Study of B_s^0 meson reconstruction in proton-proton collisions using machine learning with ALICE

Phil Lennart Stahlhut

September 2023

This project has been carried out under the supervision of
Dr. Fabio Catalano and Dr. Fabrizio Grosa

Abstract

Measurements of the production of heavy-flavor hadrons containing charm or beauty quarks in proton-proton collisions provide an important test for pertubative quantum chromodynamics and allow to study heavy-flavor hadronisation mechanisms. In Run 2, the ALICE collaboration studied the production of beauty hadrons only indirectly via D mesons originating from beauty-hadron decays. By using software triggers for the selection of events containing heavy-flavor hadrons measured by ALICE in Run 3, direct measurements of the production of beauty hadrons and the B_s^0 meson in particular are expected to be feasible.

In the course of this summer student project, the complete workflow for the reconstruction and selection of B_s^0 candidates in O^2Physics, the new analysis framework of ALICE, has been implemented. In a second step, a machine-learning model based on boosted decision trees for the separation of signal and background candidates has been developed and tested on the currently available data recorded by ALICE in Run 3.

1 Motivation

Using the data collected from the LHC Run 2, the ALICE collaboration measured the production of D mesons produced either in charm-quark hadronisation (*prompt*) or in beauty-hadron decays (*non-prompt*). From the obtained particle yields, the *fragmentation fractions* of heavy quarks to strange mesons divided by the one to non-strange mesons were measured in the charm (directly) and beauty (indirectly) sector:

$$\frac{\mathrm{D_s^+(c\bar{s})}}{\mathrm{D^0(c\bar{u}) + D^+(c\bar{d})}} \to \frac{f(\mathrm{c \to H_s})}{f(\mathrm{c \to H_u}) + f(\mathrm{c \to H_d})}$$
$$\frac{\mathrm{B_s^0(s\bar{b})}}{\mathrm{B^+(u\bar{b}) + B^0(d\bar{b})}} \to \frac{f(\mathrm{b \to H_s})}{f(\mathrm{b \to H_u}) + f(\mathrm{b \to H_d})}$$

Direct measurements of beauty-meson production will complement the existing indirect measurements of fragmentation fractions from non-prompt D mesons and provide additional information on beauty-quark hadronisation and the $\mathrm{b\bar{b}}$-production cross section.

In order to measure the production of the $\mathrm{B_s^0}$ meson, the decay

$$\mathrm{B_s^0 \to D_s^-}\pi^+ \to (\phi\pi^-)\pi^+ \to (\mathrm{K^-K^+}\pi^-)\pi^+$$

and its charge conjugated decay mode for the $\mathrm{\bar{B}_s^0}$ are considered. By selecting only decays of $\mathrm{D_s^\pm}$ via the narrow ϕ resonance, the combinatorial background coming from the misidentification of pions as kaons can be suppressed significantly.

2 Development of the $\mathrm{B_s^0}$ reconstruction workflow in $\mathrm{O^2}$Physics

The ALICE experiment will take data of Pb-Pb collisions at a rate of up to 50 kHz in Run 3 [4]. In order to collect the statistics needed for the precision measurements that ALICE aims at, a continuous readout is adopted. This requires a significant reduction of the 3.5 TB/s of raw data ([4]) coming from the detectors to 100 GB/s of data to be stored permanently ([4]). Thus, a new software framework for the Online-Offline ($\mathrm{O^2}$) processing was implemented combining the synchronous (preliminary reconstruction, calibration and compression) and asynchronous (final calibrations, physics analysis as well as simulation) processing stages [4].

In the course of this project, the workflow needed to perform a measurement of the $\mathrm{B_s^0}$ meson was implemented in $\mathrm{O^2}$Physics, the physics analysis framework of $\mathrm{O^2}$. The entire workflow starting from the reconstructed *primary vertex* (PV) and tracks of single collisions up to selecting the $\mathrm{B_s^0}$ candidates is sketched in fig. 2. It makes use of the already existing service tasks such as *bunch crossing* (BC) and event selection, track propagation and selection, track-to-collision association and *particle identification* (PID) with the *Time Projection Chamber* (TPC) and *Time-of-flight* (TOF) detector. Moreover, tasks for the pre-selection of 3-prong *secondary vertices* (SV) of heavy-flavor decay candidates (Track-index-skim creator), calculation of the heavy-flavor candidate's physical properties and decay vertex (3-prong-candidate creator) as well as selection of $\mathrm{D_s^\pm}$ decays to $\mathrm{K^-K^+}\pi^\pm$ (candidate selector $\mathrm{D_s^\pm \to K^-K^+}\pi^\pm$) are employed. The possibility of using a machine-learning selection was added to the $\mathrm{D_s^\pm}$ candidate selector in order to apply the *boosted decision tree* (BDT) models (see section 3) trained for the heavy-flavor software triggers. The tasks that were added are the $\mathrm{B_s^0}$ candidate creator, $\mathrm{B_s^0 \to D_s^-}\pi^+$ candidate selector, $\mathrm{B_s^0}$ tree creator and $\mathrm{B_s^0}$ task.

In the $\mathrm{B_s^0}$ candidate creator, the $\mathrm{D_s^\pm}$ decay vertex is recalculated at first from its extrapolated daughter tracks in order to determine the covariance matrix of the $\mathrm{D_s^\pm}$ track parameters which

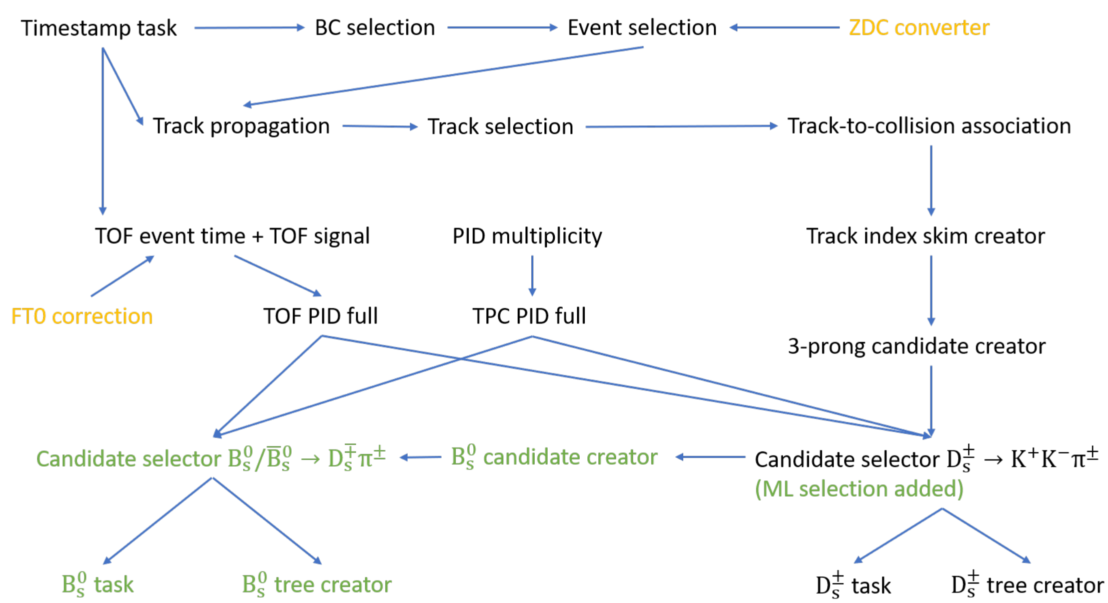

Figure 2: Sketch of the B_s^0 workflow implementation in O^{2}Physics. Tasks that were added in the course of this project are marked in green, converter tasks which are needed for running on older data sets in yellow.

has not been stored after running the 3-prong-candidate creator. Then pions which are not $D_s^{\pm}$ daughters and carry the opposite charge with respect to the $D_s^{\pm}$ are selected. From the $D_s^{\pm}$ and $\pi^{\mp}$, the 2-prong decay vertex and various physical properties of the resulting B_s^0 candidate are calculated. In addition to that, the reconstructed B_s^0 candidates from *Monte Carlo* (MC) simulations are checked if they are in fact originating from a B_s^0 decay or rather from a B^0 or partly reconstructed beauty-hadron decay:

1. $B_s^0 \to D_s^-\pi^+ \to (\phi\pi^-)\pi^+ \to (K^-K^+\pi^-)\pi^+$
2. $B^0 \to D_s^+\pi^- \to (\phi\pi^+)\pi^- \to (K^-K^+\pi^+)\pi^-$
3. Decays of B_s^0, B^0, B^+, Λ_b^0 which contain $K^-K^+\pi^-\pi^+$ as a subset of the final state particles.

If the B_s^0 candidate under consideration fulfills one of these categories, a corresponding flag is set. In addition to that, generated MC particles are checked if they are $B_s^0 \to D_s^-\pi^+$ (1.) or $B^0 \to D_s^+\pi^-$ (2.) decays. The result of the reconstructed MC particle matching for a MC data set with forced decays of charm and beauty hadrons is depicted in fig. 3. For the used MC simulation,

- a $b\bar{b}$-pair was injected in a third of the events,
- B_s^0 mesons were forced to decay to $D_s^-\pi^+$,
- a third of the B^0 mesons were forced to decay to $D_s^+\pi^-$,
- $D_s^{\pm}$ mesons were forced to decay via the $D_s^{\pm} \to \phi\pi^{\pm} \to K^-K^+\pi^{\pm}$ channel.

Since the production cross section for B^0 mesons is approximately four times larger than the one for B_s^0 mesons, but only a third of the generated B^0 mesons are forced to decay via the desired decay

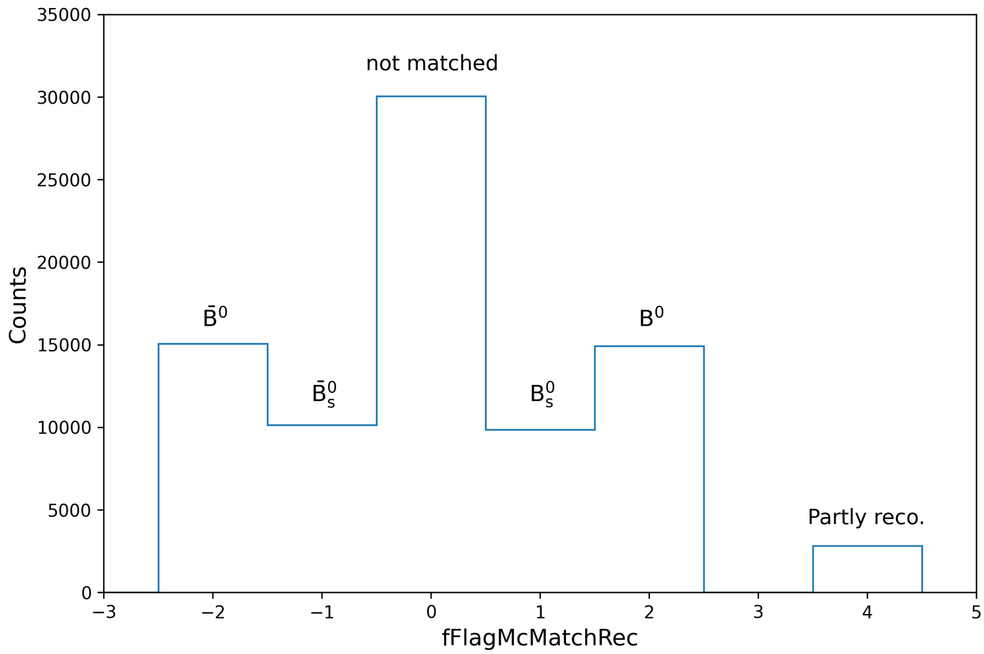

Figure 3: Matched reconstructed B^0_s candidates from MC data.

channel, the measured B^0/B^0_s ratio is expected to be 4/3. The observed value of the reconstructed candidates of approximately 3/2 is in reasonable agreement with the expectation.

The $B^0_s \to D^-_s\pi^+$ candidate selector employs a single flag were sequential bits are set from zero to one if certain selection criteria are passed: At first, it is checked whether the B^0_s candidate is flagged as a $B^0_s \to D^-_s\pi^+$ decay (skims selection). Then certain topological selections on the B^0_s candidate can be applied, including its mass, (transverse) decay length and cosine of pointing angle as well as the transverse momenta and impact parameters of its daughters (see explanation of the variables in section 3). Afterwards it is examined whether the pion candidate passes all PID criteria for pions using the PID information obtained from the TPC and TOF detectors. Finally, the B^0_s candidate selector allows to apply a machine-learning model to select the desired signal candidates.

After creating and selecting the B^0_s candidates, their physical properties can be plotted in histograms with the B^0_s task and stored in tables for further data analysis or training of machine-learning models with the B^0_s tree creator. In addition to that, the B^0_s tree creator allows to store specific tables containing the event properties or kinematic variables of the generated B^0_s particles.

3 Training a BDT model for the classification of $B^0_s \to D^-_s\pi^+$

Due to the complex topology of the $B^0_s \to D^-_s\pi^+$ decay and its rarity, a refined signal-identification procedure needs to be adopted. In contrast to the traditionally used linear selections, a candidate selection based on boosted decision trees (BDT) performing a binary classification between signal and background has been developed for this summer student project. BDTs are a family of machine-learning algorithms for classification and regression problems which have been used extensively in high-energy physics experiments. A BDT consists of multiple *decision trees* (DT) which comprise a sequence of simple tests to classify an instance depending on its feature by comparing a numeric attribute to a threshold value or a categorical attribute with a set of possible values. An individual test of the decision tree is called *node*. The training of the decision tree is an automatic procedure in which the tree is built recursively by attempting to find patterns in the training data, in which the true class of each instance is known. At each node of the tree, the feature and its value maximizing the separation between the classes is selected and the training data is partitioned accordingly. The building process is stopped when all instances in a node are of the same class or a pre-determined stopping condition is reached. Since large decision trees perform poorly on unknown data, typically small trees consisting of only a few nodes are used since they learn patterns in the data rather than the training data itself. In order to obtain a model with a good classification power, numerous decision trees are combined in a procedure called *boosting*, where the trees are constructed sequentially to compensate the shortcomings of the previous trees with the following ones [1].

For the training of the BDT, the XGBoost library was used in this summer student project. It stands for *extreme gradient boosting* since the BDT is based on the gradient boosting method: An objective function composed of the training loss measuring the difference between the predictions and the target values of the training data and a regularization term for control of the BDT complexity is defined. The machine-learning algorithm is then performing a gradient descent of this function in order to create the model that minimizes it. Once the training is completed, every tree of the BDT trained with the XGBoost algorithm assigns a score to each instance related to its input features. The overall output of the BDT is then given by the sum of all tree scores and can be normalized to values between zero and one. In the following, the optimal BDT score can be determined which maximizes the significance of the signal peak and the signal efficiency simultaneously. This value will then be chosen as the cut value above which all candidates are classified as signal and below which as background [1].

The training was performed on one half of the overall data, called *training set*, and the performance of the resulting BDT is estimated on the other half of the overall data set, the *test set*. Matched reconstructed $\mathrm{B_s^0}$ candidates from MC simulations were used as signal candidates. Background candidates were obtained by selecting $\mathrm{B_s^0}$ candidates with an invariant mass of $m < 5.180\,\mathrm{GeV}/c^2$ or $m > 5.550\,\mathrm{GeV}/c^2$ from a real data sample, which roughly corresponds to cutting all $\mathrm{B_s^0}$ candidates within the $\pm 4\sigma$ interval around the $\mathrm{B_s^0}$ peak ($m_{\mathrm{B_s^0}} = 5.366\,92\,\mathrm{GeV}/c^2$ [3]). As the used data sample is not triggered on heavy-flavor decays, contamination of the background sample with $\mathrm{B^0}$ candidates ($m_{\mathrm{B^0}} = 5.279\,66\,\mathrm{GeV}/c^2$ [3]) is assumed to be negligible. In order to reject $\mathrm{B_s^0}$ candidates with poorly reconstructed secondary vertices, only candidates with $\chi^2_{PCA} < 0.001$ are selected. Moreover, only 25% of the background candidates are kept such that there are each approximately 20000 signal and background candidates. A comparison of the signal and background distributions of various physical variables is depicted in fig. 4 with a corresponding description given below. Variables that were used for training the BDT model are printed in bold. The correlations of the training variables and the mass and transverse momentum of the $\mathrm{B_s^0}$ candidate for both signal and background samples are shown in fig. 5. The *bending* or *transverse plane*, typically labeled with XY, refers to the plane transverse to the beam axis and therefore also transverse to the magnetic field in the inner barrel detectors of ALICE in which the trajectories of charged particles are bent.

- fY: Rapidity y of $\mathrm{B_s^0}$ candidate.
- fEta: Pseudorapidity η of $\mathrm{B_s^0}$ candidate.
- fPhi: Azimuthal angle ϕ of $\mathrm{B_s^0}$ candidate.
- fPt: Transverse momentum (p_T) of $\mathrm{B_s^0}$ candidate in GeV/c^2.
- fPtProng0: p_T of $\mathrm{D_s^\pm}$ candidate in GeV/c^2.

- fPtProng1: p_{T} of $\pi^{\pm}$ candidate in GeV/c^2.
- **fDecayLength**: Decay length, i.e. distance between PV and SV, in cm.
- **fDecayLengthXY**: Decay length of $\mathrm{B_s^0}$ candidate projected into bending plane in cm.
- **fDecayLengthNormalised**: Decay length of $\mathrm{B_s^0}$ candidate divided by the error of the decay length which was calculated from the uncertainties of the PV and SV.
- **fDecayLengthXYNormalised**: Normalized decay length of $\mathrm{B_s^0}$ candidate projected into bending plane.
- **fChi2PCA**: Squared sum of the non-weighted distances of the SV to its prongs *point of closest approach* (PCA).
- **fNSigTpcPi1**: TPC $n\sigma$ separation for prong 1 with pion mass hypothesis.
- **fImpactParameter0**: (signed) impact parameter of $\mathrm{D_s^{\pm}}$ in the bending plane with respect to PV in cm.
- **fImpactParameter1**: (signed) impact parameter of $\pi^{\pm}$ in the bending plane with respect to PV in cm.
- **fImpactParProduct**: product of the impact parameters of both daughter particles.
- **fMaxNormalisedDeltaIP**: maximum normalized difference between measured and expected track impact parameters of all candidate daughters.
- **fCpa**: Cosine of pointing angle θ_p, defined as the angle between the vector connecting PV and SV and the reconstructed momentum $\vec{p}_{rec}$ of the $\mathrm{B_s^0}$ candidate.
- **fCpaXY**: Cosine of the pointing angle in the bending plane.

In order to obtain the best BDT model, the external parameters of the model, called *hyperparameters*, have been optimized. These include the number of trees that compose the BDT, the learning rate, which is a weighting factor applied to the output of newly added trees, and the maximum depth of the individual trees. For this purpose, a *Bayesian optimization* has been applied with a *5-fold cross-validation* technique: The training set is divided into five parts called *folds*. For every given combination of the hyperparameters, the model training is performed on 4 folds and the model performance is evaluated on the remaining fold. This operation is repeated five times permuting the folds every time such that the model performance is evaluated once on each of the 5 folds. The mean value of the metric which is used to quantify the performance of the model can be considered as the final estimator of the model performance with these given hyperparameters. The employed metric in this project was the ROC AUC (see below). The combinations of hyperparameters are determined with a Bayesian optimization algorithm which maximizes an unknown function describing the performance of the BDT model on a bounded set $\mathcal{X}$ of hyperparameters. A posterior distribution on functions is built to describe the unknown function starting from multivariate Gaussian distributions. It is used to determine the next point on $\mathcal{X}$ where to evaluate the unknown function and the result is used to update the posterior distribution. Thus, the posterior distribution improves with increasing number of tests and the optimization algorithm therefore explores the more interesting regions of $\mathcal{X}$, significantly reducing the computing time needed for the hyperparameter optimization. The resulting optimized BDT model consisted of 144 decision trees with a learning rate of 0.1925 and a maximum depth of 4 nodes [1].

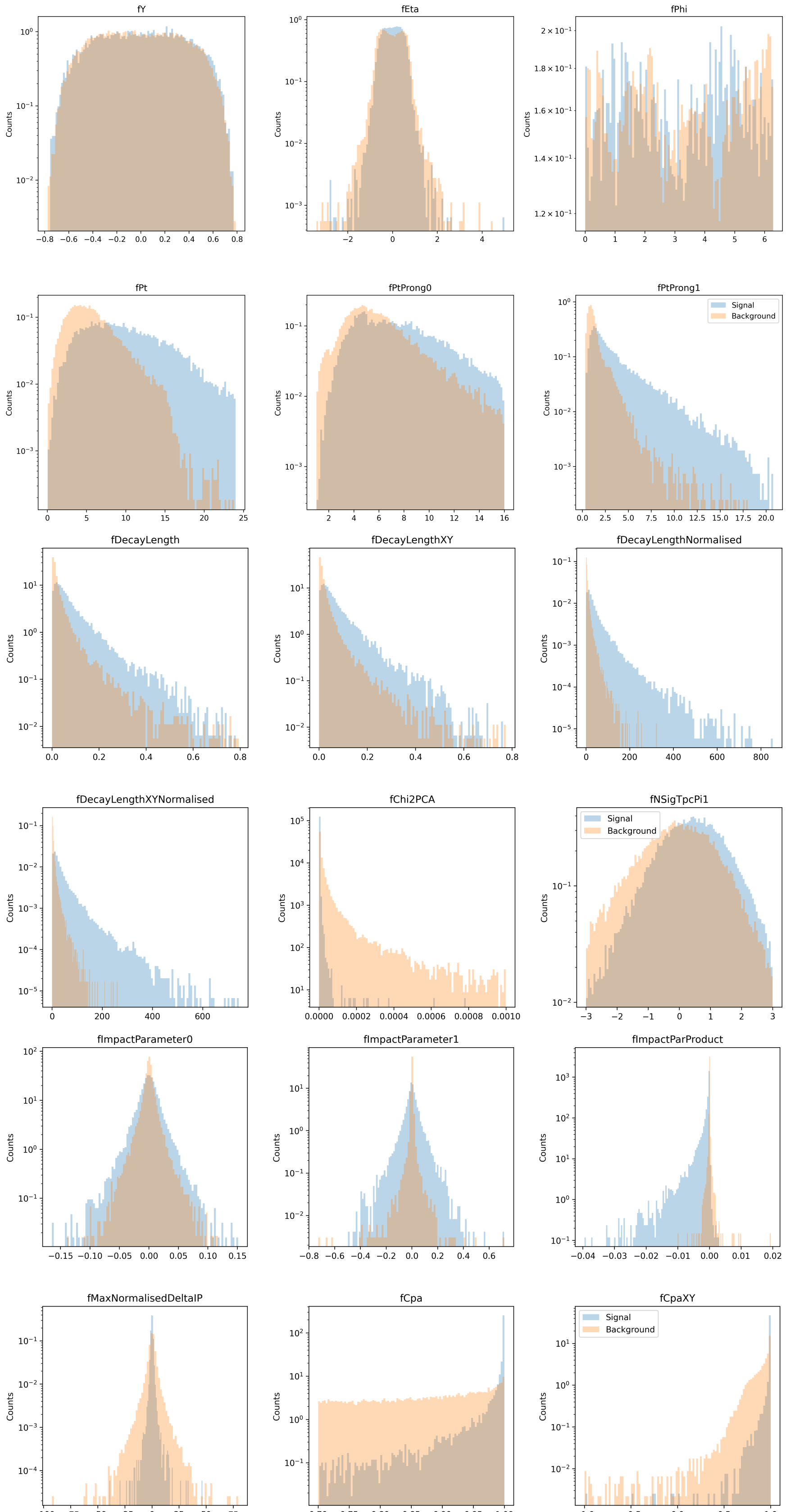

Figure 4: Comparison of the distributions of various physical variables for signal and background candidates.

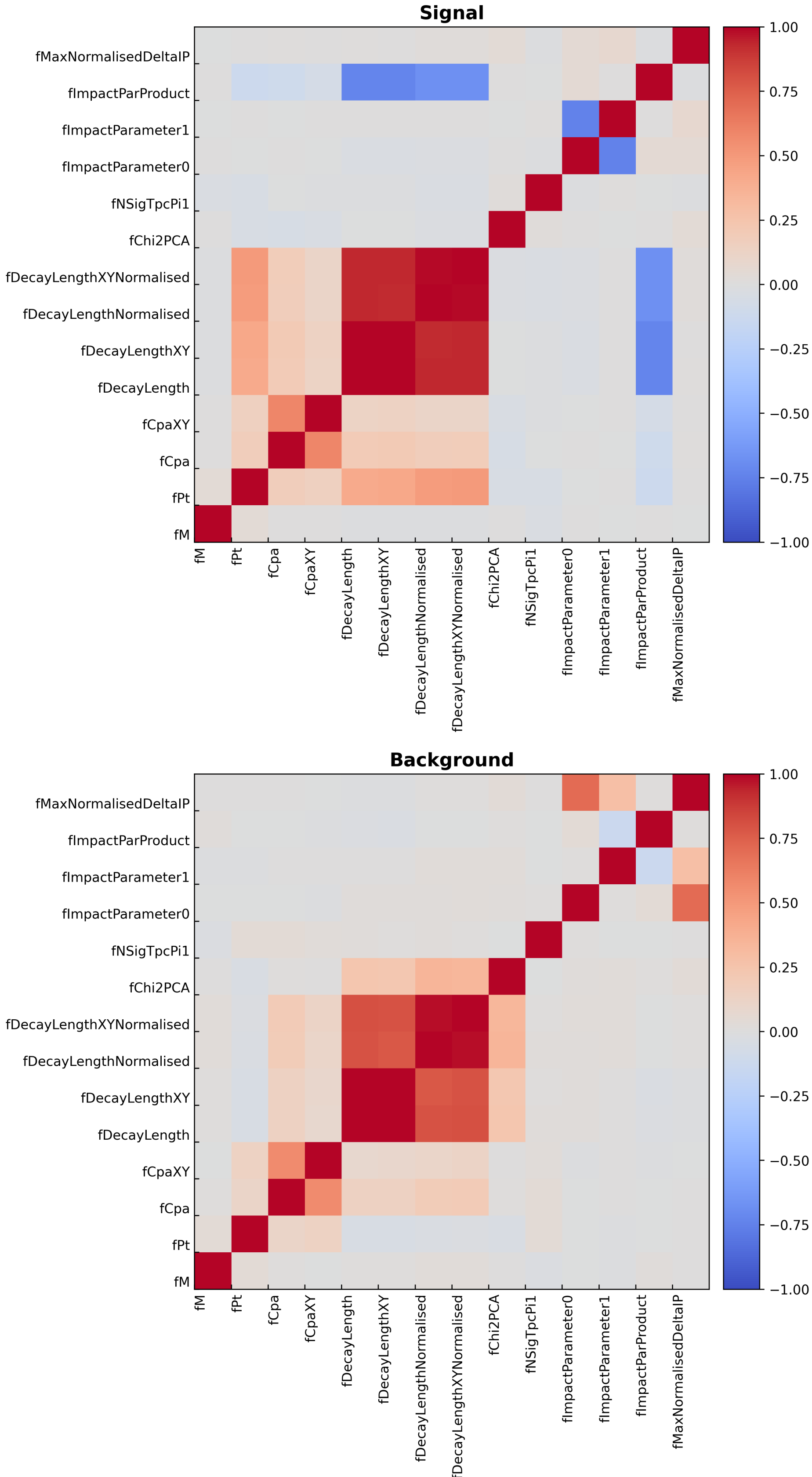

Figure 5: Correlations of various physical variables for signal and background candidates.

The distributions of the obtained BDT scores of the signal and background candidates is shown in fig. 6. For both test and training set, a clear discrimination between signal and background is visible as the BDT score distributions for signal and background peak at 1 or 0 respectively. The model appears to be slightly over-trained since the distributions of the BDT scores of signal or background candidates from the test set are slightly higher than the corresponding distributions from the training set at low or high score values respectively.

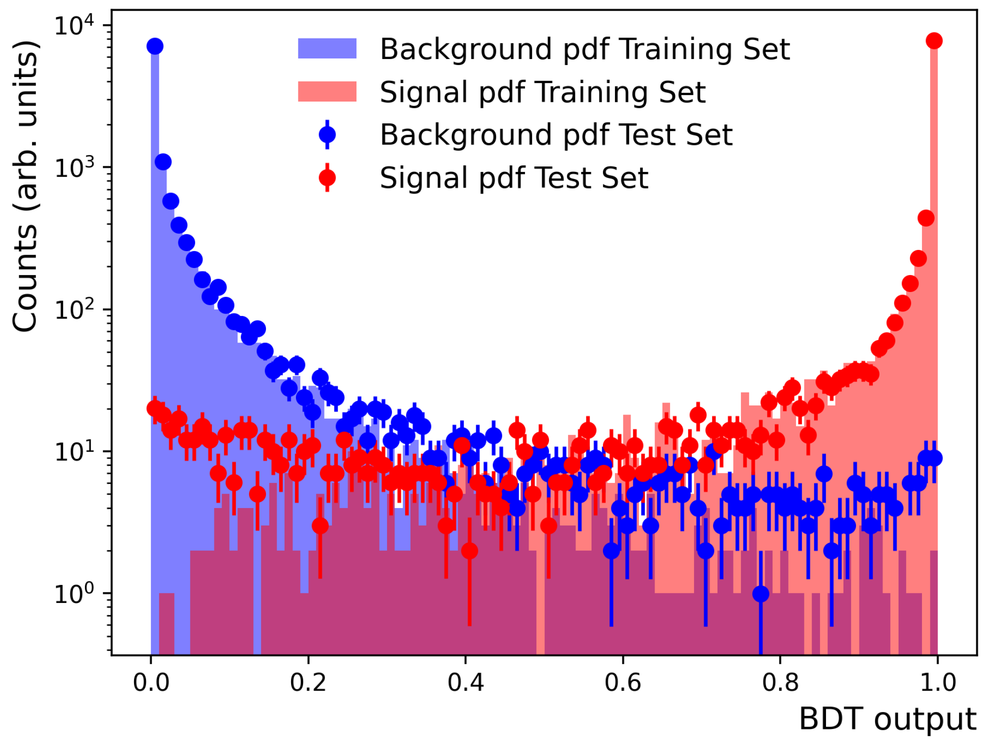

Figure 6: Distributions of the BDT scores of signal and background candidates from the training and test set respectively.

A possible approach to evaluate the performance of a BDT model for binary classification problems is to calculate the *area under the Receiver Operating Characteristic curve* (ROC AUC). The *Receiver Operating Characteristic* (ROC) is obtained by plotting the *true positive rate* (TPR) as a function of the *false positive rate* (FPR) which are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

where the definitions of the *true positive* (TP), *false positive* (FP), *true negative* (TN) and *false negative* (FN) labels can be taken from table 1. The ROC of the obtained model can be seen in fig. 7. When the BDT cut is increased from zero to one, the FPR decreases from one to zero. A ROC which is a straight line with slope of 0.5 would correspond to a model with random classification, whereas a ROC which is constant with exactly one over the entire range of the FPR would correspond to a model with perfect discrimination. Therefore, the ROC AUC values range from 0.5 to 1 respectively [1].

A good discrimination of signal and background and a slight over-training can be confirmed from the ROC AUC values reported in fig. 7, as both ROC AUC values are larger than 0.99 and the ROC AUC value for the training set is higher than the one for the test set.

	Real class P	Real class N
Predicted class P	TP	FP
Predicted class N	FN	TN

Table 1: Confusion matrix.

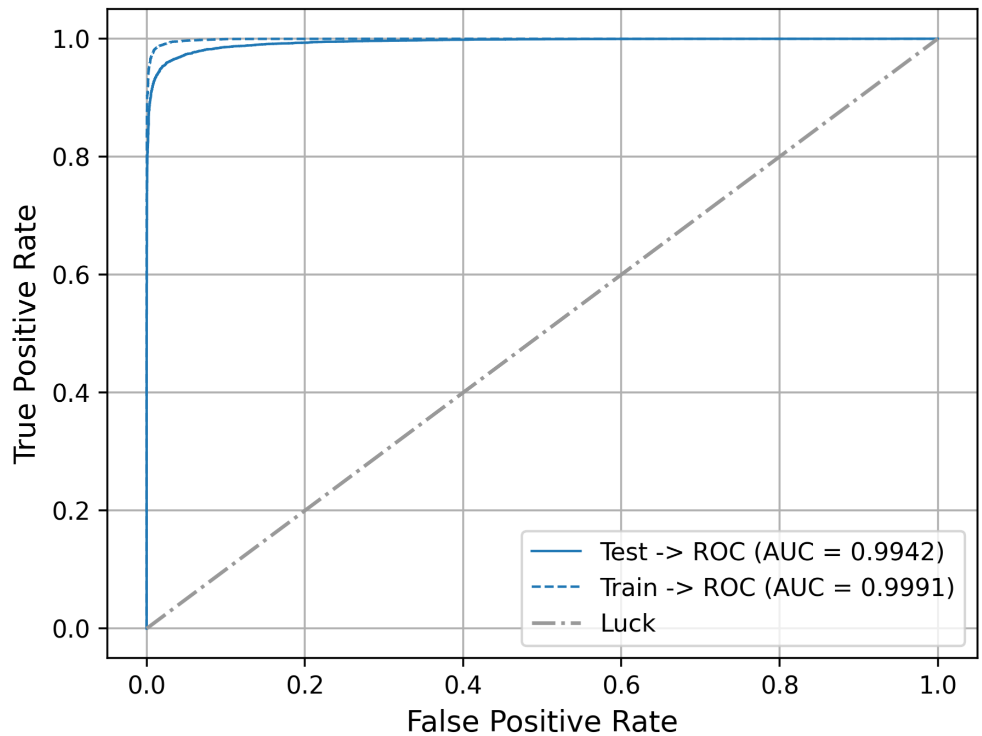

Figure 7: ROC curves for the train and test set with the corresponding AUC values.

In order to develop more refined machine-learning models for the classification of $B_s^0 \to D_s^- \pi^+$ candidates and better understand the decision process of the BDT, it is useful to study how much each of the candidate's features influences the BDT's predictions. For this study, SHAP (*SHapley Additive exPlanation*) values were used which attribute to each feature the change in the expected model prediction when conditioning on that feature. Thus, they can be used to explain how to get from the base value that would be predicted if we did not know any features to the current output of the model [2].

The top plot of fig. 8 shows the SHAP value distributions of all B_s^0 candidates for all different training variables. The width depicts the number of candidates per given SHAP value interval and the coloring indicates the feature value. For instance, high cosine of pointing angle values typically give a positive contribution to the BDT output such that the corresponding candidate is more likely to be classified as signal. On the contrary, high values of the impact parameter product tend to give a negative contribution to the BDT score and the corresponding candidate is therefore more likely to be classified as background. The absolute value of the mean of the SHAP values is given in the bottom plot of fig. 8. On one hand, the impact parameters of the daughter particles, the PID information of the pion and the cosine of the pointing angle in the transverse plane have a comparatively low impact on the model output. On the other hand, the cosine of pointing angle, the normalized decay lengths, the χ^2_{PCA} and the impact parameter product have the strongest impact on the model output. This is in accordance with the expectation from fig. 4 as the distributions of signal and background candidates differ the most for these features. However, it is known that the normalized decay length is not well represented in the MC simulations and thus the BDT model probably distinguishes candidates from MC simulations from candidates from real data rather than signal from background.

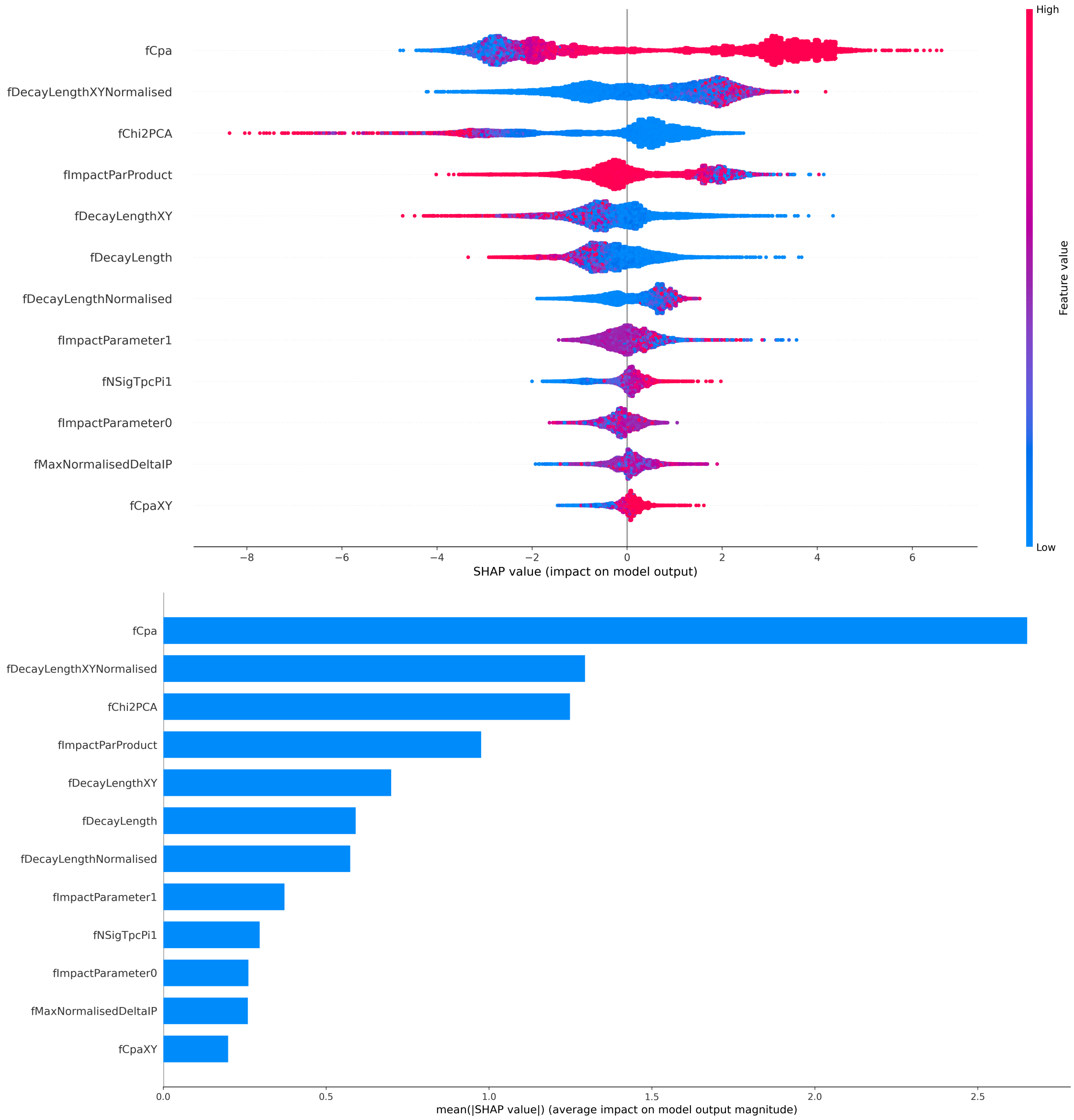

Figure 8: SHAP value distributions (top) and mean SHAP values (bottom) of the different physical variables used for the training of the BDT model.

The signal efficiency, i.e. the fraction of true B^0_s classified as signal for a given BDT cut, is plotted as a function of different BDT cut values in fig. 9. The chosen BDT cut was not precisely determined by calculations of the expected significance since the statistics in the analyzed data set were very limited. Instead, the BDT score of 0.6 was chosen as the resulting signal efficiency is still well above 90% and the background contamination should be negligible when comparing with fig. 6. The final invariant mass spectrum of a real data sample containing approximately 39.8 billion collisions recorded by the ALICE experiment is shown in fig. 10. Due to the limited statistics, there is no peak at the B_s^0 mass, but only background fluctuations visible since only a fraction of the available data has been analyzed.

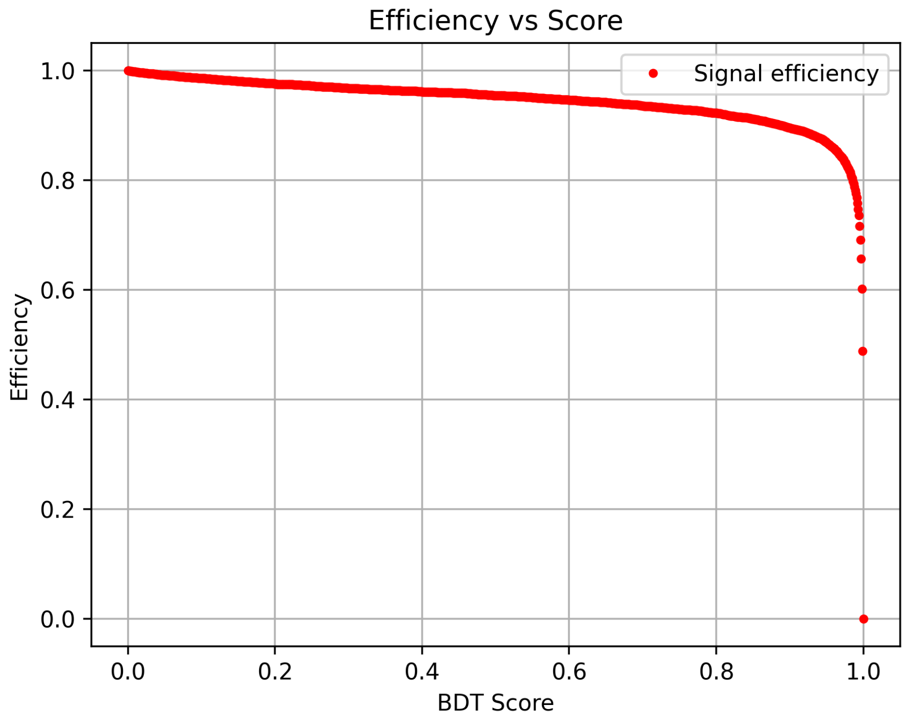

Figure 9: Signal efficiency as a function of the applied BDT score cut.

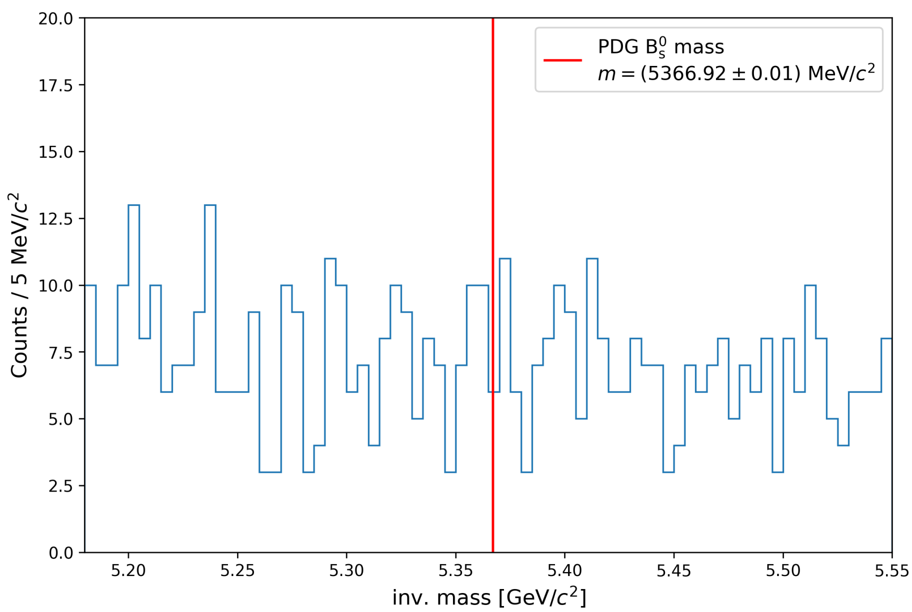

Figure 10: Obtained invariant mass spectrum of B_s^0 candidates with a BDT score > 0.6.

4 Conclusion and Outlook

In the course of this summer student project, all the necessary code to perform a measurement of the B_s^0 meson has been added to O^2Physics. The workflow has been successfully tested locally and on ALICE's computing grid. In addition to that, the first BDT model for the classification of the $B_s^0 \to D_s^- \pi^+$ decay has been trained and tested. The resulting BDT model has a good separation of signal and background. However, the currently available statistics is too low in order to see a B_s^0 peak in the invariant mass spectrum.

The next step would be to evaluate the expected performance of the currently collected luminosity. Additionally, a compact and self-consistent data format of pre-selected and linked tables for $D_s^\pm$ and corresponding pion candidates can be implemented such that the workflow can be run on a much smaller data set containing only the decay of interest.

References

[1] Fabio Catalano. "Strange D-meson measurements in pp and Pb-Pb collisions exploiting machine learning with ALICE at the LHC". PhD thesis. Politecnico di Torino, 2022. URL: `https://cds.cern.ch/record/2810132/files/CERN-THESIS-2022-046.pdf`.

[2] Su-In Lee Scott M. Lundberg. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30* (2017). URL: `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

[3] R. L. Workman et al. "Review of Particle Physics". In: *PTEP* 2022 (2022), p. 083C01. DOI: `10.1093/ptep/ptac097`.

[4] Chiara Zampolli. *ALICE data processing for Run 3 and Run 4 at the LHC*. 2020. arXiv: `2012.04391 [physics.ins-det]`.