EPJ Web of Conferences **295**, 09027 (2024) https://doi.org/10.1051/epjconf/202429509027
CHEP 2023

Muon and Pion Identification at BESIII Based on Machine Learning Algorithm

Yuncong Zhai[1], *Teng* Li[1,*], *Xingtao* Huang[1,**], and *Na* Yin[1]

[1]Institute of Frontier and Interdisciplinary Science, Shandong University, Qingdao, China

Abstract. BESIII is designed to study physics in the τ-charm energy region utilizing the high luminosity BEPCII. For collision physics experiments like the BESIII experiment, particle identification (PID) is one of the most important and commonly used tools for physics analysis. The effective μ/π identification performance is of great significance for most of BESIII physics analysis. However, due to the close masses of these two particles, as well as the intrinsic correlation between multiple detector information, traditional methods at BESIII is facing challenges in μ/π identification. In recent decades, machine learning (ML) techniques have been rapidly developed and have shown successful applications in HEP experiments. The PID based on ML provides powerful capability of combining more detection information from all sub-detectors with the data-driven approach. In this article, targeting at the μ/π identification problem at the BESIII experiment, we have developed a new PID algorithm based on the gradient boosted decision tree (BDT) model. Preliminary results show that the XGBoost classifier provides obviously higher discrimination power than traditional methods. In addition, based on the substantial amount of high-quality data taken by the BESIII detector, a method of evaluating and suppressing the systematical error of the ML model is also introduced, which is critical for applying the model to physics studies.

Keywords: Particle Identification; Machine Learning; Boosted Decision Tree; BESIII

1 Introduction

The Beijing Spectrometer III (BESIII) [1] is a HEP detector running in the τ-charm energy region at the Beijing Electron-Positron Collider II (BEPCII) [2]. BESIII is aimed to study physics topics such as the quantum chromodynamics (QCD), the weak interaction, and properties of quarks and gluons, etc.

Particle identification (PID) [3], i.e., the ability to discriminate between different particle species produced during the collision, plays a crucial role in most BESIII physics analysis. Excellent PID capability is often the crucial item to be considered in most BESIII physics analysis. Good μ/π separation is required for precision f_D/f_{D_s} measurements, where f_D and f_{D_s} are the decay constant of D and D_s mesons. Excellent electron identification will help to improve the accuracy of the CKM elements $|V_{cs}|$ and $|V_{cd}|$. The identification of hadronic ($\pi/K/p$) particles is a commonly used tool in BESIII physics analysis. Due to the intrinsic

*e-mail: tengli@sdu.edu.cn
**e-mail: huangxt@sdu.edu.cn

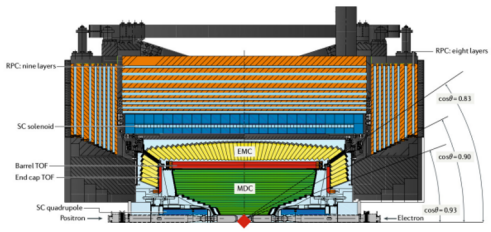


Figure 1: The BESIII spectrometer layout

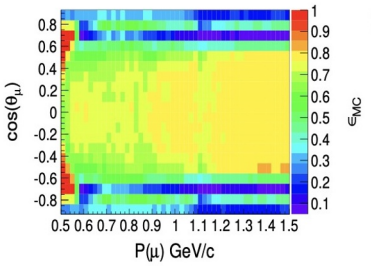


Figure 2: The traditional μ particle identification efficiency in the BESIII experiment

correlations between variables, traditional PID methods, such as the the maximum likelihood method [3, 4] is facing big challenges in complex PID tasks, since the μ/π separation power decreases in certain angle and momentum ranges at BESIII (as shown in Figure 2).

With the rapid development of artificial intelligence and machine learning (ML) techniques, they are becoming increasingly popular in the HEP field over the past decades. In HEP experiments, ML techniques are opening up new possibilities and have shown promising results. For the PID problem, a number of successful examples have emerged over the last decade, usually with higher accuracy and efficiency compared with traditional methods [5–7]. Among these examples, one of the most popular approaches is the BDT model which is able to provide good performance.

In this paper, a μ/π classifier is developed for the BESIII experiment based on the BDT model implemented with XGBoost library. The rest of this paper is organized as following. In Section 2, a brief introduction to the basic methodology in this work is presented. In Section 3, we present the data sample used in this work, as well as the feature selection and model tuning processes. Section 4 shows the performance of the classifier, as well as discusses the way on how the systematical errors are evaluated and suppressed, which is critical for applying the model to physics studies. Finally, Section 5 gives the summary and outlook of this work.

2 Methodology

2.1 Algorithms and models for machine learning

2.1.1 Boosted decision tree

Decision tree [8] is a ML model based on a hierarchical tree structure, mostly used to solve regression or classification problems. Each decision tree consists of the root node, the decision nodes, branches, splitting criterion as basic elements [9, 10]. The basic working principle of decision trees is testing features at the root node then assigning the data to a child node recursively. Since the feature engineering is crucial for the performance, the information Gain or Gini Index are commonly used as criteria for selecting input features. Information Gain can be expressed as equation 1 [11]:

$$Gain(A) = Info(A) - Info\_A(D). \tag{1}$$

Ross Quinlan's ID3 and C4.5 [12] are examples of information Gain-based calculations.

Gini Index is a metric used to measure of impurity/purity of an attribute in a dataset. It can be calculated by using the following formula [10]:

$$Gini\ Index = 1 - \sum_{j=1}^{n} p_j^2, \tag{2}$$

the p_j denotes the probability and it is utilized in the process of constructing a decision tree through the employment of the classification and regression tree (CART) [13] supervised learning algorithm.

The Boosting algorithm [14] is an integrated method used to improve the accuracy of learning algorithms. The basic idea is to integrate multiple weak classifiers into a strong classifier according to a certain weighting ratio, which can significantly improve the accuracy of the final results.

2.1.2 Extreme gradient boosting decision tree

Boosting is a broad family of algorithms that includes Adaboost, XGBoost (one of the BDT implementations), and others. XGBoost is an optimized and efficient algorithm that is known for its speed, scalability, and ability to handle large dataset [15–17].

XGBoost was developed and launched by Tianqi Chen et al. in 2016 [18], which is a tree integration model that uses CART as the base learner. XGBoost enhances the traditional gradient boosting algorithm by incorporating regularization techniques, parallel processing, and tree pruning to improve performance and reduce overfitting. For a dataset containing n samples of m-dimensional features $\mathcal{D} = \{(x_i, y_i)\}$ ($|\mathcal{D}| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}$), the output predicted by XGBoost can be expressed as the following equation:

$$\begin{aligned} \hat{y}_i &= \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}, \\ \mathcal{F} &= f(x) = \omega_{q(x)}, \end{aligned} \tag{3}$$

where f_k represents the k-th CART, and each f_k corresponds to an independent tree structure q. The predicted value of the CART for the sample outcome $\omega_{q(x)}$. In addition, K represents the total number of classification and regression trees set in the model. The objective function to be optimized is :

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t). \tag{4}$$

Formally, y_i represents the predicted value of the i-th instance at the t-th iteration. To minimize the objective function, f_t needs to be added. The regularization term, denoted as $\Omega(f_t)$ in equation 4, is also required. This term is calculated using equation 5:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2. \tag{5}$$

which is used to control the variance of the fit and in turn controls the flexibility of the learning task. By adding a regularization term to smooth the final learning weights, overfitting can be avoided. Here, γ and λ represent the coefficients of the penalty term, T is the total number of leaf nodes for the t-th tree, and ω_j is the output score of the j-th leaf node on the t-th tree, where j ranges from 1 to T.

2.2 Evaluation of systematic error

Given the data sample used to train the classifier is generated based on Monte-Carlo (MC) simulation, the difference between MC data and real data is the major source of systematic errors of the model. Since the systematic errors is critical for physics studies, cross validation between MC and real data is necessary. In this study, we perform a cross validation by collecting the high-quality real data generated by the realistic detector response to estimate the systematic errors, then comparing PID efficiency between the MC and real data test sets. The systematic error is defined as:

$$\Delta\varepsilon = \frac{\varepsilon(Data) - \varepsilon(MC)}{\varepsilon(MC)}, \tag{6}$$

where $\varepsilon(MC)$ and $\varepsilon(Data)$ denote PID efficiencies obtained from the MC and real data test sets, respectively. By evaluating the bias when dealing with different test sets (MC samples or data samples), the generalization ability of the classifier could also be understood.

3 Data sample

3.1 Data sample

To train and validate the classifier, a substantial amount of data samples are selected from MC simulation, as well as from real data collected by the detector. All the data is reconstructed based on the BESIII offline software system (BOSS) [19].

The μ and π single particle MC samples are used as the training set. To reduce potential bias, we ensure that the distribution of momentum and incident angle ($cos(\theta)$) in the single particle MC samples is flat, guaranteeing that the PID performance does not rely on the momentum or incident angle. The momentum of the single particle MC samples ranges from 0.1 to 1.5 GeV/c, while *theta* is from 0.5 to 2.6 rad. The total number of μ/π particles in the MC data set is 128,000 for each type of particle (with a balanced number of $\pi^{\pm}/\mu^{\pm}$ particles). The training, validation and test sets are randomly divided at a ratio of 8:1:1.

Besides the MC data set, we also selected high purity of μ and π particles from real data as the cross validation data set to evaluate the differences between simulated and real data. The μ and π particles are selected from multiple well-studied decay channels, such as $e^+e^- \to J/\psi \to \gamma\mu^+\mu^-$ and $J/\psi \to \pi^+\pi^-\pi^0 \to \pi^+\pi^-\gamma\gamma$. With very strict selection criteria, the purity of the μ and π samples reach 97.97% and 99.37%, respectively. In addition, 100,000 pions with momentum from 0.2 to 0.4 GeV/c and 100,000 muons with momentum from 1.4 to 1.5 GeV/c are selected as the secondary validation dataset based on the decay channel $\psi(2S) \to \pi^+\pi^- J/\psi \to \pi^+\pi^-\mu^+\mu^-$ (the purity of μ and π particles reaches 99.19% with the standard selection criteria).

3.2 Features selection

The BESIII experiment is composed of several sub-detectors, some of which measure particle properties that can be used for identification. Multiple sub-systems of BESIII detector can provide PID information, including a main drift chamber (MDC), a Time-of-Flight counter (TOF), a CsI(Tl) electromagnetic calorimeter (EMC), and a muon counter (MUC), from the inner layer to the outter layer.

- The ionization energy loss of charged tracks per unit pathlength in the MDC (dE/dx).
- The particle travel time measured by TOF detector.

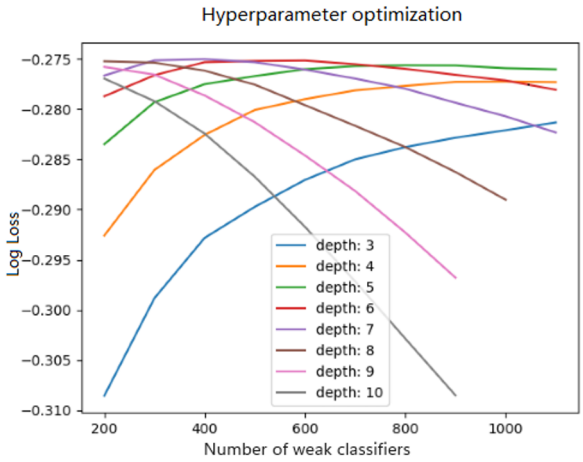
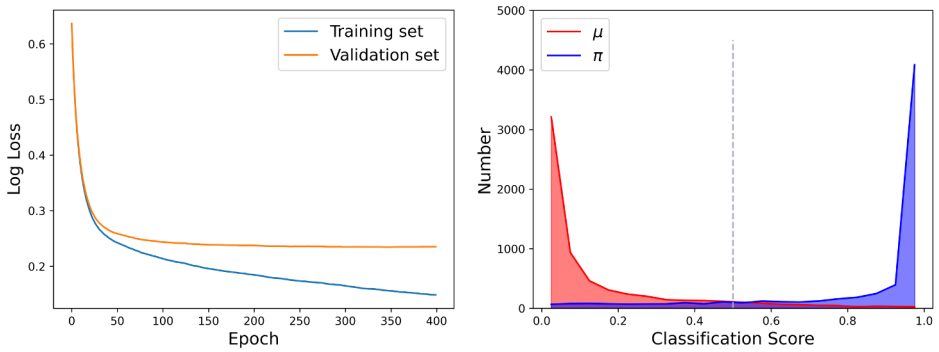


Figure 3: The optimized results of XGBoost classifier hyper-parameters.

- The information of the electromagnetic showers in the EMC, e.g., spatial coordinates of a reconstructed shower, the energy depositions of the seed crystal (E_{seed}), the energy deposition reconstructed by 3×3 crystal array ($E_{3\times3}$) and 5×5 crystal array ($E_{5\times5}$).
- Hit pattern in the MUC.



(a) (b)

Figure 4: (a) The learning curve during the training process. (b) Classification score distribution.

3.3 BDT model tuning

In order to obtain the best combination of the hyper-parameters of the BDT model (number of weak classifiers, $n_{estimators}$ and the max depth of each classifier, max_{depth}), a grid search is performed via a three-fold cross validation [20]. According to the results shown in Figure 3,

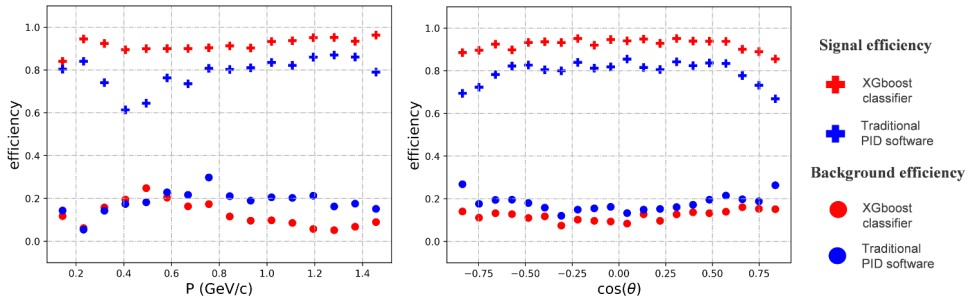
the max_{depth} of 7 and $n_{estimators}$ of 400 are chosen as the optimal combination of the hyper-parameters.

The learning curve of the optimised model is shown in Figure 4a, indicating that the model converges as expected and do not appear to over-fit the training set very much.

4 Performance and error estimation

4.1 PID efficiency

To evaluate the performance of the XGBoost classifier, we compare the signal efficiency and the background efficiency of μ and π particles, with the traditional method developed within the BOSS system. To get a fair comparison, the same MC data test set is used for both XGBoost classifier and the traditional PID software. As shown in Figure 5, it is obvious that the XGBoost classifier outperforms the traditional method in the aspect of signal efficiency and background efficiency for particles in most momentum and incident angle ranges.



(a) (b)

Figure 5: Comparison of the μ signal and background efficiency between the XGBoost classifier and the traditional method. (a) Efficiency as a function of momentum. (b) Efficiency as function of cos(θ) (the incident angle).

4.2 Cross validation between different test sets

To further validate the classifier, a secondary test set (MC and real data samples from different decay channels in Section 3.1) is used as a cross-validation. The μ and π discrimination efficiencies on the MC and real data test sets from the two decay channels are compared. Figure 6 shows the μ and π signal efficiency on the MC data test sets of two decay channels. The difference is low between the two test sets indicates that the classifier is not significantly influenced by variations in the momentum spectrum. It shows that the XGBoost classifier behaves relatively stable across different decay channels and is in particular important as a common tool for different physics studies.

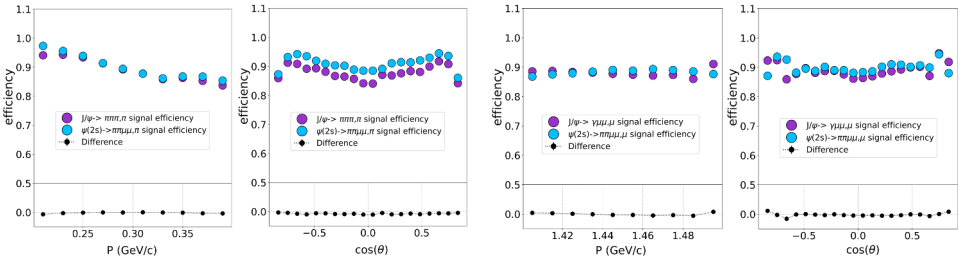
4.3 Cross validation between different MC and real data test sets

To evaluate the systematic error caused by the difference between simulated and real detector response, the performance of the XGBoost classifier on the MC and real data sample is compared. According to the comparison results shown in Figure 7, the discrepancies between the

MC and the real data are within a range of 50%, which is relatively small. This demonstrates that the generalization ability of the classifier is suitable for actual physics studies.

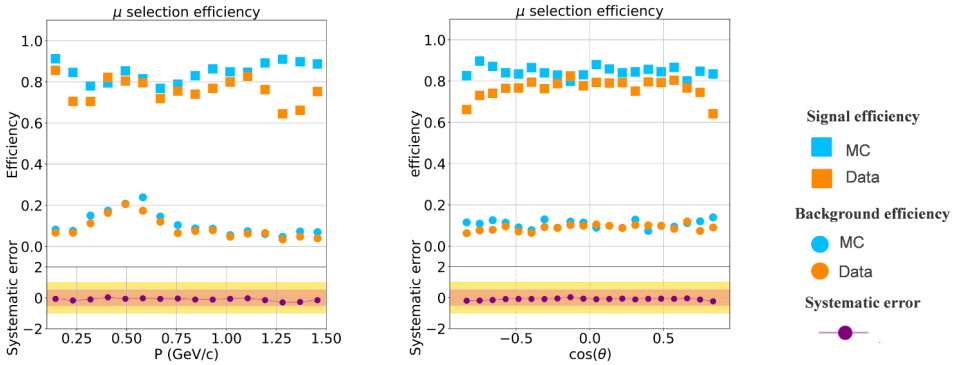
5 Conclusion

In the BESIII experiments, the precise separation of charged particles directly affects the outcomes of our experiments. After considering a large amount of experimental data and the challenges associated with identifying μ/π particles, this work develops a μ and π identification classifier at BESIII based on machine learning algorithm.



(a) (b)

Figure 6: μ/π signal efficiency with the XGBoost classifier of two different test sets. (a) π signal efficiency. (b) μ signal efficiency.



(a) (b)

Figure 7: Error estimation.

In this work, the μ and π classifier based on the BDT model is aimed at boosting performance of various physics studies in the BESIII experiment. To maximize the discrimination power and exploit the physics potential of the BESIII detector, the model utilizes information collected by multiple sub-systems, including the dE/dx measurement from MDC, the time of flight, EMC shower shapes , as well as the response of the muon detector.

By comparing with the traditional maximum likelihood method, we find strong evidence that the BDT model outperforms the traditional method, providing a significantly stronger

discrimination power between μ and π particles. To evaluate the systematic errors of the classifier, high quality control samples are carefully selected from the J/ψ data, and the difference between MC simulation sample and real data is evaluated. The results show that the BDT model has a good generalization ability, making it reliable for physics studies.

6 Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) under Contracts Nos. 2020YFA0406304, 12025502, 12105158, 12188102.

References

[1] M. Ablikim et al. (BESIII), Nucl. Instrum. Meth. A **614**, 345 (2010), 0911.4960
[2] C.A. Zhang, *Performance of the BEPC and progress of the BEPCII*, in *32nd International Conference on High Energy Physics* (2004), pp. 993–997
[3] D.M. Asner et al., Int. J. Mod. Phys. A **24**, S1 (2009), 0809.1869
[4] M. Ablikim, M.N. Achasov, P. Adlarson, S. Ahmed, M. Albrecht, M. Alekseev, A. Amoroso, F.F. An, Q. An, Y. Bai et al., Chinese Physics C **44**, 040001 (2020)
[5] D. Bourilkov, Int. J. Mod. Phys. A **34**, 1930019 (2020), 1912.08245
[6] A.N. Charan, J. Phys. Conf. Ser. **2438**, 012111 (2023), 2301.11654
[7] V. Khachatryan et al. (CMS), JINST **10**, P06005 (2015), 1502.02701
[8] L.A. Badulescu, Proceedings of Annals of University of Craiova (2007)
[9] A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, S.D. Brown, Journal of Chemometrics: A Journal of the Chemometrics Society **18**, 275 (2004)
[10] J.S. Kushwah, A. Kumar, S. Patel, R. Soni, A. Gawande, S. Gupta, Materials Today: Proceedings **56**, 3571 (2022)
[11] T.H. Lee, A. Ullah, R. Wang, Macroeconomic forecasting in the era of big data: Theory and practice pp. 389–429 (2020)
[12] I.D. Mienye, Y. Sun, Z. Wang, Procedia Manufacturing **35**, 698 (2019)
[13] L. Breiman, *Classification and regression trees* (Routledge, 2017)
[14] R.E. Schapire et al., *A brief introduction to boosting*, in *Ijcai* (Citeseer, 1999), Vol. 99, pp. 1401–1406
[15] T. Chen, C. Guestrin, *Xgboost: Reliable large-scale tree boosting system*, in *Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA* (2015), pp. 13–17
[16] S.S. Dhaliwal, A.A. Nahid, R. Abbas, Information **9**, 149 (2018)
[17] W. Chang, Y. Liu, Y. Xiao, X. Yuan, X. Xu, S. Zhang, S. Zhou, Diagnostics **9**, 178 (2019)
[18] T. Chen, C. Guestrin, *Xgboost: A scalable tree boosting system*, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794
[19] W. Li, H. Liu, Z. Deng, K. He, M. He, X. Ji, L. Jiang, H. Li, C. Liu, Q. Ma et al., *THE OFFLINE SOFTWARE FOR THE BESIII EXPERIMENT* (2006), https://api.semanticscholar.org/CorpusID:221378819
[20] Y. Shuai, Y. Zheng, H. Huang, *Hybrid software obsolescence evaluation model based on PCA-SVM-GridSearchCV*, in *2018 IEEE 9th international conference on software engineering and service science (ICSESS)* (IEEE, 2018), pp. 449–453