PROCEEDINGS OF SCIENCE

Machine learning model for separation of muons from punch-through hadrons in EAS at GRAPES-3 experiment

D. Bezboruah,[a,*] M. Chakraborty,[b] M. M. Devi,[a] S.R. Dugad ,[b] S.K. Gupta,[b] B. Hariharan,[b] Y. Hayashi,[c] P. Jagadeesan,[b] A. Jain,[b] P. Jain,[h] S. Kawakami ,[c] H. Kojima,[d] S. Mahapatra,[e] P.K. Mohanty,[b] Y. Muraki ,[f] P.K. Nayak,[b] T. Nonaka,[g] A. Oshima,[d] D. Pattanaik ,[b,e] M. Rameez,[b] K. Ramesh,[b] L.V. Reddy,[b] A. Sarker,[a] S. Shibata ,[d] F. Varsi [h] and M. Zuberi[b] [The GRAPES-3 Collaboration]

[a] *Tezpur University, Sonitpur, Assam-784028, INDIA*
[b] *Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India*
[c] *Graduate School of Science, Osaka City University, Osaka 558-8585, Japan*
[d] *College of Engineering, Chubu University, Kasugai, Aichi 487-8501, Japan*
[e] *Utkal University, Bhubaneswar 751004, India*
[f] *Institute for Space-Earth Environmental Research, Nagoya University, Nagoya 464-8601, Japan*
[g] *Institute for Cosmic Ray Research, Tokyo University, Kashiwa, Chiba 277-8582, Japan*
[h] *Indian Institute of Technology Kanpur, Kanpur 208016, India*

E-mail: dbbphy1@tezu.ernet.in, arnabs@tezu.ernet.in, devimm@tezu.ernet.in

Gamma Ray Astronomy at PeV EnergieS-phase 3 (GRAPES-3) is a cosmic ray experiment with an array of extensive air shower detectors and a muon telescope. The primary goal of the experiment is the precision study of the cosmic ray energy spectrum, its nuclear composition and also multi-TeV γ-ray astronomy. The punch-through hadrons produced near the air shower core can lead to problems in the precise estimation of the number of muons and hadrons which is an essential parameter for reconstruction. Machine learning (ML) can prove to be immensely useful in distinguishing between different particle types which will significantly improve the physics analysis of the GRAPES-3 experiment. In this work, we have tested the feasibility of using Boosted Decision Trees (BDTs) for the task of muon-hadron separation at GRAPES-3. We study the efficiency of BDTs for separating muons from hadrons in extensive air showers detected in the experiment. We have obtained 89.5 % accuracy in classifying single incoming muon and hadron.

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

*Speaker

PoS(ICRC2023)522

1. Introduction

High-energy cosmic ray particles interact with the Earth's atmosphere producing a cascade of secondary particles giving rise to Extensive Air Showers (EAS) having lateral extensions with tens of kilometres. The cascades of secondaries can be electromagnetic cascades if produced by a gamma-ray or electron, or a hadronic cascade if produced by a nucleus or a hadronic primary. The electromagnetic shower contains only electrons, positrons and γ-rays. On the other hand, hadronic showers can have three main components: the hadronic component (consists mainly of pions, kaons and nucleons), the muonic component (produced primarily by charged pions and decaying kaons) and the electromagnetic component (produced primarily by neutral pions). The muon content of the air showers plays a pivotal role in determining the mass and composition of primary cosmic rays (PCRs) and also in γ-ray astronomy. The muon multiplicity distribution (MMD) exhibits significant sensitivity in determining the composition of PCRs [1], as heavier primaries generally yield a higher number of muons. PCRs with more than 100 TeV energy can be studied only by ground-based experiments consisting of detector arrays covering a large area due to their limited statistics for ballon or satellite-based measurements. Ground-based experiments often have dedicated muon detectors along with a large array of detectors to detect the electromagnetic component. The punch-through hadrons, produced mainly near the air shower core can penetrate the muon detectors and thereby interfere with the precise estimation of the number of muons.

Nowadays, machine learning (ML) is a widely used technique for several tasks in high-energy physics like event reconstruction, particle tracking and identification etc. ML algorithms, ranging from traditional methods to deep learning architectures, have revolutionized classification tasks for many particle physics experiments. Many ground-based cosmic ray experiments have also used ML for data analysis, pattern recognition and event classification tasks and have observed substantial improvement in the results. In this preliminary work, we have explored the potential use of Boosted Decision Trees (BDTs) [2] to separate muon and hadron events in the GRAPES-3 (Gamma Ray Astronomy at PeV EnergieS s- phase 3) experiment.

2. The GRAPES-3 Experiment

GRAPES-3 is a ground-based EAS experiment located in Ooty, India. It consists of an array of 400 plastic scintillators with an inter-detector separation of 8m covering a total of 25000m^2 area [3, 4]. It also has a muon telescope (G3MT) [5, 6] with 3712 proportional counters (PCRs) covering 560m^2 of area. The schematic view of the GRAPES-3 air shower array is shown in figure 1. The green-coloured regions show the 16 muon telescope modules in G3MT. In each muon telescope module, the PRCs are arranged in 4 layers, where Layer-0 and Layer-3 are the bottom-most and top-most

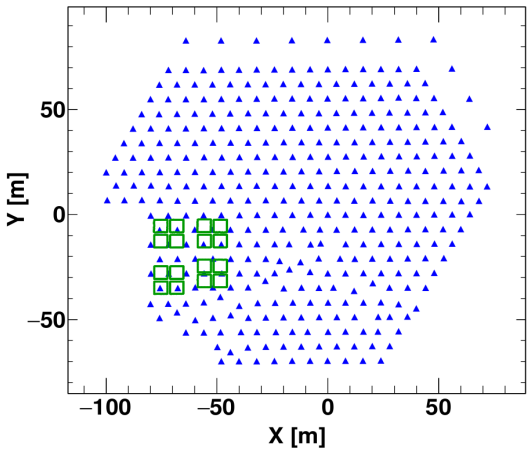


Figure 1: The GRAPES-3 air shower array.

layers respectively. The adjacent Layers are arranged orthogonal to each order with a 50 cm separation between them, which makes it possible for a 3D reconstruction of the muon tracks with an angular resolution of 4°. The 15 layers of concrete blocks above layer-0 provide an energy threshold of 1GeV$\times \sec(\theta)$ for incoming muon with zenith angle θ. The concrete is well capable of shielding most of the electromagnetic components of the EAS. However, hadrons with sufficiently high energy can interact with the concrete, producing a hadronic EAS with multiple secondaries capable of penetrating the concrete barrier and depositing energy in the PRCs forming a complex cluster of hits. The discrete hits can be reconstructed with the muon having multiple tracks, but the cluster of hits contributes to the early saturation of the muon modules. This problem of hadron punch through is significantly less (below 2%) when we consider showers having a core beyond 60 m from the centre of the muon modules. In this work, we are trying to correctly classify muon and hadron events using ML techniques to have a greater precision of the muon content in the air shower. It is expected that this will substantially improve the muon multiplicity obtained from GRAPES-3, especially at PeV energies.

3. Boosted Decision Trees

Decision trees are sophisticated supervised multivariate ML technique that has been excellent candidate for particle identification and classification from their first exploration by the MiniBooNe collaboration [7]. Single decision trees are regarded as 'weak learners' as they are very sensitive to input data. In boosting methods, many weak learners are combined to result in a robust multivariate algorithm. The ensemble of base learners (decision trees) (h_k) is generated iteratively in a way that adding the new learner will minimise the loss function of the entire ensemble of learners. In other words, successive trees are added to address the mistakes of their predecessors. At each iteration, the algorithm focuses on the data points that were misclassified in the previous iteration, giving them more weights (boosting) and adding a new tree based on these weights. As a result, successively added trees are better at classifying previously misclassified events. The final prediction is a weighted combination of the predictions of all the weak classifiers. This can be represented by the following mathematical relation,

$$F(x) = \sum_{k=1}^{N_{tree}} \alpha_k h_k(x), \tag{1}$$

where N_{tree} is the number of base trees, α_k is the weight associated with the k^{th} decision tree and x is the feature space of the event. The base learners (decision trees) are a collection of internal nodes and leaves. The root node contains all the events with their features. At each successive node, events are divided depending on whether the value of a particular input feature is above a threshold. The cut on the input features is chosen in such a way as to maximise the purity of the split or separability of the signal and background. A boosting algorithm proceeds iteratively finding the next learner of the ensemble (h_i) by minimizing a loss function $L(y_i, f(x_i))$,

$$\sum_{i=1}^{N} L_n(y_i, F_{n-1}(x_i) + \alpha_n h_n(x_i)) \longrightarrow min_{\alpha,h} \tag{2}$$

by finding appropriate value of the weights α_n, of the new learner. Here, in this equation, $F_{n-1}(x)$ is the loss function up to that point. In gradient boosting, the gradient descent algorithm is

used to minimize the loss function. Extreme gradient boosting (XGBoost) [8] is an optimized implementation of gradient boosting that incorporates additional algorithmic enhancements, such as parallel processing and regularization techniques. In this work, we have explored binary decision tree classifiers implementing the XGBoost algorithm.

4. Data and Input Features

The GEANT4 simulated events for G3MT [6] are used for model training and validation. The simulation contains EAS corresponding to proton (H), helium (He), nitrogen (N), aluminium (Al) and iron (Fe) primaries in the energy range 1 Tev to 100 PeV simulated using CORSIKA. PCR particles were traced to the GRAPES-3 observational level and then injected into the GEANT4 simulated G3MT modules and the corresponding information of each PRC hit was stored in a ROOT-based framework. In the present analysis, PCRs in the energy range 1000 TeV to 1584.89 TeV for all four modules of station 0 in G3MT is used. In order to avoid class imbalance, an equal number of muon and hadron events are chosen randomly. The input features are built using the information on the number of hits, corresponding Layer and Counter information along with the energy deposited by the particles in the PRCs. The muon/hadron classification task is defined as a binary classification problem by labelling muons as the positive class (1) and hadrons as the negative class (0).

Sl. No	Input Features	Description
1	nHits	Total number of hits in one event
2	nLayers	Total number of Layers having hit in one event
3	E_{total}	Total energy deposited in all the Layers in one event
4	FH_0	Number of hits in Layer 0 / nHits
5	FH_1	Number of hits in Layer 1 / nHits
6	FH_2	Number of hits in Layer 2 / nHits
7	FH_3	Number of hits in Layer 3 / nHits
8	dE_0	Energy deposited in Layer 0
9	dE_1	Energy deposited in Layer 1
10	dE_2	Energy deposited in Layer 2
11	dE_3	Energy deposited in Layer 3

Table 1: Features used for training the ML classification model. Here 0,1, 2,3 represent the corresponding Layer numbers as in the simulated data file.

The primary input features used are the total number of PRC hits produced by one secondary (nHits), the number of layers in which energy is deposition (nLayers) during its passage through one G3MT module and the sum of that energy (E_{total}) deposited in the PRCs. Additionally, we have also constructed some features to incorporate the behaviour at each layer basis, which are listed in table 1 with their definitions. The distribution of input features for both muon and hadron classes are shown in figure 2. In all the figures, the blue-coloured histograms represent the features for muons and the red-coloured histograms represent the corresponding feature distribution for

hadrons. For this study, we have used the XGBClassifier from the Python implementation of the XGBoost library. 70% of the data is used for training the model and the remaining 30% is used for testing the model predictions.

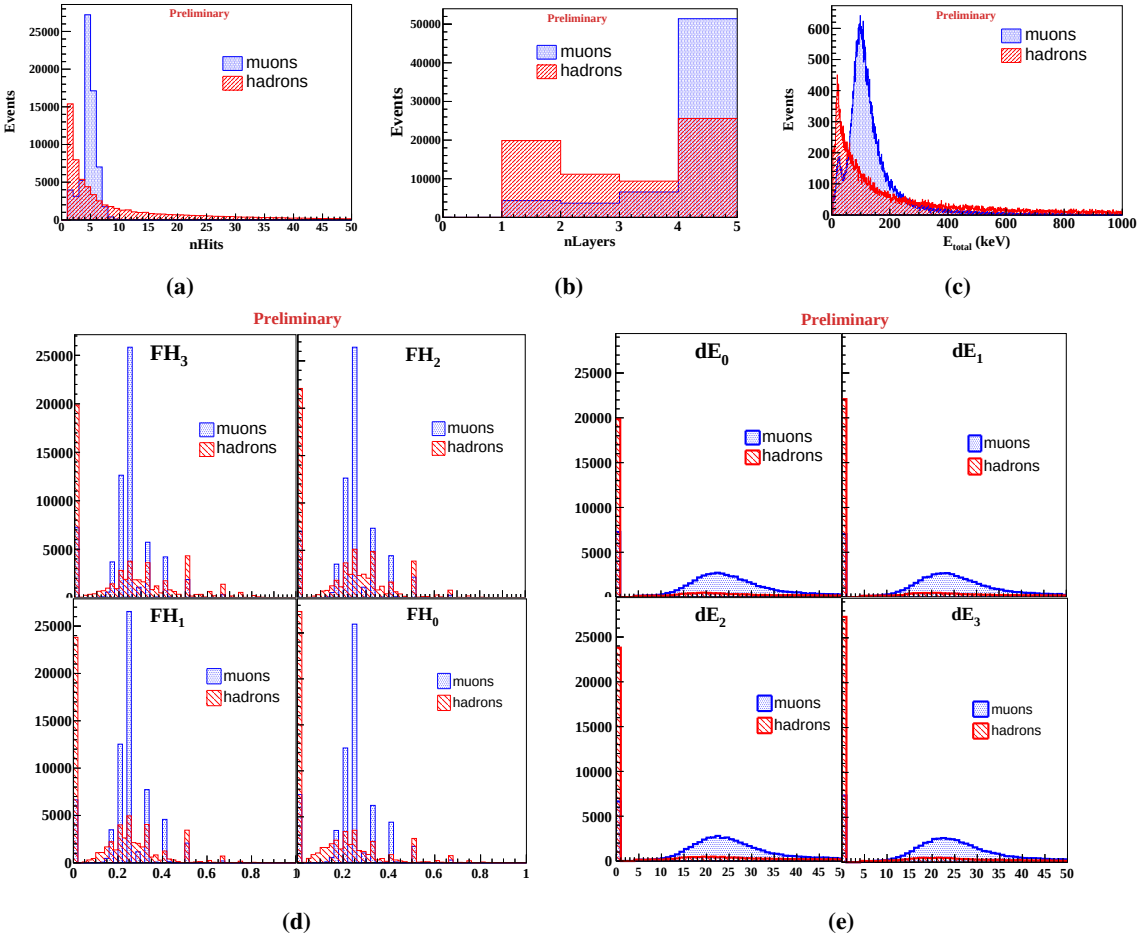
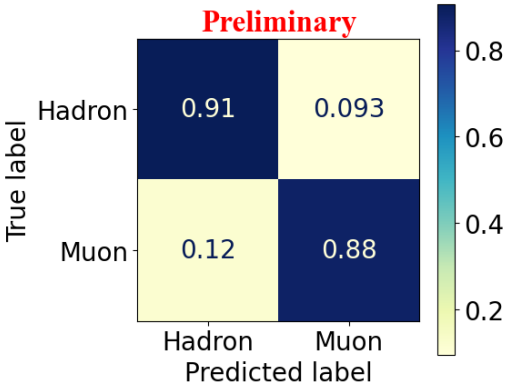


Figure 2: Distribution of input features for muon (blue) and hadron (red) class. The feature definitions are given in figure 1.

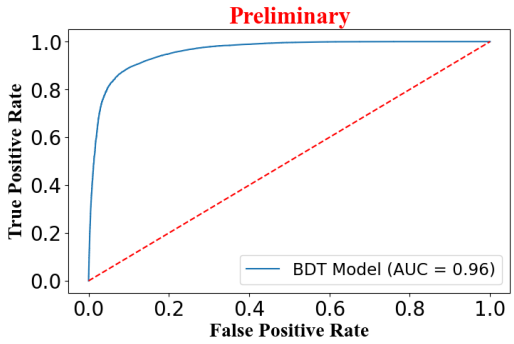
5. Results and Discussions

The primary objective of this work is to obtain the correct number of muons associated with each EAS. As the preliminary step, we are concentrating on the accuracy with which we can classify a muon or a hadron associated with the air shower, based on the information available from the detector output.

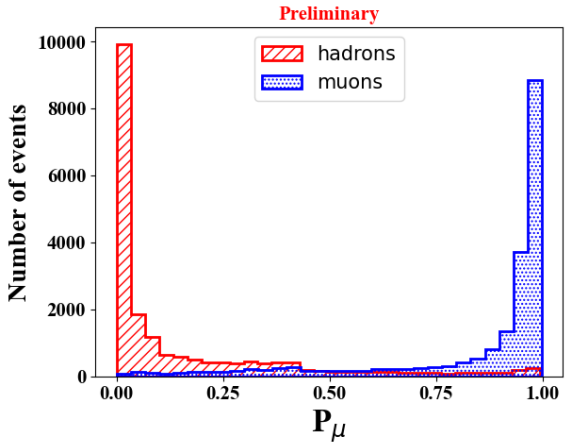
The performance of our model is summarized in the form of a confusion matrix (CM) in figure 3a. The CM is a way to visualize the performance of a classification algorithm. The instances in the actual dataset are represented by each row of the matrix, whereas each column represents the instances in a predicted class. The BDT model developed can predict 91% of true hadrons as hadrons whereas 9.3% are misclassified as muons. In the case of true muon events, it can predict



(a) Confusion Matrix for the BDT model



(b) ROC curve of the BDT model



(c) Muon prediction probability (P_μ) for muon (red) and hadron class (blue)

Figure 3: The confusion matrix (a), ROC curve (b) and muon prediction probability (c) of the BDT model using the test data.

88% as muons and the remaining 12% are misclassified as hadrons. The ROC (Receiver Operating Characteristics) curve for the model is shown in figure 3b. The ROC curve is the plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) for different candidate threshold values in the range of 0 to 1. TPR measures the proportion of true muon instances correctly identified as muons by the BDT model. Whereas, the FNR is the proportion of true hadron instances incorrectly classified as muons. The area under the curve (AUC) gives an overall performance of the model and it is a very popular metric used by the ML community to evaluate the model performance. We have obtained an AUC of 0.96 which indicates that the BDT model has satisfactory classification ability.

The distribution of muon prediction probability (P_μ) using the BDT model is shown in figure 3c. The blue histogram is for muon events and the red histogram is for hadron events. If the threshold probability for selection is considered as 0.5 (which is the default value), then the number of miss-classified muon events is more than the number of miss-classified hadron events which can also be interpreted from the confusion matrix shown in figure 3a.

In this analysis, we consider the case of a single incoming secondary muon or hadron. The BDT algorithm has achieved a classification accuracy of 89.50% in classifying these hits as either muon or hadron class.

6. Conclusion

Machine learning techniques can be used to bring significant improvement in the event reconstruction of particles in high-energy physics experiments. In this work, we have explored the feasibility of using ML techniques for muon/hadron separation in cosmic ray showers at the GRAPES-3 experiment. We are only considering one incoming secondary particle in the muon detector at a moment. The BDT model shows good classification capability for both muons and hadrons.

7. Acknowledgments

We are grateful to D.B. Arjunan, A.S. Bosco, V. Jeyakumar, S. Kingston, N.K. Lokre, K. Manjunath, S. Murugapandian, S. Pandurangan, B. Rajesh, R. Ravi, V. Santhoshkumar, S. Sathyaraj, M.S. Shareef, C. Shobana and R. Sureshkumar for their role in the efficient running of the experiment. We acknowledge the support of the Department of Atomic Energy, Government of India, under Project Identification No. RTI4002. This work was partially supported by grants from Chubu University, Japan. We also acknowledge the Science and Engineering Research Board (SERB), DST for the grant CRG/2021/002961. DB acknowledges DST for providing INSPIRE FELLOWSHIP to support this work. AS acknowledges the fellowship received from CSIR-HRDG (09/0796 (12409)/2021-EMR-I). The authors also acknowledge the DST FIST grant SR/FST/PSI-211/2016(C) received by the Department of Physics, Tezpur University.

References

[1] H. Tanaka *et al.*, "Studies of the energy spectrum and composition of the primary cosmic rays at 100-TeV - 1000-TeV from the GRAPES-3 experiment," *J. Phys. G*, vol. 39, p. 025201, 2012.

[2] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[3] S. K. Gupta *et al.*, "GRAPES-3: A high-density air shower array for studies on the structure in the cosmic-ray energy spectrum near the knee," *Nucl. Instrum. Meth. A*, vol. 540, pp. 311–323, 2005.

[4] P. Mohanty, S. Dugad, U. Goswami, S. Gupta, Y. Hayashi, A. Iyer, N. Ito, P. Jagadeesan, A. Jain, S. Karthikeyan, *et al.*, "Measurement of some eas properties using new scintillator detectors developed for the grapes-3 experiment," *Astroparticle Physics*, vol. 31, no. 1, pp. 24–36, 2009.

[5] Y. Hayashi *et al.*, "A large area muon tracking detector for ultra-high energy cosmic ray astrophysics: The GRAPES-3 experiment," *Nucl. Instrum. Meth. A*, vol. 545, pp. 643–657, 2005.

[6] F. Varsi *et al.*, "A GEANT4 based simulation framework for the large area muon telescope of the GRAPES-3 experiment," *JINST*, vol. 18, no. 03, p. P03046, 2023.

[7] B. P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor, "Boosted decision trees, an alternative to artificial neural networks," *Nucl. Instrum. Meth. A*, vol. 543, no. 2-3, pp. 577–584, 2005.

[8] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, aug 2016.