EPJ Web of Conferences **337**, 01194 (2025) https://doi.org/10.1051/epjconf/202533701194
CHEP 2024

Di-Muon cocktail reconstruction using Machine Learning technique in CBM experiment at FAIR

Pawan Kumar Sharma[1,2,*], *Raktim* Mukerjee[1,], *Abhishek* Sharma[3,], *Apar* Agarwal[1,], *Partha Pratim* Bhaduri[1,], *Anand Kumar* Dubey[1,], and *Subahashish* Chattopdhyay[1,]

[1]Variable Energy Cyclotron Centre (VECC) Kolkata, India
[2]Homi Bhabha National Institute (HBNI) Mumbai, India
[3]Aligarh Muslim University, Aligarh, India

Abstract. The CBM experiment at FAIR will investigate strongly interacting matter at high baryon density and moderate temperature. One of proposed key observables is the measurement of the low mass vector mesons (LMVMs), which can be detected via their di-lepton decay channel. Di-leptons are clean probes for hot and dense matter, since they only interact electromagnetically, they escape the medium nearly unperturbed, thus allowing unique access to the properties of the medium. The Muon Chamber (MuCh) detector system is being built to identify the muon pairs originating from the direct and Dalitz decay of low mass vector mesons, in a background mostly populated by muons from weak decay of pions and kaons produced in the collisions. In the future, CBM experiment at FAIR, will add to existing experimental data with new results from the intermediate energy range, probing the di-lepton emission from the high net baryon density region.

We report, simulation results for the reconstruction of di-muon spectra for 8AGeV , where A is the mass number of the nucleus being accelerated, central AuAu collisions using machine learning (ML) techniques for selection of muon track candidates. Various ML algorithms like Gradient boosted decision trees (BDTG), K-Nearest neighbour (KNN), Multi-layer Preceptron (MLP), HMatrix etc. from the TMVA class have been employed for the present study. The results from different ML models have been compared with the traditional selection manual cuts based method for reconstruction of omega (ω), eta (η), phi (ϕ) mesons and full freeze-out di-muon cocktail spectra and improvements in di-muon performance with ML classifiers are reported. For comparable S/B ratio, the pair reconstruction efficiency and significance is observed to be increased significantly for omega (ω), eta (η), phi (ϕ) mesons using ML techniques.

*e-mail: pk.sharma@vecc.gov.in

1 Introduction

The Compressed Baryonic Matter (CBM) experiment at FAIR is designed to investigate the QCD phase diagram at high baryon densities with interaction rates up to 10 MHz using triggerless free-streaming data acquisition system[1]. One of the proposed key observables is the measurement of the Low Mass Vector Mesons (LMVM), i.e. ρ, ω, ϕ, which can be detected via their di-lepton decay channel. As the decayed leptons leave the dense and hot fireball without further interactions, they can provide unscathed information about the fireball produced in energetic nuclear collisions. Measurement of vector mesons decaying into lepton pairs require efficient background suppression, high interaction rates to have sufficient statistics and good pair reconstruction efficiency. In order to select events containing the rare observables, the tracks of each collision have to be reconstructed and filtered online with respect to their physical signatures.

Fig. 1 depicts the CBM experimental setup employing a suite of specialized detector subsystems such as Silicon Tracking System (STS), Muon Chamber (MuCh), Transition Radiation Detector (TRD), Time of Flight (ToF) etc to measure particles produced in heavy-ion collisions. These subsystems work together to track particles, identify hadrons, and reconstruct collision events under extreme conditions. The experimental challenge for muon measurements in heavy-ion collisions at FAIR energies is to identify low-momentum muons in an environment of high particle densities. The CBM concept is to track the particles through a hadron absorber system, and to perform a momentum-dependent muon identification. This concept is realized by segmenting the hadron absorber in several layers, and placing triplets of tracking detector planes in the gaps between the absorber layers. The MuCh detector system is placed downstream of the STS.

The experimental challenge in di-lepton measurements is to suppress the huge combinatorial background of uncorrelated lepton pairs. In the case of muon measurements, the muon background is generated by weak decays of pions and kaons, by mismatches of hadrons upstream and muons downstream the hadron absorber, and by hadrons punching through the absorber. In the beam energy range between 2 and 40 A GeV, no di-leptons have been measured so far, in heavy ion collisions. The CBM collaboration will systematically measure both di-electrons and di-muons in p+p, p+A and A+A collisions as function of beam energy and size of the collision system. The di-electron and di-muon high-precision data will complement each other, and will provide a complete picture on di-lepton radiation of dense baryonic matter.

In this report, we discuss the reconstruction of di-muon spectra originating from the direct and Dalitz decay of low mass vector mesons for 8 A GeV central Au-Au collisions using machine learning techniques as available within ROOT software for selection of muon track candidates. We compared the results from various machine learning models with manual selection cuts for reconstruction of omega (ω), eta (η), phi (ϕ) mesons and for full di-muon cocktail spectra.

We check how optimal some of the most common algorithms for signal-background discrimination, such as Boosted Decision Trees, Neural Networks. We compared the overall performance of different algorithms in simulated CBM data. Gradient Boosted Decision Trees (BDTG) are an advanced form of boosted decision trees where each new tree is trained to correct the errors of the previous trees using gradient descent. Instead of just focusing on misclassified points, It minimize a loss function by adding trees that predict the gradients.

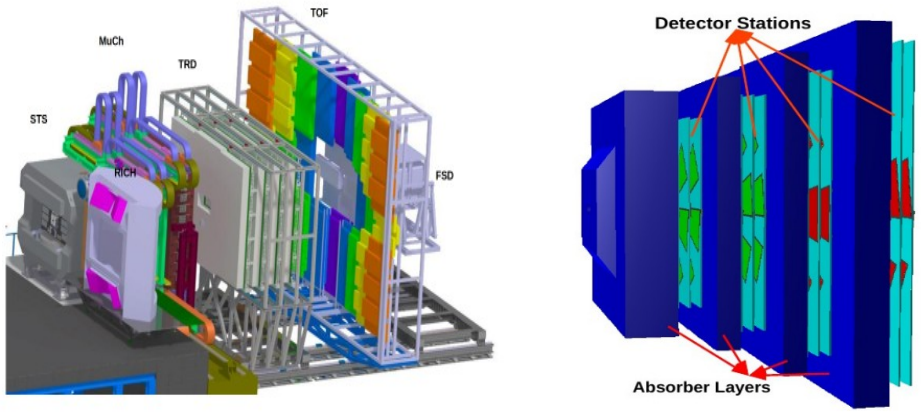

Figure 1. CBM experimental setup having various detector subsystem (left panel) and Schematics of muon chamber setup (right panel)

2 Simulation procedure

Using the latest version of muon chamber (MuCh) detector geometry, we have attempted to reconstruct LMVM (ω, η, ϕ) in the event by event mode following standard CBM reconstruction software, CBMROOT [2]. One event means one Au+Au collision, after analyzing complete particle information from one collisions, particles from next event are processed. Background of central Au-Au collisions at 8 AGeV was generated using UrQMD [3] event generator, whereas for low mass vector mesons signals PLUTO [4] event generator was used. Single signal meson decaying into $\mu^+ + \mu^-$ was embedded into each background event. The particles are then transported through the CBM muon detector setup using the GEANT3 transport engine. The energy deposited in the active volume of the detectors is converted to digital signals (digis), which when clustered give hits. Hits from various detector subsystem are used for the track reconstruction. Muon track candidates can be selected from the manual cuts or using ML techniques. Oppositely charged muon track candidates are used for calculation of di-muon invariant mass spectra. The full freeze-out di-muon cocktail spectra include the di-muon decay channel of ω, η, ρ, ϕ and dalitz decay of omega ($\omega \rightarrow \mu^+ + \mu^- + \pi^0$) , eta ($\eta \rightarrow \mu^+ + \mu^- + \gamma$) leading to continuum in pair mass distribution.

3 Machine learning techniques

The motivation behind using Machine learning (ML) is to improve the di-muon reconstruction performance over the existing procedure based on manual selection cuts. ML can be a very useful tool for identifying muon tracks coming through decay of LMVMs.

Each machine learning model has unique characteristics and performs differently on varying datasets. In our case, BDTG outperforms other models. Despite the rise of deep learning models like MLPs, tree-based models remain highly successful in handling tabular data [5]. Decision trees are built using a top-down, recursive partitioning approach, where the dataset is split into subsets until a stopping criterion is met. Decision trees are a compelling choice due to their simplicity, interpretability, and flexibility. Unlike linear models, decision trees can capture non-linear relationships without assuming specific data distributions or requiring feature scaling. They handle both numerical and categorical features well, are robust

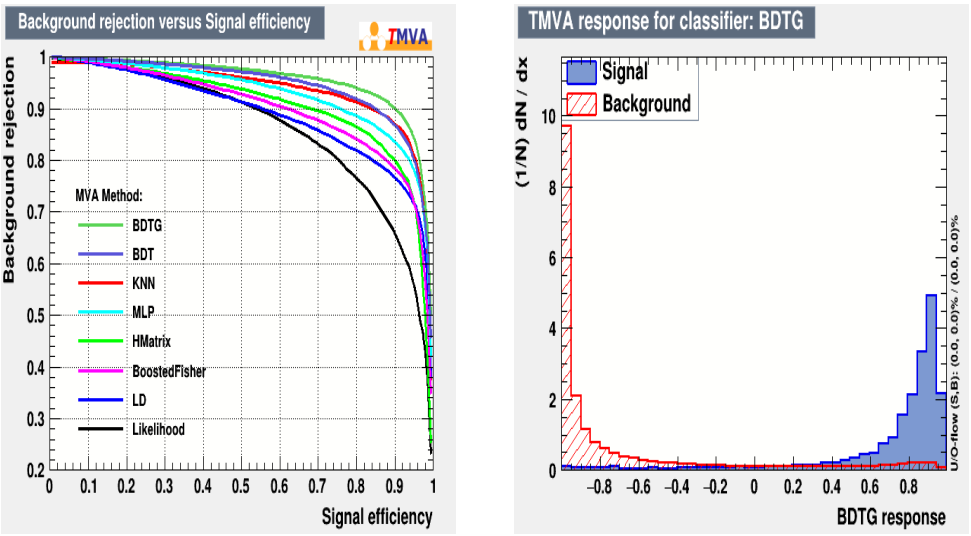

Figure 2. Background rejection versus signal efficiency plot or ROC curve for different models(left panel) and Separation of signal & background using BDTG model for cocktail data (right panel)

to outliers, and can deal with missing data effectively. Boosted decision trees (BDTs) have gained popularity due to these characteristics, as well as their robustness. Various boosting techniques, such as adaptive boosting [6] and gradient boosting [7], along with XGBoost [8], have made multivariate models widely applicable. By focusing on feature splits that maximize information gain, decision trees perform well on high-dimensional datasets and inherently capture feature interactions. These attributes have made them extensively utilized in domains like high-energy physics, where intricate tabular data is abundant, demonstrating remarkable effectiveness [9] [10] [11]. Decision trees also serve as the foundation for ensemble methods like Random Forests and Gradient Boosted Trees, which further improve generalization and reduce overfitting. However, standalone decision trees can overfit, which is why ensembles are often used to enhance performance. Random Forest and Boosting are ensemble methods that use multiple decision trees to improve predictive accuracy. Boosting is a way of enhancing the classification and regression performance (and increasing the stability with respect to statistical fluctuations in the training sample) of typically weak MVA methods by sequentially applying an MVA algorithm to reweighted (boosted) versions of the training data and then taking a weighted majority vote of the sequence of MVA algorithms thus produced.

Various ML algorithms like Gradient Boosted Decision Trees (BDTG), K-nearest Neighbor (KNN), Multi-Layer Perceptron (MLP), HMatrix etc. from the TMVA class have been employed for the present study. Hits from various detector subsystems of CBM are used for the track reconstruction. Tracks reconstructed in STS are propagated using the Kalman Filter technique to pass through the MuCh layers. MuCh hits located around the propagation point are taken as the candidates of the propagated track. For final analysis, Chi^2 of track fitting of detector subsystem, number of STS, TRD, TOF, and MuCh layers associated with the propagated tracks are taken. These ML models are trained using variables such as number of MuCh Hits, number of STS Hits, number of ToF Hits, number of TRD Hits, χ^2_{MuCh}, χ^2_{Vertex}, χ^2_{STS}, ToF momentum and ToF mass associated with a global track in order to find suitable muon track candidates. The Receiver Operating Characteristic (ROC) curve is a fundamental

tool for evaluating the performance of binary classification models in machine learning (fig.2 (left panel)). Di-muon cocktail data include contributions from direct and/or Dalitz decays of η, ρ, ω, ϕ mesons. Among all classifiers, BDTG has been found to perform best. Fig.2 (right panel) shows separation of signal and background using BDTG classifier as the function of the classifier output response or score.

4 Result and Analysis

Traditionally, in previous CBM muon simulations [12] the following manual selection cuts are used to filter suitable muon track candidates for calculation of di-muon inavariant mass distribution: Number of MUCH hits ≥ 11, number of STS hits ≥7, number of TRD hits≥1, number of ToF hits ≥1, $\chi^2_{MuCh} \leq 3$, $\chi^2_{STS} \leq 2$, $\chi^2_{Vertex} \leq 3$, $\sigma_{ToF} = 2$, applied in sequence. In order to compare the reconstruction performance, for ML based analysis, we have used a single cut on ML classifier score or output variable to obtain di-muon invariant mass spectra for the two body decay of resonances. For each case the combinatorial background is calculated using super event (SE) analysis [13]. In this method, one muon track candidate is combined with all the other oppositely charged muon track candidates of all the events, to calculate the uncorrelated invariant mass distribution. The advantage of using the super event technique is that due to the large combination of the muon pairs, the statistical uncertainties associated with the large mass bins are rather small.

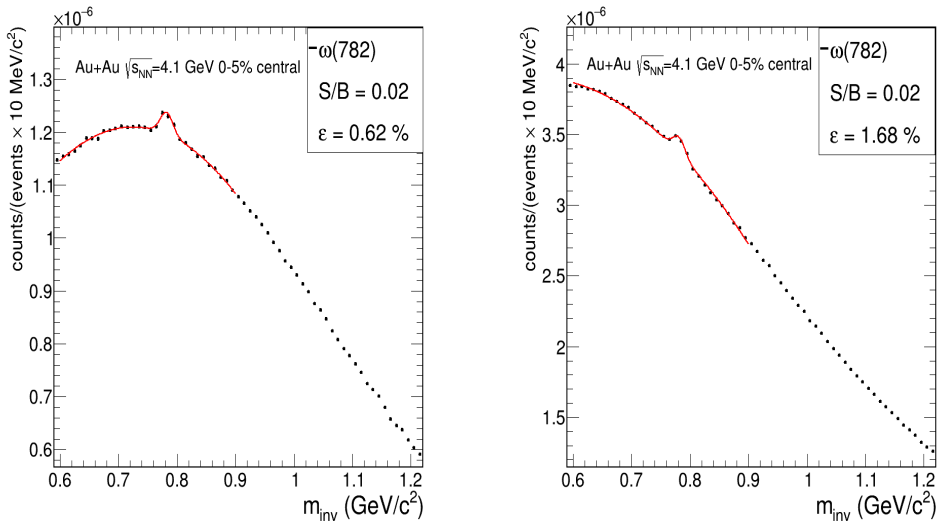

Figure 3. Invariant mass spectra of $\omega \rightarrow \mu^+ + \mu^-$ meson using manual selection cuts (left panel) and using BDTG model with score cut at 0.7 (right panel)

Fig. 3 (left panel) shows the invariant mass distribution of omega($\omega \rightarrow \mu^+\mu^-$) for Au+Au central collision at 8AGeV beam energy in which muon track candidates were selected via manual selection cuts, and fig. 3 (right panel) is the invariant mass distribution of omega where muon track candidates were selected with a single cut on the BDTG score as 0.7. For similar S/B ratio (0.02), reconstructed efficiency increases 2.71 times for BDTG classifier case while comparing to manual selection cuts. The cut on BDTG score 0.7 is chosen as to

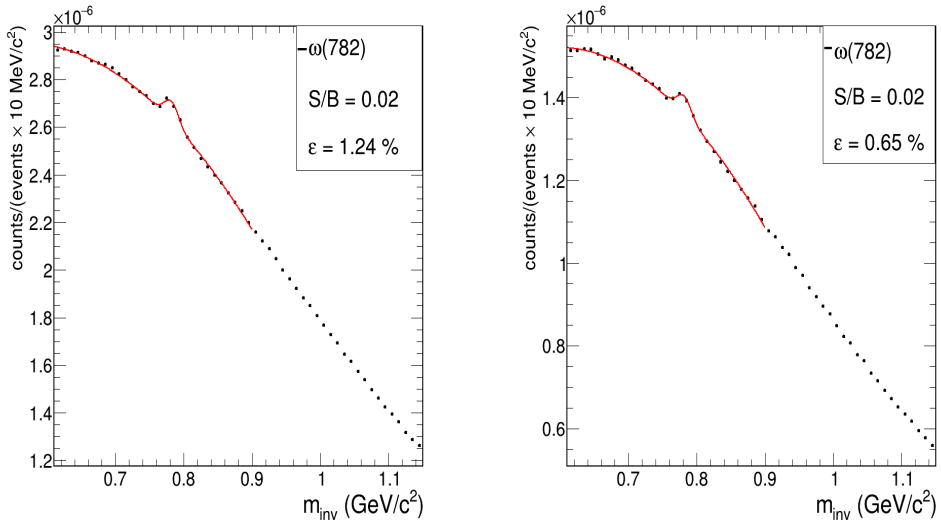

Figure 4. Invariant mass spectra of $\omega \rightarrow \mu^+ + \mu^-$ meson using KNN with score cut at 0.88 (left panel) and using HMatrix model with score cut at 0.22 (right panel)

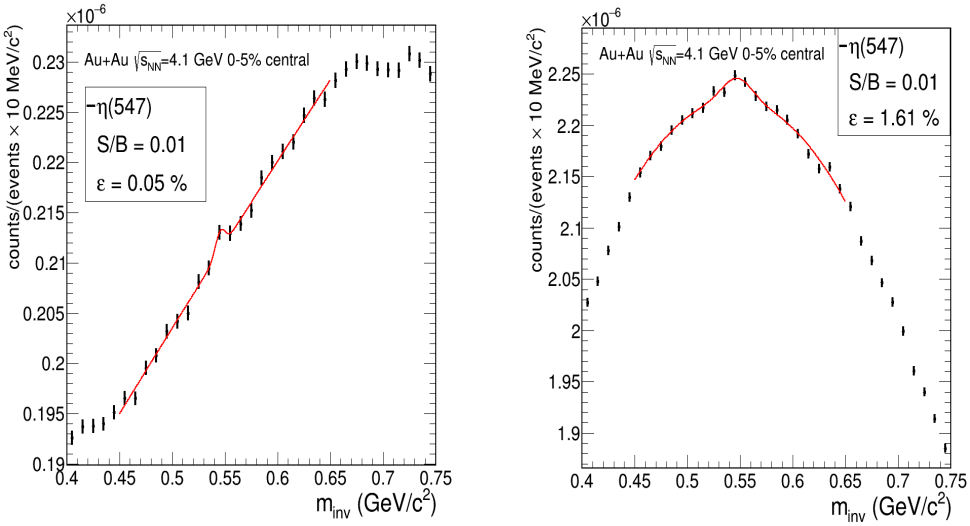

Figure 5. Invariant mass spectra of $\eta \rightarrow \mu^+ + \mu^-$ meson using manual selection cuts (left panel) and using BDTG model with score cut at 0.7 (right panel)

comparing the performance of machine learning classifier with manual cut-based method for similar S/B (signal-background) ratio. The reconstructed efficiency of meson increases with the ML classifier score but S/B ratio decreases. The S/B ratio is calculated using a combined fit to the invariant mass spectrum, which models signal as a Gaussian and the background as a second-order polynomial.

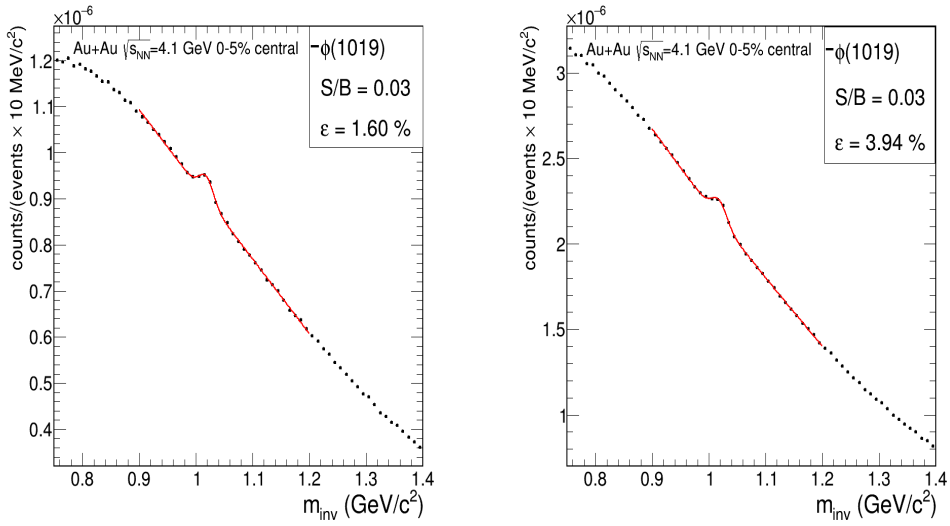

Figure 6. Invariant mass spectra of $\phi \rightarrow \mu^+ + \mu^-$ meson using manual selection cuts (left panel) and using BDTG model with score cut at 0.7 (right panel)

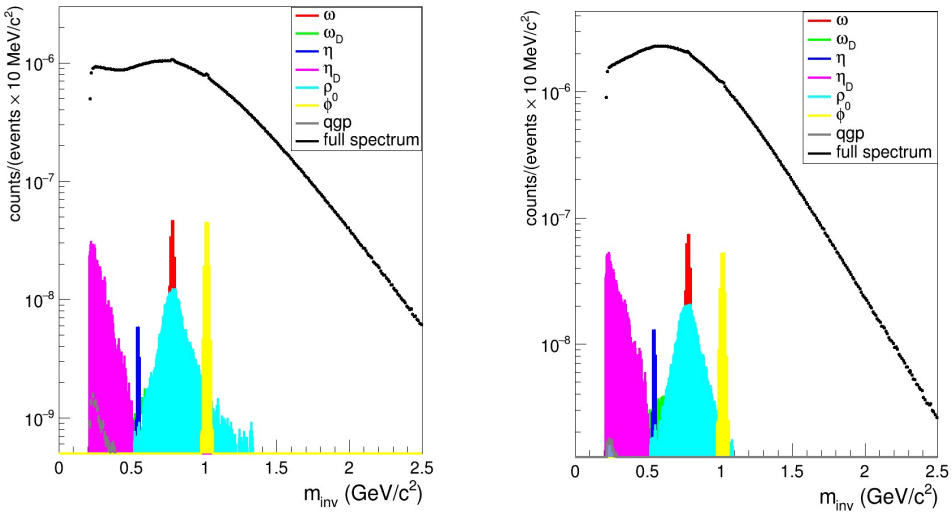

Figure 7. Invariant mass spectra of full di-muon cocktail using manual selection cuts (left panel) and using BDTG model with score cut at 0.7 (right panel)

Fig. 4 shows the invariant mass spectra of ω meson using KNN model with score cut at 0.88 (left panel) and using HMatrix model with score cut at 0.22 (right panel). Among the ML classifiers, the improvement in reconstructed efficiency for similar S/B ratio, is better with BDTG than HMatrix and KNN classifiers. Fig. 5, 6, 7 shows the invariant mass spectra of η, ϕ and full freeze-out di-muon cocktail respectively, using manual cuts (left panel) and using BDTG classifier (right panel). The improvement in di-muon performance using ML al-

Particle	method	S/B ratio	Efficiency (%)	Normalized Significance
ω	manual cuts	0.02	0.62	1
ω	BDTG@0.7	0.02	1.68	1.83
ω	KNN@0.88	0.02	1.24	1.57
ω	HMatrix@0.22	0.02	0.65	1.23
η	manual cuts	0.006	0.05	1
η	BDTG@0.7	0.006	1.61	2.12
ϕ	manual cuts	0.03	1.60	1
ϕ	BDTG@0.7	0.03	3.94	1.52

Table 1. Comparison of ML with manual cuts for LMVMs.

gorithm can be seen from the above figures and is summarized in table 1. The significance is normalized with the significance calculated using the manual cut-based method. For a comparable S/B ratio, the efficiency and significance of pair reconstruction is seen to increase significantly for ω, η, ϕ mesons. The full spectrum of di-muon cocktail sums the di-muon decay channel of ω, η, ρ, ϕ and dalitz decay of omega ($\omega \rightarrow \mu^+ + \mu^- + \pi^0$) , eta ($\eta \rightarrow \mu^+ + \mu^- + \gamma$), thermal radiation from QGP and combinatorial background.
Work is under progress for further optimization of the ML procedure for reconstruction of the di-muon continuum contributions from both from freeze-out cocktail and the thermal radiations from a hot and dense fireball. There are also plans to carry out muon simulations for CBM using the GEANT4 transport engine, as it offers better handling of hadronic interactions. However, in the current study, GEANT3 was used to reduce computational time, as the primary goal was a comparative study between a manual cut-based selection method and an ML-based approach.

References

[1] V. Klochkov, The compressed baryonic matter experiment at fair, Nuclear Physics A **1005**, 121945 (2021), the 28th International Conference on Ultra-relativistic Nucleus-Nucleus Collisions: Quark Matter 2019. https://doi.org/10.1016/j.nuclphysa.2020.121945

[2] V. Friese, for the CBM Collaboration, The high-rate data challenge: computing for the cbm experiment, Journal of Physics: Conference Series **898**, 112003 (2017). 10.1088/1742-6596/898/11/112003

[3] M. Bleicher et al., Relativistic hadron hadron collisions in the ultrarelativistic quantum molecular dynamics model, J. Phys. G **25**, 1859 (1999), hep-ph/9909407. 10.1088/0954-3899/25/9/308

[4] I. Frohlich, T. Galatyuk, R. Holzmann, J. Markert, B. Ramstein, P. Salabura, J. Stroth, Design of the Pluto Event Generator, J. Phys. Conf. Ser. **219**, 032039 (2010), 0905.2568. 10.1088/1742-6596/219/3/032039

[5] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, in *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Curran Associates, Inc., 2022), Vol. 35, pp. 507–520, https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf

[6] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences **55**, 119 (1997). https://doi.org/10.1006/jcss.1997.1504

[7] Y. Lou, M. Obukhov, BDT: Gradient Boosted Decision Tables for High Accuracy and Scoring Efficiency, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, USA, 2017), KDD '17, p. 1893–1901, ISBN 9781450348874, https://doi.org/10.1145/3097983.3098175

[8] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, CoRR **abs/1603.02754** (2016), 1603.02754.

[9] P. Englert, Improved Precision in $Vh(\to b\bar{b})$ via Boosted Decision Trees (2024), 2407.21239.

[10] Y. Rana, A.K. Dubey, An Analytical Comparison Among Various Multivariate Methods Used for Particle Discrimination, Springer Proc. Phys. **304**, 974 (2024). 10.1007/978-981-97-0289-3_258

[11] Coadou, Yann, Boosted decision trees and applications, EPJ Web of Conferences **55**, 02004 (2013). 10.1051/epjconf/20135502004

[12] 228172, Tech. Rep. CBM Progress Report 2019, Darmstadt (2020), https://repository.gsi.de/record/228172

[13] P. Senger, V. Friese (CBM Collaboration), Tech. Rep. CBM Progress Report 2021, Darmstadt (2022), https://repository.gsi.de/record/246663