Discrimination of Gamma and Hadron Showers in Atmospheric Cherenkov Telescope Data using Machine Learning

Dr. J. Sravanthi[1*], Thadagoni Shireesha[2], Jamandla Srilekha[2], Gundeti Pavan[2], Shalva Sathwika[2]

[1]Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (Data Science), Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana.

*Corresponding Email: sravanthi.jataboina@gmail.com

ABSTRACT

Gamma or Hadron discrimination in ground-based gamma-ray observatories at the sub-TeV energy range is challenging as traditional muon-based methods become less effective at lower energies. Further, it is a challenge to distinguish hadron-initiated air showers from those initiated by primary gamma photons in detecting high energy gamma point sources using ground-based particle detector array. Therefore, this project presents an integrated machine learning application designed for the discrimination of gamma and hadron showers recorded by atmospheric Cherenkov telescopes. The system features an interactive graphical user interface (GUI) developed using Tkinter, which enables users to perform a comprehensive workflow including dataset uploading, preprocessing, exploratory data analysis (EDA), model training, evaluation, and prediction. The data preprocessing module handles missing values, label encoding, normalization, and train-test splitting. Various EDA visualizations such as histograms, box plots, scatter plots, correlation heatmaps, violin plots, count plots, and KDE plots—aid in understanding the underlying data distributions and relationships. Two classification approaches are implemented and compared: a Logistic Regression model and a Decision Tree classifier enhanced with AdaBoost. Model persistence is achieved via joblib to enable efficient reuse of trained models. Performance evaluation shows that while the Logistic Regression model attains moderate results, the DTC with AdaBoost classifier significantly outperforms it with near-perfect accuracy, precision, recall, and F1-score. The application also supports real-time prediction on new datasets, providing users with immediate insights into the classification of gamma and hadron events.

1. INTRODUCTION

Imaging Atmospheric Cherenkov Technique (IACT) is a ground-based gamma ray observation technique, used in γ ray astrophysics to detect very high energy γ ray photons [1]. The Earth's atmosphere serves as a giant local calorimeter for gamma ray detection. When a very high energy gamma ray enters the atmosphere, it interacts with atmospheric nuclei and an e+e− pair is produced. The produced e+ and e− from the pair will make collisions with atmospheric nuclei and are subjected to energy loss via multiple Coulomb scattering. In these collisions the charged particles will be accelerated and emit electromagnetic radiation (bremsstrahlung). This process continues till the threshold for the physical processes is involved is reached and an electromagnetic cascade is produced, also known as air shower [1]. The charged particles in this air shower are travelling faster than the phase velocity of light in the atmosphere and will emit Cherenkov radiations at Cherenkov angle, which is inversely proportional to the velocity of the particle and the refractive index [2]. A very narrow cone (1.0°at 8 km above sea level) of Cherenkov radiation generated by the ultrarelativistic charged particle in the cascade penetrates to the ground level [3]. The produced Cherenkov light is spread over a larger area of several hundred square meter around the shower axis and can be detected using ground-based telescopes working on the principle of IACT. An arrangement of large focusing mirrors reflects the Cherenkov light onto an array of photo-multiplier tubes which captures high speed images of these very short Cherenkov radiation flashes (5-20 ns) [4]. The direction and energy of the primary gamma rays

doi: 10.48047/ijiee.2025.15.4.18

can be reconstructed from the shower image using different techniques. An overwhelming background of hadronic cascade is initiated by the cosmic rays entering into Earth's atmosphere.

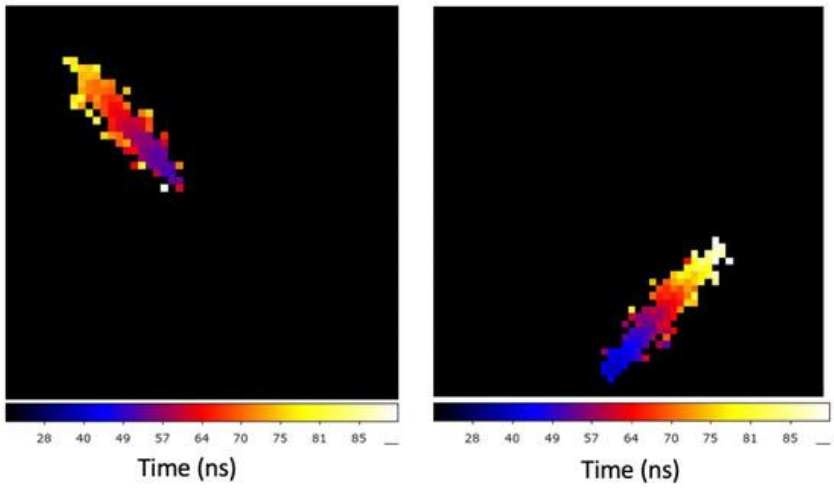

Fig. 1: Simulated Cherenkov shower images as observed by an ASTRI Mini-Array telescope, illustrating the differences in time evolution between different types of events. The image covers the entire ASTRI field of view (each side spanning approximately 10 deg). The color scale indicates the time when the pixel is triggered by the incoming Cherenkov photons, with the black, indicating that the pixel has not been triggered. The time (in ns) increases from blue to white. In this example, the gamma-initiated event (left) evolves slowly than the hadron-initiated event (right).

To study high energy gamma rays of astrophysical origin, the intense isotropic background of hadronic cascade from cosmic rays must be rejected with high efficiency without losing much of the primary gamma ray signal. Traditionally, a direct selection method is used for the γ -hadron separation in Cherenkov telescope data analysis, in which direct cuts are made on the image parameters [5]. The γ -hadron separation is a multivariate binary classification of the events into two classes: signal (gamma) and background (hadrons) using different shape features of the shower image. A huge class imbalance is present in the Cherenkov telescope data due to the dominant hadronic background.

2. LITERATURE SURVEY

The study of gamma rays is crucial for understanding extreme astrophysical events and probing fundamental physics. Gamma rays in the sub-TeV to TeV range can provide insights into active galactic nuclei, gamma ray bursts, and potential new physics beyond the standard model, including dark matter research [6]. Muon identification remains one of the main methods for gamma/hadron separation at TeV energies [7]. However, the muon content of hadronic showers is scarce at sub-TeV energies. Alternatively, one may analyze the shower footprint patterns on the ground to infer features of the shower development [8].

Hadron-induced showers, unlike pure electromagnetic showers, generate particles with high transverse momentum, causing the shower to spread more laterally and form clusters. Initial studies have shown great potential in using these ground patterns for gamma/hadron discrimination, but further work is needed to fully understand and optimize their effectiveness against noise [9].Machine learning (ML) has recently become an innovative instrument in the field of physics, especially within astroparticle physics, enabling advancements in numerous domains that require extensive data analysis. In modern cosmic and gamma-ray observatories, ML techniques have shown significant potential to enhance gamma/hadron separation [10], event reconstruction [11], and even allow neutrino identification with water Cherenkov detectors (WCDs) [12].

doi: 10.48047/ijiee.2025.15.4.18

These breakthroughs are analogous to significant advances in related domains, such as the IceCube detection of neutrinos from the galactic plane [13], investigations on the mass composition and Xmax estimation of ultra-high-energy cosmic rays by the Pierre Auger Observatory [14], and the search of new physics with the Large Hadron Collider [15]. Among ML-based techniques, transformers, which use attention mechanisms for data analysis, have emerged as particularly effective. Their initial applications in event reconstruction for astroparticle experiments have demonstrated promising results [16].

In this work, we demonstrate that state-of-the-art pretrained vision transformers (ViTs) have significant potential to accurately discriminate between gamma and hadron-induced air showers. The footprint image, created from signals detected by the individual WCDs in the detector array, is used as input for the model. The method's robustness against noise, including atmospheric muons and low-energy proton showers, is also assessed [17]. In this study, the state-of-the-art pretrained ViT model for classification "google/vit-base-patch16-224-in21k" has been used. This model is a BERT-like transformer encoder that includes an extra linear layer applied to the classification token for executing classification tasks.

It was pretrained on a large dataset called ImageNet-21k [18], which is composed of 14 million images at a resolution of 224×224 pixels and has 21843 classes. It is crucial to note that using a pretrained ViT considerably lowers the computational expense involved in training from scratch, while taking advantage of the rich feature representations obtained from large datasets. For the sake of reproducibility, we employed a ViT model [19]. Most of these advances ˇ use Monte Carlo simulations of the showers to train and assess supervised models. Simulations are useful since they can be generated in any desired proportion, number of events, and different configurations. But once real data becomes available, an adaptive strategy is needed to train the models further [20].

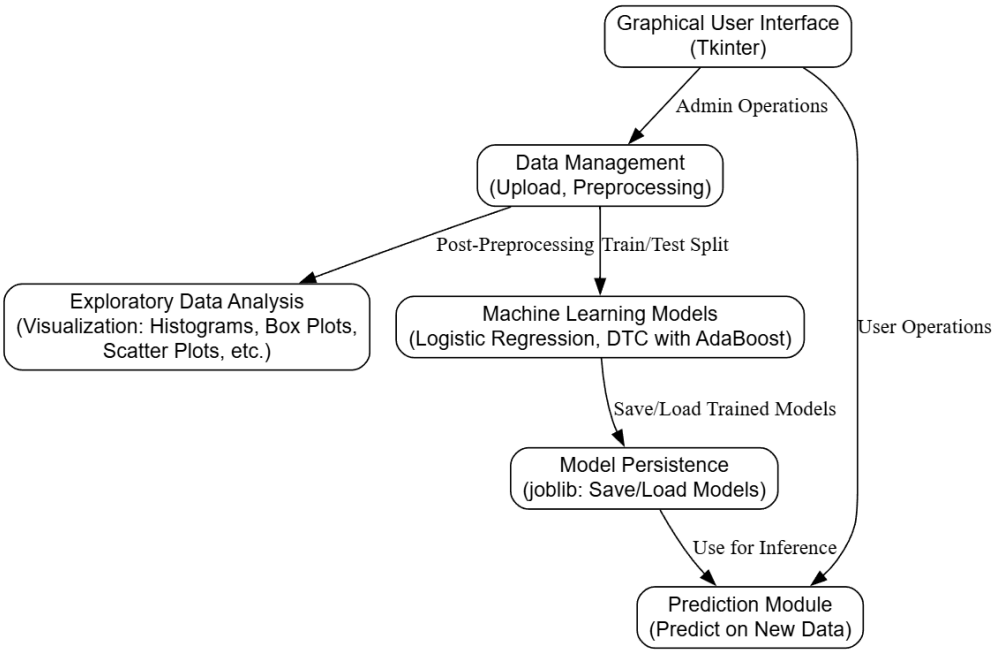

Fig. 2: Workflow of proposed system.

3. PROPOSED METHODOLOGY

This research is a complete machine learning application built with a graphical user interface (GUI) that helps users analyze atmospheric Cherenkov telescope data to distinguish between Gamma and Hadron showers.The application uses Python's Tkinter library to create an interactive window. Depending on

doi: 10.48047/ijiee.2025.15.4.18

the user's role (ADMIN or USER), different functionalities are provided through buttons. ADMIN users can upload datasets, perform preprocessing, conduct exploratory data analysis (EDA), and train models, while USERs can upload new data for predictions. Users can upload CSV files containing telescope data. The project cleans the data by filling in missing values, encodes categorical labels into numerical ones, and normalizes features using standard scaling. It then splits the data into training and testing sets.

The application includes multiple functions to visualize the data. Users can generate various plots such as histograms, box plots, scatter plots, heatmaps, violin plots, count plots, and KDE plots to better understand the distribution and relationships within the data. Two machine learning approaches are implemented: Logistic Regression, which either loads pre-trained weights or trains a new classifier to predict the class labels, and a Decision Tree with AdaBoost, which uses a decision tree as the base estimator within an AdaBoost framework. It similarly checks for saved model weights or trains a new model if needed. Both models are evaluated using metrics like accuracy, precision, recall, and F1 score, and their performance is visualized using confusion matrices. The application allows users to upload a new dataset for which the trained classifier predicts the outcome (Gamma or Hadron). Each prediction is displayed alongside the corresponding data row for clarity.

3.1 Data Preprocessing

Data preprocessing is a crucial step to ensure the dataset is clean, structured, and suitable for machine learning models. The dataset used in this project contains records of gamma and hadron showers with various numerical features. Before feeding this data into a machine learning model, several preprocessing steps are applied:

Step-1: Handling Missing Values: Missing or null values can introduce bias and affect model performance. In this project, missing values are replaced with zero (0) to maintain dataset consistency.

Step-2: Label Encoding: The target variable (gamma or hadron) is a categorical label. To make it interpretable by machine learning models, it is converted into numerical values using Label Encoding (Gamma = 0, Hadron = 1).

Step-3: Feature Selection and Normalization: The dataset contains multiple numerical features. To ensure that features with larger numerical ranges do not dominate those with smaller ones, StandardScaler is applied to standardize all feature values. This transformation scales the features to have a mean of zero and a standard deviation of one, improving model convergence and performance.

3.2 Proposed DTC with AdaBoost model

DTC (Decision Tree Classifier) with AdaBoost is a machine learning ensemble method that combines multiple weak learners (Decision Trees) to form a strong predictive model. AdaBoost (Adaptive Boosting) is used to improve the performance of a simple Decision Tree model by sequentially training multiple trees and adjusting their weights to focus on misclassified samples. This approach helps in reducing variance and bias while increasing model robustness.

Step-1: Initialize Data Weights: Each sample is initially assigned equal weight.

Step-2: Train a Weak Learner (Decision Tree): A small-depth Decision Tree (weak learner) is trained on the dataset.

Step-3: Evaluate Model Performance: Misclassified samples receive higher weights, increasing their importance for the next model iteration.

Step-4: Train Additional Weak Learners: New trees are sequentially trained with adjusted weights to correct previous mistakes.

doi: 10.48047/ijiee.2025.15.4.18

Step-5: Combine Weak Learners: The final model aggregates all weak learners' outputs using weighted voting to make a robust classification decision.

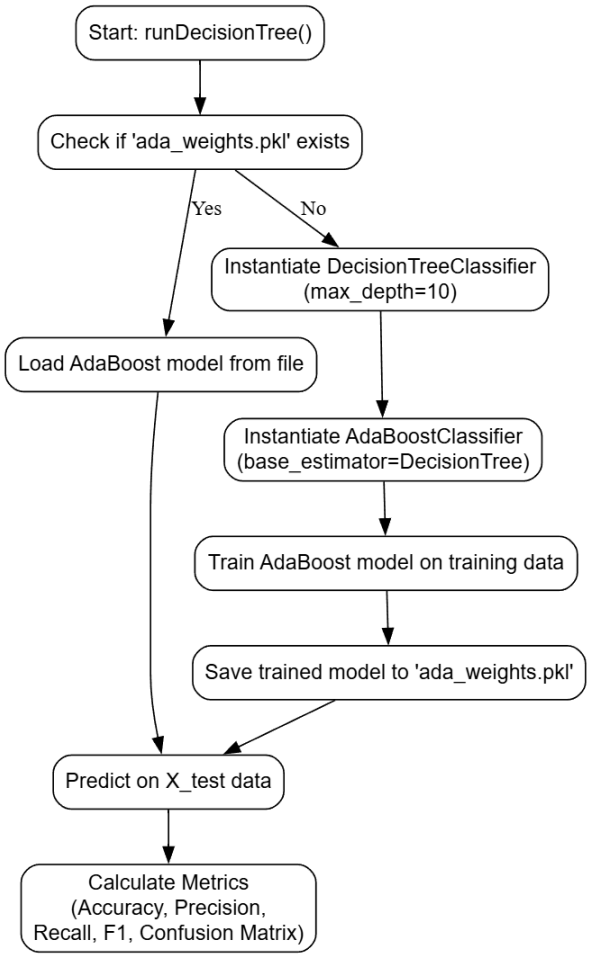

Fig. 4: Proposed DTC with AdaBoost Classifier architectural layer diagram.

4. RESULTS AND DESCRIPTION

4.1 Dataset description

The data are MC generated to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique. Cherenkov gamma telescope observes high energy gamma rays, taking advantage of the radiation emitted by charged particles produced inside the electromagnetic showers initiated by the gammas, and developing in the atmosphere. This Cherenkov radiation (of visible to UV wavelengths) leaks through the atmosphere and gets recorded in the detector, allowing reconstruction of the shower parameters. The available information consists of pulses left by the incoming Cherenkov photons on the photomultiplier tubes, arranged in a plane, the camera. Depending on the energy of the primary gamma, a total of few hundreds to some 10000 Cherenkov photons get collected, in patterns (called the shower image), allowing to discriminate statistically those caused by primary gammas (signal) from the images of hadronic showers initiated by cosmic rays in the upper atmosphere (background).

Typically, the image of a shower after some pre-processing is an elongated cluster. Its long axis is oriented towards the camera center if the shower axis is parallel to the telescope's optical axis, i.e. if the telescope axis is directed towards a point source. A principal component analysis is performed in the camera plane, which results in a correlation axis and defines an ellipse. If the depositions were

doi: 10.48047/ijiee.2025.15.4.18

distributed as a bivariate Gaussian, this would be an equidensity ellipse. The characteristic parameters of this ellipse (often called Hillas parameters) are among the image parameters that can be used for discrimination. The energy depositions are typically asymmetric along the major axis, and this asymmetry can also be used in discrimination. There are, in addition, further discriminating characteristics, like the extent of the cluster in the image plane, or the total sum of depositions.

The dataset consists of multiple features representing the characteristics of atmospheric particle showers, specifically distinguishing between gamma and hadron events

4.2 Result description

This research implements a graphical user interface (GUI) for data exploration, preprocessing, machine learning model training, evaluation, and prediction, specifically aimed at discriminating between Gamma and Hadron showers in Atmospheric Cherenkov Telescope data. Fig. 5 class Distribution using count plot depicts the frequency of each class (gamma vs. hadron), giving a clear picture of the target variable distribution.

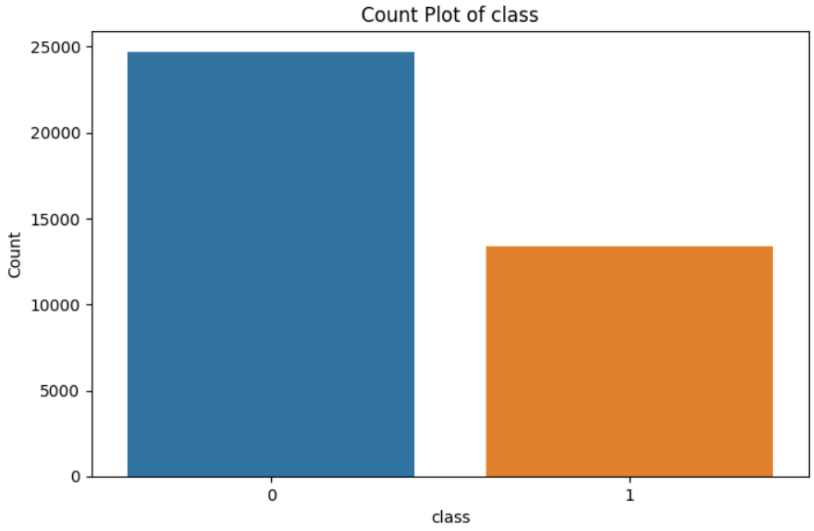

Fig. 5: Count plot of target variable.

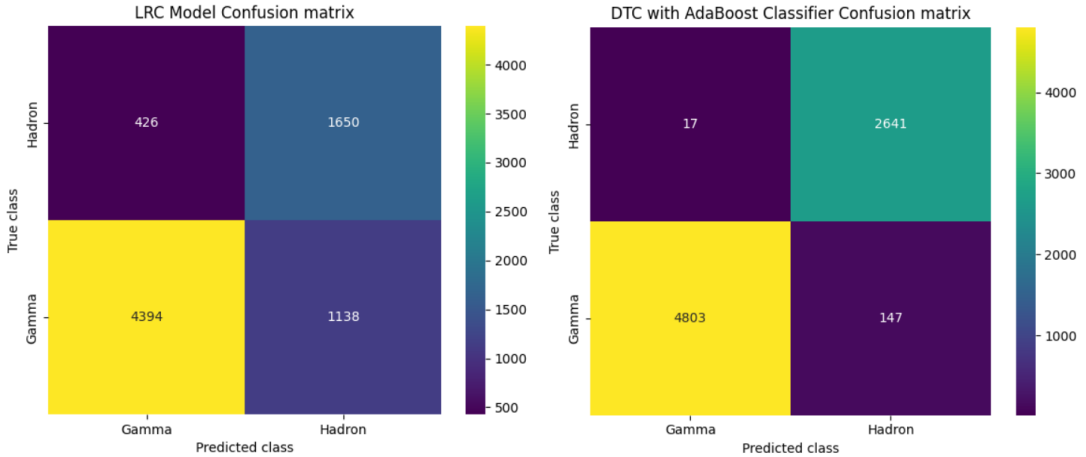

Fig. 6: Confusion matrix obtained using LRC model (left). Proposed DTC with AdaBoost (right).

Fig. 6 (left) illustrates the evaluation of the LRC model in terms of accuracy, precision, recall, and F1-score. The classification scatter plot visualizes the separation of gamma and hadron classes based on the model's predictions. The model achieves an accuracy of 79.21%, showing its ability to classify

doi: 10.48047/ijiee.2025.15.4.18

shower events with moderate precision. The scatter plot highlights areas of misclassification, reflecting the model's limitations in capturing complex decision boundaries. Fig. 6(right) provides an assessment of the proposed DTC with AdaBoost, including key performance metrics. The model achieves an accuracy of 98.94%, demonstrating a significant improvement over LRC model. The scatter plot exhibits a well-defined separation between gamma and hadron events, indicating the effectiveness of boosting techniques in refining decision boundaries. The improved precision and recall scores confirm the model's robustness in accurately classifying shower events.

Fig. 7: Sample prediction on test data.

Fig. 7 displays an example of the prediction results. After uploading a test dataset, the system processes the data and applies the trained classifier to predict whether each event corresponds to a gamma or hadron shower. The output is presented in a readable format, where each row of the input data is paired with its predicted outcome. This clear mapping allows users to verify predictions and gain insights into the model's performance on unseen data.

Table. 1: Performance comparison quality metrics obtained using LRC model, and proposed DTC with AdaBoost classifier.

Algorithm Name	Accuracy	Precision	Recall	f1-score
Logistic Regression	79.20	74.48	78.15	75.65
DTC with AdaBoost Classifier	98.93	98.63	99.01	98.82

Table.1 compares the performance metrics of multiple classification models, including Logistic Regression and DTC with AdaBoost. The bar chart or line Table highlights differences in accuracy, precision, recall, and F1-score across models. The comparison validates the superior performance of the proposed approach, demonstrating the impact of ensemble learning in improving classification effectiveness.

5. CONCLUSION

The comprehensive evaluation of the ML-driven approach for discriminating between gamma and hadron showers in atmospheric Cherenkov telescope data demonstrates the effectiveness of the

doi: 10.48047/ijiee.2025.15.4.18

proposed Decision Tree Classifier (DTC) with AdaBoost. While traditional models such as Logistic Regression provided baseline performance, the integration of DTC with AdaBoost significantly enhanced classification accuracy, ensuring more precise differentiation between gamma-ray signals and hadronic background noise. The robust preprocessing techniques, coupled with exploratory data analysis (EDA) and feature scaling, contributed to improved model reliability. Additionally, the comparative analysis of classification algorithms highlighted the superiority of the proposed approach in handling complex spatial and intensity-based features of shower images. The successful implementation of this methodology establishes its applicability in astrophysical research, enabling more accurate and efficient identification of gamma-ray sources.

REFERENCES

[1] Idan Shilon, Manuel Kraus, Matthias Büchele, Kathrin Egberts, Tobias Fischer, Tim Lukas Holch, Thomas Lohse, Ullrich Schwanke, Constantin Steppa, and Stefan Funk. Application of deep learning methods to analysis of imaging atmospheric cherenkov telescopes data. Astroparticle Physics, 105:44–53, 2019.

[2] EB Postnikov, AP Kryukov, SP Polyakov, DA Shipilov, and DP Zhurov. Gamma/hadron separation in imaging air cherenkov telescopes using deep learning libraries tensorflow and pytorch. In Journal of Physics: Conference Series, volume 1181, page 012048. IOP Publishing, 2019.

[3] Aaron Baughman, Daniel Bohm, Micah Forster, Eduardo Morales, Jeff Powell, Shaun McPartlin, Raja Hebbar, and Kavitha Yogaraj. Large scale diverse combinatorial optimization: Espn fantasy football player trades. arXiv preprint arXiv:2111.02859, 2021.

[4] Richard P Feynman. Simulating physics with computers. In Feynman and computation, pages 133–153. CRC Press, 2018.

[5] Vojt ech Havlí cek, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. Nature, 567(7747):209–212, 2019.

[6] A. Viana, H. Schoorlemmer, A. Albert, V. de Souza, J. P. Harding, and J. Hinton, Searching for dark matter in the galactic halo with a wide field of view TeV gamma-ray observatory in the Southern Hemisphere, J. Cosmol. Astropart. Phys. 12 (**2019**) 061.

[7] P. Assis, A. Bakalová, U. Barres de Almeida, P. Brogueira, R. Conceição, A. De Angelis, L. Gibilisco, B. González, A. Guillén, G. La Mura et al., The Mercedes water Cherenkov detector, Eur. Phys. J. C **82**, 899 (2022).

[8] R. Conceição, B. S. González, A. Guillén, M. Pimenta, and B. Tomé, Muon identification in a compact single-layered water Cherenkov detector and gamma/hadron discrimination using machine learning techniques, Eur. Phys. J. C **81**, 542 (2021).

[9] S. Kunwar, H. Goksu, J. Hinton, H. Schoorlemmer, A. Smith, W. Hofmann, and F. Werner, A double-layered water Cherenkov detector array for gamma-ray astronomy, Nucl. Instrum. Methods Phys. Res., Sect. A **1050**, 168138 (2023).

[10] I. Shilon, M. Kraus, M. Büchele, K. Egberts, T. Fischer, T. L. Holch, T. Lohse, U. Schwanke, C. Steppa, and S. Funk, Application of deep learning methods to analysis of imaging atmospheric Cherenkov telescopes data, Astropart. Phys. **105**, 44 (2019).

[11] J. Glombitza, V. Joshi, B. Bruno, and S. Funk, Application of graph networks to background rejection in imaging air Cherenkov telescopes, J. Cosmol. Astropart. Phys. 11 (**2023**) 008.

[12] X. Wang et al. (LHAASO Collaboration), Gamma hadron separation using traditional single parameter method and multivariate algorithms with LHAASO-WCDA experiment, Proc. Sci. ICRC2019 (**2019**) 820.

doi: 10.48047/ijiee.2025.15.4.18

[13] F. Aharonian et al. (LHAASO Collaboration), Performance of LHAASO-WCDA and observation of the Crab Nebula as a standard candle, Chin. Phys. C **45**, 085002 (2021).

[14] C. Jin, S.-z. Chen, H.-h. He et al. (LHAASO Collaboration), Classifying cosmic-ray proton and light groups in LHAASO-KM2A experiment with graph neural network, Chin. Phys. C **44**, 065002 (2020).

[15] T. Capistrán et al. (HAWC Collaboration), Use of machine learning for gamma/hadron separation with HAWC, Proc. Sci. ICRC2021 (**2021**) 745 [arXiv:2108.00112].

[16] I. Watson et al. (HAWC Collaboration), Deep learning for the HAWC observatory, Proc. Sci. ICRC2023 (**2023**) 927.

[17] A. Dosovitskiy et al., An image is worth words: Transformers for image recognition at scale, arXiv:2010.11929.

[18] A. Dosovitskiy et al., Hugging Face: vit-base-patch16-224-in21k., https://huggingface.co/google/vit-base-patch16-224-in21k (2021) (Accessed May-2023).

[19] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, Visual transformers: Token-based image representation and processing for computer vision, arXiv:2006.03677.

[20] Grespan, P., Jacquemont, M., Lopez-Coto, R., Miener, T., ´ Nieto-Castano, D., and Vuillaume, T. (2021). Deep- ˜ learning-driven event reconstruction applied to simulated data from a single Large-Sized Telescope of CTA. arXiv:2109.14262 [astro-ph].

doi: 10.48047/ijiee.2025.15.4.18