SECOND GENERATION MACHINE LEARNING BASED ALGORITHM FOR LONG-LIVED PARTICLES RECONSTRUCTION IN UPGRADED LHCb EXPERIMENT*

Sabin Hashmi

AGH University of Science and Technology, Kraków, Poland

Received 25 April 2022, accepted 29 June 2022, published online 9 September 2022

The paper presents the developments and preliminary results related to the implementation of a Machine Learning based Algorithm for reconstruction of the long-lived particles in an upgraded LHCb experiment. The analysis is based on a Monte-Carlo simulation prepared for LHC Run 3 data-taking conditions. Studied tracks are reconstructed with an official LHCb software application Moore in configuration that is very close to the one that will be operated as a part of the final software trigger system.

DOI:10.5506/APhysPolBSupp.15.3-A36

1. Introduction

Large Hadron Collider beauty (LHCb) is one of the four large experiments operating currently at the Large Hadron Collider (LHC) and is designed to search for new physics phenomena in the heavy flavor quark sector and perform precise measurements of CP symmetry violation in beauty and charm quarks sector. At present, the detector is undergoing a major upgrade. To filter out the data produced by each proton–proton interaction, a robust and efficient trigger system is required. In the upgraded system, the hardware trigger that worked based on information from Calorimeters and Muon Systems is completely removed instead a flexible fully-software trigger will be used. The Machine Learning based long-lived particle reconstruction algorithm is a part of this new upgraded system. It will apply a cascade of filters to remove fake tracks and improve both the efficiency and purity of the reconstructed tracks. The Machine Learning models are trained with simulated samples that may not reproduce all of the properties of collision data. This will require careful monitoring and appropriate updates of the

* Presented at the 28$^{\text{th}}$ Cracow Epiphany Conference on *Recent Advances in Astroparticle Physics*, Cracow, Poland, 10–14 January, 2022.

models with re-tuned Monte-Carlo events. The LHCb experiment collected during Run 1 and Run 2 a data sample corresponding to the integrated luminosity of 9 fb^{-1}, whilst after Run 3 and Run 4, the integrated luminosity should reach at least 50 fb^{-1}. Thus, applying a flexible and configurable software trigger is vital for the LHCb upgrade by enabling us to collect far larger data samples without compromising the performance of the hardware channels imposed by the previously operated hardware trigger.

2. Large Hadron Collider

Large Hadron Collider (LHC) is currently the largest and most powerful particle accelerator in the world. LHC is designed to accelerate protons and heavy ions. The nominal centre-of-mass collision energy for protons is 13 TeV. The LHC particle accelerator is not a single ring but uses several other machines that accelerate the colliding beams in stages. The cascade includes both linear and circular accelerators. The final beams of protons or ions can cross at four points along the LHC tunnel providing data for all experiments.

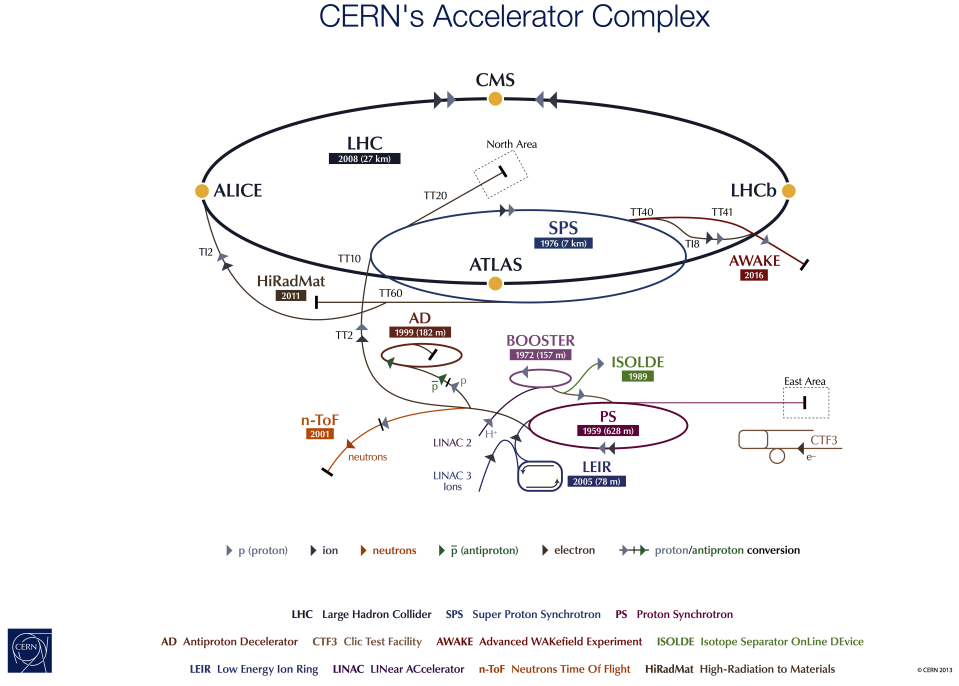


Fig. 1. LHC complex.

In Fig. 1, the LHC complex is shown. The acceleration chain starts with the hydrogen which is ionized to produce protons for the Liniac2 linear accelerator. Next, the beam is injected into the Proton Synchrotron Booster and then into Proton Synchrotron (PS) to raise its energy up to 25 GeV. The proton beam is then moved to Super Proton Synchrotron (SPS) and accelerated to 450 GeV. Later on, in the final stage, the proton beam is injected into the LHC beam pipes for further acceleration.

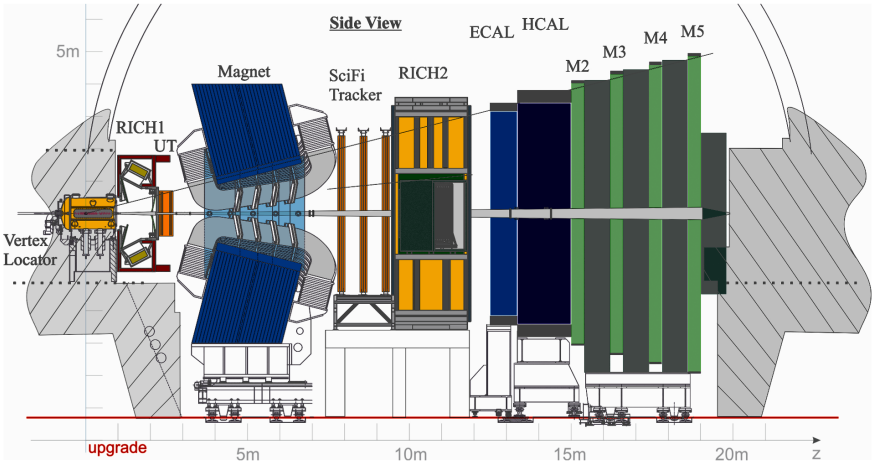


Fig. 2. LHCb Tracking Systems.

3. LHCb tracking and high-level trigger systems

LHCb is a forward general-purpose detector with a unique coverage of $2 < \eta < 5$ detecting 40% of all heavy quarks produced by the proton–proton collisions.

There are three main sub-detectors (Fig. 2) designed for the reconstruction of charged particles emerging from the collisions. They perform track reconstructions with a momentum resolution 0.5% for the particle momenta $p < 20$ GeV/c and 1.0% at 200 GeV/c with 95% efficiency.

3.1. LHCb Upgrade I

The LHCb experiment completed the first decade of data collection and analysis in Run 1 and 2. Currently, LHCb is undergoing a final commissioning in preparation for Run 3 which starts in 2022. There are significant upgrades in LHCb, including the implementation of full software trigger and replacing the previous tracking detectors completely [1]. This upgrade helps to overcome the bottleneck in the read-out system. Significant improvement is instantaneous luminosity that increased from 11 fb^{-1} to 25 fb^{-1}.

3.2. LHCb upgrade tracking system

Vertex Locator (VELO) is the detector situated the closest to the collision point with active elements reaching approximately up to 6 mm from the proton beams. Upstream Tracker (UT) is the second tracking system placed after VELO and before bending magnets. Finally, the Scintillating Fibre Tracker (SciFi) detector is placed after the magnet. Most studies at LHCb are based on so-called long tracks that are registered by all tracking detectors. Long tracks are normally generated from the decay products of short-lived particles (*e.g.*, B or D mesons). In this paper, we concentrate on studying so-called downstream tracks that are created by the daughter charged particles of long-lived composite particles such as K^0_S and Λ that decay outside VELO, thus leaving no signals at VELO but only in UT and SciFi detectors.

3.3. High level trigger

At each collision, numerous particles are produced what corresponds to a vast volume of generated data. Most of the information is not relevant for the LHCb analyses so, the software trigger is designed to store only the necessary high-level objects needed for reconstruction the desired decays. The remaining raw data will be discarded to the safe storage and push the number of accepted events as high as possible. This novel approach is called the LHCb Turbo Stream and will allow to study extremely rare decay modes. The requirements on the tracking quality are very hard since after the data will be stored, no re-processing will be possible. A part of the HLT processing chain is the long-lived particles reconstruction algorithm described in the next section.

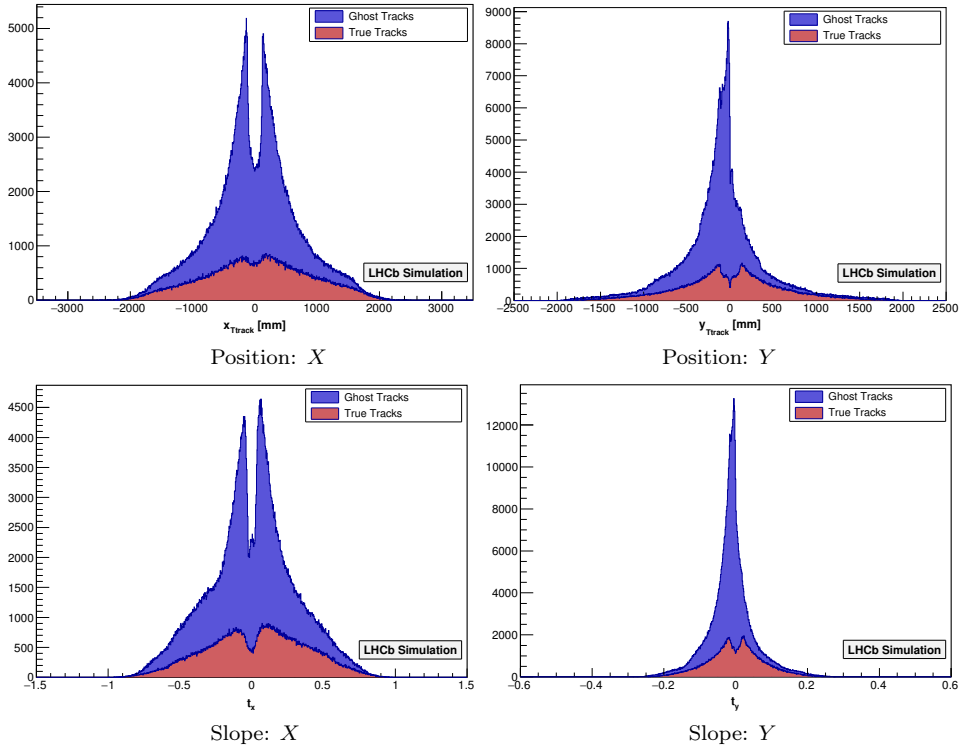
4. Long-lived track reconstruction algorithm with Machine Learning

The general idea of enhancing the currently used long-lived tracking procedure is to apply a sequence of Machine Learning based classifiers that act as filters. This approach can lead to enriching the purity of reconstructed track sample while keeping as high efficiency as possible. In this paper, we report on the first stage filter for classifying the SciFi track segments. In order to perform the training of selected Machine Learning models, a large set of simulated physics events is needed. Using appropriate extrapolation to the Run 3 conditions, signal samples have been produced and processed by the same software that will be used in the final HLT system. Using the simulated samples, we can perform the best case scenario the so-called "cheated analysis" where we have access to all the information regarding the reconstructed objects (such as true particle identification or true momentum).

With this information, we can work out the true label for each event in the training data set that is divided into true and fake SciFi tracks[1]. A True Track label is assigned to a SciFi seed when it is associated with an MC-particle that has hits in the UT detector and no hits in the VELO. A veto on the electron PID is also imposed for the True Tracks. A Ghost Track label, on the other hand, is assigned to tracks that have no MC particle associated [2].

The simulated data set used in the training of the SciFi tracks classifier was creating with the following signal samples: $K^0_S \to \pi^+\pi^-$ and $\Lambda \to p\pi^+/\Lambda \to p\pi^-$.

After processing the data, approximately 10 million tracks are obtained each of which is described by 12 kinematical variables (features in Table 1) that reflect the characteristics of the tracks. The pool of tracks contains approximately 2 million True Tracks. The distribution of the kinematical variables is presented in Fig. 3 and Fig. 4.

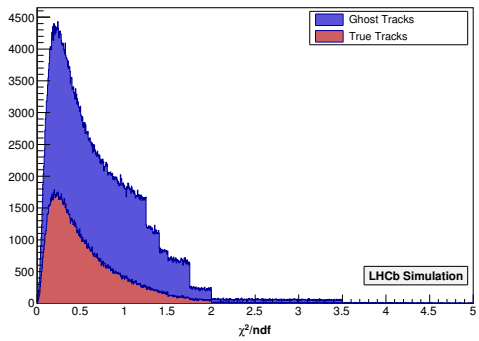


Position: X Position: Y

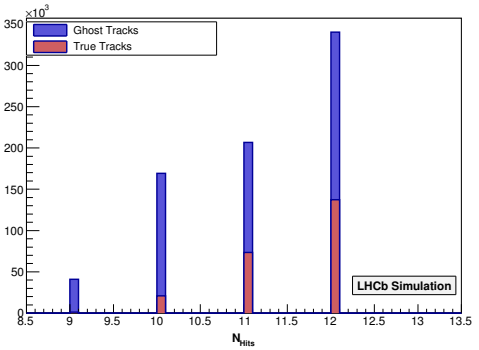
Slope: X Slope: Y

Fig. 3. Kinematic variables.

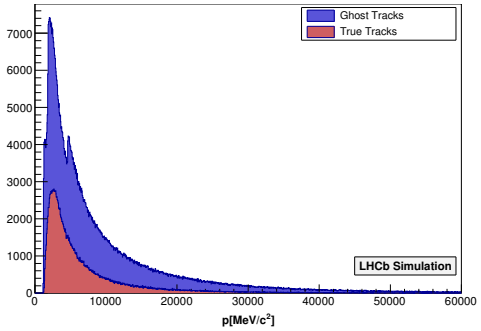
[1] In the first stage of these studies, we consider only the SciFi segments that are the seeding part of full downstream tracks. We are going to use this simplification throughout the text and designate the SciFi seeds as tracks.



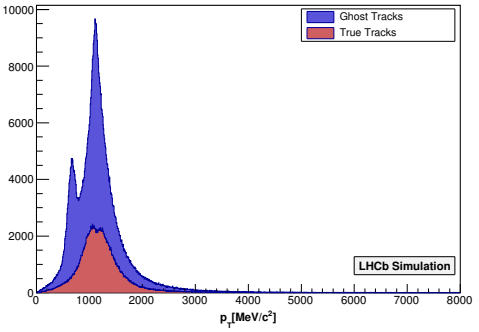
χ^2/n.d.f.



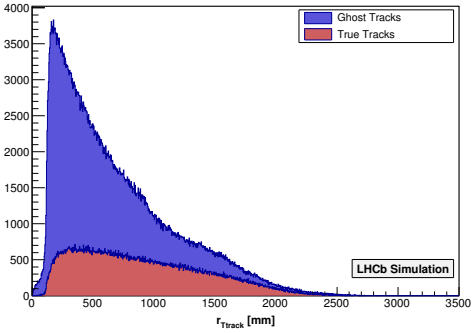
N_{hits}



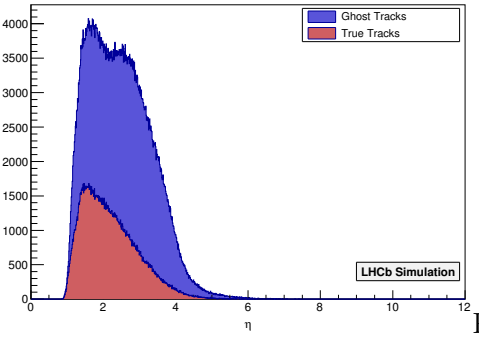
Momentum



Transverse momentum



Position: R



Pseudo-rapidity

Fig. 4. Kinematic variables.

Table 1. The kinematic variables.

Variable	Description
χ^2/n.d.f.	The variable determined by Pat-Seeding Algorithm
N_{hits}	Number of hits constructing a seed
Momentum	Momentum of the track
Position X	The X-position of track's first state
Position Y	The Y-position of track's first state
Position R	The distance to track's first state from beam line
T_x	Slope of track in X–Z plane
T_y	Slope of track in Y–Z plane
η	Pseudo-rapidity
p_{T}	Transverse momentum

5. Initial results

From the processed data, we can build a Machine Learning based binary classifier to increase the purity of tracks by flagging the data into True Tracks and Ghost Tracks. There are many Machine Learning/Deep Learning Algorithms available, out of which a few most popular ones were studied to build a reliable classifier for the used set of track data.

The Run 2 long-lived track reconstruction algorithm has evolved to a stage where a set of two classifiers were used to discriminate the fake tracks. It was integrated into the HLT processing chain and showed significant improvements compared to the Run 1 implementation. Since Run 3 conditions are significantly different from Run 2 ones, the previous version of the algorithm has to be replaced with a new one and improved Machine Learning models, trained with new simulated events, for better performance.

After several computational analyses, boosting based algorithm, Catboost Classifier has been selected to proceed with the final algorithm for the SciFi seeds filter, due to its novelty and measured performance (see Fig. 5). For more thorough evaluation of the model quality, two main metrics are used: Accuracy Score (Eq. (1)) and F1 Score (Eq. (2)). Due to the imbalance in the evaluation set, accuracy of the model will not be a good matrix to evaluate the quality of the model

$$\text{Accuracy Score} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \tag{1}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2}$$

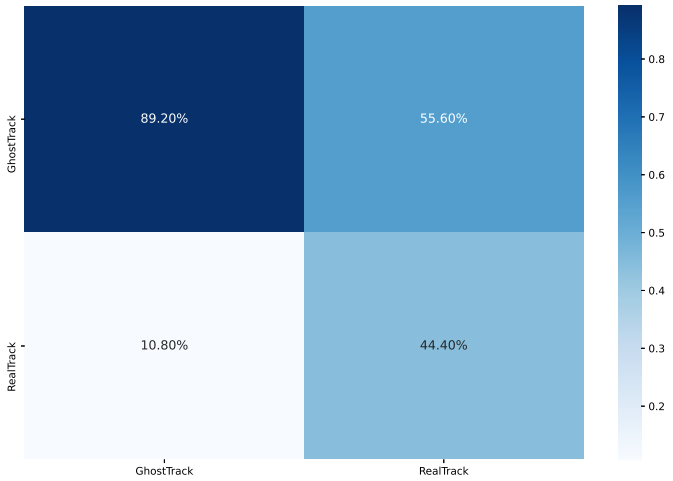


Fig. 5. Normalised confusion matrix for ML model track predictions.

6. Future plans and conclusions

The upgraded LHCb experiment will be the leading experiment in new physics searches and precise CP violation measurements. The significant increase in integrated luminosity of data to be taken during both Run 3 and Run 4 will bring the statistical uncertainties to the level of theoretical ones. A long-lived track reconstruction algorithm using Machine Learning approach may play a significant role in studying new possible exotic particles with lifetimes large enough to avoid detection in the vertex detector.

The future plans of the research include the implementation of the second stage filtering for the full downstream tracks, feature engineering and search for the optimal hyper-parameters, study the performance in terms of the track efficiency and purity, verifying the model performance in real-time with real data and, finally, integration of the model with the existing downstream algorithm for better track quality and efficiency.

We acknowledge the support from the Polish Ministry of Science and Higher Education and the National Science Centre, Poland (NCN), UMO-2018/31/B/ST2/03998.

REFERENCES

[1] Framework TDR for the LHCb Upgrade: Technical Design Report (CERN-LHCC- 2012-007; LHCb-TDR-12).

[2] PatLongLivedTracking: a tracking algorithm for the reconstruction of the daughters of long-lived particles in LHCb (LHCb-PUB-2017-001; CERN-LHCb-PUB-2017-001).