

Training a machine to find the rare decay $B^0 \rightarrow K^*(892)^0 e^+ e^-$ at the Belle experiment

Cyrille Praz (Summer Student)

Supervised by Dr. Simon Wehle

September 6, 2017



Abstract

The Belle experiment was designed to allow for precise measurements of B mesons decays. This document presents the steps needed to train a machine to find the rare decay $B^0 \rightarrow K^*(892)^0 e^+ e^-$ and suppress the background on Monte Carlo (MC) simulated data. The background suppression strategy, based on classifiers and multivariate statistics, is tested and cross-validated by the extraction of the branching ratios $\mathcal{B}(B^0 \rightarrow J/\psi K^*(892)^0)$ and $\mathcal{B}(B^0 \rightarrow \psi(2S) K^*(892)^0)$ on two MC data sets.

Contents

1	Introduction	3
2	The Belle detector and the rare decay $B^0 \rightarrow K^*(892)^0 e^+ e^-$	3
3	Background suppression with classifiers	4
3.1	Preliminary cuts for particle and event selection	5
3.2	Training variables	6
3.2.1	Non continuum suppression	7
3.2.2	Continuum suppression	8
3.3	Training and testing	9
3.3.1	Receiver operating characteristics	10
3.3.2	Figure of merit	10
4	Cross-validation by extraction of branching ratios	11
5	Conclusion and outlook	15

1 Introduction

The Belle experiment was designed to allow for precise measurements of B mesons decays. It was located along the assymetric-energy e^+e^- collider KEKB in Tsukuba, Japan, and acquired data from 1999 to 2010. The KEKB was able to produce B^+B^- and $B^0\bar{B}^0$ pairs by operating at the $\Upsilon(4S)$ resonance energy. This experiment can be used to test very finely the predictions of the Standard Model of particle physics (SM) and look for new physics. For example, deviations in the branching ratios of rare B mesons decays could be interpreted as new particles contributing to the Feynman diagrams.

This document presents the steps needed to train a machine to find the rare decay $B^0 \rightarrow K^*(892)^0 e^+ e^-$ and suppress the background on Monte-Carlo (MC) simulated data. In the first part, the Belle experiment is very briefly described as well as the decay $B^0 \rightarrow K^*(892)^0 e^+ e^-$. In the second part, one builds and tests a background suppression strategy based on classifiers and multivariate statistics. Finally, the procedure is cross-validated by an extraction of the branching ratios $\mathcal{B}(B^0 \rightarrow J/\psi K^*(892)^0)$ and $\mathcal{B}(B^0 \rightarrow \psi(2S) K^*(892)^0)$ on two MC data sets.

2 The Belle detector and the rare decay $B^0 \rightarrow K^*(892)^0 e^+ e^-$

As stated above, the KEKB collider produces B mesons pairs by operating at the $\Upsilon(4S)$ resonance energy. It collides 8.0 GeV e^- and 3.5 GeV e^+ to reach the resonance:

$$M_{\Upsilon(4S)} = (10.5794 \pm 0.0012) \text{ GeV} \approx \sqrt{s} = 2\sqrt{8 \cdot 3.5} \text{ GeV}, \quad (1)$$

where \sqrt{s} is the total center-of-mass energy. Then, $\Upsilon(4S)$ decays into B mesons pairs with a probability $> 96\%$ [3]. The daughter particles are detected by the Belle detector. Figure 1 shows a schematic view of the detector and table 1 lists its main components. The details of its working principles go beyond the framework of this report.

Detector component	Main purpose
Silicon Vertex Detector (SVD)	tracking, vertex locator
Central Drift Chamber (CDC)	tracking, momentum and energy loss measurement
Time Of Flight counter (TOF)	velocity measurement
Aerogel Cherenkov Counter (ACC)	velocity measurement
Thallium doped Cesium Iodine (CsI)	energy measurement
Extreme Forward Calorimeter (EFC)	energy measurement
K_L^0 and μ detection system (KLM)	particle identification

Table 1: Summary of the main components of the Belle detector. The combination of the outputs allows for Particle IDentification (PID).

This analysis aims to find the rare decay $B^0 \rightarrow K^*(892)^0 e^+ e^-$, whose branching ratio is $(1.03^{+0.19}_{-0.17}) \times 10^{-6}$ [3]. This very low value reflects the fact that the SM does not allow this decay on tree-level. Indeed, the Z^0 boson does not couple different generations of quarks: figure 2. The lowest order Feynman diagrams for this decay are either penguin

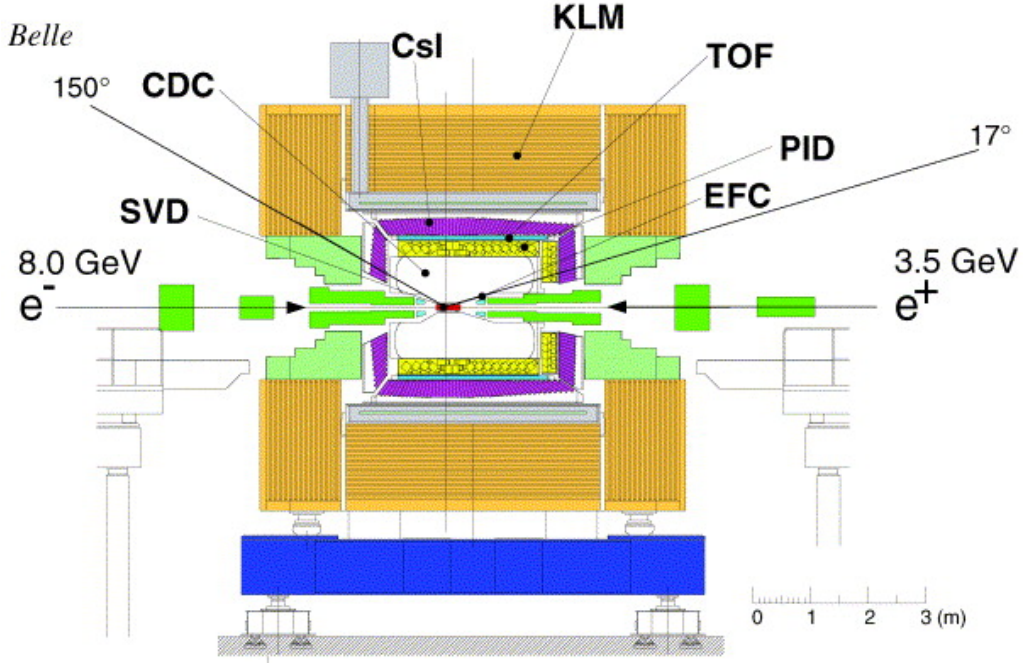


Figure 1: The Belle detector. [4]

diagrams (figure 3) or box diagrams. As a consequence, new particles would enhance or suppress this decay with a relatively large contribution.

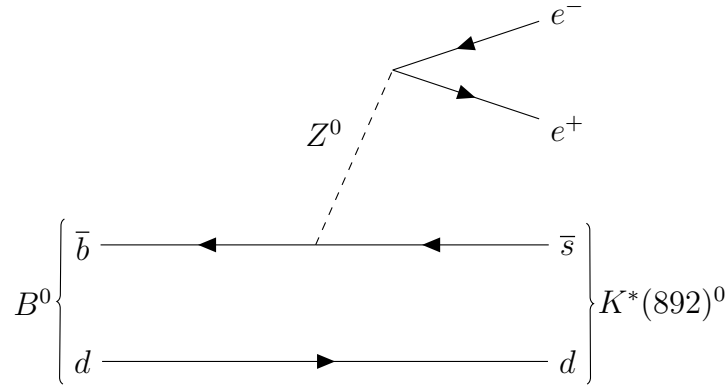


Figure 2: Feynman diagram *forbidden* by the Standard Model.

3 Background suppression with classifiers

In this section, one develops a background suppression strategy based on classifiers. One uses the scikit-learn python library, which provides several classifiers. The physical background is simulated by the Belle generic MC. Its different components are summarised in table 2.

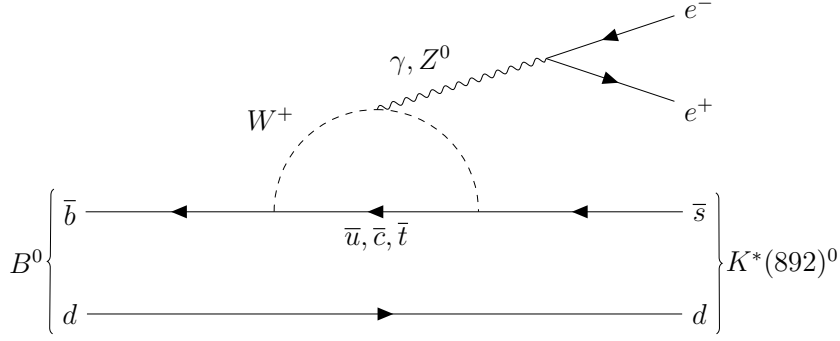


Figure 3: Feynman diagram allowed by the Standard Model.

Name	Description
uds	continuum $e^+e^- \rightarrow u\bar{u}, d\bar{d}, s\bar{s}$
charm	continuum $e^+e^- \rightarrow c\bar{c}$
mixed	$\Upsilon(4S) \rightarrow B^0\bar{B}^0$, with generic B^0 decay
charged	$\Upsilon(4S) \rightarrow B^+B^-$, with generic B^+ decay

Table 2: Description of the events generated by the Belle generic MC. [1]

3.1 Preliminary cuts for particle and event selection

The first stage of the reconstruction is made by the detector itself. By cutting on the particle identification system (PID) outputs it is possible to select the following daughter particles: e^\pm , μ^\pm , π^\pm , K^\pm , K_S^0 , π^0 and γ (see [1] for more details). The next step is to make an event selection. Figure 4 compares the signal and background normalised distributions of the dilepton invariant mass $M_{e^+e^-}$. Three sharp peaks can be observed in the background distribution. The first one comes from $\gamma \rightarrow e^+e^-$, $\pi^0 \rightarrow e^+e^- \gamma$ and non-physical low energy events. The other two peaks correspond to the $J/\psi \rightarrow e^+e^-$ and $\Upsilon(2S) \rightarrow e^+e^-$ decays respectively. In order to suppress these sources of background, three preliminary cuts, summarised in table 3, are applied.

Background source	Selection
$B^0 \rightarrow K^{(*)0}(J/\psi \rightarrow e^+e^-)$	$-0.45 < M_{e^+e^-(\gamma)} - M_{J/\psi} < +0.08 \text{ GeV}/c^2$
$B^0 \rightarrow K^{(*)0}(\Upsilon(2S) \rightarrow e^+e^-)$	$-0.20 < M_{e^+e^-(\gamma)} - M_{\Upsilon(2S)} < +0.08 \text{ GeV}/c^2$
$\gamma \rightarrow e^+e^-$ and $\pi^0 \rightarrow e^+e^- \gamma$	$M_{e^+e^-} > +0.14 \text{ GeV}/c^2$

Table 3: Background sources and corresponding selections. The values for $M_{J/\psi}$ and $M_{\Upsilon(2S)}$ are given by the PDG [3].

On top of that, two other cuts are applied on two variables defined in the $\Upsilon(4S)$ rest frame: the energy difference ΔE and the beam constrained mass M_{bc} defined by

$$M_{bc} = \sqrt{E_{\text{beam}}^2 - p_B^2} \quad (2)$$

and

$$\Delta E = E_B - E_{\text{beam}} \quad (3)$$

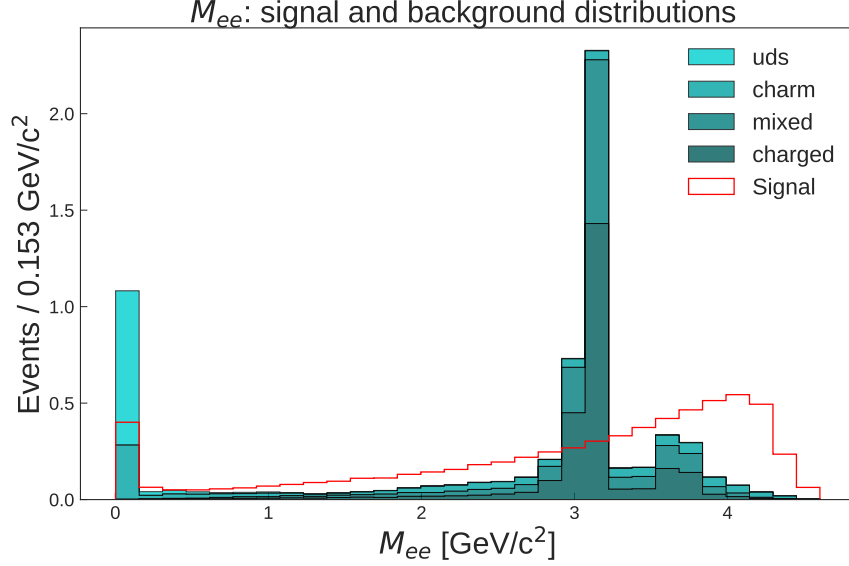


Figure 4: Dilepton invariant mass $M_{e^+e^-}$ normalised distributions. The different sources of background are defined in table 2.

where E_{beam} is the beam energy, E_B the B meson energy and p_B the B meson momentum, all of them defined in the $\Upsilon(4S)$ rest frame. Preliminary cuts on these two variables are given in table 4.

Variable	Selection
M_{bc}	$5.22 < M_{bc} < 5.89 \text{ GeV}/c^2$
ΔE	$-0.15 < \Delta E < +0.15 \text{ GeV}/c^2$

Table 4: Beam constrained variables and corresponding cuts.

3.2 Training variables

A variable is chosen to train the classifier if its distribution allows for a good distinction between signal and background. Figure 5 displays two such variables: ΔE , already defined above, and R_2 , which will be introduced in section 3.2.2. Another point needed to be considered is the possible correlations between the training variables and M_{bc} .

If one or several training variables are correlated to M_{bc} , the classifier may learn in which region of M_{bc} the signal is located and may cause a peak in the M_{bc} background

Variable	Description
P_{tot}	Total momentum of the B^0 candidate
$\cos \theta_B$	Cosine of the angle θ_B between the B^0 candidate and the beam
$\cos \theta_L$	Cosine of the angle θ_L between the direction of e^- and the direction of the e^+e^- system in the B^0 candidate rest frame

Table 5: Description of a set of variables causing an artificial background peaking.

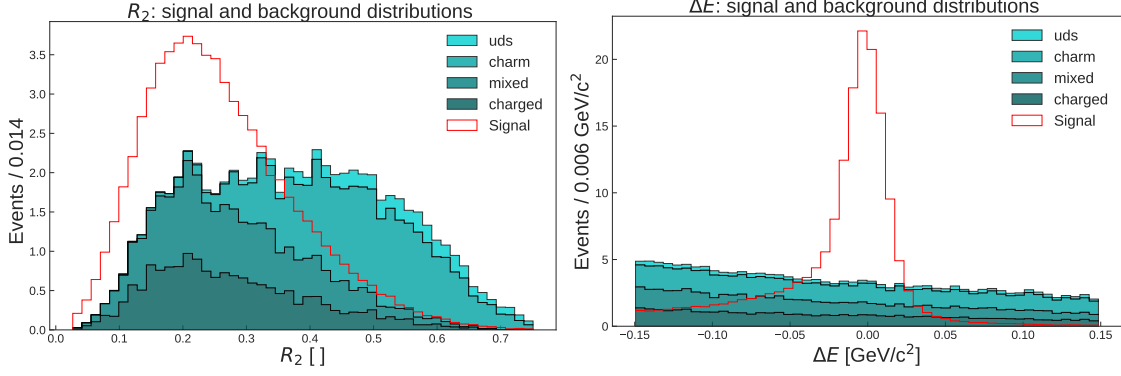


Figure 5: Comparison between signal and background normalised distributions. The different sources of background are defined in table 2.

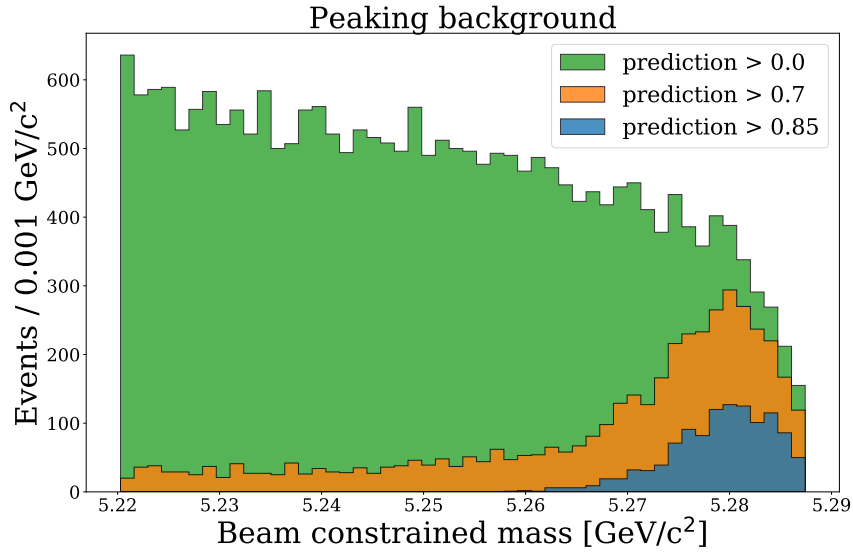


Figure 6: M_{bc} background distribution for 3 cuts based on the outputs (predictions) of a Gradient Boosted Decision Tree trained with a variable correlated to M_{bc} .

distribution. Figure 6 shows this effect with a Gradient Boosted Decision Tree (GBDT) trained on the variables P_{tot} , $\cos\theta_B$ and $\cos\theta_L$ defined in table 5. As the cut on the classifier output goes close to 1, the background exhibits a peak around the B^0 mass. Here, the most problematic training variable is P_{tot} , because it is strongly correlated to M_{bc} , as expected from the energy-momentum relation $E^2 = p^2c^2 + m^2c^4$.

There are two families of training variables depending on which source of background is considered (see again table 2).

3.2.1 Non continuum suppression

In total, 10 variables are used for the non continuum background suppression. They are listed in table 6. Figure 7 shows that there is no linear correlation between these variables and M_{bc} . Higher order correlations could exist, but the results indicate that, if any, it does not cause background peaking.

Variable	Description
ΔE	Energy difference defined in equation (3)
$\cos \theta_B$	Cosine of the angle θ_B between the B^0 candidate and the beam
Ch0_m	$K^*(892)^0$ mass
FT_LL	Result of a Neural Network flavor tagging, see [1]
FT_NN	Result of a Maximum Likelihood flavor tagging, see [1]
χ^2	χ^2 of the vertex fit of the candidate
Δz_{ll}	Distance between the two leptons in the beam direction
Kp_lm_invMass	Invariant mass of $K^+\ell^-$
mom_dir_dev	Momentum direction deviation
dist_to_IP	Distance to the interaction point

Table 6: Description of the training variables used for the non continuum background suppression.

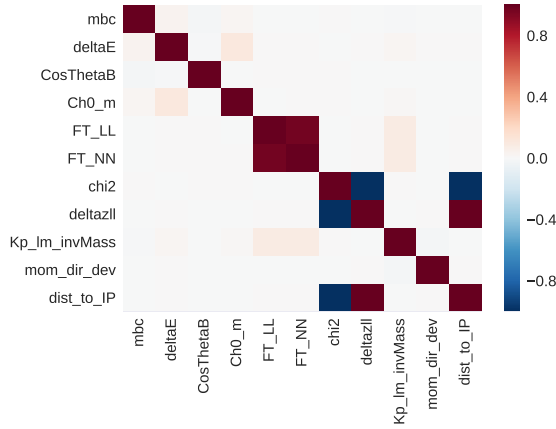


Figure 7: Linear correlation coefficients among the non continuum suppression variables and M_{bc} .

3.2.2 Continuum suppression

The continuum suppression variables look at the angular shape of the events, because the light quarks pairs coming from the annihilation of an electron and positron form back to back jet-like structures. For this purpose, one introduces the Fox-Wolfram moments H_k [5] and the ratio R_k :

$$H_k = \sum_{i,j=1}^N \frac{|\vec{p}_i| |\vec{p}_j| P_k(\cos \theta_{ij})}{E_{visible}^2} \quad (4)$$

and

$$R_k = \frac{H_k}{H_0}, \quad (5)$$

where one sums over all the particles in the considered event. $E_{visible}$ is the total visible energy of the event, p_i the momentum of the i th particle, P_k the k th Legendre polynomial and θ_{ij} the angle between the i th and the j th particle. One constructs also Super Fox-Wolfram moments [6] and Cleo Cones variables [7].

In total, 30 continuum suppression variables are selected for the training. Again, figure

8 shows that there is no linear correlation between these variables and M_{bc} .

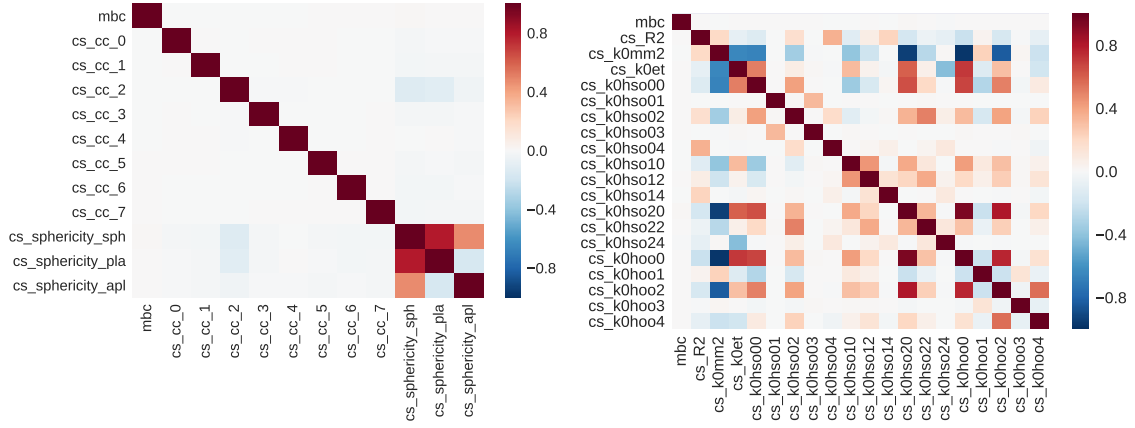


Figure 8: Linear correlation coefficients among the continuum suppression variables and M_{bc} .

3.3 Training and testing

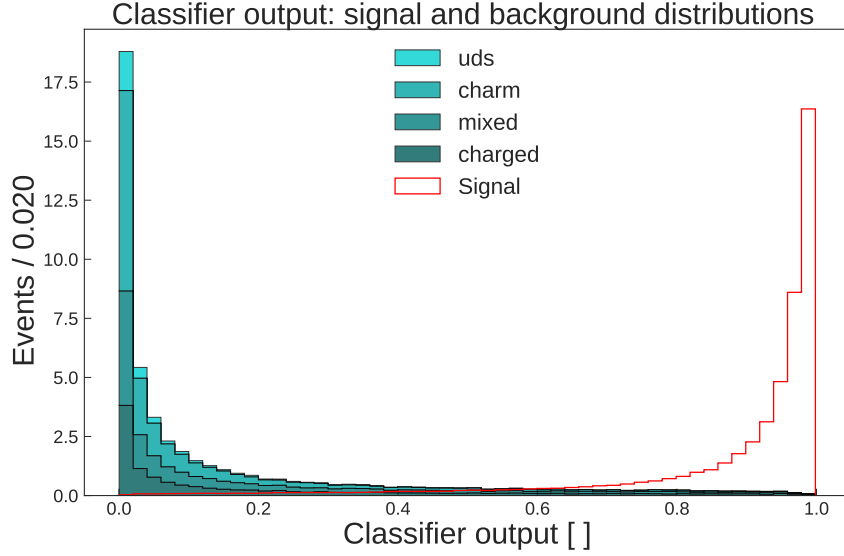


Figure 9: Signal and background separation obtained by applying a GBDT. The distributions are normalised.

After having chosen the training variables, the considered classifiers are trained on a MC data set (the train set) and tested on another data set (the test set) to remain unbiased. Figure 9 displays the signal and background separation obtained from the outputs (predictions) of a GBDT trained with the continuum and non continuum suppression variables together. One can see a good separation between signal and background. In this subsection, one introduces several quantities to evaluate and compare the classifiers.

3.3.1 Receiver operating characteristics

In order to evaluate the quality of a classification, two quantities are defined: the efficiency and the purity. They are given by the two following relations:

$$\text{efficiency} = \frac{n_T}{N_T} \quad (6)$$

and

$$\text{purity} = \frac{n_T}{n_T + n_F}, \quad (7)$$

where n_T (n_F) is the number of true (false) selected events and N_T is the total number of true events.

When the efficiency and the purity are computed as a function of cuts on the classifier output, one obtains curves called receiver operating characteristics (ROC). Figure 10 displays the ROC curves for 3 classifiers. The area is simply the area under the curve, while the score is a value returned by the scikit-learn classifiers. These 3 classifiers have been chosen because of their good scores and their robustness against overfitting, an effect which may occur if, by example, the maximal depth of a decision tree is chosen too high. Figure 11 illustrates a typical case of overfitting: the decision tree achieves a perfect separation between signal and background when applied to the training data set, but shows poor results on the testing data set. Here, the maximal depth of the decision tree was set to 2000 (see [9] for details).

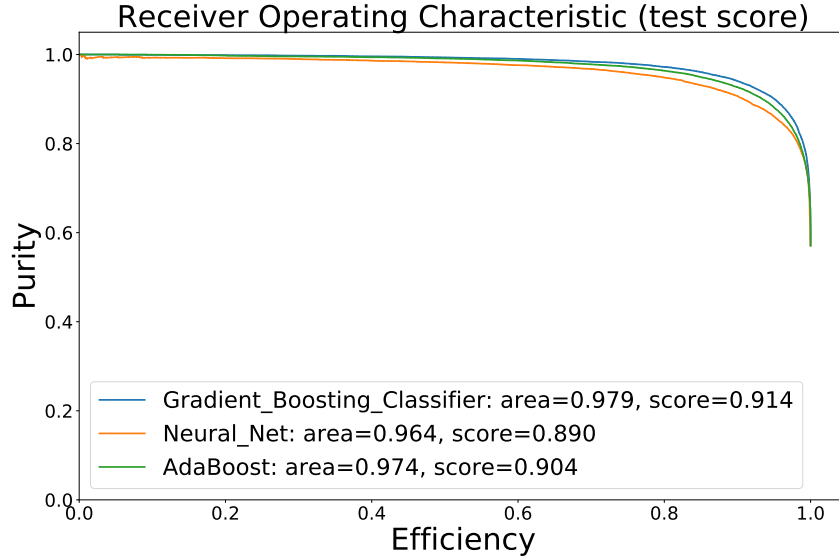


Figure 10: ROC curves for 3 classifiers applied to the test data set.

3.3.2 Figure of merit

Another quantity used to compare the classifiers and choose a right cut on their outputs is the Figure of Merit (FOM), defined by

$$\text{FOM} = \frac{n_{\text{sig}}}{\sqrt{n_{\text{sig}} + n_{\text{bkg}}}} \Big|_{5.27 \text{ GeV}/c^2 < M_{bc} < 5.29 \text{ GeV}/c^2}, \quad (8)$$

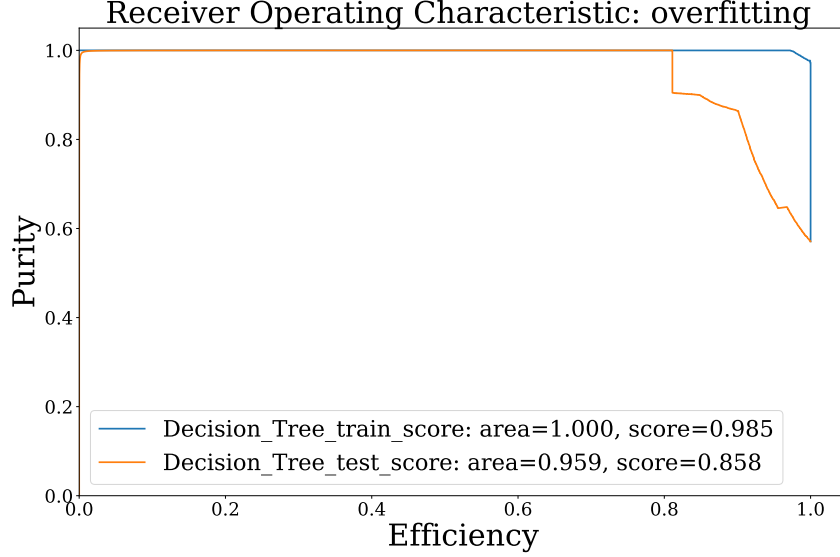


Figure 11: Comparison between 2 ROC curves obtained by applying a Decision Tree on the test and on the train data sets respectively.

where n_{sig} (n_{bkg}) is the expected number of signal candidates in the Belle data after the selection. This expected number of signal candidates is given by

$$n_{\text{sig}} = \varepsilon \cdot \mathcal{B}^{\text{PDG}}(B^0 \rightarrow K^*(892)^0 e^+ e^-) \cdot 2 \cdot N_{B\bar{B}} \cdot \mathcal{B}^{\text{PDG}}(\Upsilon(4S) \rightarrow B^0 \bar{B}^0), \quad (9)$$

where ε is the efficiency, $N_{B\bar{B}}$ the number of B mesons pairs produced at Belle, which is $(772 \pm 11) \times 10^6$ [1], and the \mathcal{B}^{PDG} are the PDG values of the branching ratios, given by $(1.03^{+0.19}_{-0.17}) \times 10^{-6}$ for $B^0 \rightarrow K^*(892)^0 e^+ e^-$ and $(48.6 \pm 0.6)\%$ for $\Upsilon(4S) \rightarrow B^0 \bar{B}^0$ [3].

Figure 12 shows the FOM as a function of cuts on three classifiers outputs. These curves consist of 200 points uniformly distributed among the outputs of each classifier. The reason why the Ada Boost curve does not extend to 0 and 1 is that all of its outputs are close to 0.5.

The cuts providing the highest FOM are listed in table 7. The best FOM is obtained by combining two GBDT trained on the continuum and non continuum suppression variables separately. The signal and background distributions obtained after applying these last cuts are displayed in figure 13. The signal distribution was rescaled such that its area corresponds to n_{sig} . In particular, one can see that there is no background peaking in the signal region.

4 Cross-validation by extraction of branching ratios

It is possible to cross-validate the procedure described above by extracting the two branching ratios $\mathcal{B}(B^0 \rightarrow J/\psi K^*(892)^0)$ and $\mathcal{B}(B^0 \rightarrow \psi(2S) K^*(892)^0)$ on generic MC data sets. If everything is correct, one should obtain the Event Generator (EvtGen) parameters used to simulate these data. To extract these branching ratios, one applies the classifier in parallel to the signal and the generic MC data sets after a cut on $M_{e^+e^-}$ which vetoes everything outside the considered charmonium peak (see again figure 4). Then, the

Classifier	Classifier output cut(s)	FOM	efficiency [%]	n_{sig}	n_{bkg}
Gradient Boosting	0.95 (ncs) & 0.49 (cs)	4.85	3.89	31	10
Gradient Boosting	0.94	4.32	4.90	39	43
Ada Boosting	0.52	4.11	5.40	43	67
Neural Network	0.95	3.04	4.42	35	100

Table 7: Maximum FOM and corresponding cuts, efficiencies and expected numbers of signal (n_{sig}) and background (n_{bkg}) events. The best result is obtained by training two GBDT on the continuum suppression (cs) and non continuum suppression (ncs) variables separately.

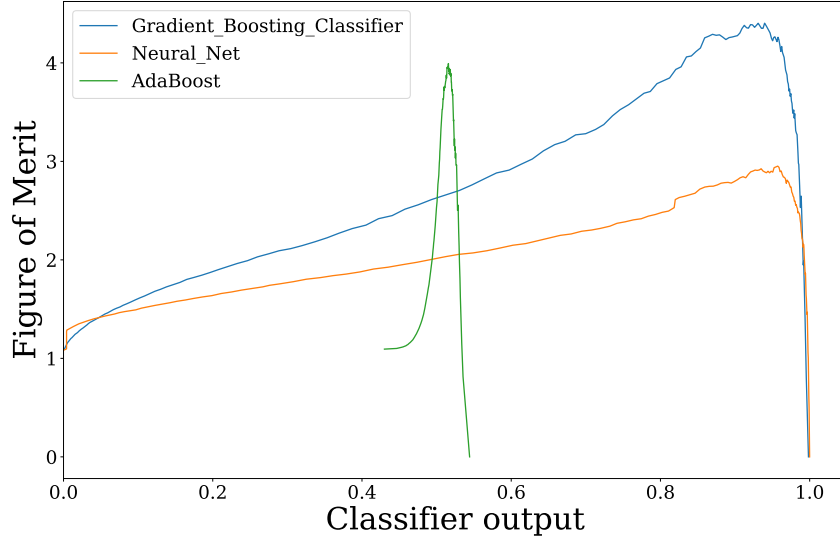


Figure 12: FOM, defined in equation (8), as a function of cuts based on the outputs of 3 classifiers.

branching ratio is given by

$$\mathcal{B}(B^0 \rightarrow X_{c\bar{c}} K^*(892)^0) = \frac{N_{obs}}{\varepsilon \cdot \mathcal{B}^{PDG}(X_{c\bar{c}} \rightarrow e^+ e^-) \cdot 2 \cdot N_{B\bar{B}} \cdot \mathcal{B}^{PDG}(\Upsilon(4S) \rightarrow B^0 \bar{B}^0)}, \quad (10)$$

where N_{obs} is the number of generic MC candidates remaining after the selection, $X_{c\bar{c}} = J/\psi$ or $\psi(2S)$, $N_{B\bar{B}}$ the number of B mesons pairs produced at Belle and ε the reconstruction efficiency, defined with its uncertainty by the system

$$\begin{cases} \varepsilon = \frac{N_{rec}}{N_{gen}} \\ \sigma_\varepsilon = \sqrt{\frac{N_{rec}(N_{gen} - N_{rec})}{N_{gen}^3}}, \end{cases} \quad (11)$$

where N_{rec} (N_{gen}) is the number reconstructed (generated) events. N_{obs} is extracted after the selection by fitting the M_{bc} distribution on the generic MC data set. A Crystall Ball model [2], depending on three parameters (m_0 , σ and α), is used to fit the signal:

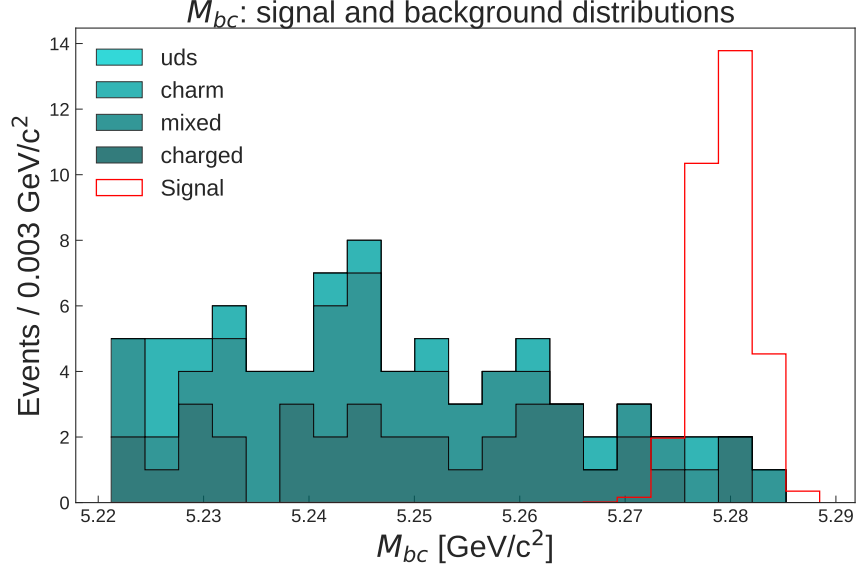


Figure 13: Comparison between signal and background distributions after the cuts providing the best FOM. The signal distribution was rescaled such that its area corresponds to n_{sig} defined in equation (9). The different sources of background are defined in table 2.

$$P^{CB}(M_{\text{bc}}, m_0, \sigma, \alpha, n) = \begin{cases} e^{-\frac{(M_{\text{bc}} - m_0)^2}{2\sigma^2}} & \text{if } M_{\text{bc}} > m_0 - \alpha\sigma \\ \left(\frac{n}{\alpha}\right)^n \cdot e^{-\frac{\alpha^2}{2}} \cdot \left(\frac{m_0 - M_{\text{bc}}}{\sigma} + \frac{n}{\alpha} - \alpha\right)^{-n} & \text{if } M_{\text{bc}} \leq m_0 - \alpha\sigma. \end{cases} \quad (12)$$

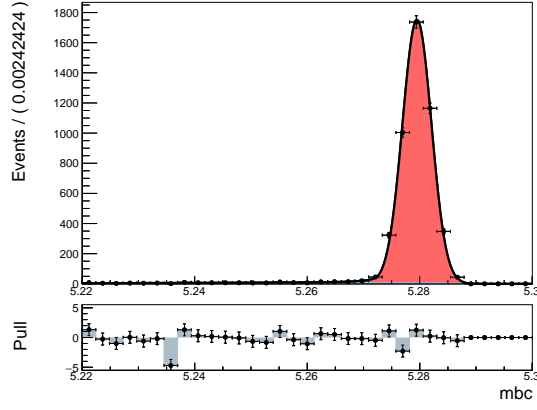
The background is fitted with a model introduced by the Argus Collaboration [8] and depending on two parameters (m_0 and α):

$$P^{ARGUS}(M_{\text{bc}}, m_0, \alpha) = M_{\text{bc}} \cdot \sqrt{1 - \left(\frac{M_{\text{bc}}}{m_0}\right)^2} \cdot e^{-\alpha(1 - (M_{\text{bc}}/m_0)^2)}. \quad (13)$$

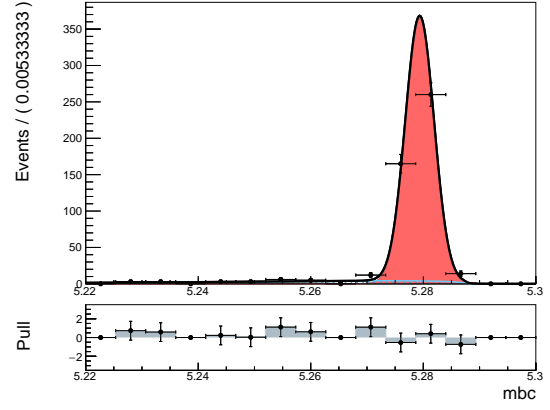
Two generic MC data sets were used. Figure 14 displays the fits and table 8 lists the corresponding results. Here, only the statistical errors are taken into account. The branching ratio $\mathcal{B}(B^0 \rightarrow J/\psi K^*(892)^0)$ measured on the two data sets is in good agreement with the EvtGen parameter used to simulate these data. In the case of $\mathcal{B}(B^0 \rightarrow \psi(2S)K^*(892)^0)$, two values are mentioned in the EvtGen script. One of them is in agreement within the statistical errors of the measurement; the other value could indicate the presence of a systematic error.

	$X_{c\bar{c}} = J/\psi [\times 10^{-3}]$	$X_{c\bar{c}} = \psi(2S) [\times 10^{-4}]$
$\mathcal{B}^{\text{First}}(B^0 \rightarrow X_{c\bar{c}}K^*(892)^0)$	1.31 ± 0.05	8.3 ± 0.5
$\mathcal{B}^{\text{Second}}(B^0 \rightarrow X_{c\bar{c}}K^*(892)^0)$	1.36 ± 0.04	8.0 ± 0.5
$\mathcal{B}^{\text{EvtGen}}(B^0 \rightarrow X_{c\bar{c}}K^*(892)^0)$	1.31	7.2

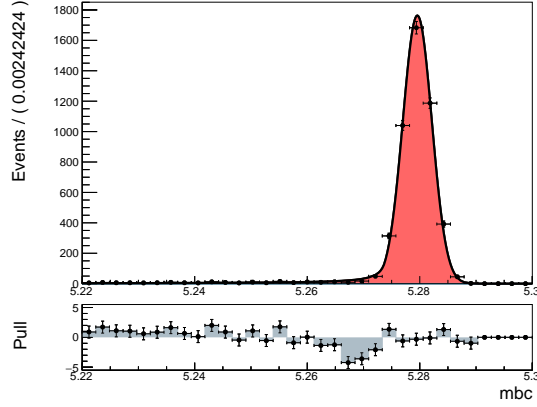
Table 8: Extracted branching ratios from two generic MC data sets (First and Second) and comparison with the EvtGen parameters used to simulate the data. In the case of $\mathcal{B}(B^0 \rightarrow \psi(2S)K^*(892)^0)$, the value 8.0 is also mentioned in the EvtGen script.



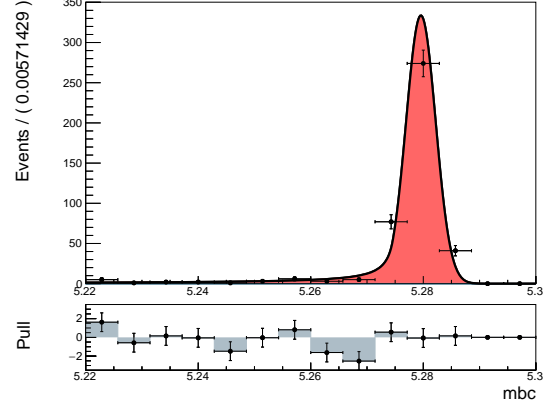
(a) First data set, $M_{e^+e^-}$ cut around $M_{J/\psi}$, M_{bc} is given in GeV/c^2 .



(b) First data set, $M_{e^+e^-}$ cut around $M_{\psi(2S)}$, M_{bc} is given in GeV/c^2 .



(c) Second data set, $M_{e^+e^-}$ cut around $M_{J/\psi}$, M_{bc} is given in GeV/c^2 .



(d) Second data set, $M_{e^+e^-}$ cut around $M_{\psi(2S)}$, M_{bc} is given in GeV/c^2 .

Figure 14: M_{bc} fits needed to extract $\mathcal{B}(B^0 \rightarrow J/\psi K^*(892)^0)$ and $\mathcal{B}(B^0 \rightarrow \psi(2S) K^*(892)^0)$ from MC data sets. The models used to fit the signal and the background are defined in equations (12) and (13) respectively.

5 Conclusion and outlook

The Gradient Boosting Classifier was proven robust against overfitting and allowing for a higher FOM than several other classifiers. The best FOM was obtained by combining two of them trained on the continuum and non continuum suppression variables separately.

The method was cross-validated by extracting the two branching ratios $\mathcal{B}(B^0 \rightarrow J/\psi K^*(892)^0)$ and $\mathcal{B}(B^0 \rightarrow \psi(2S)K^*(892)^0)$ from generic MC data sets. The results were in agreement with the EvtGen parameters within statistical errors. However, in the latter case, more tests are necessary to determine which EvtGen value was used to simulate the data.

In any case, this study should be completed by looking for differences between MC and real data and other sources of systematic uncertainties. The Belle II experiment, planned to be started in 2018 and aiming to reach a higher luminosity, will allow for much more precise measurements.

Acknowledgements

I would like to express my deep gratitude to DESY for having given me the opportunity to participate in this Summer Student Programme. I would also like to thank Dr. Simon Wehle, my supervisor, for his guidance and support. His willingness to give his time so generously has been very much appreciated.

References

- [1] S. Wehle, *Angular Analysis of $B \rightarrow K^* \ell \ell$ and Search for $B \rightarrow K^* \tau \tau$ at the BELLE Experiment*, DESY-THESIS-2016-025 (2016).
- [2] T. Skwarnicki, *A Study of the radiative cascade transitions between the upsilon-prime and upsilon resonances*, DESY-F31-86-02 (1986).
- [3] C. Patrignani *et al.* (Particle Data Group), *Chin. Phys. C*, **40**, 100001 (2016).
- [4] A. Abashian *et al.*, *The Belle detector*, *Nucl. Instrum. Methods* **A479**, 1 (2002).
- [5] Geoffrey C. Fox *et al.*, *Observables for the Analysis of Event Shapes in e^+e^- Annihilation and Other Processes*, *Phys. Rev. Lett.* **41**, 1581 (1978).
- [6] S. H. Lee *et al.*, *Evidence for $B^0 \rightarrow \pi^0 \pi^0$* , *Phys. Rev. Lett.* **91**, 261801 (2003).
- [7] D. M. Asner *et al.*, *Search for exclusive charmless hadronic B decays*, *Phys. Rev. D* **53**, 1039 (1996).
- [8] H. Albrecht *et al.*, *Measurement of the polarization in the decay $B \rightarrow J/\psi K^*$* , *Phys. Lett. B* **340**, 217-220 (1994).
- [9] Pedregosa *et al.*, *Scikit-learn: Machine Learning in Python*, *JMLR* **12**, 2825-2830 (2011).