PROCEEDINGS OF SCIENCE

Reconstruction of multiple calorimetric clusters in the LHCf experiment with machine learning techniques

Giuseppe Piparo[a,b,*] **for the LHCf collaboration**

[a] *INFN Catania section,*
Via S. Sofia 64, Catania (CT), Italy

[b] *Department of Physics and Astronomy "Ettore Majorana", University of Catania,*
Via S. Sofia 64, Catania (CT), Italy

E-mail: giuseppe.piparo@ct.infn.it

One of the major challenges in the Large Hadron Collider forward (LHCf) experiment is the accurate reconstruction of calorimetric clusters when multiple particles hit the same detector tower simultaneously. Traditional reconstruction methods struggle with overlapping signals, especially in events involving more than two particles or a combination of photons and neutrons. This paper presents the development of machine learning (ML) techniques to improve the reconstruction efficiency of such complex events. We discuss the motivations for integrating ML into the LHCf reconstruction pipeline, outline the ML approach and dataset preparation, and compare the performance of ML models with standard methods. The results demonstrate a significant improvement in reconstructing multi-hit events, which is essential for analyses involving π^0, η, K_s^0 mesons, and Λ^0 baryon. Finally, we explore future prospects for ML applications in the LHCf experiment.

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

[*] Speaker

PoS(ICHEP2024)1016

1. Introduction

The LHCf experiment measures very forward neutral particles produced in proton-proton and proton-ion collisions at the LHC, providing data for calibrating and testing hadronic interaction models (HIM) used in simulating extended air showers (EASs) produced by ultra-high energy cosmic rays (UHECR) in Earth atmosphere [1]. The apparatus consists of two independent detectors, Arm1 and Arm2, located at zero degrees relative to the LHC beam line, about 141.5 meters from Interaction Point 1 (IP1). Each detector has two sampling calorimeters, the Small Tower (ST) and the Large Tower (LT), covering different rapidity regions near the beam pipe. Both Arm1 and Arm2 feature calorimeters composed of 16 GSO scintillator layers interleaved with 17 tungsten layers, featuring about 44 radiation lengths (X_0) and 1.6 interaction lengths (λ_I). The GSO and tungsten layers are independent for the two towers, enabling simultaneous, separate energy measurements in the ST and LT. Position determination is achieved using position-sensitive layers: Arm1 uses plastic scintillating fibers (SciFi), while Arm2 employs silicon micro-strip detectors. The position resolutions are about 200 μm for Arm1 and 40 μm for Arm2 in the case of electromagnetic showers [2]. A significant challenge is accurately reconstructing multiple calorimetric clusters when several particles hit the same detector tower simultaneously. Overlapping signals in both the scintillators and position-sensitive layers complicate the reconstruction. Traditional methods are effective for single-hit events but lose accuracy as overlapping particles increase. This issue is particularly relevant in analyses focusing on particles like π^0, η, K_s^0, and Λ^0, whose decay products may hit the detector simultaneously and in close proximity. Recent advancements in machine learning (ML) offer promising solutions [3]. ML techniques can handle complex data patterns, making them suitable for improving reconstruction efficiency in multi-hit events. By leveraging detailed detector information, ML models can better deconvolute overlapping signals and accurately determine the energy and position of individual particles. In this paper, we present the development of ML approaches for the LHCf experiment, focusing on events involving two hits in the same detector tower. We discuss the limitations of traditional methods and the benefits of integrating ML into the reconstruction pipeline.

2. Machine Learning Approach

2.1 Motivations for Machine Learning in LHCf

The primary motivation for including ML techniques in the LHCf reconstruction pipeline stems from the need to enhance the accuracy of cluster reconstruction in multi-hit scenarios. Traditional methods excel in single-hit events but face limitations when dealing with multiple particles hitting the same calorimetric tower. In such cases, the total energy deposited must be correctly shared among the incident particles, and accurate position reconstruction becomes challenging due to overlapping signals. These challenges are particularly significant in analyses involving:

- **Type II π^0 and η meson analysis**: Events where two photons from π^0 and η meson decays hit the same calorimetric tower. Traditional methods based on peak height ratios in the transverse profile of position detectors can lead to a broadening of the invariant mass peak and reduced reconstruction efficiency [4–7].

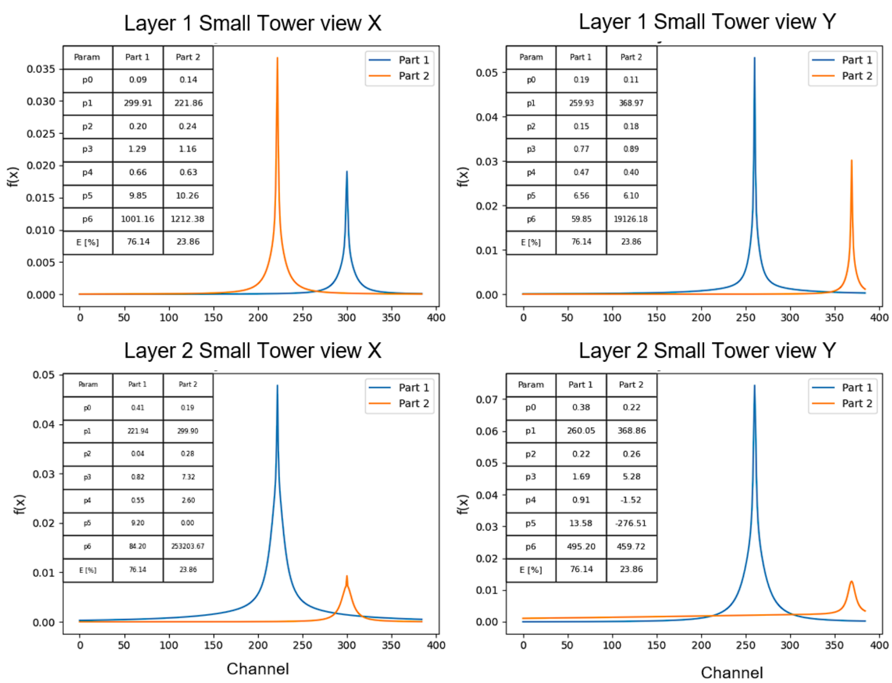

Figure 1: Example of Lorentzian fits to the transverse profiles in the silicon layers for a two-hit event in the Small Tower. The fit parameters serve as input variables for the ML models.

- K_s^0 **analysis**: Forward production of K_s^0 mesons, decaying into two π^0 and resulting in four photons, often requires reconstructing three or four photons in a single tower. Accurate reconstruction is crucial for improving HIMs and understanding the strange quark production [8].

- Λ^0 **analysis**: Studying forward Λ^0 production, decaying into a neutron and a π^0, involves identifying complex event topologies with neutrons and photons overlapping in the same tower. Accurate neutron identification and multi-particle reconstruction are essential for this analysis.

Integrating ML techniques can address these challenges by providing more sophisticated tools for handling overlapping signals and improving energy sharing and position reconstruction in complex events.

2.2 Dataset Preparation

The ML models were trained to predict energy sharing between particles in two-hit events. We generated a dataset using a full detector Monte Carlo (MC) simulation of proton-proton collisions at $\sqrt{s} = 13$ TeV with the QGSJETII-04 event generator, focusing on the Arm2 detector. Separate models were developed for the two towers. For each event, the transverse profiles from the silicon layers were fitted with a three-component Lorentzian function to extract fit parameters for each particle in both X and Y views, resulting in 56 input variables (seven fit parameters per particle per view for the first two silicon layers). Figure 1 shows an example of the fitted profiles. The dataset

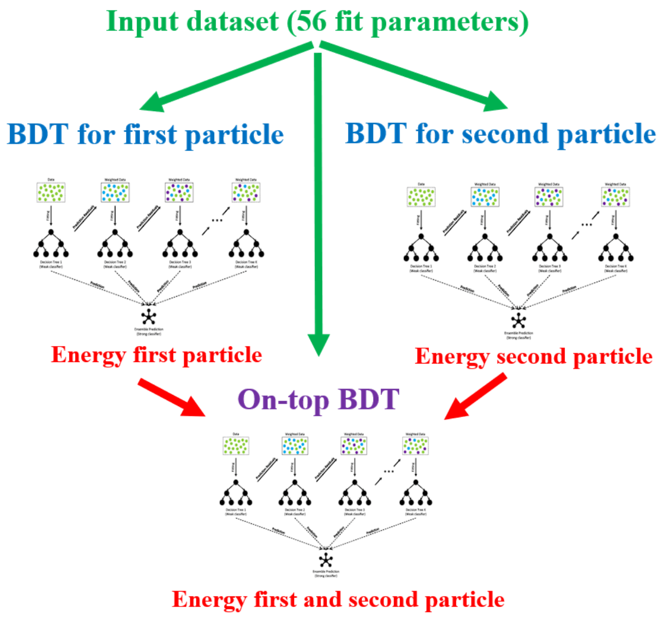

Figure 2: Schematic of the ML model architecture for energy reconstruction in two-hit events.

comprised approximately 130,000 events for the Small Tower and 60,000 for the Large Tower, split into 70% for training and 30% for testing.

2.3 Model Architecture

The ML pipeline utilizes ensemble methods based on gradient-boosting decision trees (BDTs). The architecture includes two first-level BDTs that predict the energy of each particle individually using input variables from the Lorentzian fits. An on-top BDT then combines the predictions from the first-level BDTs with the original input variables to infer the total event energy (see Figure 2). We employed two popular libraries, XGBoost [9] and CatBoost [10], to implement the BDT models, optimizing hyperparameters to minimize the Root Mean Square Error (RMSE) between predicted and true energies.

3. Results and Future Prospects

The performance of the ML models was evaluated by comparing them with the standard LHCf energy-sharing method, which uses the ratio of peak heights from the transverse profiles. The RMSE was used as the evaluation metric. Figures 3 and 4 show scatter plots of the predicted versus true energies for the Small Tower and Large Tower, respectively. The ML models significantly outperformed the traditional method, reducing RMSE values substantially. The ML models provided much closer agreement between predicted and true energies for both towers, while the standard method showed larger deviations, especially at higher energies. The comparable performance between XGBoost and CatBoost indicates the effectiveness of gradient-boosting decision trees for this application. These results demonstrate that ML techniques enhance energy reconstruction accuracy in two-hit events, offering significant improvements over traditional methods and indicating

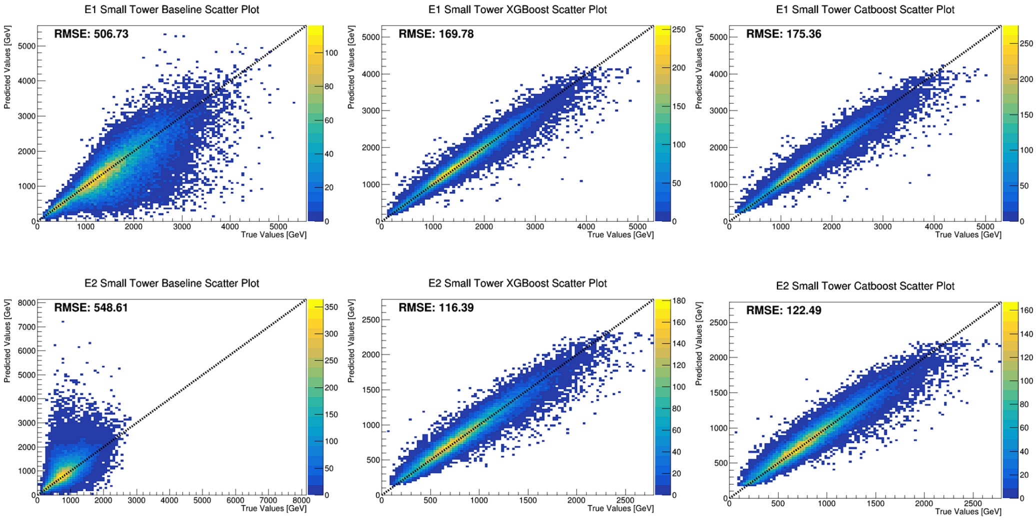

Figure 3: Predicted vs. true energies for two particles in the Small Tower using the standard method, XGBoost, and CatBoost models.

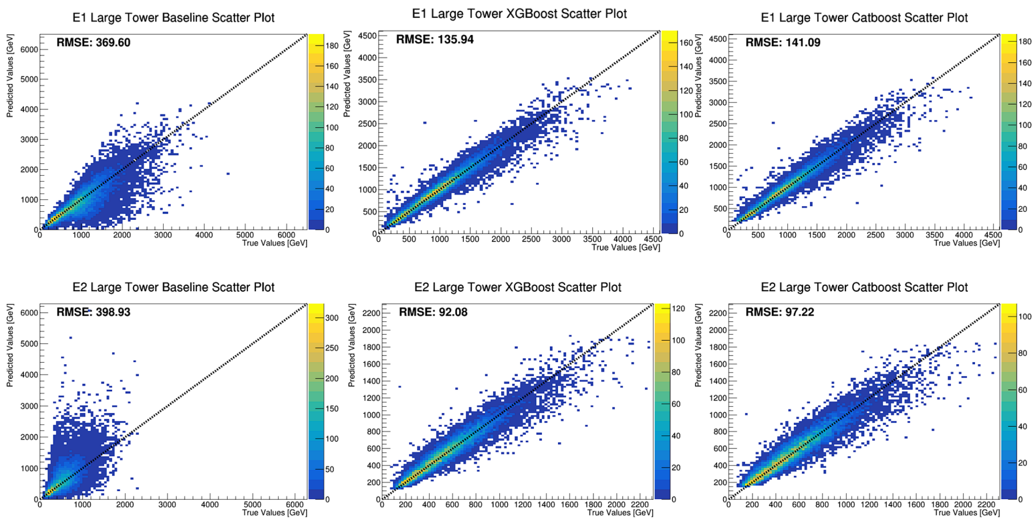

Figure 4: Predicted vs. true energies for two particles in the Large Tower using the standard method, XGBoost, and CatBoost models.

their potential to improve analyses of complex events in the LHCf experiment. Future work will extend these techniques to events with three or more overlapping particles, essential for analyses involving K_s^0 and Λ^0 mesons. Preliminary tests with three-hit datasets are encouraging, though further refinement and larger datasets are needed. Additionally, using raw energy deposition data from silicon detectors, instead of fitted parameters, may enhance the models ability to deconvolute overlapping signals and improve position reconstruction. Other potential ML applications in LHCf include automating peak identification and classification in energy distributions and training models for neutron identification based on calorimetric signatures. Scaling the models with larger, more diverse datasets and exploring advanced ML architectures will further optimize the reconstruction pipeline, contributing to more accurate measurements and enabling new exotic analysis channels in the LHCf experiment.

Acknowledgments

This work is (partially) supported by ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU-CUP - F77G22000120006.

References

[1] O. Adriani *et al.*, "The LHCf Detector at the CERN Large Hadron Collider," *Journal of Instrumentation*, vol. 3, no. 08, S08006, 2008, doi:10.1088/1748-0221/3/08/S08006.

[2] Y. Makino *et al.*, "Performance Study for the Photon Measurements of the Upgraded LHCf Calorimeters with Gd_2SiO_5 (GSO) Scintillators," *Journal of Instrumentation*, vol. 12, P03023, 2017, doi:10.1088/1748-0221/12/03/P03023.

[3] A. Albertsson *et al.*, "Deep Learning for Calorimeter Energy Reconstruction," *Journal of Instrumentation*, vol. 13, P12003, 2018, doi:10.1088/1748-0221/13/12/P12003.

[4] O. Adriani *et al.*, "Measurement of Forward Neutral Pion Transverse Momentum Spectra for $\sqrt{s} = 7$ TeV Proton-Proton Collisions at LHC," *Physical Review D*, vol. 86, 092001, 2012, doi:10.1103/PhysRevD.86.092001.

[5] O. Adriani *et al.*, "Transverse-Momentum Distribution and Nuclear Modification Factor for Neutral Pions in the Forward-Rapidity Region in Proton-Lead Collisions at $\sqrt{s_{NN}} = 5.02$ TeV," *Physical Review C*, vol. 89, no. 6, 065209, 2014, doi:10.1103/PhysRevC.89.065209.

[6] O. Adriani *et al.*, "Measurements of Longitudinal and Transverse Momentum Distributions for Neutral Pions in the Forward-Rapidity Region with the LHCf Detector," *Physical Review D*, vol. 94, no. 3, 032007, 2016, doi:10.1103/PhysRevD.94.032007.

[7] O. Adriani *et al.*, "Measurement of the Forward η Meson Production Rate in p-p Collisions at $\sqrt{s} = 13$ TeV with the LHCf-Arm2 Detector," *Journal of High Energy Physics*, vol. 2023, no. 169, 2023, doi:10.1007/JHEP10(2023)169.

[8] O. Adriani *et al.*, "LHCf - Technical Proposal for the LHC Run3," CERN Technical Report CERN-LHCC-2019-008; LHCC-P-014, 2019.

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016, doi:10.1145/2939672.2939785.

[10] L. Prokhorenkova *et al.*, "CatBoost: Unbiased Boosting with Categorical Features," in *Advances in Neural Information Processing Systems*, vol. 31, pp. 6638–6648, 2018, doi:10.48550/arXiv.1706.09516.

Full Authors List: LHCf Collaboration

O. Adriani[1,2], E. Berti[2], P. Betti[1,2] L. Bonechi[2], M. Bongi[1,2], R. D'Alessandro[1,2], S. Detti[2], E. Gensini[1,2], M. Haguenauer[3], C. Isseverc[4], Y. Itow[5,6], K. Kasahara[7], K. Kinoshita[5], H. Kobayashi[5], C. Leitgeb[4], Y. Matsubara[8], H. Menjo[5], Y. Muraki[5], P. Papini[2], G. Piparo[9,10] S. Ricciarini[2,11], T. Sako[12], M. Sakuma[5], N. Sakurai[13], M. Scaringella[2], Y. Shimizu[14], T. Tamura[14], A. Tiberio[1,2], S. Torii[15], A. Tricomi[9,10,16], W. C. Turner[17], and K. Yoshida[7]

[1] University of Firenze, Firenze, Italy
[2] INFN Section of Firenze, Firenze, Italy
[3] Ecole-Polytechnique, Palaiseau, France
[4] Humboldt University of Berlin, Berlin, Germany
[5] Institute for Space-Earth Environmental Research, Nagoya University, Nagoya, Japan
[6] Kobayashi-Maskawa Institute for the Origin of Particles and the Universe, Nagoya University, Nagoya, Japan
[7] Faculty of System Engineering, Shibaura Institute of Technology, Japan
[8] Center for Muon Science and Technology, Chubu University, Kasugai, Japan
[9] University of Catania, Catania, Italy
[10] INFN Section of Catania, Catania, Italy
[11] IFAC-CNR - Florence, Italy
[12] Institute for Cosmic Ray Research, University of Tokyo, Kashiwa, Japan
[13] Faculty of Design Technology, Osaka Sangyo University, Daito City, Japan
[14] Kanagawa University, Kanagawa, Japan
[15] RISE, Waseda University, Shinjuku, Tokyo, Japan
[16] CSFNSM, Catania, Italy
[17] LBNL, Berkeley, California, USA