Machine Learning based Global Particle Identification Algorithms at the LHCb Experiment

Denis Derkach[1,*], *Mikhail* Hushchyn[1,2,**], and *Nikita* Kazeev[1,2,3,***] on behalf of the LHCb collaboration

[1]National Research University Higher School of Economics
[2]Yandex School of Data Analysis
[3]Università degli Studidi Roma "La Sapienza"

Abstract. One of the most important aspects of data processing at flavor physics experiments is the particle identification (PID) algorithm. In LHCb, several different sub-detector systems provide PID information: the Ring Imaging Cherenkov detectors, the hadronic and electromagnetic calorimeters, and the muon chambers. The charged PID based on the sub-detectors response is considered as a machine learning problem solved in different modes: one-vs-rest, one-vs-one and multi-classification, which affect the models training and prediction. To improve charged particle identification for pions, kaons, protons, muons and electrons, neural network and gradient boosting models have been tested. This paper presents these models and their performance evaluated on Run 2 data and simulation samples. A discussion of the performances is also presented.

1 Introduction

Particle identification (PID) plays a crucial role in LHCb [1] analyses. The LHCb PID system is composed of a tracking system, two ring-imaging Cherenkov detectors (RICH), electromagnetic (ECAL) and hadron (HCAL) calorimeters and a series of muon chambers. Combining information from these subdetectors allows one to distinguish between various species of long-lived charged particles. Advanced machine learning techniques are employed to obtain the best PID performance and control systematic uncertainties in a data-driven way.

2 Global Particle Identification

Global PID at LHCb identifies the charged particle type associated with a given track. There are six PID hypotheses: electron, muon, pion, kaon, proton, and ghost track. Ghost tracks are charged tracks that do not correspond to real particles which passed through the detector, they are errors of the tracking algorithm. Different particle types have different responses in the LHCb subdetectors as illustrated in Fig 1.

*e-mail: Denis.Derkach@cern.ch
**e-mail: mikhail.hushchyn@cern.ch
***e-mail: nikita.kazeev@cern.ch

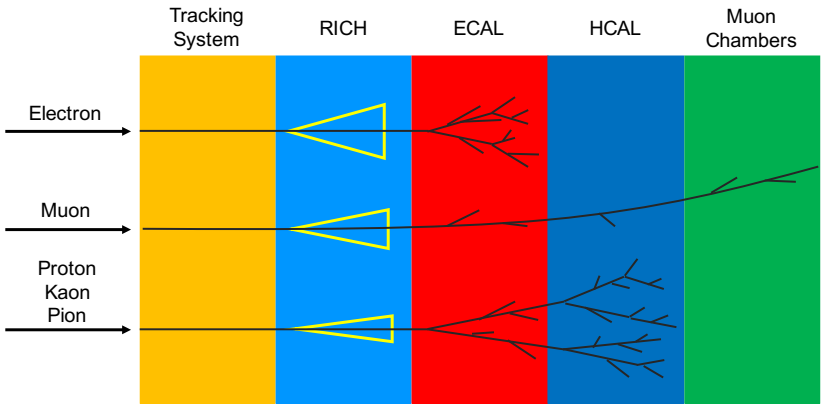

Figure 1. Illustration of different particle type responses in the LHCb systems.

Global PID is a multiclassification problem in machine learning. Information from the LHCb tracking system, RICHs, calorimeters and muon chambers are used as inputs for classifiers to estimate a track type. Several different classifiers are considered. ProbNN [1] (baseline) is Global PID algorithm currently used at LHCb. In this paper it is taken as a baseline solution. ProbNN is based on six binary one-layer artificial neural networks. Each of these networks corresponds to one particle type and is trained to separate this particle type from all others. ProbNN uses neural networks implemented in the TMVA library.

Deep NN is a Global PID model based on deep neural network of Keras library [2]. The network has three hidden layers with 300, 300 and 400 neutrons in each layer, respectively. It is trained in multiclassification mode to separate all six particle types in the same time.

CatBoost [3] is a gradient boosting over oblivious decision trees classifier used for PID. As well as the ProbNN model, there are six CatBoost classifiers to separate one particle type from all others.

There are 60 observables from the LHCb systems available for PID models. CatBoost and Deep NN models use all these observables as inputs. ProbNN takes different subsets of the observables for each of the PID hypotheses. These subsets are selected based on different physical reasons.

The classifiers are trained on a MC sample containing all of the different charged particle types. The sample has one million labeled tracks for each particle type. Calibration samples, containing particles that can be identified based only on kinematic properties, are used to estimate the classifier performance on real data. The samples contain the following decays:

- $J/\psi \to \mu^+\mu^-$,
- $B^+ \to J/\psi(e^+e^-)K^+$,
- $D^{*+} \to D^0(K^-\pi^+)\pi^+$,
- $\Lambda^0 \to p\pi^-$.

The quality of the Baseline (ProbNN), Deep NN and CatBoost models are compared for the following particle pairs: μ-vs-π, K-vs-π, p-vs-π, p-vs-K, e-vs-π and e-vs-K. The first particle in a pair is considered as signal, the second is considered as background. The PID performance of each classifier is shown in Fig 2.

Electrons have very different responses in the RICHs compared with other particles and have no responses in a hadron calorimeter and a muon system. In result, all PID models demonstrate the best PID quality for electrons as shown in Fig 2. Using all observables helps CatBoost and Deep NN models to achieve better results than for the ProbNN model. Pions,

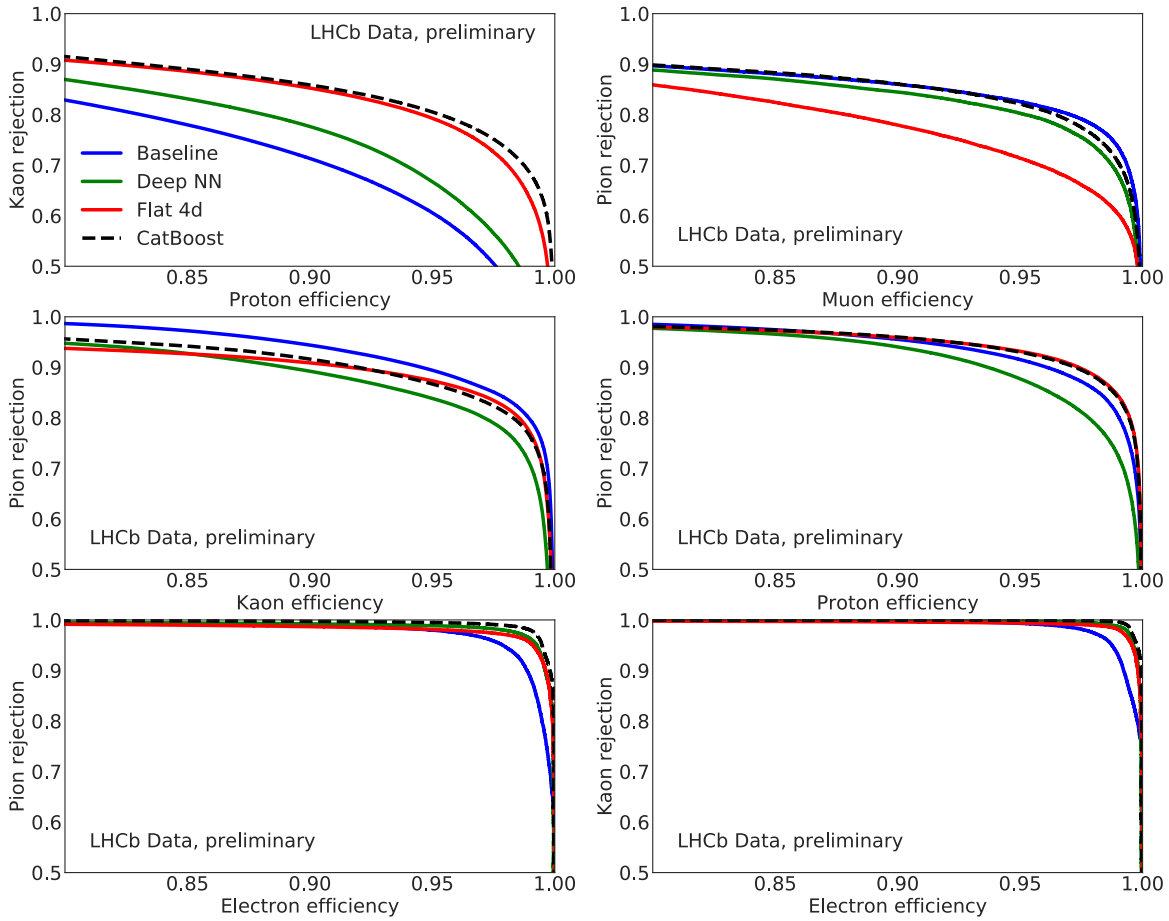

Figure 2. Dependences between background rejection and signal efficiency for six particle pairs.

protons and kaons have similar responses in the calorimeters and separated mainly based on the RICHs. All three PID models have similar results for K-vs-π and p-vs-π pairs. Pions and muons are hardly distinguished by the RICHs but have very different responses in the calorimeters and the muon system. This makes quality of all PID models similar to each other for μ-vs-π pair. Separation of kaons and protons is the most difficult. They have similar responses in all systems and require all information from the RICHs, the calorimeters and the muon system for separation. This allows CatBoost and Deep NN models to outperform the ProbNN model for p-vs-K pair.

3 Flat PID Model

The PID information strongly depends on the kinematic variables. This relationship leads to strong dependency between PID efficiency and kinematic variables as shown in Fig 3. To reduce the systematic uncertainty arising from the use of PID efficiencies in certain physics measurements, it is also beneficial to achieve a flat dependency between efficiencies and spectator variables such as particle momentum. For this purpose, the Flat 4d model based on the boosted decision trees that guarantee the flatness property for efficiencies have also been developed. Relative to the ProbNN model, the Flat 4d model has a flatter PID efficiency as a function of particle p, p_T, η and $nTracks$ (event multiplicity) observables. The classifier achieves this flatness using a modified loss function [4]. The PID performance of the classifier is shown in Fig 2. Flat 4d models uses all 60 observables from all the LHCb systems as well as

CatBoost and Deep NN models. However, requirement of the flatter PID efficiency decreases the PID quality of the model.

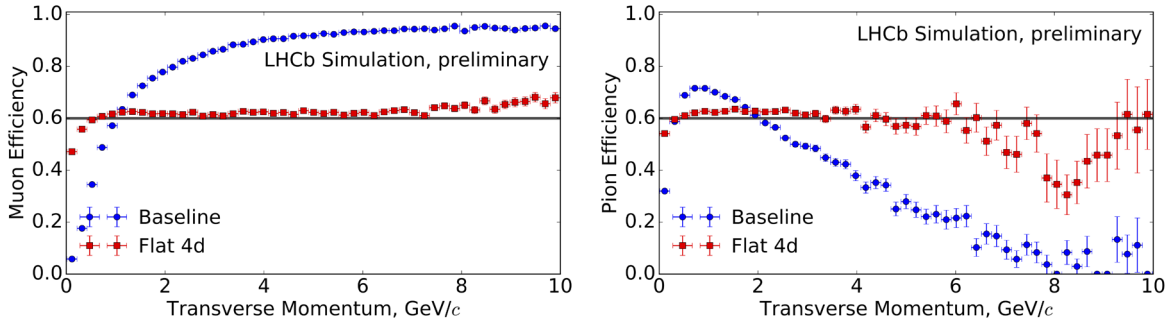

Figure 3. Dependence between Flat 4d model efficiencies and particle transverse momentum for each particle type. The curves correspond to the same global signal efficiency of 60%.

4 Conclusions

Several particle identification algorithms have been developed based on machine learning models. These models efficiently combine PID information from the LHCb tracking system, ring-imaging Cherenkov detectors, electromagnetic and hadron calorimeters, and muon chambers achieving high quality of global charged particle identification. The modified loss function of the classifiers provides better PID efficiency flatness in particle momentum, transverse momentum, pseudorapidity and number of tracks in the event.

5 Acknowledgments

The research leading to these results has received funding from Russian Science Foundation under grant agreement $n \circ 17 - 72 - 20127$.

References

[1] LHCb collaboration, LHCb detector performance, Int. J. Mod. Phys. V30 (2015) 153–22.
[2] F. Chollet, Keras, https://github.com/fchollet/keras (2015).
[3] A.-V. Dorogush, A. Gulin, G. Gusev, N. Kazeev, L. Ostroumova Prokhorenkova, A. Vorobev, Fighting biases with dynamic boosting (2017) arXiv:1706.09516.
[4] A. Rogozhnikov, A. Bukva, V. Gligorov, A. Ustyuzhanin, M. WilliamsNew, New approaches for boosting to uniformity Journal of Instrumentation, V10 N03 (2015) 03–2
[5] M. Calvo, E. Cogneras, O. Deschamps, M. Hoballah A tool for γ π^0 separation at high energies LHCb-PUB-2015-016 (2015)