LIP-STUDENTS-23-22

Investigating the Flavour Anomalies with Machine Learning at LHC

João Bernardo[1,a] and Madalena Ferreira[2,b]

[1] *Instituto Superior Técnico, Lisboa, Portugal*
[2] *Universidade de Aveiro, Aveiro, Portugal*

Project supervisors: A. Boletti and N. Leonardo

November 17, 2023

Abstract.
While the Standard Model is very successful to describe the properties of the elementary particles and their interactions, a set of experimental measurements of B-hadron decays is found to be in tension with its predictions. These are the so called Flavour Anomalies. One of the central processes involved in these anomalies consists of a bottom quark decaying into a strange quark and two charged leptons ($b \to s\, l^+ l^-$). This article focuses on the study of the $B^0 \to K^{*0} J/\Psi$ resonant channel, using the data collected by the Compact Muon Solenoid experiment, at the Large Hadron Collider, during Run 2. A binned likelihood fit is performed to the mass spectrum of the B^0 meson candidates in order to obtain the signal and background yields. Afterwards, single- and multi-variate analysis methods (neural networks and boosted decision trees) are applied to further discriminate the signal and background events, and improve the background rejection power of the selection criteria. The boosted decision trees algorithm proves to be efficient, reducing the background yield by $\sim 20\%$ while rejecting less than 1% of the signal relative to pre-selection. The framework here developed can be used in further studies for the non-resonant channel $B^0 \to K^{*0}\, l^+ l^-$, where the new physics might lie.

Keywords: CMS, Heavy-flavour Physics, Machine Learning, New Physics, LFU

1 Introduction

The Standard Model (SM) predicts that the different charged leptons (electron, muon, tau) have identical electroweak interaction strengths. This is referred to Lepton Flavour Universality (LFU). However, previous experimental measurements have shown to be in tension with this principle. If the discrepancy is large enough ($\sim 5\sigma$ deviation between the SM predictions and measured observables), the violation of LFU would imply physics Beyond the Standard Model (BSM), Fig. 1, such as a new fundamental interaction between quarks and leptons [1].

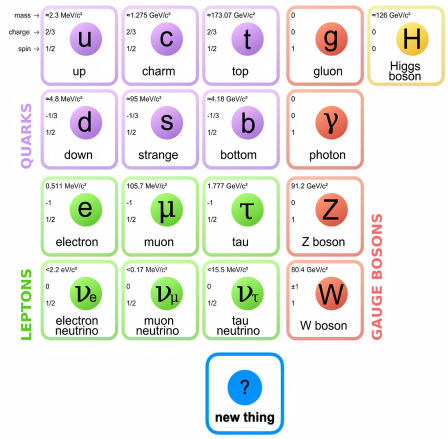

Figure 1. Standard Model and the search for new BSM particles.

One of the most interesting types of measurement for the LFU study consists in the comparison of the probability of a B-hadron decaying in channels involving different leptons flavours. Also, the rare decay of a bottom quark to a strange quark and two charged leptons $b \to s\, ll$, sensitive to New Physics (NP), is one of the most promising decays for the study of LFU. Focusing on the B^0 mesons, this corresponds to the decay channel $B^0 \to K^{*0}\, ll$.

1.1 The $B^0 \to K^+\pi^-\mu^+\mu^-$ decay

The article focuses on the B^0 meson decay to a $K^+\pi^-\mu^+\mu^-$ final state particles in two processes, when the muons are produced directly or through a *charmonium* ($c\bar{c}$) resonance.

The resonant channel studied in this analysis is $B^0 \to K^{*0} J/\Psi \to K^+\pi^-\mu^+\mu^-$, where the K^{*0} decays to a $K^+\pi^-$ and the J/Ψ to a pair of muons with opposite charge. This decay corresponds to a well known SM process, described at lowest order by the Feynman diagram shown in Fig. 2.

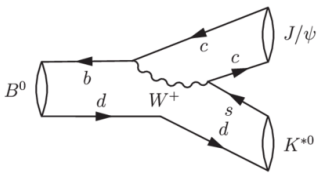

Figure 2. Tree-level Feynman diagram for the decay $B^0 \to K^{*0} J/\Psi$.

The non-resonant channel corresponds to $B^0 \to K^{*0}\mu^+\mu^- \to K^+\pi^-\mu^+\mu^-$. The SM lowest order Feynman diagrams exist only as a single loop-level (Figs. 3.a and 3.b). On the other hand, NP-mediated interactions, such as new gauge bosons Z' or leptoquarks, can contribute also at tree-level, to the same non-resonant decay (Figs. 3.c and 3.d).

[a] e-mail: joaobernardosilva@tecnico.ulisboa.pt
[b] e-mail: madalenablanc@ua.pt

1.2 Flavour Anomalies (b → sll)

The search for NP can be done through direct or indirect searches. The later aims at precise measurements of SM processes and compare them with the theoretical predictions. It also has the advantage of being sensitive to high energy scales, well beyond $O(10)$ TeV, since the contributions can be virtual. Lastly, it is model independent, encompassing multiple contributing NP-scenarios connecting initial and final state particles (Fig. 3). This is done with the aid of Effective Field Theory (EFT) models, extending the SM Lagrangian: $\mathcal{L}_{SM-EFT} = \mathcal{L}_{SM} + \sum_i C_i O_i$, where C_i are the Wilson coefficients and O_i the respective operators [2]. As such, it plays a large role in the flavour anomalies study.

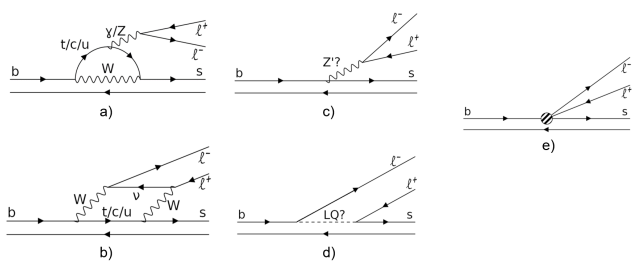

Figure 3. SM penguin (a) and box (b) Feynman diagrams. BSM Feynman diagrams showing the Z' (c) and leptoquark (d) interactions. Model independent EFT (e), allowing both SM and BSM processes for the decay $b \to s\,ll$.

Several observables can be measured for the $b \to s\,ll$ decay. One good observable is the Branching Fraction (BF), i.e. the probability of a decay to occur. Fig. 4 shows the latest measurements of the $b \to s\,\mu\mu$ for the normalized differential branching fractions [3].

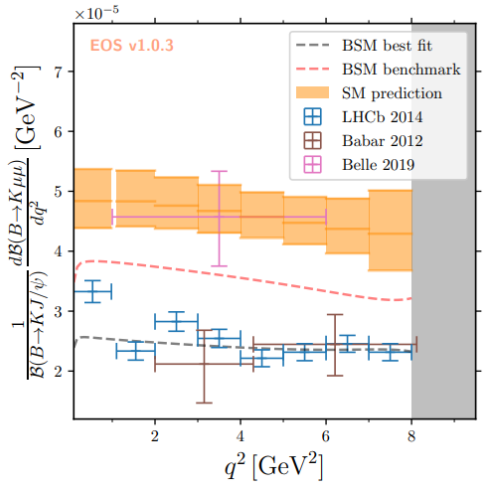

Figure 4. Normalized differential branching ratios as a function of the dimuon invariant mass q^2.

Clearly, a notable distinction exists between experimental and simulated data. The forthcoming experiments aim to reduce the experimental uncertainty, potentially unveiling new physics BSM or, alternatively, attributing the observed gap to statistical fluctuations or an overlooked systematic effect.

2 $B^0 \to K^{*0}\,J/\Psi$ decay channel

The article focuses in the resonant channel $B^0 \to K^{*0} J/\Psi \to K^+\pi^-\mu^+\mu^-$, by applying Single (SVA) and Multivariate Analysis (MVA) to discriminate signal, S, (final state particles coming from the B^0 meson) from background, B, (final state particles coming from other processes), and comparing the performance and the figure of merit (FOM) obtained. The FOM corresponds to the signal significance (Z), given by

$$FOM = \frac{S}{\sqrt{S+B}} \quad . \tag{1}$$

The B^0 candidate's mass corresponds to the invariant mass of the final state particles $m(K^+\pi^-\mu^+\mu^-)$. Fig. 6 shows the respective histogram for the events with an invariant mass $m \in [5, 5.6]$ GeV.

A proper MVA model shall reduce significantly the background while maintaining most of the signal, relative to the existing pre-selection, which will be demonstrated further in the article (Sec. 8).

3 The CMS detector

The CMS detector (Fig. 5), short for Compact Muon Solenoid, is a general-purpose experiment at the Large Hadron Collider (LHC) facility at CERN. It boasts a distinctive cylindrical design, comprising several subdetectors.

The beamspot, i.e. the region where the collisions between LHC beams occur, is located in the center of the detector. Following this core region is a silicon tracker, precisely designed to trace the trajectories of charged particles. The measured curvature allows to infer their momentum and charge.

Moving outward, there's the electromagnetic calorimeter (ECAL). This subdetector specializes in registering the energy deposition of photons and electrons, from the resulting electromagnetic showers.

Adjacent to the electromagnetic calorimeter is the hadronic calorimeter (HCAL), dedicated to capturing the energy deposited by hadrons.

Further along the detector, the superconducting solenoid takes center stage, generating a 4 T magnetic field, essential for particle trajectory analysis.

Lastly, positioned outside the solenoid, the muon chambers lie. These chambers consist of up to four stations of gas-ionization muon detectors strategically placed amidst the layers of the steel return "yoke".

For the current study of $B^0 \to K^{*0} J/\Psi \to K^+\pi^-\mu^+\mu^-$, the essential CMS subdetectors are: the silicon trackers (responsible for tracking the paths of charged particles, contributing to the measurement of their momenta, and accurate vertex reconstruction, measuring the flight length of

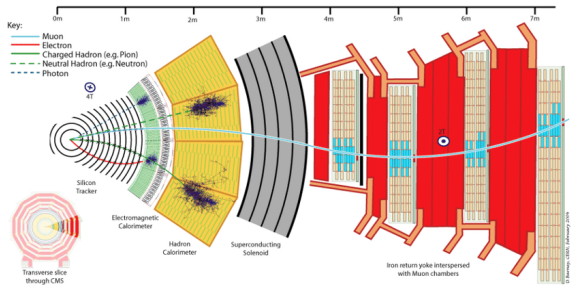

Figure 5. Schematic transverse view of the CMS detector.

the B^0 meson) and the muon chambers (trigger and muon identification). A list of the main variables measured in the detector is provided in Sec. 6.

4 Binned likelihood fit

The data collected by the CMS experiment, at the LHC, during Run 2 is analysed, in order to extract the signal and background yields for the $B^0 \to K^{*0} J/\Psi$ channel. To do so, it's performed a binned likelihood fit. The signal Probability Density Function (PDF) is given by the combination of a Gaussian distribution f_G (3) and a Crystal Ball (CB) function f_{CB} (4):

$$\mathcal{P}_S = C\, f_G + (1 - C)\, f_{CB} \ , \tag{2}$$

where C corresponds to relative importance of the Gaussian function,

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \ . \tag{3}$$

The CB function consists of a Gaussian core portion and a power-law low-end tail, below a certain threshold,

$$f(x;\alpha,n,\mu,\sigma) = \mathcal{N}_{CB} \cdot \begin{cases} \exp\left\{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right\}, & \frac{x-\mu}{\sigma} > \alpha \\ A \cdot \left(B - \frac{x-\mu}{\sigma}\right)^{-n}, & \frac{x-\mu}{\sigma} \leq \alpha \end{cases} , \tag{4}$$

where $\mathcal{N}_{CB}$ is a normalization factor and A and B are functions of the parameters α, n and σ.

The background PDF is simply given by an exponential function:

$$\mathcal{P}_B \equiv f(x;\lambda) = \mathcal{N} \cdot \exp(\lambda x) \ . \tag{5}$$

Lastly, in order to extract the signal (Y_S) and background (Y_B) yields a global PDF is used ($\mathcal{P}_G$), combining both signal and background PDFs:

$$\mathcal{P}_G = Y_S \mathcal{P}_S + Y_B \mathcal{P}_B \ . \tag{6}$$

The global fit (blue) adjusts well to the data points (black), Fig. 6, obtaining signal and background yields of $Y_S = (1806.8 \pm 4.5) \times 10^3$ and $Y_B = (1520.3 \pm 4.5) \times 10^3$ events, where the background (signal) corresponds to the number of events below (above) the green curve.

In the signal fit, it's obtained a $C = 0.485 \pm 0.006$, which means the signal PDF is approximately defined by a 48.5% Gaussian distribution and a 51.5% Crystal Ball function.

The fitted mean parameter is $\mu = 5275.1 \pm 0.1$ MeV, where in the Particle Data Group (PDG) the B^0 meson mass is $m = 5279.66 \pm 0.12$ MeV [4]. This discrepancy on the fitted B^0 mass with respect to the PDG does not include systematic errors in the estimate.

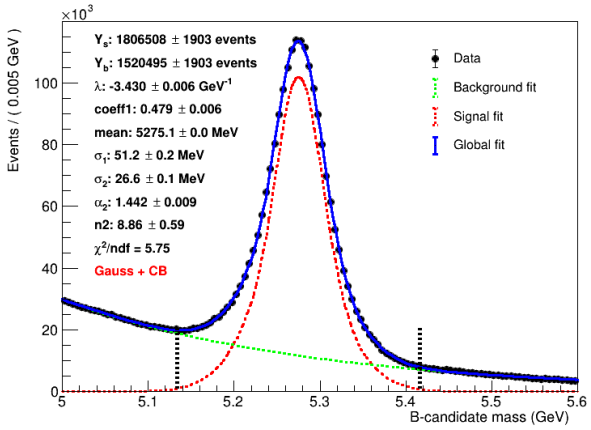

Figure 6. Binned likelihood fit of the dataset collected by the CMS during Run 2. Left and right black dashed vertical lines define the limits of the sideband regions.

5 Signal and background samples

The next step, in order to train the machine learning models, is to obtain a pure sample of signal and background.

5.1 Background sample

The background is estimated from data using the two regions above and below the signal peak, called sidebands. To define the limit of the sidebands, and control the amount of signal leaking into them, we define an effective resolution of the signal peak.

The "effective" standard deviation is calculated, using the gaussian and CB standard deviations (σ_1 and σ_2), and the coefficient C:

$$\sigma_{eff} = \sqrt{C\,\sigma_1^2 + (1 - C)\,\sigma_2^2} \ . \tag{7}$$

The black dashed vertical lines, represented in Fig. 6, correspond to a $3.5\,\sigma_{eff}$ around the *mean*. Hence, $\sim 0.3\%$ of the signal PDF (red) is located in the sideband region, with the left and right sidebands corresponding to the background delimited by m∈[5, 5.134] GeV and m∈[5.416, 5.6] GeV.

The pure background sample will then be composed of the right and left sidebands of the data sample ($B^0 \to K^{*0} J/\Psi$ channel collected with dimuon triggers in 2018).

5.2 Signal sample

A second dataset, containing exclusively Monte Carlo (MC) simulations of the decay $B^0 \to K^{*0} J/\Psi$ is used. This dataset corresponds to a pure signal sample. In the present study only the peak region, delimited by the left and right sideband edges, is chosen (Fig. 7).

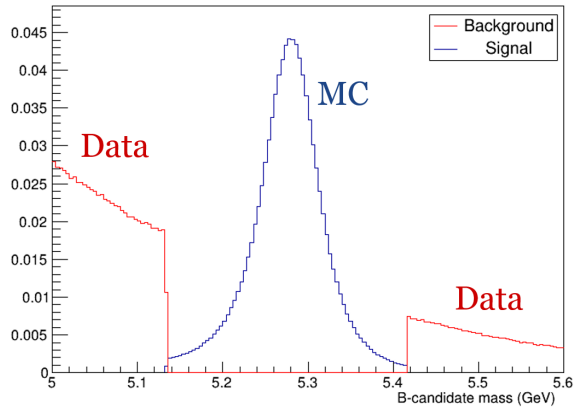

Figure 7. Normalized histogram containing pure signal (MC) and background (Data) samples, delimited by the left and right sideband edges.

In the MC simulations, it's possible to generate as many signal events ($B^0 \to K^{*0} J/\Psi$) as desired, only limited by the computation power. The size of the data sample is determined by the associated luminosity. Also, the background in the peak region must be taken into account. Hence, the FOM (1) is scaled accordingly (10) with the following scale factors for signal and background:

$$f_s = \frac{S^{data}}{S^{MC}} \ , \tag{8}$$

with S^{data} (S^{MC}) corresponding to the signal in the data (MC) sample, and:

$$f_s = \frac{R_3}{R_1 + R_2} \ , \tag{9}$$

with R_1, R_2 and R_3 corresponding to the left sideband, the right sideband and the peak region background (background delimited by both sidebands).

6 B^0 meson variables

The eleven variables used in the selection are listed below, accompanied by Fig. 8, to enhance the comprehension of the underlying physics:

- Flight length: distance between the primary vertex (beamspot) and the secondary vertex (B^0 decay)
- Flight length significance: ratio between the flight length and its error
- Cos(α): cosine of α (angle between the flight direction and the reconstructed B^0 meson momentum)
- Vertex confidence level: probability that the four trajectories ($h^+h^-\mu^+\mu^-$) are originated in a common point
- Negative (positive) track DCA from beamspot: shortest distance from the BS to the negative (positive) hadron continued trajectory (d_0)
- Leading (trailing) muon p_T: highest (lowest) transverse momentum of the two muons
- Negative (positive) track p_T: transverse momentum of the negative (positive) hadron
- B-candidate tag: binary tag separating B^0 from $\bar{B}^0$

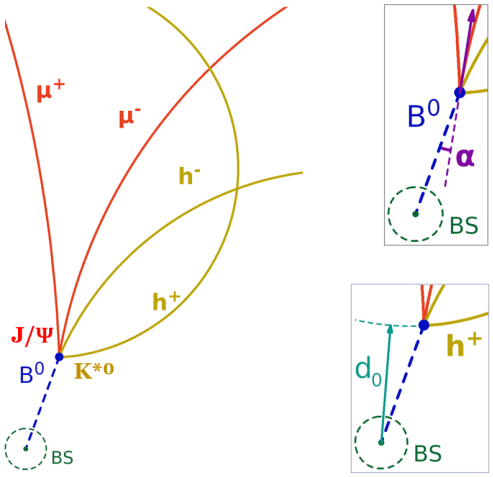

Figure 8. Schematic view and highlighted variables for the production and decay [5] of $B^0 \to K^{*0} J/\Psi$.

The $J/\Psi \to \mu^+\mu^-$ and $K^{*0} \to K^+\pi^-$ decay almost instantaneously, relative to the precision of the detectors, having no measured flight lengths.

6.1 Variable distributions

After obtaining signal and background samples, expressed in Fig. 7, it's presented some of the chosen variables distributions (Figs. 9 and 10).

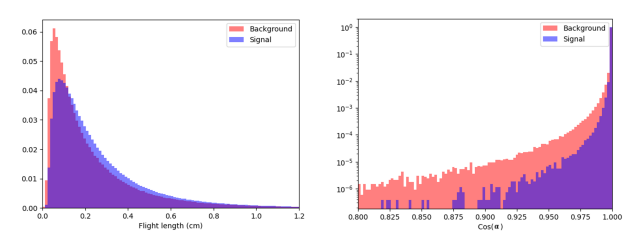

Figure 9. Normalized distributions of the Flight length (left) and Cos(α) (right).

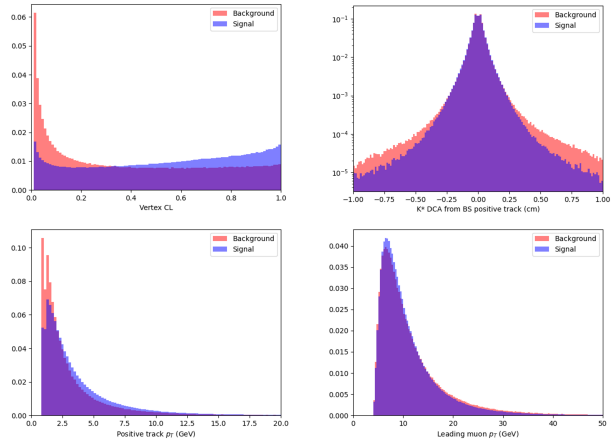

Figure 10. Normalized distributions of the Vertex confidence level (top left), Positive track DCA from BS (top right), Positive track p_T (bottom left) and Leading muon p_T (bottom right).

6.2 Single variate analysis

Prior to the multivariate analysis, a single variate analysis is performed. To measure the quality of the cut, it's calculated the FOM for each variable, by applying vertical cuts along the *x-axis* of the distributions, in order to maximize the signal obtained compared to the background. As explained above, the scaling factors, f_s and f_b, from the B^0 meson fit, need to be applied:

$$FOM_{scaled} = \frac{S \cdot f_s}{\sqrt{S \cdot f_s + B \cdot f_b}} . \quad (10)$$

Figure 11 shows an example of the FOM for the flight length variable. No maximum is obtained beyond preselection.

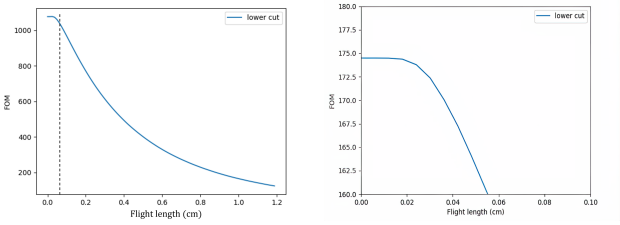

Figure 11. Flight length's FOM (left), "zoomed in" the range [0, 0.1] cm (right). .

In an ideal scenario, the FOM plot would have a maximum, indicating an optimal value for the cut. Regarding the flight length variable (Fig. 11), there is no visible value that optimizes the separation between signal and background. This happens to all the variables chosen, showing that no single variable adds enough information to improve the separation between signal and background.

7 Multivariate Analysis

Since the single variate analysis shows no clear-cut results, machine learning methods are applied: Neural Networks (with PyTorch) and Boosted Decision Trees (with TMVA).

The code used to implement both of these methods can be found in Ref [6]. All the eleven variables represented in Sec. 6 are used in the Neural Networks. For the Boosted Decision Tree, the flight length significance is removed (found to give a worst performance for the chosen hyperparameters).

7.1 Neural networks

The first method uses a Feedforward Neural Network, adapted from the code in [7] to perform a binary classification on the set of variables described in Sec. 6. It is called feedforward because the information flows in one direction from input to output, through any hidden layers, without any cycles or loops (i.e., the network is acyclic), distinguishing it from recurrent or convolutional neural networks which have cyclical connections or spatially aware layers, respectively (Fig. 12).

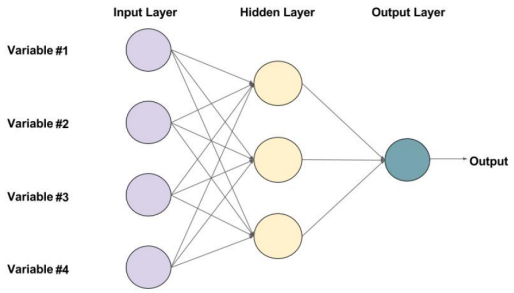

Figure 12. Feedforward Neural Network Architecture with one hidden layer containing 3 neurons [8].

To train this neural network it's used the default settings of 3 hidden layers (64 neurons per layer), with *ReLU* being used as the activation function between hidden layers, and the output layer with one node.

After training for 50 epochs the final accuracy obtained is 0.80 and the Area Under the Curve (AUC), a measure of the model's ability to distinguish between classes, is 0.629 (Fig. 13).

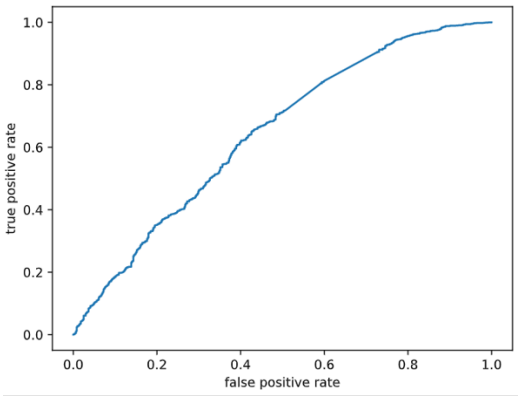

Figure 13. ROC Curve for True Positive Rate Vs False Positive Rate.

7.2 Boosted decision trees

A Boosted Decision Tree (BDT) is a machine learning algorithm that combines multiple Decision Trees (DT) to improve predictive accuracy and reduce overfitting (Fig. 14).

Each DT has a root node, decision (internal) nodes and leaf (terminal) nodes, where the last corresponds to the endpoints of the tree and represents the final predictions. A Boosted Decision Tree (BDT) combines multiple decision trees by training them sequentially, with each tree attempting to correct the errors of its predecessor, thereby progressively improving the model's predictive performance.

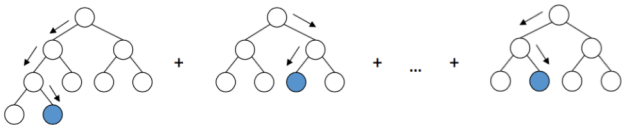

Figure 14. Boosted Decision Tree architecture. Each tree has a root node, decision nodes and leaf nodes.

For the BDT, a *ROOT* library for MVA analysis is employed (TMVA) [9]. As for the hyperparameters, 250 trees are used, with a maximum depth of 5 layers and a minimal node size of 2.5% (minimum percentage of training events required in a leaf node). For the split, 70% of the signal/background are used for training and 30% for testing.

The BDT response using the signal and background samples is shown in Fig. 15.

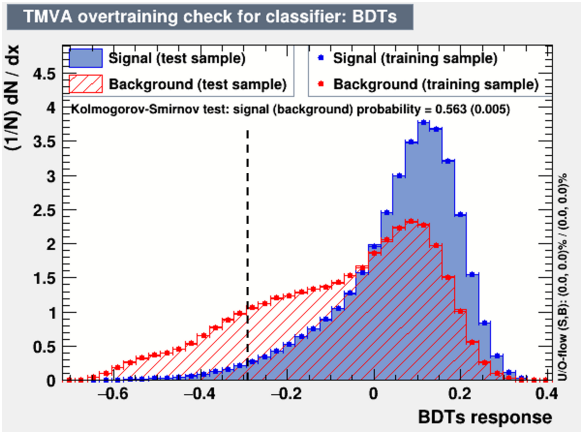

Figure 15. BDT score for the signal and background samples. The black dashed line represents the best cut at -0.296.

As can be seen in Fig. 15, the normalized probability density functions for the test and training samples overlay or are very close to each other across the spectrum of BDT response values, reflecting a lack of overfitting.

The ROC-curve is represented in Fig. 16, showing an Area Under the Curve (AUC) of 0.701.

Lastly, a FOM is obtained for the BDTs response, with a maximum at a BDT score = - 0.296. This is represented schematically by the black dashed line (Figs. 15 and 17).

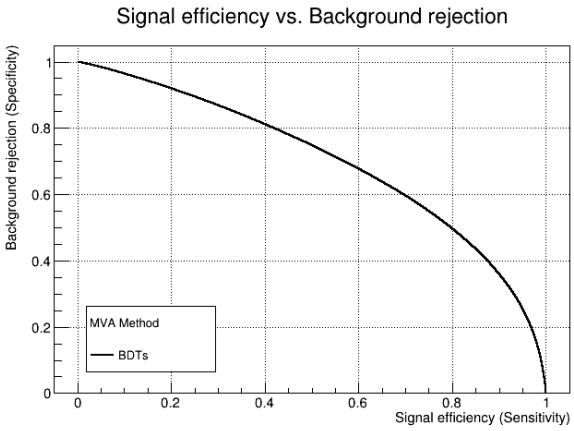

Figure 16. ROC curve for the Background rejection Vs. Signal efficiency.

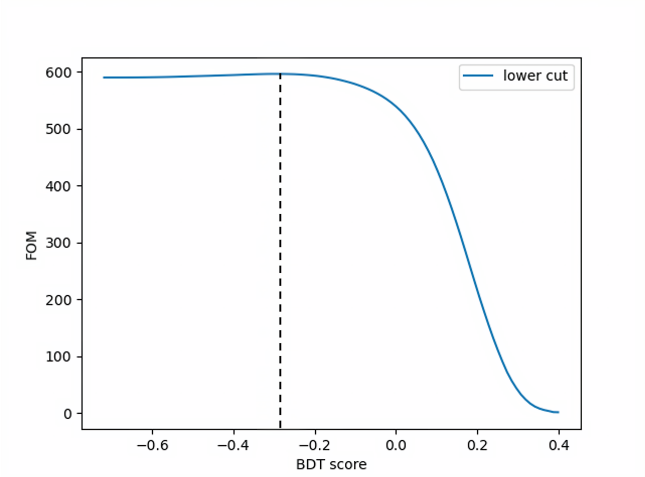

Figure 17. BDT's FOM with the black dashed line at -0.296.

8 Results

In sec. 4 the binned likelihood fit was performed to the dataset (collected by CMS during Run 2), extracting both signal and background yields.

An optimal cut identified in section 7.2 is applied on the data. Only the events with a BDT score > - 0.296 will pass the selection. The *survivors* will be submitted to a binned likelihood fit (Fig. 18), extracting, once again, the signal and background yields (Tab. 1).

Y_S [$\times 10^3$]	Y_B [$\times 10^3$]
1806.8 ± 4.5	1520.3 ± 4.5
1794.7 ± 2.0	1233.2 ± 2.0

Table 1. Signal and background yields before (top) and after (bottom) applying BDT.

As can be seen in the Tab. 1, after BDT there is a background reduction of ~ 20%, while maintaining most of the signal (signal reduction of ~ 1%). This shows the power

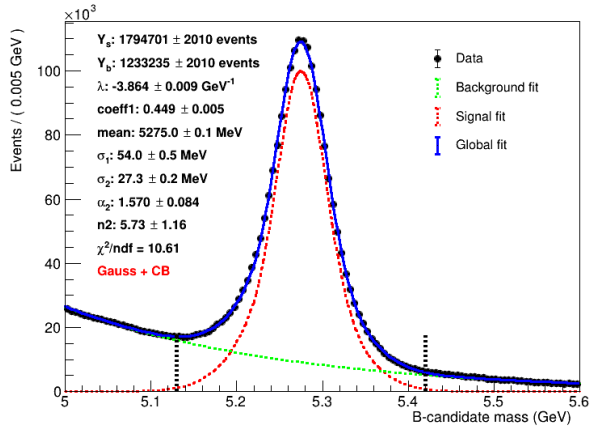

Figure 18. Binned likelihood fit after BDT.

of MVA analysis in optimizing the rejection of background events.

9 Conclusions and next steps

The decay channel $B^0 \rightarrow K^{*0} J/\Psi$ has been extensively studied, choosing eleven key variables in order to discriminate signal from background. For the single variate analysis (SVA) there is no feature that could maximize the FOM, beyond what was achieved with pre-selection. For the MVA analysis, NN and BDT algorithms were explored. Of these, a cut on the BDT output showed improvements in the FOM. This result is then used to reduce the background events while maintaining most of the signal events. As for the NN, the results present a similar predicting capability compared to the BDT (seen in the similar AUC scores).

In the future, the developed tools here created can be applied to the rare $B^0 \rightarrow K^{*0} \mu^+ \mu^-$ decay channel for exploring the flavour anomalies in dedicated datasets collected by CMS.

Acknowledgements

We are indebted to our supervisors, Dr. Alessio Boletti and Prof. Nuno Leonardo, for all the advice, patience and knowledge shared throughout the course of three months, as well as for the insightful comments to the present article. We'd also like to express our gratitude to Simão Costa for the assistance with machine learning tools. Lastly, we thank the LIP organization for the 2023 Summer Internship Program and the continuous science outreach.

References

[1] Phys. Rev. Lett. **131**, 111802 (2023), 2302.02886

[2] A. Greljo, J. Salko, A. Smolkovič, P. Stangl, JHEP **05**, 087 (2023), 2212.10497

[3] N. Gubernari, M. Reboud, D. van Dyk, J. Virto, JHEP **09**, 133 (2022), 2206.03797

[4] *Particle Data Group*, https://pdglive.lbl.gov/Viewer.action

[5] *Heavy-flavour physics @LHC and flavour anomalies*, https://indico.lip.pt/event/1418/contributions/4657/attachments/3910/6129/FlavourAnomalies.pdf

[6] *LFU Summer Internship 2023*, https://github.com/LIP-Flavour-Anomalies-studies/SummerLIP23

[7] *Simão's Machine Learning tutorials*, https://github.com/simao14/2023SummerLIP_tutorials

[8] *Feedforward Neural Networks*, https://learnopencv.com/understanding-feedforward-neural-networks/

[9] *TMVA - Toolkit for Multivariate Data Analysis*, https://doi.org/10.48550/arXiv.physics/0703039