Krakow, 31.12.2025.

Prof. Jerzy Smyrski
Faculty of Physics, Astronomy and Applied Computer Science
Jagiellonian University

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE

Instytut Fizyki
im.
Mariana Smoluchowskiego

Review of the doctoral thesis of M.Sc. Sabin Hashmi

under the title „Real-Time Machine Learning-Enhanced Reconstruction of Long-Lived Tracks in LHCb High-Level Trigger and Calibration Analysis of the UT Detector"

Sabin Hashmi's doctoral thesis, submitted for review, is related to the LHCb experiment at the LHC collider at CERN. The main goals of this experiment are to measure CP symmetry violation and search for signatures of New Physics beyond the Standard Model. The research conducted so far by the LHCb collaboration has led to many important discoveries, including the observation of a new type of hadron consisting of five quarks (pentaquarks) and the observation of CP symmetry violation in the decays of charm mesons. The modernization of the LHCb detector, carried out in 2019 and 2020, aimed to adapt it to operate at higher luminosities and achieve greater sensitivity in the search for physical phenomena beyond the Standard Model. Among other things, the straw tube Outer Tracker was replaced by the scintillation fiber SciFi detector, and the hardware trigger was replaced by software-based trigger. The aim of Mr. Hashmi's doctoral thesis was to investigate the applicability of machine learning algorithms for finding ghost tracks in SciFi and UT detector data and to integrate these algorithms with the software trigger to improve its performance.

Mr. Hashmi's supervisor is Professor Tomasz Szumlak from the Department of Particle Interactions and Detection at the Faculty of Physics and Applied Computer Science, AGH University of Science and Technology. A group of physicists from this department is involved in the LHCb experiment and has made significant contributions to both the hardware and software components of the experiment, as well as to the analysis of experimental data.

ul. St. Łojasiewicza 11
PL 30-348 Kraków
tel. +48(12) 664-47-03
fax +48(12) 664-49-06
e-mail: fizyka@uj.edu.pl

The dissertation is divided into eight chapters. The first chapter provides a brief introduction to the Standard Model and the LHC accelerator. It also presents the LHCb experiment, with particular emphasis on the three charged particle trackers: the Vertex Locator (VELO), the Upstream Tracker (UT), and the Scintillating Fiber Tracker (SciFi), as well as software-based trigger system, designed for operation at higher luminosities in Run 3 measurements. This trigger system consists of two main components: a Low-Level Trigger (LLT), which pre-selects events using certain conditions, such as the presence of high-transverse momentum tracks in the muon detectors, and a High-Level Trigger (HLT). The latter consists of two stages: HLT1, which performs partial event reconstruction and pre-selection, and HLT2, which performs full event reconstruction.

Chapter 2 presents a classification of charged particle tracks based on the type of tracker in which the particle was detected, with particular emphasis on so-called "downstream tracks" detected in UT and SciFI but invisible in VELO. These tracks may originate from long-lived particles, such as neutral kaons or lambda hyperons. The LHCb software frameworks for online and offline data analysis are also presented. The "Hybrid Seeding Algorithm" algorithm for finding track candidates in SciFi is presented in more detail. In the horizontal plane, it searches for hits from SciFi vertical detection modules consistent with the selected track model—a straight line for high-momentum tracks and a parabola for low-momentum tracks. Hits from inclined modules are added to the track candidates using a method based on the Hough transform. Details of the transform used are not provided. An algorithm for reconstructing downstream tracks, using candidates found in SciFi and interpolating them into UP, is also presented. After finding hits in the UP that match the candidate, the track fitting is performed.

Chapter 3 introduces the basic concepts and methods of machine learning. It discusses the main categories of supervised and unsupervised learning. It also presents the basic structure of neural networks and the key parameters used to evaluate the performance of machine learning models.

Chapter 4 presents a comparison of different classifiers based on machine learning methods for distinguishing true SciFi tracks from ghost tracks in track seeds. For training the classifiers, Monte Carlo simulated tracks are used. As input variables to the classifier, in addition to the chi2 value and the number of hits included in the track seed, also geometric parameters of the track, e.g. x and y positions, are selected.

It's a pity that the work doesn't present examples of true and ghost tracks, e.g. in the form of graphics showing projections of hits on horizontal and vertical planes. This would allow the reader to gain an intuition about the difference between the two categories of tracks.

Due to its simplicity, a linear logistic regression model was used as the reference model. Several characteristics describing the effectiveness of this model were determined, including the ROC curve, the confusion matrix and distributions of SHAP values. This model reasonably discriminates true tracks from ghosts, as evidenced by the AUC value for ROC curve of 95%.

In the next step, the classification of the SciFi track seeds by eight different machine learning models, including the linear logistic regression, was compared. The Catboost model was chosen as optimal for the SciFi track classification due to its better performance compared to alternative models, as well as other properties including scalability and GPU support.

SHAP analysis of the Catboost model tests showed that the variables with the greatest impact on the classification of tracks are the number of SciFi hits, chi-squared value and x coordinate value. The author does not comment on this result, although understanding it may be important from the point of view of classifying the SciFi tracks. It is worth noting that in the case of the logistic regression model, the most important parameter next to the number of SciFi hits and the chi-squared value was pseudo-rapidity and not the x coordinate value.

Chapter 5 presents the test results of selected machine learning models for distinguishing between true and ghost downstream tracks registered in the SciFi and UT. Selection is performed for track candidates selected by the LHCb Hybrid seeding algorithm with ghost tracks in SciFI rejected by the developed procedure. Model training was performed using the same simulation data as used for the SciFi track classification, but the number of model input variables was increased from 8 to 11 by adding the momentum value, the transverse momentum value, and the number of UT hits. The best results in track classification were obtained with the Catboost model. The hyperparameters of this model were optimized using Optuna software. It is not stated whether hyperparameter optimization was also applied to other tested models. For track classification with the Catboost, an AUC value of 85% was achieved.

Chapter 6 presents the technical aspects of integrating the two developed track selection models into the LHCb track reconstruction software and assessing the impact of these models on track reconstruction. Including the models slightly reduces track reconstruction efficiency but significantly removes the ghost tracks. It does not introduce significant changes in shape of distributions of kinematical variables such as momentum or pseudo-rapidity.

The impact of the new models on the physics performance was investigated by reconstructing the invariant mass of the short-lived neutral kaons decaying into a charged pion pair. No effect of including the models on the invariant mass distribution was observed, although, perhaps naively, one could expect a reduction in the kaon peak width or a reduction in the background below the peak. The author explains that the main advantage of using the models in this case is a significant reduction in the time spent by the trigger software on processing ghost tracks. It is a pity that this explanation was not supported by providing corresponding processing time values.

Chapter 7 presents two tools developed to support calibration of pedestals in the readout electronics of the UT silicon strip detector. Pedestals can change during data taking and correcting them is important for effective noise discrimination while maintaining high

detection efficiency. The first tool uses Kullback-Leibler divergence to compare two calibration runs. The second tool uses Long Short-Term Memory neural network for forecasting pedestal values for the forthcoming calibration runs based on previous calibration runs. Figure 73 shows comparison of predicted and true pedestal values. The author does not comment on this comparison, but in my opinion, there is no clear correlation between the prediction and the measured values.

Chapter 8, the last one, summarizes the work.

In my assessment of the dissertation, I find its methodology and obtained results very interesting and valuable. The presented work and analyses were conducted with insight and care. The PhD student's expert knowledge of both machine learning methods and LHCb software is evident. Fully functional software for identification of ghost tracks was developed. Several of its key characteristics were analyzed and its effectiveness was demonstrated using simulation data. Directions for further research were identified. The work represents a significant contribution to the research conducted by the LHCb Collaboration. The significant scope of the analyses performed, their competent description and critical interpretation of the obtained results demonstrate the doctoral student's ability to conduct independent scientific work.

I found a number of minor defects in the reviewed thesis:

- In the introductory part of the thesis, presenting the author's contributions to the described research, only the title of author's publication is given, and the journal is not indicated.

- Page 9, Eq. 1: instead of N it should be N1 x N2 –where N1, N2 are the number of particles per bunch.

- Page 9, Eq. 2: N is the interaction rate (not number of particles per bunch crossing).

- Page 9: "This increased L .." – What is "This" directed at?, and the next sentence "This directly increases" – What is "This" directed at?

- Page 36, Fig. 14 caption: is "Logit Function", should be "Sigmoid function".

- Page 42, the chapter title is „Perceptrons to Deep Neural Networks Architecture". Shouldn't it be „From Perceptrons to Deep Neural Networks Architecture"?

- Page 53, Fig. 21: It is overkill to use an x-y graph to give the values of two fractions; they can be given in one short sentence. The same remark applies to figure 40.

- Page 54, Fig. 22: The symbols used in the drawings - blue rectangles with lines, black circles are not explained.

- Page 58, The title of the section - "Weight Co-efficients" should probably be "Weight Coefficients".

- Page 60, Table 3 and 4: It is not explained what the columns named "True Track (Raw)", "True Track (Scaled)" and "Weighted input" contain. Furthermore, there are no comments regarding the numerical values given in the tables. For example, it is unclear why the number of True Tracks (Raw) is 12 and the number of Ghost Tracks (Raw) is 10.

- Page 63, first sentence after Table 6: it is not clear what "ProkhorenkovaCatboost2017" means.

- Page 73, Table 8 is identical to Table 2. There is no need to repeat it, especially since it contains technical data that is not necessary for understanding the issues presented.

- Page 75, Fig. 42: It is not stated for what position in the z direction the presented x-y distribution was prepared (position of UT?).

- Page 83, 3rd line: "Table[??]".

- Page 101, Fig.64: The legends in the figures are unreadable due to too small font.

- Page 102, Fig. 65: Fitted curves do not describe the data, so they do not provide a basis for drawing conclusions about the data distribution.

- Page 112, Fig. 68: The caption indicates that absolute pedestal values are shown, but the figure shows both positive and negative values.

- Page 113, Eq. 24:

  left side of the equation - P should be inside the round brackets,

  right side of the equation – the dot before log is unnecessary,

  the text does not explain what variable x and index n mean,

  inside the sum on the right side of the equation there is no index n over which the summation is performed; shouldn't it be $x=n$?

The above defects do not significantly affect the value of the work, which I consider to be good.

In conclusion, I declare that the reviewed dissertation meets the requirements specified in Article 187 of the Act of 20 July 2018, the Law on Higher Education and Science (as amended), and I request that it be admitted to the next stages of the procedure for the award of a doctoral degree in the field of exact and natural sciences, in the discipline of physical sciences.