Department of Physics and Astronomy
University of Heidelberg

Bachelor Thesis in Physics
submitted by

Dominik Wilken

born in Meppen (Germany)

2021

Optimizing event selection for the search of the very rare $B^0 \to \mu^+\mu^-$ decay

This bachelor thesis has been carried out by Dominik Wilken at the
Physikalisches Institut in Heidelberg
under the supervision of
Prof. Dr. Stephanie Hansmann-Menzemer

Abstract

The very rare decay $B^0 \to \mu^+\mu^-$ is one of the channels studied for potential signs of new physics at the LHCb experiment at CERN. So far the decay has not been observed and due to the rarity of the decay it is necessary to understand the background contributions and be able to separate those events from the signal events. One dominant source of background consists of hadronic decays of the for $B^0_{(s)} \to hh'$ where the hadrons in the final state are misidentified as muons. To achieve a better separation between signal and the hadronic background a machine learning based analysis is performed. This classifier is trained on a combination of topological decay variables and particle identification variables. This classifier is able to reduce the amount of expected background significantly, compared to previous versions of the $B^0 \to \mu^+\mu^-$ analysis.

Zusammenfassung

Der seltene Zerfall $B^0 \to \mu^+\mu^-$ ist einer der Zerfälle, die am LHCb Experiments des CERN untersucht werden um nach Anzeichen für neue Physik zu suchen. Die Seltenheit des Zerfalls macht es nötig den Untergrund gut zu verstehen und zu identifizieren. Eine der dominanten Quellen von Untergrund sind hadronische Zerfälle der Form $B^0_{(s)} \to hh'$, in denen die Hadronen als Myonen falsch identifiziert werden. Eine von maschinenellem Lernen gestützte multivariate Analyse wird verwendet um die Separation von Signal und hadronischem Untergrund zu verbessern. Als Lernmaterial für die Kategorisierung dienen eine Kombination aus topologischen Zerfallsvariablen sowie Variablen der Teilchenidentifikation. Die Menge an erwarteten Untergrund Ereignissen lässt sich so stark verringern, im Vergleich zum vorherigen Stand der $B^0 \to \mu^+\mu^-$ Analyse.

Contents

1 Introduction **4**

2 Theoretical Background **5**

2.1 The Standard Model 5

2.1.1 Quarks 5

2.1.2 Leptons 6

2.1.3 Exchange Bosons 6

2.2 Rare Decays 7

2.2.1 Purely Leptonic Decays 7

2.2.2 Sensitivity for New Physics 9

3 The LHCb Detector **10**

3.1 The LHC 10

3.2 The LHCb Experiment 10

3.2.1 Tracking System 10

3.2.2 Particle Identifcation 12

3.2.3 Combined Particle Identification 12

3.2.4 Trigger System 13

4 Analysis and Tools **15**

4.1 Background Contributions 15

4.2 Used Monte Carlo Samples 16

4.3 Punzi Figure of Merit 17

4.4 Multivariate Classifier via Machine Learning 18

4.4.1 Boosted Decision Trees 18

4.4.2 XGBoost 19

4.4.3 Training a Multivariate Classifier 19

4.4.4 Evaluating the Classifier 20

5 Preselection **21**

6 Reducing Background via a MVA **23**
6.1 Performance of ProbNN PID 23
6.2 Training the MVA 25
6.2.1 Input Variables for the MVA 25
6.2.2 Setup of the Training 31
6.2.3 Performance 32

7 Conclusion & Outlook **36**

Chapter 1

Introduction

The Standard Model (SM) of particle physics is one of the most thoroughly tested theories. Yet it is known that the SM does not completely describe the universe. There are several phenomena that can so far not be explain with the Standard Model like dark matter. Other phenomena such as the discovery of non-zero neutrino masses straight up contradict the current SM. Not even to mention the incompatibility with general relativity.

Therefore it has long been known that the SM is incomplete and many theories of physics beyond the Standard Model have been put forward. Yet so far most predictions made by these theories have either been ruled out by experiments or occur on energy scales which cannot be reached with current technology.

However it is possible to probe these potential New Physics phenomena at lower, currently accesible energies in precision measurement. The analysis that is of interest in this thesis concerns the very rare decay of the neutral B meson into two muons. The main challenge - besides producing a sufficient number of collisions - in measuring this decay is the reduction of background. This thesis focuses on the hadronic background of the form $B^0_{(s)} \to hh'$ where the hadrons are misidentified as muons. It aims to use machine learning methods to improve the rejection of background events by using a combination of topological and particle identification information.

The decay in question is quite similar to the decay of the strange neutral B meson into two muons which while also being very rare occurs far more often than for the regular B meson and has therefore already been observed. The observation was originally confirmed in 2013 by LHCb and CMS.

Chapter 2

Theoretical Background

2.1 The Standard Model

The Standard Model (SM) of particle physics is so far the best working theory to describe all known elementary particles and their interactions. The quantum field theory describes three of the four fundamental forces known to date: the electromagnetic, strong, and weak force - notably excluding gravity.

The Standard Model consists of 24 particles, which are divided into 12 fermions, that have a half integer spin and 12 bosons, that have an integer spin. The bosons are responsible for mediating the fundamental interactions, except for the Higgs bosons, which is responsible for giving mass to elementary particles.

2.1.1 Quarks

There are two categories of fermions, one being quarks and the other being leptons. The quarks exist in three generations of which each consist of an up-type and a down-type quark. As shown in Table 2.1 quarks hold an electric charge of either $+\frac{2}{3}$ e or $-\frac{1}{3}$ e.

		1. Generation	2. Generation	3. Generation
Up-type	Particle	up quark u	charm quark c	top quark t
	Charge		$+\frac{2}{3}$ e	
	Mass	0.0022 GeV	1.28 GeV	173.1 GeV
Down-type	Particle	down quark d	strange quark s	bottom quark b
	Charge		$-\frac{1}{3}$ e	
	Mass	0.0047 GeV	.096 GeV	4.18 GeV

Table 2.1: Quarks

Besides an electric charge quarks also carry a colour charge, allowing them to interact via the strong force. As a consequence quarks only appear in bound states that are in total colour neutral, i.e. as baryons consisting of three quarks (like neutrons and protons) or as mesons consisting of a quark anti quark pair (like pions or B mesons).

2.1.2 Leptons

		1. Generation	2. Generation	3. Generation
Neutrinos	Particle	e neutrino ν_e	μ neutrino ν_μ	τ neutrino ν_τ
	Charge	0		
	Mass	$< 2.2 \cdot 10^{-9}$ GeV	$< .00017$ GeV	$< .018$ GeV
Charged Leptons	Particle	electron e	muon μ	tau τ
	Charge	-1 e		
	Mass	0.000511 GeV	0.105 GeV	1.7768 GeV

Table 2.2: Leptons

Similarly to the quarks, the leptons are also categorized in three generations, each consisting of a charged lepton and a corresponding neutrino. In contrast to the quarks they can only interact via the weak and electromagnetic force.

2.1.3 Exchange Bosons

	Electromagnetic	Strong Force	Weak Force	
Particle	photon γ	gluons g	W bosons $W^\pm$	Z bosons Z
Mass	0	0	80.39 GeV	91.19 GeV
Charge	0	0	± 1 e	0

Table 2.3: Exchange bosons

Electromagnetic Force

Electromagnetism is described by the theory of Quantum Electrodynamics. All particles with an electric charge take part in electromagnetic interactions, which are mediated by the photon. This kind of interaction is the basis for the formation of atoms and molecules.

Strong Force

The strong force is mediated by 8 gluons and couples only to particles having a color charge, those being quarks and gluons. Therefore the strong force is responsible for holding together all hadrons, notably including nucleons and B mesons.

Weak Force

The weak force is exchanged by charged $W^{\pm}$ bosons and neutral Z bosons and couples to both quarks and leptons - especially also to neutrinos which thereby is the only force that interacts with them.

Also weak interactions are the only way for quarks to change flavour as in the nuclear β decay, where a d quark is turned into a u quark (therefore a neutron turns into a proton). The mechanism of quark flavour change is described by the Cabibbo-Kobayashi-Maskawa Matrix (CKM matrix for short) V_{CKM}. Notably those flavour changing transitions are only possible if also an electric charge is being transmitted. The CKM matrix is unitary and the transition probabilities are proportional to the absolute square of its entries $|V_{qq'}|^2$.

$$\begin{bmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{bmatrix} = \begin{bmatrix} 0.974901 \pm 0.00011 & 0.22650 \pm 0.00048 & 0.00361^{+0.00011}_{-0.00009} \\ 0.22636 \pm 0.00048 & 0.97320 \pm 0.00011 & 0.04053^{+0.00083}_{-0.000035} \\ 0.00854^{+0.00023}_{-0.00016} & 0.03978^{+0.00082}_{-0.00060} & 0.999172^{+0.000024}_{-0.000035} \end{bmatrix}$$

As can be seen above transitions within the same generation are strongly preferred. [13]

2.2 Rare Decays

Rare decays are - as the name suggests - particle decays that only occur at a small rate, that are prohibited in the SM on tree level, but are possible through loop processes and therefore significantly suppressed.

2.2.1 Purely Leptonic Decays

One kind of rare decays are the decays of B mesons into a purely leptonic state, i.e. charged leptons. While the decay into neutrinos would theoretically also be possible, they would be virtually undetectable and even rarer.

In the Standard Model the branching ratio is given by

	$\mathcal{B}(B_s \to \ell^+\ell^-)$	$\mathcal{B}(B_d \to \ell^+\ell^-)$
$\ell^\pm = e^\pm$	$(8.24 \pm 0.36) \times 10^{-14}$	$(2.63 \pm 0.32) \times 10^{-15}$
$\ell^\pm = \mu^\pm$	$(3.52 \pm 0.15) \times 10^{-9}$	$(1.12 \pm 0.12) \times 10^{-10}$
$\ell^\pm = \tau^\pm$	$(7.46 \pm 0.30) \times 10^{-7}$	$(2.35 \pm 0.24) \times 10^{-8}$

Table 2.4: Branching ratios predicted by the Standard Model

$$\mathcal{B}(B_q^0 \to \ell^+\ell^-)_{SM} = \tau_{B_q} \frac{G_F^2 m_W^4 \sin^4\theta_W}{16\pi^2} f_{B_q}^2 m_\ell^2 m_{B_q} \sqrt{1 - \frac{4m_\ell^2}{m_{B_q}^2}} |V_{tb}V_{tq}^*|^2 |C_{10}^{SM}|^2 \qquad (2.1)$$

Here V means the CKM matrix element, which describes the amplitudes for flavour changing quark interactions

The decay is strongly helicity suppressed as the decay goes from the pseudoscalar $B_{(s)}$ to two spin $\frac{1}{2}$ leptons resulting in a $\propto m_\ell^3$ proportionality leading to a smaller branching ratio for lighter leptons, thereby giving branching ratios for the decay into electrons that are currently not detectable. On the other hand the decay into τs is challenging to detect as well, due to the short lifetime of the τ and the subsequent decay containing at least one neutrino.

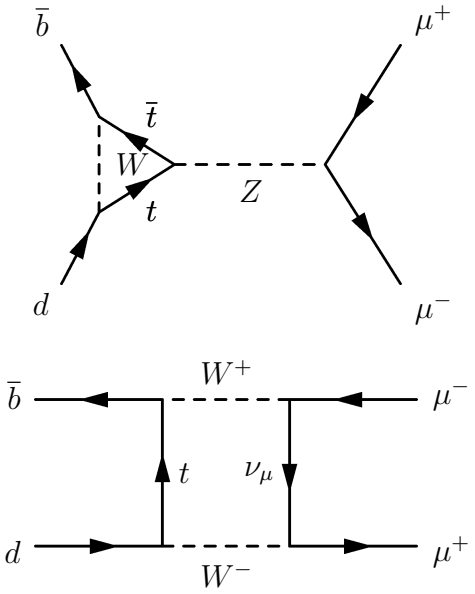


Figure 2.1: Main SM contributions to $B^0 \to \mu^+\mu^-$

2.2.2 Sensitivity for New Physics

A deviation in the branching ratio, or to be more precise in the fraction $\frac{\mathcal{B}(B^0 \to \mu^+\mu^-)}{\mathcal{B}(B_s^0 \to \mu^+\mu^-)}$ would be a hint for so far unobserved physics. One example would be models with an extended Higgs sector, like the minimal supersymmetric standard model (MSSM), which hypothesizes two more heavier neutral Higgs bosons (H^0, A^0, the SM Higgs boson being denoted as h^0) along with two charged Higgs bosons ($H^\pm$). This leads to new contributions shown in Fig. 2.2

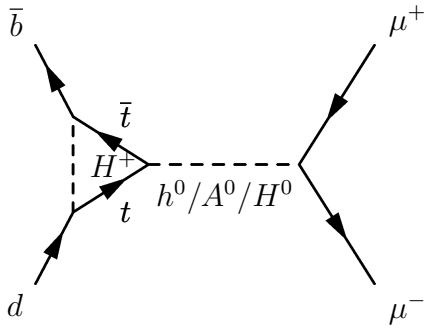


Figure 2.2: MSSM contributions to $B^0 \to \mu^+\mu^-$

Chapter 3

The LHCb Detector

3.1 The LHC

LHC [1] stands for **L**arge **H**adron **C**ollider, which is currently the largest particle accelerator and collider. It is located at the European Organisation for Nuclear Research in Geneva, Switzerland, though due to its size it reaches into France.

The LHC is a circular accelerator with a circumference of 26.7 km that initially allowed a centre-of-mass energy of 7 TeV and was by 2016 upgraded to 13 TeV.

There are four points along the ring at which the particles beams can be set to collide. Currently 7(-9) different experiments are operating at the LHC, of which the largest ones are `ALICE`, `ATLAS`, `CMS` and `LHCb`.

3.2 The LHCb Experiment

LHCb stands for Large Hadron Collider **b**eauty which is a reference to the b-quark that is one the main research focus of the LHCb experiment. Among its goals is the test of CP violation and the examination of rare decays to possibly detect discrepencies with the Standard Model, which would potentially hint at new physics. So far data has been taken in two runs, run 1 from 2011 to 2012 and run 2 from 2015 to 2018. As of now the detector is in the process of being upgraded for a third Run, that is scheduled to start in spring 2022.

3.2.1 Tracking System

In contrast to the other big detectors at the LHC, the LHCb detector is not a 4π-detector, but instead it is built in forward direction over a length of ~ 21 m.

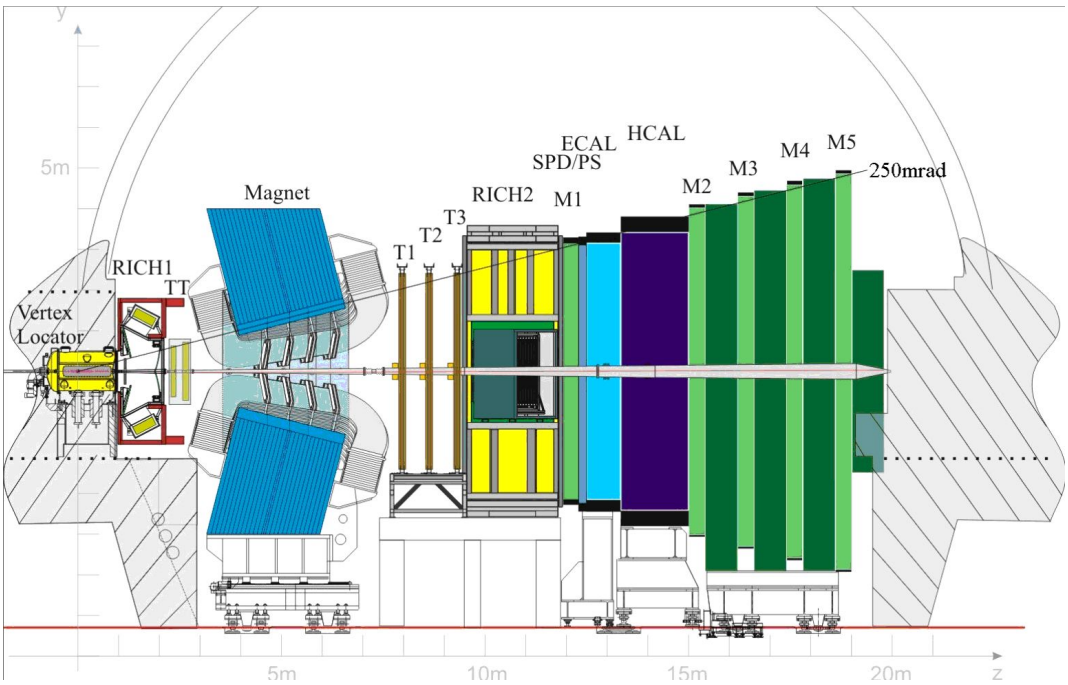


Figure 3.1: The LHCb detector [10]

Vertex Locator (VELO)

As the name suggests the **VE**rtex **LO**cator is used to reconstruct the primary vertex of the collisions, by tracking the produced particles in close vicinity to the collision point. It consists of 42 silicon detector modules which are placed along the z axis (defined as in Figure 3.1).

To reduce radiation damage the distance between the modules and the beam pipe can be varied, from 3.5 cm during injection of the beam, when it is more defocused, up to 5 mm with a focused beam.

Tracking Stations

The main tracking system is made up of four tracking stations, first the Trigger Tracker (TT), which is located in between the magnet and RICH1, and secondly the other three stations numbered T1-T3, which can be found behind the magnet before the RICH2. The three latter stations can further be subdivided into an Inner Tracker and an Outer Tracker each. Two technologies are employed for the detectors:

1. The silicon trackers, which consist of silicon microstrip detectors, which allow for a very good spatial resolution. Silicon detectors are rather expensive and therefore only used for smaller areas. The entire TT and the Inner Tracker of T1-T3 are made out of these.

2. The Outer Tracker of T1-T3 instead uses straw-tube drift chambers.

The Magnet

As already alluded to above inside the detector a magnetic field is present. This is needed in order to measure the momentum of charged particles, as well as to determine the sign of the charge.

The magnet consists of two coils, with a mass of 27 tons each and each consisting of $\sim 3000\,\mathrm{m}$ aluminium cable.

3.2.2 Particle Identifcation

Ring Imaging Cherenkov detectors (RICH)

As can be seen in Figure 3.1 LHCb has two RICH detectors. They are used for particle identification and work on the basis of Cherenkov radiation, emitted by the traversing particles. Cherenkov radiation occurs when a particle travels through a medium with a velocity higher than the speed of light in that medium. RICH1 is situated right behind the VELO in order to catch lower energy particles (with a momentum of $1 - 60\,\mathrm{GeV}$) which would otherwise be deflected out of the detector by the magnetic field.

Calorimeters

To measure the energy of the produced particles two calorimeter systems are used. One electromagnetic calorimeter (ECAL) which is sensitive to light particles like electrons and photons and one hadronic calorimeter (HCAL) in which hadrons mostly deposit their energy.

Muon System

As muons are traversing the calorimeters without significant energy loss, an additional detector system is employed to allow for a better detection and measurement of muons. This system consists out five stations (M1-M5), of which the first is placed behind RICH2 and before the calorimeters, while the rest (M2-M5) are placed behind the calorimeters. Each of these stations then consists out of four regions. Their granularity is shaped according to the particle density.

3.2.3 Combined Particle Identification

Particle identification already happens at the individual subdetectors (i.e. the RICH, the calorimeters and the muon systems), as each of those determines a likelihood for

each particle hypothesis. These subdetector likelihoods are then further combined as the sum of the individual logarithms relative to the pion hypothesis, into the Delta Log Likelihood (DLL).

Another type of combined PID variable is called ProbNN, which is a pseudo-probability determined by a Neural Network by taking into account the correlations between the subdetector likelihoods and additional information from the tracking system.

Furthermore for muon identification an algorithm called `isMuon` is employed, that analyses how many consecutive hits in the different muon stations can be detected, while also accounting for the momentum of the muon candidate. The requierements to pass `isMuon` are shown in Table 3.1.

momentum	associated hits in
$3 < p < 6\,\text{GeV}$	M2+M3
$6 < p < 10\,\text{GeV}$	M2+M3 + (M4 or M5)
$p > 10\,\text{GeV}$	M2+M3+M4+M5

Table 3.1: Required associated hits in the different stations for different momenta needed to pass isMuon

The `isMuon` efficiency for identifying muons is around 97% while the probability of a pion being missidentified as a muon is about 1 - 3 %

3.2.4 Trigger System

When in operation, there are around 40 million collisions happening every second at the interaction point, of which about 10 million are within the acceptance of the detector. To save all this data would be technically challenging and requiring a large amount of storage. Thus in order to make the data collection more feasible a multitude of triggers are used to preemptively decide which events are worth saving for later analysis.

On the upside the rate of events containing a B decay is only $\sim$ 15 kHz of the previous 10 MHz. Yet the rate at which the events can be written to storage is limited to 2 kHz therefore the triggers are designed to filter out particularly interesting decays out of the 15 kHz of B decays.

Level Zero (L0)

The L0 trigger's purpose is to reduce the data flow from 10 MHz to 1 MHz. This is done by using the momentum information provided by the calorimeters and the muon system.

A B decay has a larger momentum perpendicular to the beam axis (i.e. the transversal momentum p_T) compared to those stemming directly from the primary intersection. Furthermore the VELO it is able to conduct a simplified vertex reconstruction, allowing the rejection of events with several proton-proton interactions, as it is far more difficult to reconstruct B decays in these events.

High Level Trigger (HLT)

The HLT actually consists of two trigger levels named HLT1, which reduces the data rate to order of 10 kHz, and HLT2, which delivers the above mentioned 2 kHz of data that are ultimately recorded. HLT1 works mostly in the region of candidate direction to confirm high p_T candidate particles in the software reconstruction. Another measure used to identify B decays is the high impact parameter to the primary collision vertex, due to the relatively long lifetime of B mesons, allowing them to travel ~ 1 cm from the primary vertex before they decay.

HLT2 can now perform a complete reconstruction of the remaining events, allowing to search for reconstructed decay vertices that are displaced from the collision point thereby hinting at a B meson.

Chapter 4

Analysis and Tools

4.1 Background Contributions

The decay of interest in this analysis, $B^0 \to \mu^+\mu^-$ is very rare. Thus the number of non-signal dimuon candidates in the B^0 mass window vastly outnumbers the number of signal candidates. Therfore it is of outermost importance to understand all sources of backgrounds.

Firstly, there is the so-called **combinatorial *B* background**, from $b\bar{b} \to \mu^+\mu^- X$ events. As the $b\bar{b}$ cross-section in pp collision as well as the branching ratio of semi-muonic B decays is large, there is a sizeable possibility for two muons from two different B hadrons be to reconstructed as belonging to the same decay. The main way this background is dealt with is by analysing the vertex quality, displacement of the muons and their isolation.

The next largest source of background is the **hadronic *B* background**. This refers to decays of the form $B_{(s)} \to h^+h'^-$ where $h^\pm = \pi^\pm, K^\pm$. This background is less abundant due to the need for both hadrons to be misidentified as muon, yet given rarity of the signal those still matter. The way to deal with this background is by applying stronger PID requirements on both hadrons.

background contributions	prevalence relative to (expected) $B^0 \to \mu^+\mu^-$
$b\bar{b} \to \mu^+\mu^- X$	$\sim 10^8$
$B^0_{(s)} \to h^+h'^-$	$10^4 \sim 10^5$
$H_b^0 \to h^\pm\mu^\mp\nu_\mu$	$\sim 10^5$
$B_c^+ \to J/\Psi(\to \mu^+\mu^-)\mu^+\nu_\mu$	$\sim 10^5$
$B^{(0/+)} \to \pi^{(0/+)}\mu^+\mu^-$	$\sim 10^2$

Table 4.1: Processes contributing to background for $B^0 \to \mu^+\mu^-$

Further sources for background are **semileptonic B hadron decays** of the form $H_b^0 \to h^{\pm}\mu^{\mp}\nu_\mu$ with $(H_b^0, h) = (B^0, \pi^-), (B_s^0, K^-), (\Lambda_b^0, p)$ in which the hadron is once again misidentified as a muon and the neutrino is not detected. These decays are about five times more common than the hadronic background and also require only one hadron to be misidentified. However, given that these decays involve three particles the reconstructed mass will often fall outside the mass window of the signal, which helps rejecting those. Further PID requirements and the requirement that the reconstructed momenta of the B candidates point back to the primary vertex are efficient for rejecting this class of backgrounds.

Lastly, the decays $B_c^+ \to J/\Psi(\to \mu^+\mu^-)\mu^+\nu_\mu$ and $B^{(0/+)} \to \pi^{(0/+)}\mu^+\mu^-$ contribute to possible background when either one muon and the neutrino are not reconstructed or the pion is missed, respectively. The invariant mass of these background candidates also falls more rarely in the probed mass window. It can further be sorted out through requirements on muon isolation and in case of the B_c decay with a J/Ψ veto.

In this thesis the focus is on the hadronic B background, i.e. the decay channels $B^0 \to \pi^+\pi^-$, $B^0 \to K^+\pi^-$, $B_s^0 \to \pi^+K^-$ and $B_s^0 \to K^+K^-$ and the use of PID to reject more of these events.

4.2 Used Monte Carlo Samples

To train a multivariate classifier Monte Carlo (MC) simulations are used, to have clearly tagged signal and background candidates. A larger number of events is needed in order to provide sufficient training data. Therefore Monte Carlo files of the background channels mentioned in Section 4.1 are used, along with a Monte Carlo sample of the signal decay $B \to \mu^+\mu^-$. The samples are generated for each data-taking periode seperately, namely for the years 2011, 2012, 2016, 2017 and 2018. They are configured such that they reflect the conditions of the detector in the given year. It has to be

	2011	2012	2016	2017	2018
$B^0 \to \pi^+\pi^-$	388813	2399734	1019672	1022588	1021286
$B^0 \to \pi^+K^-$	193559	1989498	987079	988435	1011752
$B_s^0 \to K^+\pi^-$	2404390	2404390	1016587	1015762	1013321
$B_s^0 \to K^+K^-$	373574	2299449	957775	959374	968044
$B^0 \to \mu^+\mu^-$	170813	153711	658342	679188	351922

Table 4.2: Number of Events in each MC

noted that the number of events for the $B_s^0 \to K^+\pi^-$ sample is the same for the years

2011 and 2012. The reason for this is that no Monte Carlo data set for this channel for 2011 is available, therefore the same as for 2012 is used.

4.3 Punzi Figure of Merit

The selection of the signal candidates has to be optimized to keep an as high as possible efficiency and to reject at the same time as much as possible background. The goal is to tune the cuts on the selection variable such that the significance of the measurement is maximal. To evaluate the significance of a cut the so-called Punzi Figure of Merit, or Punzi FoM, is used. It is defined as:

$$\text{punzi FoM} = \frac{\varepsilon_S}{\frac{3}{2} + \sqrt{B}} \tag{4.1}$$

Here ε_S refers to the signal efficiency which is defined as the ratio between the total number of signal candidates in the sample without any cut and the number of signal candidates surviving the selection cuts. Similarly B refers to the total number of background events surviving the selection cuts.

Normalisation to $\mathbf{B^+ \to J/\Psi(\to \mu^+\mu^-)K^+}$

So far only the number of events in the simulations are known, which do not correspond to the actual number of events. The expected number of events can be determined by using the number of B mesons that are produced and the branching ratios of the background channels. Yet to calculate the number of produced B mesons the luminosity would be needed, which is not precisely known and error prone. Instead the amount of detected events in a normalisation Channel, in this case $B^+ \to J/\Psi(\to \mu^+\mu^-)K^+$, is used by computing the ratio of events in the given decay channel and the normalisation channel

$$\frac{N_{B^0_{(s)}\to X}}{N_{B^+\to J/\Psi(\to\mu^+\mu^-)K^+}} = \frac{\epsilon_{B^0_{(s)}\to X} \times \mathcal{B}(B^0_{(s)} \to X) \times f_{d/s}}{\epsilon_{B^+\to J/\Psi K^+} \times \mathcal{B}(B^+ \to J/\Psi(\to \mu^+\mu^-)K^+) \times f_u} \tag{4.2}$$

Here $\mathcal{B}$ is the branching fraction of the given decay, ϵ_x is the total detection efficiency which has been determined in previous analyses. $\frac{f_{d/s}}{f_u}$ is the ratio of the hadronization fractions for the given quark and it describes the amount of $\mathrm{B}^0_{(d)}$ or B^0_s mesons relative to $B^+_{(u)}$ mesons produced. As u and d quarks are produced at very similar rates $\frac{f_d}{f_u} \simeq 1$, while $\frac{f_s}{f_u} \simeq \frac{f_s}{f_d} = 0.244 \pm 0.012$

Multiplying by $N_{B^+\to J/\Psi(\to\mu^+\mu^-)K^+}$ and writing $\beta = \frac{N_{B^+\to J/\Psi(\to\mu^+\mu^-)K^+}}{\epsilon_{B^+\to J/\Psi K^+} \times \mathcal{B}(B^+\to J/\Psi(\to\mu^+\mu^-)K^+)}$ the number of events in a given Background Channel is given by

$$N_{B^0_{(s)}\to X} = \beta \times \epsilon_B \times \mathcal{B}(B_{(s)} \to hh') \times (\frac{f_d}{f_s}) \tag{4.3}$$

4.4 Multivariate Classifier via Machine Learning

Machine learning is a practice that employs algorithms that are capable of improving themselves through the input of data. The process of improving is called learning.

The main approaches to machine learning are called supervised and unsupervised learning. The main difference between these approaches is that in the case of supervised learning the algorithm is provided a set of training data along with the desired output for this training data, while for unsupervised learning it is left to the algorithm to find own structures in the given training data.

What is used here is supervised learning, i.e. the training data is given in the form of the Monte Carlo simulation for which it is known which event is a signal event and which belongs to the background. The way machine learning is used in this thesis is the classification of events into signal events and background events by analysing multiple decay parameters as an input.

4.4.1 Boosted Decision Trees

There are many machine learning techniques to categorize data like here into Background and Signal.

One rather simple method to categorize data is decision tree learning. Given a set of training data consisting of tuples ($\mathbf{x}$,Y) where $\mathbf{x}$=($x_1, \ldots, x_n$) is the vector that contains the input variables and Y is the category of the event, here whether the event belongs to the background or signal. A set of consecutive selection cuts on the input variables is employed to best differentiate between the categories. These selection cuts can be represented as nodes in a tree where the final nodes, or leafs, are the categories into which the data is sorted.

The selection cuts are chosen to achieve the best distinction between background and signal. A common metric to quantify the gain in separation is called Gini impurity, which is then recursively maximised in the training process to optimize the selection cuts.

Instead of using a single decision tree the process can be expanded to include a large number of separate decision trees, which is called decision tree boosting. The outputs from the individual trees are taken and combined into a final classification.

4.4.2 XGBoost

One specific implementation of decision tree boosting is called XGBoost [8]. As a measure how well the model fits the training data an **objective function**

$$\text{obj}(\vec{\theta}) = L(\vec{\theta}) + \Omega(\vec{\theta}) \tag{4.4}$$

consisting of a training loss function L and a regularization term Ω, is used. The parameter θ here refers to the coefficient assigned to an input variable x_i from which a prediction $\hat{y} = \sum_i \theta_i x_i$ is determined. For a single tree it looks like

$$L(\theta) = \sum_i (y_i - \hat{y}_i)^2 \tag{4.5}$$

where y_i refers to the target value.

The regularization term describes the complexity of a tree. For this the tree can be expressed as $f_t(x) = w_{q(x)}$, where q is a function assigning a leaf to a data point x and w is the score on each leaf.

With the number of leaves being T the complexity of a tree is then defined as

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{4.6}$$

γ and λ here refer to free regularization parameters.

The output is a value ranging from 0 (Background) and 1 (Signal).

4.4.3 Training a Multivariate Classifier

For training the classifier takes two inputs. First an array of decay variables that are supposed to differentiate background and signal, and secondly the target, meaning if the respective event is supposed to be part of signal or part of the background. There are a number of parameters to tweak how the learning algorithm behaves. These are shown in Table 4.3

Model Parameter	Explanation
`iterations`	Number of individual trees
`depth`	Maximum number of nodes on each tree
`eta`	Learning rate
`colsample_bytree`	Fraction of how many of the variables are used for each tree
`subsample`	Fraction of events used in each tree
`gamma`	Minimum loss reduction required to make a further partition on a leaf node. A larger value corresponds to a more conservative algorithm
`reg_alpha` `reg_lambda`	Regularization parameters that penalize overcomplexity

Table 4.3: Model Parameters

4.4.4 Evaluating the Classifier

To test how a classifier performs one can take a look at the **R**eceiver **O**perating **C**haracteristic curve (ROC curve). It shows the portion of background events that are falsely classified as signal (false negative) against the portion of signal events events being classified correctly. As the classifier does not produce a binary output, but a score X between 0 and 1 one can now vary the threshold T for which $X > T$ to be classified as a signal event and plot the points on a curve. In case for a random classification one would now expect all events to lie on the identity line. A classifier working better than random classification would therefore produce a ROC curve where the points lie above this line. To quantify the performance one can now use the **A**rea **U**nder the **C**urve (auc), for which a larger value corresponds to a better performing classifier.

Chapter 5

Preselection

At first a preselection, found in Table 5.1, is applied that sorts out events with very poor reconstruction.

The reconstructed mass is restricted to a window from 4.9 GeV to 6 GeV. To further ensure a reasonable vertex quality a maximum value for the vertex χ^2 and the impact parameter χ^2 is required along with a minimal flight distance χ^2.

For the muons a maximum momentum and transverse momentum in a certain range is required. Also a minimal impact parameter χ^2 and a not too large track χ^2/ndof is required. Also a J/Ψ Veto, sorting out events in which one of the muons can be reconstructed to a dimuon mass close to the J/Ψ mass. Furthermore the probability of the track being a ghost should not be too large. The exect cuts are shown in Table 5.1

Further a cut is placed on a BDT variable (in the following simply referred to as BDT), that was trained on kinematic topological and isolation information and is scaled to give a roughly uniform distribution for signal events and to peak at zero for background. By demanding BDT > 0.5 events with low signal quality should be excluded. This BDT is not to be confused with the BDT based classifier that is developed in the core of this thesis.

On	Selection
B	Reconstructed Mass is between 4.9 and 6.0 GeV
	Reconstructed lifetime is smaller than 1.3248×10^{-11} s ($\sim 9\times$ lifetime of the B meson)
	`BDTS_DOCA` < 0.3
	Endvertex $\chi^2 < 9$
	Impact Parameter $\chi^2 < 25$
	Flight distance $\chi^2 > 225$
	BDTS> 0.05
μ^+/μ^-	Momentum is smaller than 500 GeV
	Transversal momentum is between 0.25 and 40 GeV
	Impact parameter $\chi^2 > 25$
	Track $\chi^2 < 4$
	Probability of the track being a ghost track< 0.4
	`InMuonAcc` $== 1$
	J/Ψ veto: Reconstructed B mass for one particle differs more than $30\,MeV$ from the J/Ψ mass ($m_{J/\Psi} = 3096.9$ GeV)

Table 5.1: Preselection Cuts

Chapter 6

Reducing Background via a MVA

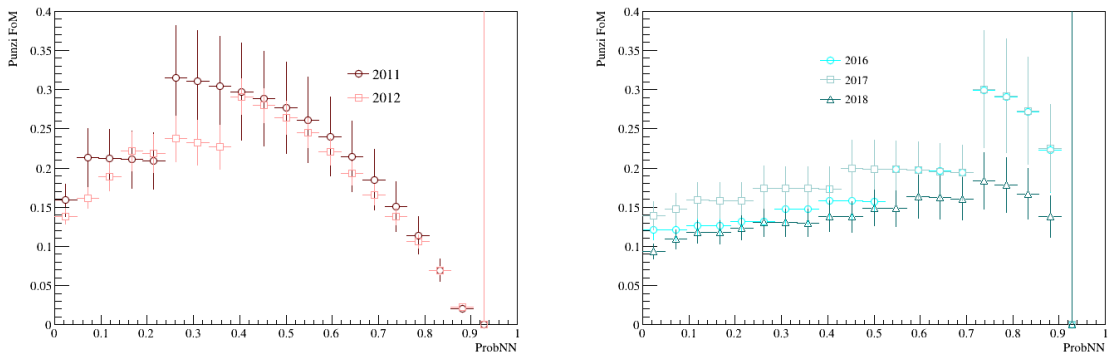
The goal here is to train a BDT to improve the rejection of hadronic $B^0_{(s)} \to hh'$ background.

6.1 Performance of ProbNN PID

In previous analyses the ProbNN_x PID variables were used for background rejection, and a combination of three of those ProbNN_x variables - simply called ProbNN - proved to perform best.

$$\text{ProbNN} = \text{ProbNN}_\mu \times (1 - \text{ProbNN}_K) \times (1 - \text{ProbNN}_p) \tag{6.1}$$

Figure 6.1 shows the Punzi FoM calulated after several cuts on ProbNN in a range from 0 to 1, each year is treated separately. For the background the $B^0_{(s)} \to hh'$ MCs and for the signal the $B^0 \to \mu^+\mu^-$ MCs as introduced in Section 4.2. Beyond slight differences between the years in each run one mainly notices the difference between the years in run 1 and run 2, as for run 2 the optimal cut off value is at around 0.8 while for run 1 it is around 0.4, which is in good accordance with previous findings.



(a) run 1 (b) run 2

Figure 6.1: Punzi FoM for different cut values of ProbNN

Applying the ProbNN > 0.8 cut on the run 2 MCs and ProbNN > 0.4 on run 1 to determine the number of expected detected events in each given year leads to Table 6.1. For signal events the fraction of events that survive the cut are given. The reason these percentages are below 50% is that due to requiring BDT> 0.5 roughly half of the events are already filtered out. It can be seen that the largest amount of Background comes from misidentified pions, while the rejection of kaons already works quite well.

	run 1		run 2		
	2011	2012	2016	2017	2018
$B^0 \to \pi^+\pi^-$	1.9 ± 1.01 (1e-05 ± 5.1e-06)	3.7 ± 0.92 (8e-06 ± 1.9e-06)	5.0 ± 1.73 (9e-06 ± 2.9e-06)	4.5 ± 1.67 (8e-06 ± 2.8e-06)	7.0 ± 2.28 (1e-05 ± 3.1e-06)
$B^0 \to K^+\pi^-$	0.9 ± 0.9 (5e-06 ± 5.2e-06)	0.6 ± 0.4 (1.5e-06 ± 8.7e-07)	4.5 ± 1.65 (8e-06 ± 2.9e-06)	1.7 ± 1.05 (3e-06 ± 1.8e-06)	1.4 ± 1.02 (2e-06 ± 1.4e-06)
$B_s^0 \to \pi^+K^-$	0.12 ± 0.13 (2.5e-06 ± 1e-06)	0.3 ± 0.23 (2.5e-06 ± 1e-06)	0.9 ± 0.49 (6e-06 ± 2.4e-06)	0.31 ± 0.29 (2e-06 ± 1.4e-06)	0.9 ± 0.54 (5e-06 ± 2.2e-06)
$B_s^0 \to K^+K^-$	0.0 ± 0.31 (0.0 ± 1.3e-06)	0.24 ± 0.28 (4.3e-07 ± 4.3e-07)	1.5 ± 1.09 (2e-06 ± 1.5e-06)	0.7 ± 0.79 (1e-06 ± 1e-06)	2.7 ± 1.64 (3e-06 ± 1.8e-06)
$B^0 \to \mu^+\mu^-$	0.4568 ± 0.002	0.4503 ± 0.0021	0.46 ± 0.001	0.4617 ± 0.001	0.46 ± 0.0014

Table 6.1: Expected number of events in each channel for ProbNN> 0.4 (run 1) or ProbNN> 0.8 (run 2). Shown in parentheses below the number of events is the fraction of events that survive the cut. For the signal channel only the fraction of surviving events is given

6.2 Training the MVA

6.2.1 Input Variables for the MVA

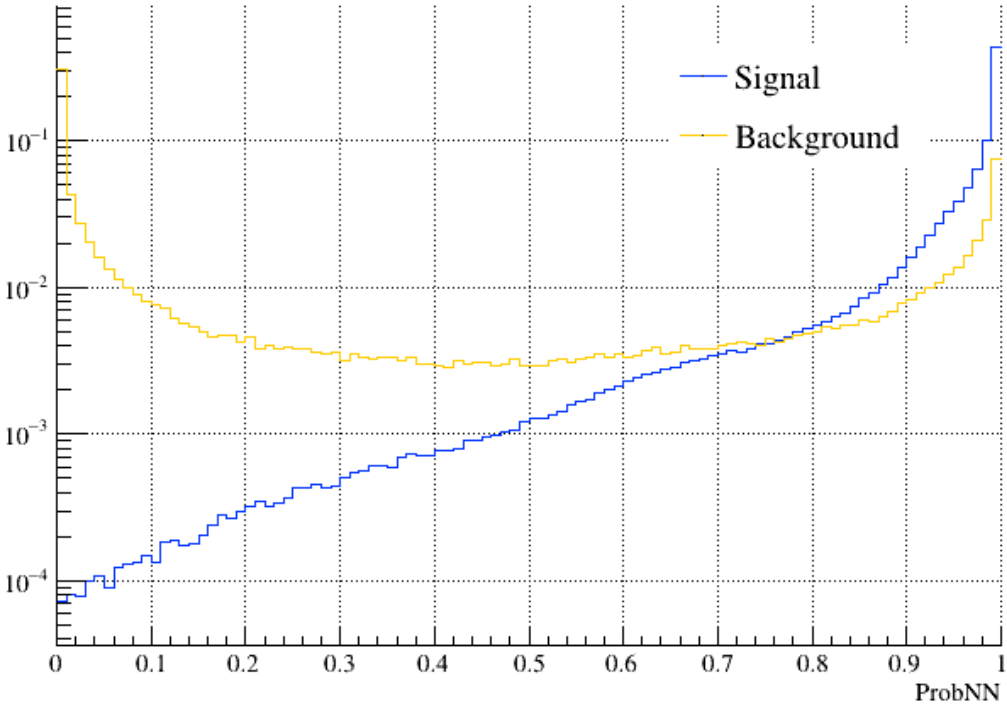
First we need to have a look at which input variables the BDT should use. The distribution of these input variables should differ between signal events and background events and therefore allow a differentiation between the two. Also the events are restricted to only those for which at least one of the as muons detected particles passes `isMuon`.

Figures 6.2-6.7 show the distributions of the stated variable for background and signal events. The events are weighted according to the $B^+ \to J/\Psi K^+$ decay as discussed in Section 4.3. Finally the distributions are rescaled to be normalized.

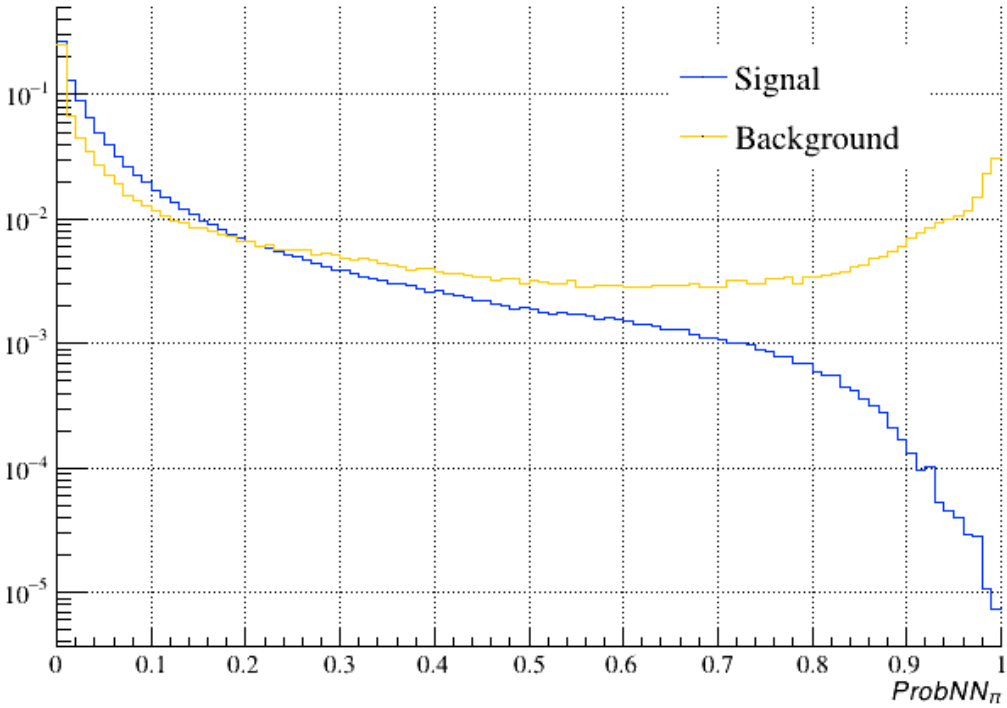
ProbNN and ProbNN$_\pi$

As ProbNN already provides a very good particle identification it is also included for the new classifier. In order to reduce overtraining not the raw ProbNN value is used, as those can often contain large negative values, which is a consequence of only demanding `isMuon`for one of the tracks, as a track that does not satisfy `isMuon`returns a ProbNN$_\mu$ value of -1000. Therefore instead the maximum value of ProbNN for the two tracks is used instead.

Further as ProbNN so far does not consider pion misidentification ProbNN$_\pi$ is also included to achieve better pion rejection.



(a) ProbNN probability distribution



(b) ProbNN_{π} probability distribution

Figure 6.2

Opening Angle

The angle between the both muon tracks can also be used to gain information on particle identification. As the angle is not directly recorded it has to be calculated by using the angular information of the muon tracks, i.e. the pseudorapidities $\eta_{\mu^\pm}$ and polar angles $\varphi_{\mu^\pm}$:

$$\measuredangle_\mu = \arccos\left(\frac{\cos\varphi_{\mu^+}\cdot\cos\varphi_{\mu^-} + \sin\varphi_{\mu^+}\cdot\sin\varphi_{\mu^+}}{\cosh\eta_{\mu^+}\cdot\cosh\eta_{\mu^-}} + \tanh\eta_{\mu^+}\cdot\tanh\eta_{\mu^-}\right) \tag{6.2}$$

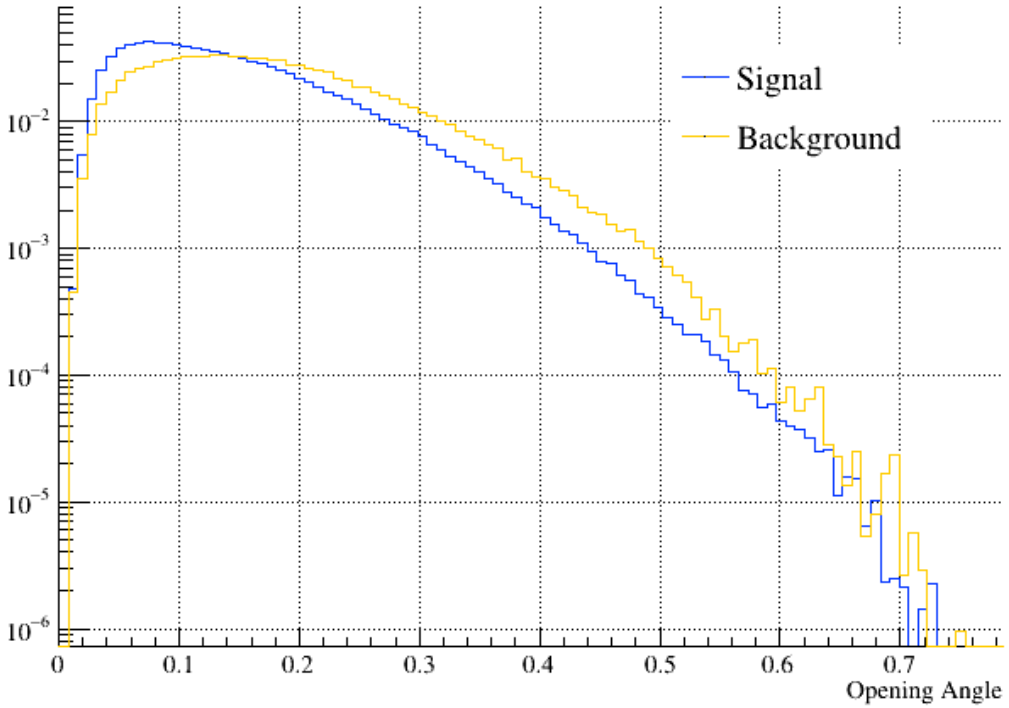


Figure 6.3: Probability distribution of the opening angles

Track χ^2

In case a pion or kaon might decay in flight into a muon the differentiation power of the ProbNN variable suffers. Yet through the decay the flight path is slightly altered which can then negatively impact the reconstruction quality of the track i.e. a larger χ^2 in the fit. Therefore using the track χ^2 information is included as an input variable for the classifier.

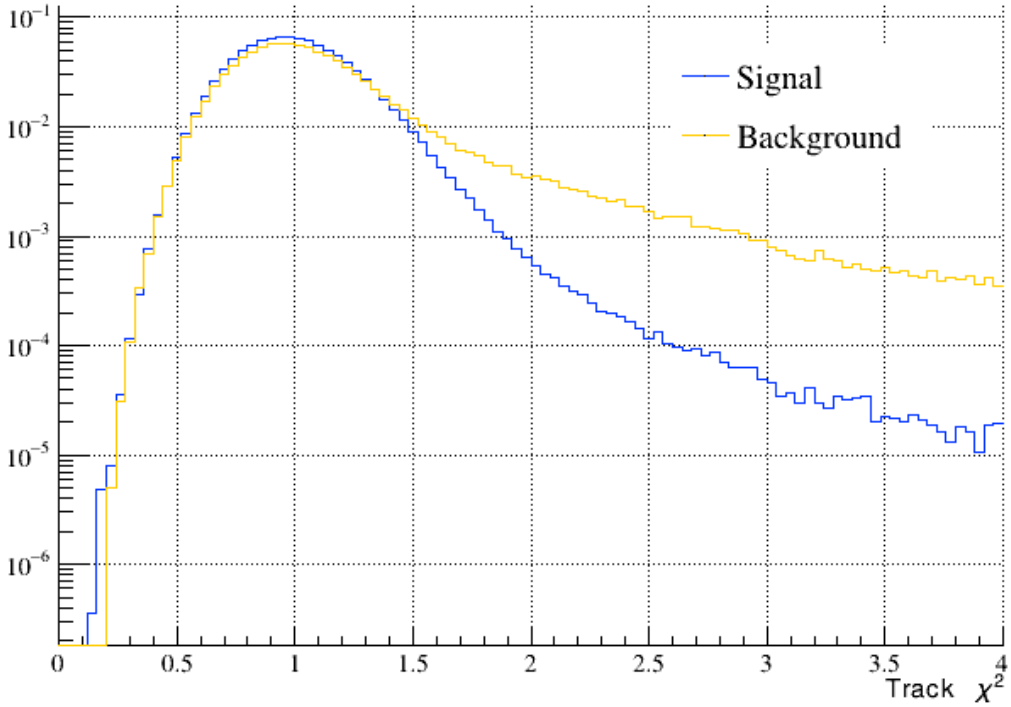


Figure 6.4: Track χ^2 probability distribution

Figure 6.4 shows the distribtution of the track χ^2 for both muon candidates.

Flight Distance χ^2

The next parameter to consider is the flight distance χ^2, meaning the quality of the reconstructed flight distance of the B meson. As can be seen in Figure 6.5 this variable is not suitable to differentiate between signal and background. It is included as the reconstruction quality of an event should be taken into account for the certainty of a classification.

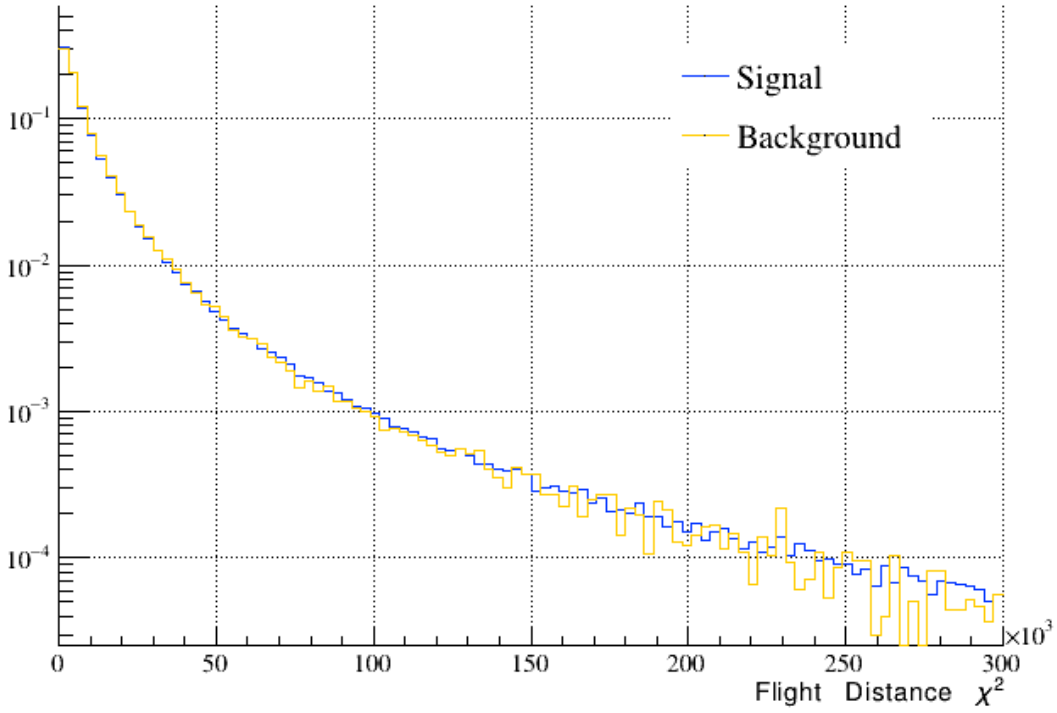


Figure 6.5: Flight distance χ^2 probability distribution

Mass Error

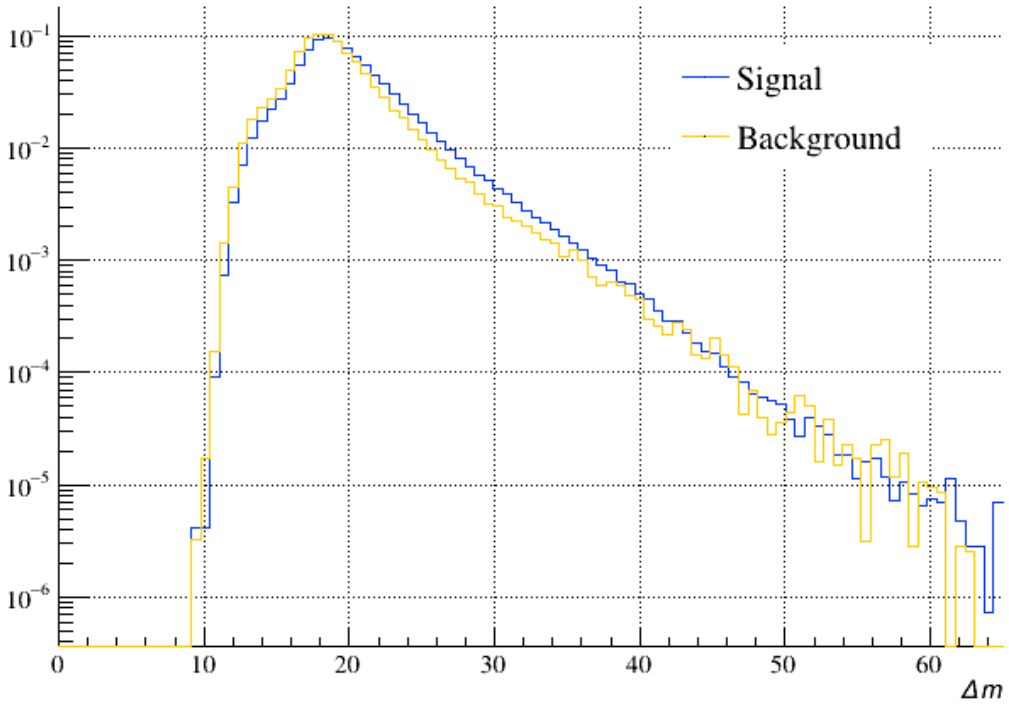


Figure 6.6: Mass Error probability distribution

As can be seen in Figure 6.6 the difference between signal and background is not great, but it is still included for training. Yet information about the quality of the B candidate can be extracted.

BDT

Lastly the BDT information as introduced in the preselection is also used as an input as it gives further information on signal quality, while also adding distinguishing power.

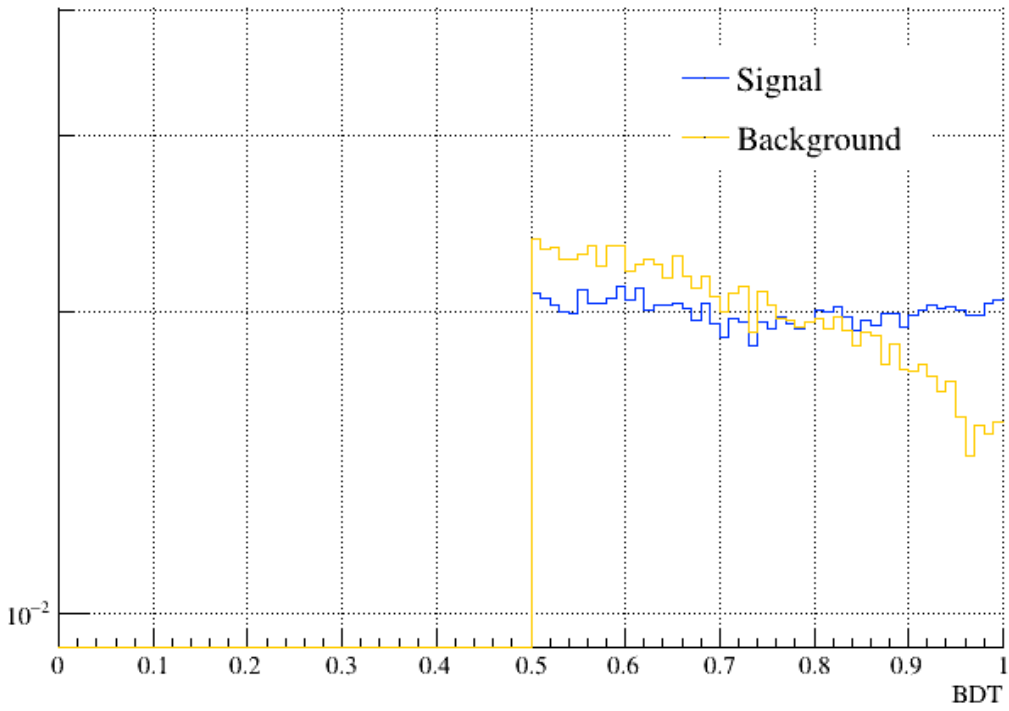


Figure 6.7: BDT probability distribution

6.2.2 Setup of the Training

The Monte Carlo data is split into two sets, each containing half of all signal and half of all background events. One set of event is used for training the classifier while the other set is used to test how the classifier performs. The events are randomly distributed into both sets. Due to having more signal events than background events a weight w_j is assigned to each event, so that $\sum_{\text{Background Events}} w_j = \sum_{\text{Signal Events}} w_i$.

The events in the training set are then used to train a XGBoost classifier with the training parameters given in Table 6.2, as prior introduced in Table 4.3. Subsequently this classifier is then evaluated by applying it on the events from the second set. From this a ROC curve is produced, which is shown in Figure 6.8. As can be seen it performs quite well, though given that ProbNN itself already sorts out a lot of background it is not that surprising. Yet the difference between the classifier performance on training set in contrast to the evaluation set is significant. For one it can be expected that the classifier works better on the same events is trained on, but too large a difference is also a sign of overtraining.

Parameter	value
`iterations`	300
`depth`	15
`eta`	1
`colsample_bytree`	1
`subsample`	1
`reg_alpha`	0
`gamma`	0
`reg_lambda`	1

Table 6.2: Parameters used in the training

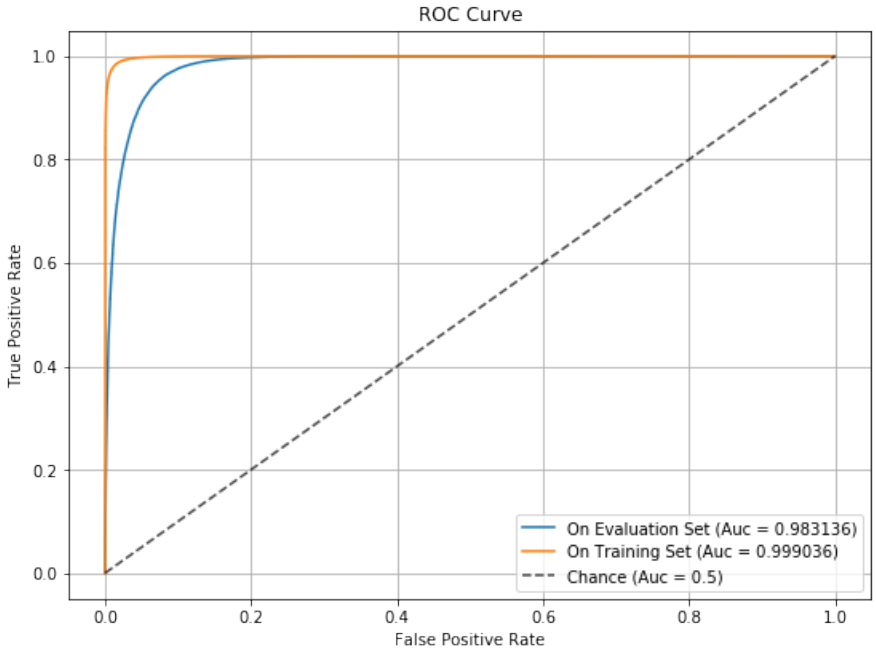


Figure 6.8: ROC Curve for the hadronic BDT classifier

6.2.3 Performance

Now that it is clear that the classifier works it is applied to the entirety of the MCs. Then a series of cuts is applied on the new classifier variable and again the Punzi FoM is calculated each time, the results are shown in Figure 6.9. One can once again see striking differences between run 1 and run 2, especially as the performance is closer to that of the ProbNN cut, though these cuts do not outperform the prior cut.

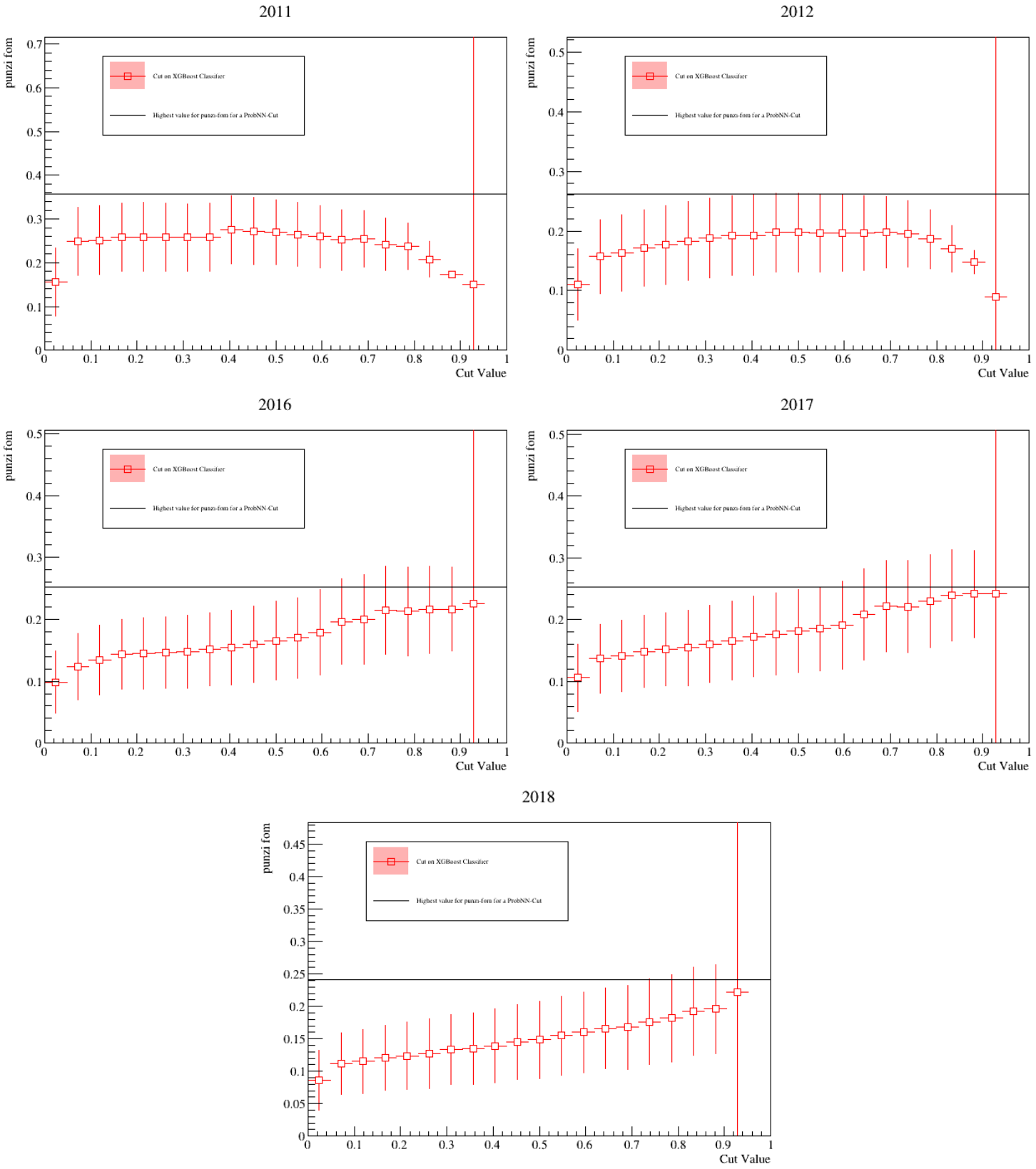


Figure 6.9: Punzi FoM for cuts on the new classifier, for comparison the highest value that can be achieved by cutting on ProbNN is also shown

To see if an improvement can be achieved by combining a cut on the new classifier with a cut ProbNN, the previous procedure to determine the performance of a cut is repeated while varying the cut values for both variables. The results are shown in Fig. 6.10, where in order to see how well the cut does in comparison to to a simple cut on ProbNN the highest Punzi FoM value for a ProbNN cut is subtracted from the newly determined figure. Therefore only cuts which outperform a simple ProbNN cut are visible.

It can be seen that improvements can be reached for all years, but especially for those in run 2.

As before the number of expected events in each channel can be calculated, shown in Table 6.3 for a ProbNN > 0.6 and XGBoost> 0.6 cut. It can be clearly seen that background rejection is improved while Signal retention stays high, except for the run 1 sets for which the signal retention is actually reduced. That this cut performs worse on run 1 is already visible in Figure 6.10.

		run 1		run 2		
		2011	2012	2016	2017	2018
$B^0 \to \pi^+\pi^-$	ProbNN + BDT	0.5 ± 0.5 (3e-06 ± 2.6e-06)	0.7 ± 0.41 (1.7e-06 ± 8.3e-07)	2.2 ± 1.15 (4e-06 ± 2e-06)	2.8 ± 1.32 (5e-06 ± 2.2e-06)	3.5 ± 1.61 (5e-06 ± 2.2e-06)
	ProbNN	1.9 ± 1.01 (1e-05 ± 5.1e-06)	3.7 ± 0.92 (8e-06 ± 1.9e-06)	5.0 ± 1.73 (9e-06 ± 2.9e-06)	4.5 ± 1.67 (8e-06 ± 2.8e-06)	7.0 ± 2.28 (1e-05 ± 3.1e-06)
$B^0 \to K^+\pi^-$	ProbNN + BDT	0.0 ± 0.44 (0.0 ± 2.6e-06)	0.21 ± 0.23 (5e-07 ± 5e-07)	1.1 ± 0.83 (2e-06 ± 1.4e-06)	2.3 ± 1.21 (4e-06 ± 2e-06)	0.0 ± 0.35 (0.0 ± 4.9e-07)
	ProbNN	0.9 ± 0.9 (5e-06 ± 5.2e-06)	0.6 ± 0.4 (1.5e-06 ± 8.7e-07)	4.5 ± 1.65 (8e-06 ± 2.9e-06)	1.7 ± 1.05 (3e-06 ± 1.8e-06)	1.4 ± 1.02 (2e-06 ± 1.4e-06)
$B_s^0 \to \pi^+K^-$	ProbNN + BDT	0.02 ± 0.054 (4e-07 ± 4.2e-07)	0.05 ± 0.092 (4e-07 ± 4.2e-07)	0.15 ± 0.2 (9.8e-07 ± 9.8e-07)	0.15 ± 0.21 (9.8e-07 ± 9.8e-07)	0.8 ± 0.49 (4e-06 ± 2e-06)
	ProbNN	0.12 ± 0.13 (2.5e-06 ± 1e-06)	0.3 ± 0.23 (2.5e-06 ± 1e-06)	0.9 ± 0.49 (6e-06 ± 2.4e-06)	0.31 ± 0.29 (2e-06 ± 1.4e-06)	0.9 ± 0.54 (5e-06 ± 2.2e-06)
$B_s^0 \to K^+K^-$	ProbNN + BDT	0.0 ± 0.31 (0.0 ± 1.3e-06)	0.0 ± 0.12 (0.0 ± 2.2e-07)	0.7 ± 0.77 (1e-06 ± 1e-06)	0.0 ± 0.37 (0.0 ± 5.2e-07)	0.9 ± 0.95 (1e-06 ± 1e-06)
	ProbNN	0.0 ± 0.31 (0.0 ± 1.3e-06)	0.24 ± 0.28 (4.3e-07 ± 4.3e-07)	1.5 ± 1.09 (2e-06 ± 1.5e-06)	0.7 ± 0.79 (1e-06 ± 1e-06)	2.7 ± 1.64 (3e-06 ± 1.8e-06)
$B^0 \to \mu^+\mu^-$	ProbNN + BDT	0.3616 ± 0.0017	0.3364 ± 0.0017	0.463 ± 0.001	0.466 ± 0.001	0.464 ± 0.0014
	ProbNN	0.4568 ± 0.002	0.4503 ± 0.0021	0.46 ± 0.001	0.4617 ± 0.001	0.46 ± 0.0014

Table 6.3: Expected number of events in each channel after a cut on the new classifier. Values for previous ProbNN cut are also shown for comparison as the lower number. In parantheses shown is the fraction of surviving events in the background channels

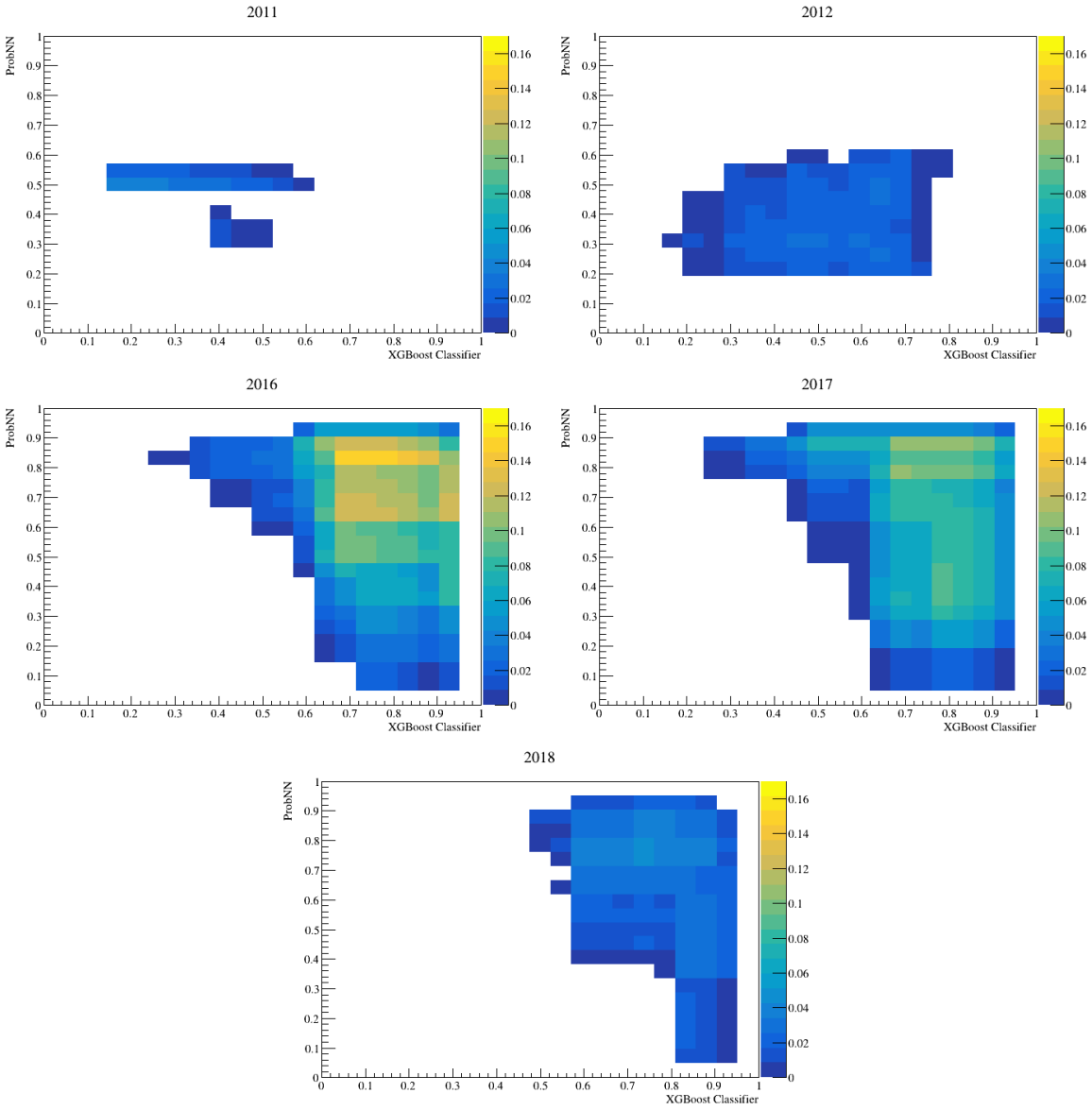


Figure 6.10: Combination of a cut on the new classifier and a cut on ProbNN. To compare to the previous best performing cut the difference in the Punzi FoM is shown as the third dimension

Chapter 7

Conclusion & Outlook

One of the most important background sources of the rare $B^0_{(s)} \to \mu^+\mu^-$ decay are the hadronic $B^0_{(s)} \to hh'$ decays. In this thesis I first evaluate the efficiency in terms of background rejection through the use of standard LHCb particle identification variables (ProbNN). The resulting optimal cut values are in good accordance to previous results.

Then I trained a BDT classifier that combines the PID information with further topological information. By combining this new classifier with the ProbNN information I achieve a better rejection efficiency for the hadronic background, while retaining most signal events.

There are undoubtedly many ways the training of the BDT can be improved. For one the selection for events that are used for training should be more elaborate. So far the events have been randomly partitioned with no regard to the prevalence of the decay channel or the year which the MC has been produced for.

Additionally the branching ratios of the different background decays could be included in the weighting of the events. Moreover a larger amount of events could be produced to have a larger set of training data, as imposing the `isMuon` requirement already eliminates a large portion of events ($\sim 98\%$).

Bibliography

[1] *LHC Machine*, Lyndon Evans and Philip Bryant, 2008 JINST **3** S08001

[2] *Rare B decays*, Thomas Blakeet al., Progress in Particle and Nuclear Physics Volume 92, January 2017, Pages 50-91

[3] *Measurement of the* $B_s^0 \to \mu^+\mu^-$ *Branching Fraction and Search for* $B^0 \to \mu^+\mu^-$ *Decays at the LHCb Experiment*, R. Aaij et al., Phys. Rev. Lett. **111**, 101805

[4] *The essence of rare beauty*, Mick Mulder, *PhD Thesis, 2020, Rijksuniversiteit Groningen*

[5] *Measurement of the CKM matrix elements* $|V_{ub}|/|V_{cb}|$ *from semileptonic* B_s *decays*, Svende Annelies Braun, *PhD Thesis, 2020, Ruperto-Carola-University of Heidelberg*

[6] *Sensitivity of searches for new signals and its optimization*, Giovanni Punzi, PHYSTAT-2003-MODT002

[7] *XGBoost: A Scalable Tree Boosting System*, Tianqi Chen, Carlos Guestrin ,KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, Pages 785–794

[8] *Introduction to Boosted Trees*, `https://xgboost.readthedocs.io/en/latest/tutorials/model.html`

[9] *The LHCb Detector at the LHC*, The LHCb Collaboration et al 2008, JINST 3 S08005

[10] *LHCb status and early physics prospects*, Monica Pepe Altarelli, `https://arxiv.org/abs/0907.0926`

[11] *Observation of the rare* $B_s^0 \to \mu^+\mu^-$ *decay from the combined analysis of CMS and LHCb data*, CMS Collaboration & LHCb Collaboration, *Nature* volume **522**, 68–72 (2015)

[12] *Flavour-changing neutral currents making and breaking the standard model*, F. Archilli et al., *Nature* volume **546**, 221–226 (2017)

[13] *Review of Particle physics*, Particle Data Group, Progress of Theoretical and Experimental Physics, Volume 2020, Issue 8, August 2020, 083C01

Acknowledgements

At last I want to thank Professor Dr. Hansmann-Menzemer for allowing me the opportunity to write my thesis in her group. Further thanks go to Dr. Flavio Archilli and Giulia Frau for introducing me to the analysis tools and supporting me throughout the process, especially during all the Covid restrictions.

Erklärung

Ich versichere, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Meppen, den 09.07.2021,