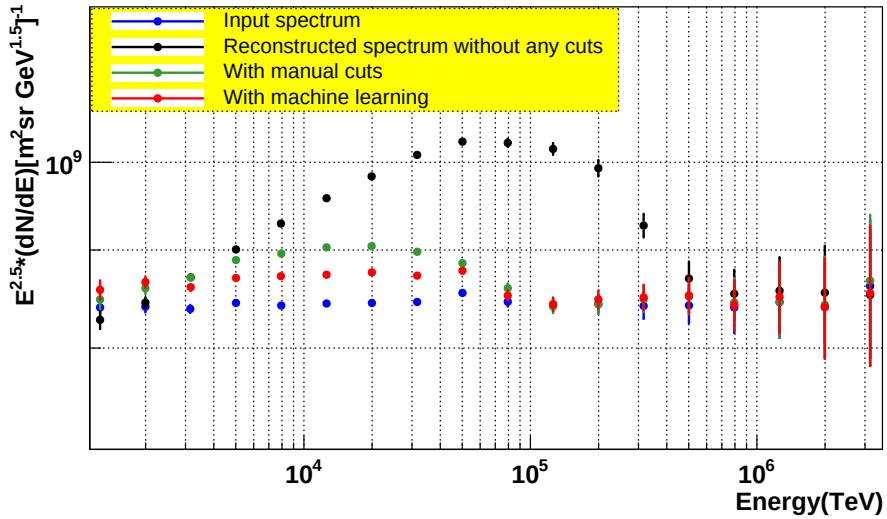

Figure 10: The reduction in bias present in energy spectrum measurements due to contaminated showers.

References

[1] S.K. Gupta et al., Nucl. Instr. Meth. A 540 (2005) 311

[2] Y. Hayashi et al., Nucl. Instr. Meth. Phys. A 545 (2005) 643

[3] H. Tanaka et al 2012 J. Phys. G: Nucl. Part. Phys. 39 025201

[4] B. HariHaran et. al., Exp Astron 50, 185–198 (2020)

[5] M. Chakraborty et. al., Proc. of Sci. (ICRC 2021)394

[6] A. Hoecker et al. TMVA - Toolkit for Multivariate Data Analysis, arXiv:physics/0703039