ETP-Bachelor-KA/2022-06

Karlsruher Institut für Technologie

Sensitivity Study for $B^+ \to K^+\alpha$ Decays at Belle II with XGBoost Classifier

Bachelor Thesis

by

Alexander Avocone

submitted to the Department of Physics
Institute of Experimental Particle Physics

Reviewer: Prof. Dr. Torben Ferber
Advisor: Dr. Slavomira Stefkova

Erklärung der selbständigen Anfertigung der Dissertationsschrift

Hiermit erkläre ich, dass ich die Dissertationsschrift mit dem Titel

»*Sensitivity Study for $B^+ \to K^+\alpha$ Decays at Belle II with XGBoost Classifier*«

selbständig und unter ausschließlicher Verwendung der angegebenen Hilfsmittel angefertigt habe.

Alexander Avocone
Karlsruhe, den 12. Dezember 2022

Preface

The research described in this thesis was conducted under the supervision of S. Stefkova (ETP, KIT) and submitted to the Department of Physics, Karlsruhe Institute of Technology, and is part of the Belle II collaboration. The simulation samples that were described in chapter 2 and analyzed throughout the thesis were prepared by S. Stefkova. The model training in chapter 3 and the subsequent analysis is my original work, except where the references are made to previous work. Furthermore, all figures were created by me unless otherwise noted.

Contents

CHAPTER 1

Introduction

With the development of human history, the physical understanding of the world also evolved. Both the theories and the experiments became more and more elaborate. Nowadays, the vast amount of data generated every day makes computers an indispensable part of data processing. Particle colliders generate millions of collisions per second on each run resulting in a vast amount of complex initial data. Data processing systems such as triggers and event filters cut the initial data down to a more manageable yet still large share. To analyze this output sophisticated statistical models are used for particle detection. One approach is called Machine Learning (ML) which mimics the learning behavior of humans. By utilizing statistical methods, the ML algorithms are trained to make predictions or classifications based on the given data.
Many ML models have been developed for specific tasks but in general, three different types of ML algorithms emerged. Depending on the learning method, the models can be divided into supervised learning, unsupervised learning, and reinforcement learning. Supervised learning uses a set of predefined data inputs and outputs to produce the desired output and is often used in classification and regression problems. Unsupervised learning, on the other hand, trains the model only on the input set to identify structures and patterns. Reinforcement learning models use rewards to achieve desirable outcomes. Due to the simple comprehensibility compared to the other two methods, supervised learning algorithms are good for exploring precisely simulated New Physics (NP) models. But what is NP? While the Standard Model (SM), first formulated in the 1970s, is the best description for the interaction of the three fundamental forces (electromagnetism, weak interaction, and strong interaction) and the corresponding particles, it fails to explain some experimental and theoretical anomalies such as dark matter, matter anti-matter symmetry, and the baryon asymmetry. This forces physicists to look beyond the SM and develop new hypotheses for these phenomena resulting in the search for NP. The Belle II experiment uses high-precision measurements to probe for NP and to find further sources of CP violation.

This thesis focuses on the new search strategy for the Belle II experiment proposed for the two-body decay of $B^+ \to K^+\alpha$ [1] where α is the source of missing particle with a supervised ML algorithm (XGBoost). To be precise, the search for the invisible

Axion-like particle (ALP) with a theoretical mass of $m_a < m_B - m_K$ was performed using simulations based on the Monte Carlo method. The simulation sample for each mass hypothesis consists of a signal sample with 100 000 generated events and a background sample corresponding to an integrated luminosity of 100 fb^{-1}. Chapter 2 deals with this topic in more detail. Chapter 3 covers the XGBoost algorithm and the gradient boosting fundamentals. Additionally, an overview of the hyperparameters and training approach will be presented. Chapter 4 outlines the performance evaluation of the trained models by calculating the projected sensitivity for the best selection cut using the Punzi figure of merit. Lastly, chapter 5 presents us with the conclusion and the future outlook for this thesis.

1.1 Experiment

The Belle II experiment consists of the asymmetric e^+e^- SuperKEKB collider and the Belle II detector. SuperKEKB is the successor of the KEKB collider and achieves a collision rate 30 times higher than its predecessor by increasing the current flow and implementing the nano-beam scheme. The beam size at the collision point is reduced to 50 nm and the collision angle between e^+ and e^- is set to 83 mrad increasing the collision cross section. To study the time-dependent CP violation of the B-meson decays, the SuperKEKB operates at the $\Upsilon(4(S))$ resonance with $E(e^+) = 4.0\,\text{GeV}$ and $E(e^-) = 7.0\,\text{GeV}$. The accelerator is also able run below the $\Upsilon(1S)$ (9.46 GeV) resonance and above the $\Upsilon(6S)(11.24\,\text{GeV})$ resonance. This $\Upsilon(4(S))$ setup yields a $B\bar{B}$ decay in 96% of the time [2].
The Belle II detector and its sub-detectors are optimized for this asymmetric e^+e^- collision.

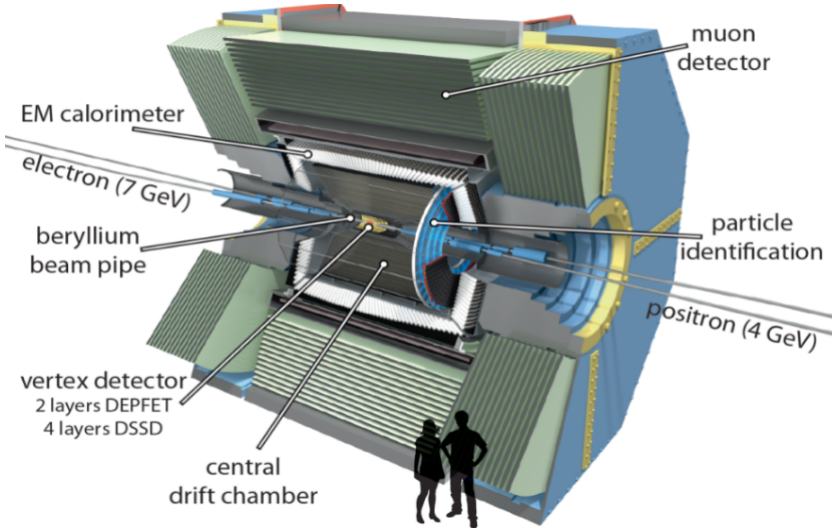


Figure 1.1: Structure of the Belle II detector with the different sub-detectors [3].

1.1.1 Belle II Detector

The Belle II detector consists of multiple sub-detectors. These are layered around the interaction point (IP) in the middle of the beryllium beam pipe. Detectors closest to the interaction region provide very high vertex resolution. Next is the central drift chamber (CDC) for trajectory and momenta tracking of charged particles. One of the particle identification detectors is layered between the CDC and the EM calorimeter (ECL) and the other is at the end-cap. The ECL provides the energy of the particle by measuring the deposits of electrons and photons. Between the ECL and the last layer is a superconducting solenoid to enable a homogeneous magnetic field for the inner detectors.

The magnetic flux return is placed in the last layer, the K_L and muon detector (KLM).

Vertex Detector

The vertex detector consists of two detector types: Pixel Vertex Detectors (PXD) and Silicon Vertex Detectors (SVD). With two layers right around the beam pipe at r = 14 mm and r = 22 mm the PXD is made up of over 7 million pixels providing a high vertex resolution. The SVD consists of 4 layers at r = 38 mm, 80 mm, 115 mm, and 140 mm.

Central Drift Camber

The CDC is made up of 56 layers of small drift cells with two different orientations alternating between layers for measuring the trajectories and momenta of charged particles to assess the energy loss. The individual cells are filled with a 50:50 mixture of He and C_2H_6 and amount to an overall radius of 1130 mm.

Particle Identification

While the CDC provides some information for particle identification, the heavy lifting is done by the time-of-propagation counter (TOP) in the barrel and the aerogel ring-imaging Cherenkov (ARICH) detector in the forward end-cap. Both use the Cherenkov effect to distinguish between pions and kaons. 16 TOP surround the CDC to cover the barrel while the ARICH is located at the end-cap.

EM Calorimeter

The ECL layer is made up of 8736 thallium-doped CsI crystals and covers all detector regions consisting of the barrel and the two end-caps. Incoming particles will produce bremsstrahlung and electron-positron pairs in the crystals resulting in an electromagnetic shower. The ECL's main task is measuring the gamma ray deposit and the distinction between electrons and hadrons, in particular π^0.

K_L and μ Detector

The KLM consists of multiple layers of alternating 4.7 cm iron plates, and detector elements. The iron plates function as magnetic flux return for the solenoid between ECL and KLM and as shower material for the long-living K_L meson to shower hadronically.

1.2 Theory

Inspired by the Axion, first proposed as a possible solution to the strong CP problem in quantum chromodynamics, the ALPs are new pseudo scalars that interact with SM particles and are predicted in many beyond SM theories as a solution to the spontaneous CP violation. Until 1956 parity and charged symmetry were believed to be fundamental

conservation laws. This changed with the parity violation discovered by Chien-Shiung Wu [4]. Although the P symmetry was broken, the combined CP symmetry was still in place. 1964 the team of James Cronin and Val Fitch [5] discovered the CP violation in the K_L^0 decays. For neutral mesons, the CP violation can occur in decay or via particle anti-particle oscillation while for charged mesons it only occurs in decay. This thesis will only focus on the CP violation and its respective ALPs for charged particles in particular $B^+ \to K^+\alpha$.
In the previously mentioned search strategy, the ALP is a pseudo Nambu-Goldstone boson with a coupling component to the SM fermions and a coupling component to the weak gauge bosons. Depending on the coupling the ALPs will either decay into leptons $\alpha_{fermion} \to l^+l^-$ or photons $\alpha_{gauge} \to \gamma\gamma$ [1, 6]. While BaBar searched for displayed lepton decay and a leptophilic scalar for α [7, 8] we will focus on ALPs which are not detectable in the Belle II detector, hence invisible. With the extensive angular detector coverage and excellent reconstruction efficiency for K^+, the missing energy α can be detected. Although $B^+ \to K^+\alpha$ is a two-body decay, it is similar to the $B^+ \to K^+\nu\bar{\nu}$ [9] for its leptons are likewise undetectable at Belle II.

CHAPTER 2

Simulation

Employing supervised learning models in search of invisible ALPs requires the use of simulation for the model training. In physics, a simulation refers to the imitation of real-world processes. This is realized by first generating the particles and then simulating the interaction with the detector. The event generator allows us to specify the properties and quantity of the particle according to the desired physics model while the detector simulation imitates the interaction between the particle and the detector, such as ionization or Cherenkov radiation. By reconstructing the particles from the raw detector output the simulation is more suitable for analysis. But real-world reconstruction has an intrinsic error rate due to background noise in each detector. This has to be accounted for in a simulation. But for obvious reasons, we also want to have the correct particle reconstruction in the simulation resulting in the classification between erroneous reconstruction and the perfect reconstruction referred to as MC truth.

Signal Samples

The signal samples are produced with EVTGEN [10] by generating 100 000 $\Upsilon(4S) \to B^+B^-$ where one B decays into $K^+\alpha$. Three different MC truth matched mass hypotheses $m_a \in \{0.005, 3.0, 4.6\}$ GeV are used for model training.

Background Samples

The 100 fb^{-1} background samples at the $\Upsilon(4S)$ resonance can be split into two different types of backgrounds. In the case of the background, the MC Truth matching was of no concern.

- **continuum background** stems from the decay of $e^+e^- \to q\bar{q}$ ($q = u, d, s, c$) and is generated with KKMC generator [11].
- **mixed** and **charged background** contains all $b \to u, d, s, c$ decays resulting from the $\Upsilon(4S) \to B^0\bar{B}^0$ and $\Upsilon(4S) \to B^+B^-$ process. Those events are generated using EVTGEN.

2.1 Variables

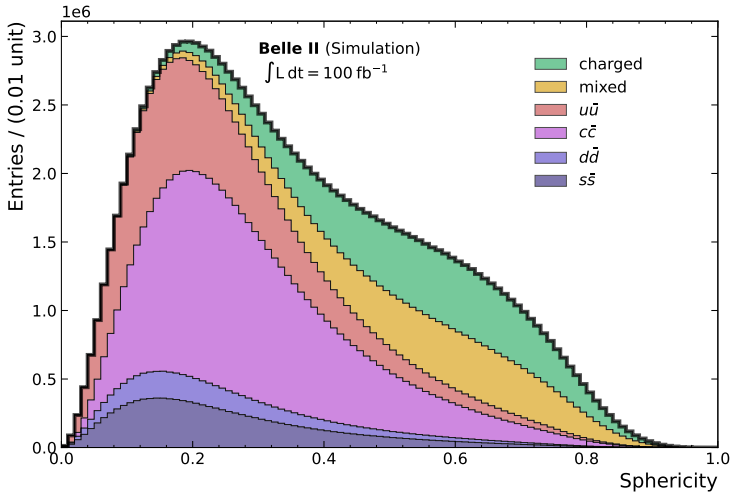


Figure 2.1: Sphericity histogram of the different background sources with an integrated luminosity of 100 fb^{-1} stacked on top of another. The continuum background ($q\bar{q}$) shows a jet-like structure while the event topology of the other two backgrounds is more spherical.

The Belle II experiment uses a variety of engineered variables to discriminate between the B-meson and the continuum background. These variables can also be used for the K-meson.

Topology

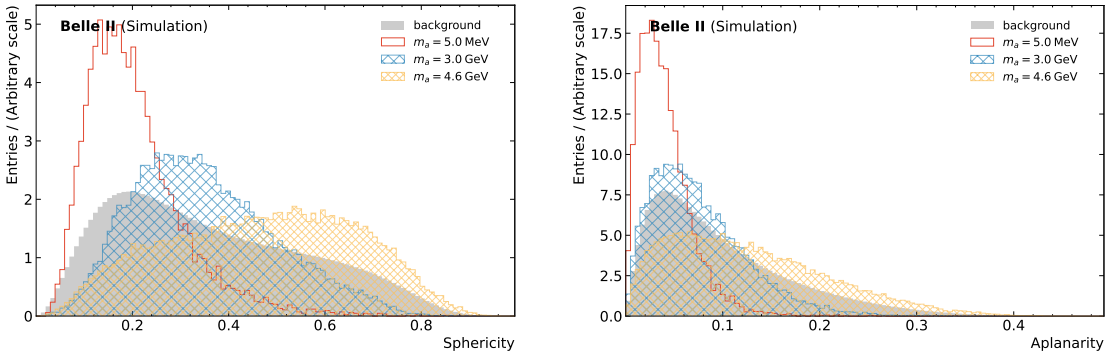


Figure 2.2: Normalized histogram comparison between different ALP hypotheses and background for the event topology. The right-hand side depicts the sphericity while the left-hand side shows the aplanarity.

The event topology can be split into two examples, sphericity and aplanarity [12]. The distribution of background topology is dominated by the jet-like structure of $q\bar{q}$ events

as seen in figure 2.1. For large ALP masses, the $B^+ \to K^+\alpha$ decay manifests a more isotropic distribution resulting in higher sphericity. Decays with low m_a result in faster Kaons displaying a less spherical distribution due to the higher momentum. The same discrimination can be used for the aplanarity. Figure 2.2 clearly shows a difference depending on the ALP hypotheses. Smaller m_a results in a flatter, less spherical jet-like topology just like the background.

Thrust

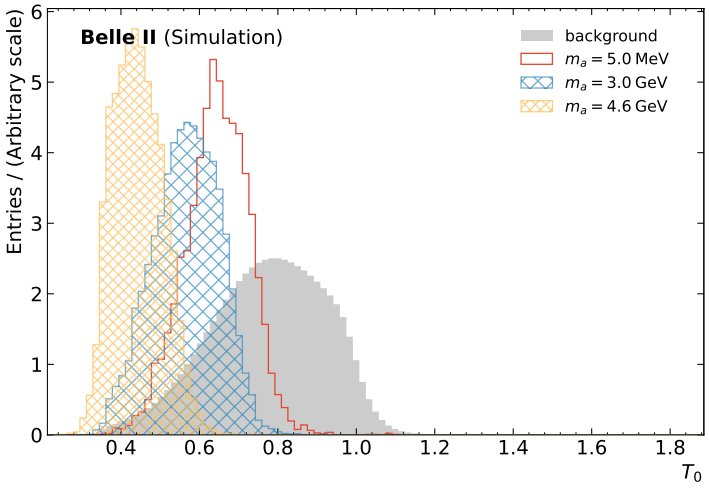


Figure 2.3: Normalized histograms of the harmonic moment of the thrust T_0 for background and different ALP hypothesis.

The thrust scalar is defined as

$$T = \frac{\sum_{i=1}^{N} \left|\vec{T} \cdot \vec{p_i}\right|}{\sum_{i=1}^{N} |\vec{p_i}|}. \tag{2.1}$$

for N particles with $\vec{p_i}$ (i = 1 to N) momenta and is one of the best classification variables. Heavy ALPs lead to Kaons with low momentum resulting in an overall lower thrust compared to the background.

Thrust angle θ_K

θ_K is the angle between the thrust axis of the Kaon $\vec{T}_K$ and the rest of the events $\vec{T}_{ROE}$. The isotropic $B^+ \to K^+\alpha$ decay for large m_a leads to a more symmetrical distribution while the histogram for lighter ALPs and the background is more asymmetric as seen in figure 2.4.

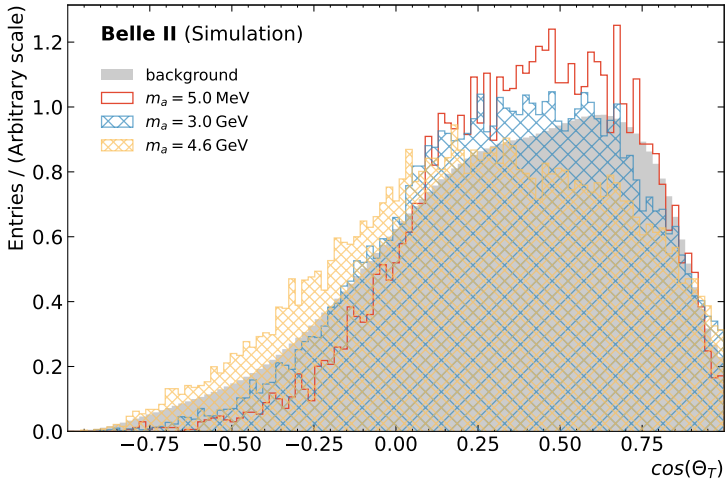


Figure 2.4: Normalized histograms of the thrust angle $\cos(\theta_T)$ for background and different ALP hypothesis.

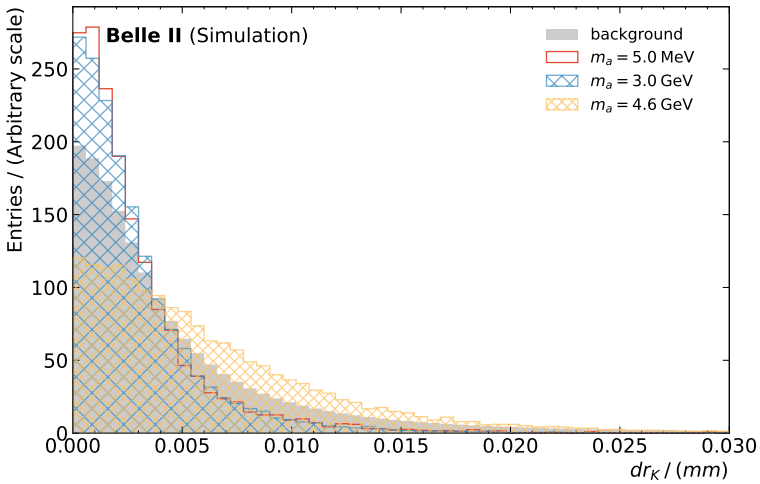


Figure 2.5: Normalized histograms of the distance between IP and the K-meson for background and different ALP hypothesis.

Distance between K and IP

This variable (dr) is defined as the distance between the IP and the closest approach of the K^+ track. The distance increases with larger m_a as seen in figure 2.5.

CLEO Cone 3

At an interval of 10°, the CLEO cones wrap around the thrust axis and measure the sum of the absolute momenta for all particles [13] in both directions. There are a total of 9 pairs of cones. This discrimination is caused by the jet-like event topology of the

continuum background and the random distribution for $\vec{T}_B$. K^+ on the other hand has a similar distribution as the background pictured in figure 2.6.

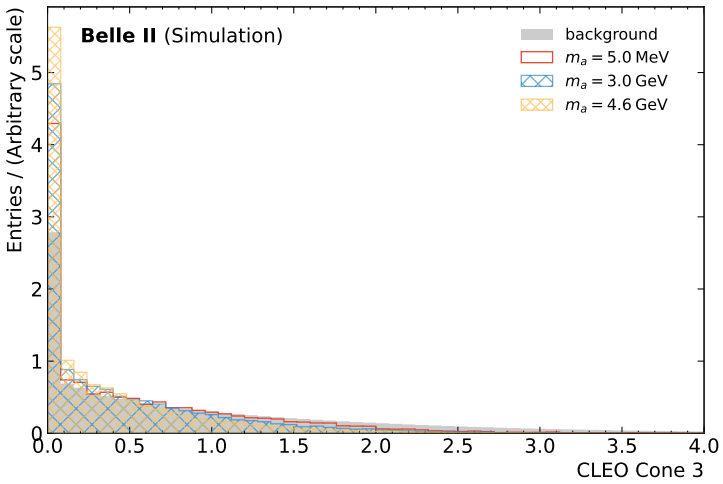


Figure 2.6: Normalized histogram comparison between the different ALP hypothesis and background for the CLEO cone 3.

Fox Wolfram Moment

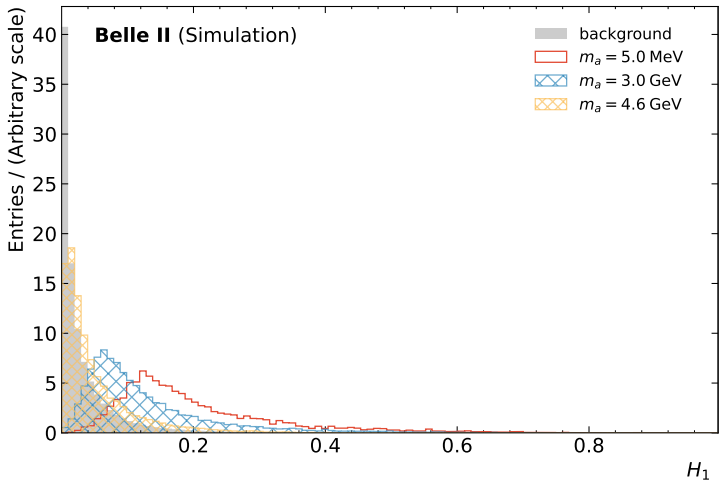


Figure 2.7: Normalized histogram comparison between the different ALP hypothesis and background.

The Fox-Wolfram moment H_l [14] is defined as

$$H_l = \sum_{i,j}^{N} |p_i|\,|p_j|\,P_l\left(\cos\theta_{i,j}\right) \tag{2.2}$$

for N particles with p_i momenta and an angle $\theta_{i,j}$ between p_i and p_j. P_l denotes the Legendre polynomial. It is a great discriminator for small m_a shown in figure 2.7.

CHAPTER 3

XGBoost

Classification

Classification algorithms use statistical methods to categorize a given input. There are many classifiers with completely different approaches. One of the easiest classifiers would be logistic regression. It uses the sigmoid function

$$p = \frac{1}{1 + e^{-X}} \tag{3.1}$$

to categorize the new data point X into 1 or 0 depending on whether the probability $p \in [0, 1]$ is greater or less then the given threshold. Logistic regression performs well for binary classification but has problems with more than 2 categories. K-nearest neighbors (KNN) on the other hand is a simple non-parametric learning algorithm that can handle multiple categories. The new data points are classified by a plurality survey of its $\mathrm{K} \in \mathbb{N}$ nearest neighbors. XGBoost uses a decision tree-based classifier with a probability output p between 0 and 1, just like the logistic regression. The predefined threshold is 0.5.

XGBoost

Extreme gradient boosting is an optimized gradient boosting algorithm specializing in system and algorithmic optimization such as regularization. When ML models train too extensively on a given training set they will fit exactly against the training set resulting in poor performance on unseen testing sets. This is called overfitting. XGBoost has multiple types of regularization options to prevent models from overfitting. Other optimizations are listed below.

System Optimization

- **Parallelization:** Weak learner consists of an outer and inner loop. The inner loop calculates the tree's features while the outer loop lists the leaf nodes. For the outer loop to start, the inner loop must be completed first. XGBoost parallelized implementation allows swapping the order of both loops resulting in a performance increase.

- **Hardware optimization:** XGBoost allocates internal buffers in every thread to store the gradient statistics. "Out of core" computing uses disk space to process data frames that cannot fit inside the core memory.

Algorithmic Improvements

- **Cross-validation:** XGBoost has a built in cross-validation method.
- **Regularization:** XGBoost has in-build methods to prevent the overfitting of models.
- **Tree pruning:** The "max_depth" hyperparameter restricts the individual tree depth. After the tree has reached its given depth, the tree pruning starts from the bottom up. Leaves with the lowest entropy gain are cut off.

3.1 Gradient boosting

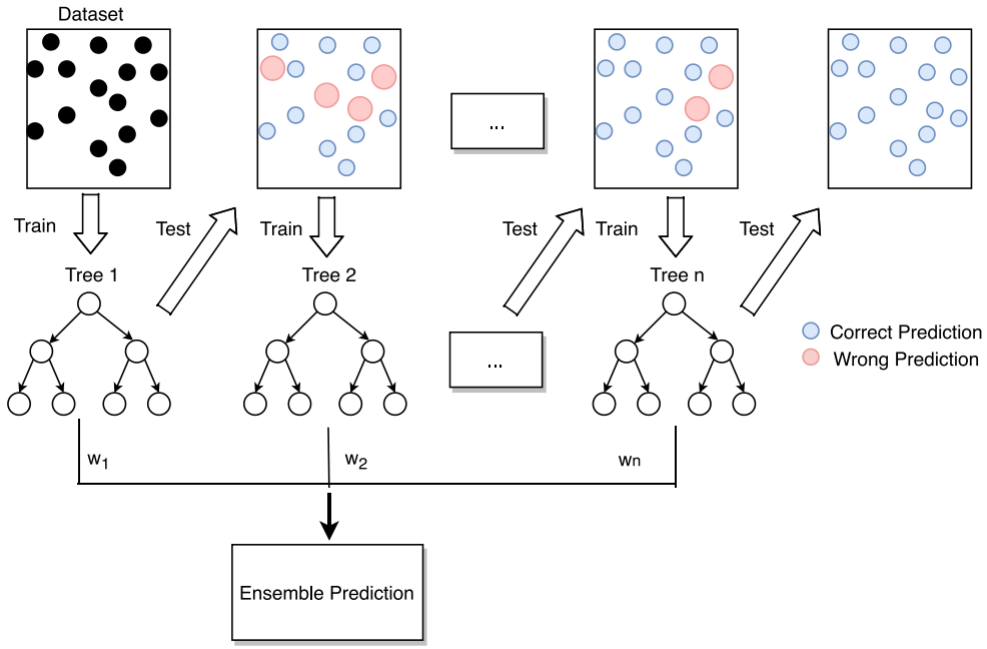


Figure 3.1: Flowchart of gradient boosting algorithm [15]. The final classifier consists of an ensemble of sequentially generated weak learners.

As seen in Figure 3.1 the gradient boosting algorithms create their first weak learner $T_1(X)$ from the data sample to generate the first prediction model $\hat{y}^{(1)}$.

$$\hat{y}^{(1)} = T_1(X) \tag{3.2}$$

Using one of the many different loss functions L we can evaluate how well this tree does. The second tree has to reduce the loss function of $\hat{y}^{(1)}$.

$$\begin{aligned} L(\hat{y}^{(1)}) &> L(\hat{y}^{(2)}) \\ \text{with } \hat{y}^{(2)} &= T_1 + \eta T_2 \end{aligned} \tag{3.3}$$

This can be done by using the derivative of L with respect to $\hat{y}^{(1)}$

$$\begin{aligned} T_2 &= -\frac{\partial L(\hat{y}^{(1)})}{\partial \hat{y}^{(1)}} \\ \hat{y}^{(2)} &= \hat{y}^{(1)} - \eta \frac{\partial L(\hat{y}^{(1)})}{\partial \hat{y}^{(1)}} \end{aligned} \tag{3.4}$$

resulting in $L(\hat{y}^{(1)}) > L(\hat{y}^{(2)})$. The constant η is called the learning rate.
This model is called gradient boosting because we used the gradient of the loss function to boost the new prediction model. The iterative function for the gradient boosting is thus defined as

$$\hat{y}^{(n)} = \hat{y}^{(n-1)} - \eta \frac{\partial L(\hat{y}^{(n-1)})}{\partial \hat{y}^{(n-1)}} \tag{3.5}$$

3.2 Hyperparameters

For model training, the XGBClassifier function was used and four specific hyperparameters were selected to find the best model. Table 3.1 shows the final values of the selected parameters.

Table 3.1: Hyperparameter values used in the model training

hyperparameter	value
n_estimators	500
learning_rate	0.2
eval_metric	log loss
scale_pos_weight	$\approx$ 7000

- **n_estimators:** Defines the number of weak learners used for the final model. n=500 yields the best results in terms of overfitting for the different ALP hypotheses.
- **learning_rate:** The learning rate η determines how fast the loss function approaches its minimum. If η is too big the model will always overshoot the minimum, and for η too small, the model training takes significantly longer.
- **eval_metric:** The evaluation metric is the loss function mentioned earlier. For our purpose, we used the Logistic loss

$$L(y_i) = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right) \cdot \log\left(1 - p\left(y_i\right)\right) \tag{3.6}$$

where $p(y_i)$ is a likelihood function maximized by the maximum-likelihood estimation.

- **scale_pos_weight:** The scale positive weight allows the user to adjust classification thresholds by scaling the signal to an arbitrary number for better performance with imbalanced data samples. The reconstruction efficiency was largely responsible for the different ratios between signal and background samples for the different mass hypotheses leading to equation 3.7

$$weight = \frac{N_{background}}{N_{signal}} \approx 7000 \tag{3.7}$$

3.3 Training approach

The model training was done using three different hypotheses $\mathrm{m_a} = 0.005\,\mathrm{GeV}$, 3.0 GeV and 4.6 GeV with 100 000 generated signal samples each. The combined signal and background set was split into three equal parts for the training, testing, and validation set. To determine the best value for the number of trees n, the logistic loss values of the train set and the validation set would usually be used to prevent overfitting but the imbalanced simulation samples ($\mathrm{N_{signal}} << \mathrm{N_{background}}$) required another method. By comparing the histograms for the train set and test set from the same model, we can evaluate the overfitting. Although XGBoot's regularization algorithm prevents the discrepancy from getting too large, there is still an evident difference resulting in equation 3.8 to evaluate the disparity. By summing up the difference between the bin counts of the train set $B_{train,i}$ and the test set $B_{test,i}$ for all bins N and dividing the result by the total number of signal events ν for the train set we get the first evaluation metric.

$$\sigma_{bins} = \sqrt{\frac{\sum_i^N (B_{train,i} - B_{test,i})^2}{\nu}} \tag{3.8}$$

Another value used for the evaluation is the signal samples' true positive rate TP. While the background classification improves with larger n, the classification accuracy for the signal samples decreases as seen in Figure 3.3. By combining σ_{bins} and TP we get

$$G = \frac{\sigma_{bins}}{TP} \tag{3.9}$$

Table 3.2: Comparing the different evaluation parameters for $m_a = 3.0\,\mathrm{GeV}$.

Number of trees	σ_{bins}	TP	G
200	56.11 ± 0.20	0.985 ± 0.0002	57.03 ± 0.20
500	53.07 ± 0.33	0.939 ± 0.0003	56.54 ± 0.34
800	53.65 ± 0.13	0.923 ± 0.0004	58.12 ± 0.13
1000	53.47 ± 0.12	0.914 ± 0.0003	58.51 ± 0.13
1500	59.71 ± 0.19	0.906 ± 0.0004	65.91 ± 0.19
2000	67.23 ± 0.14	0.915 ± 0.0007	73.51 ± 0.12

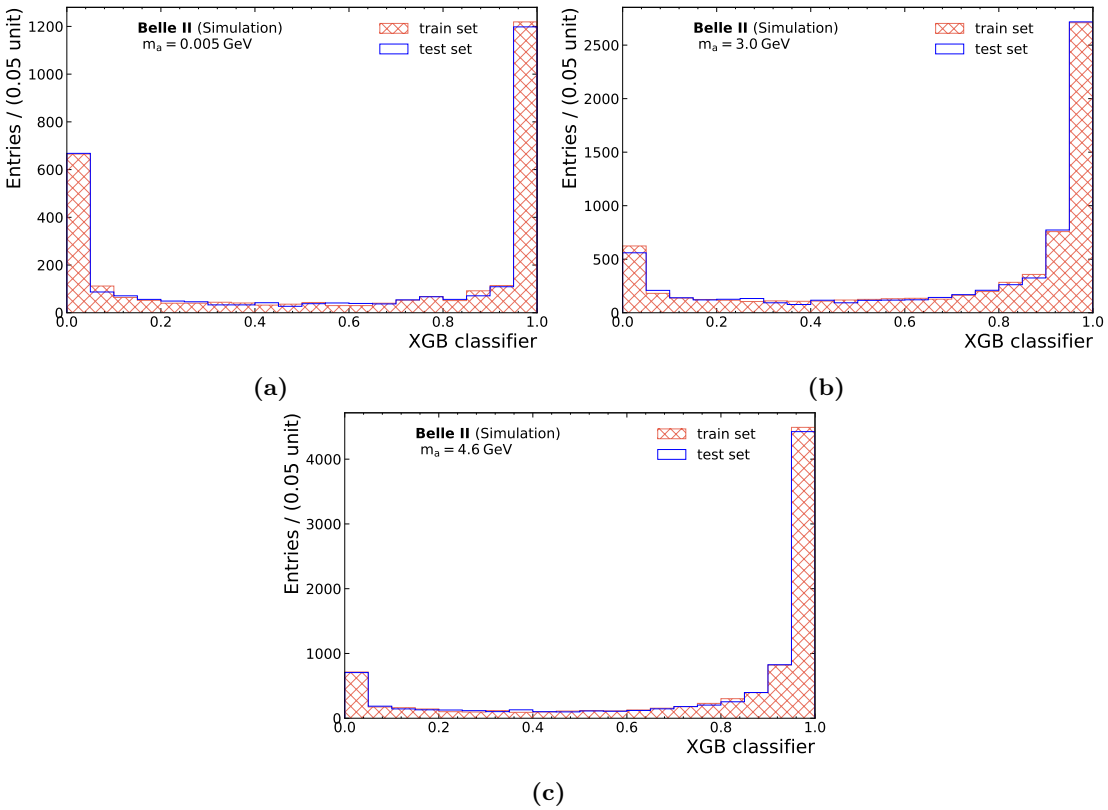
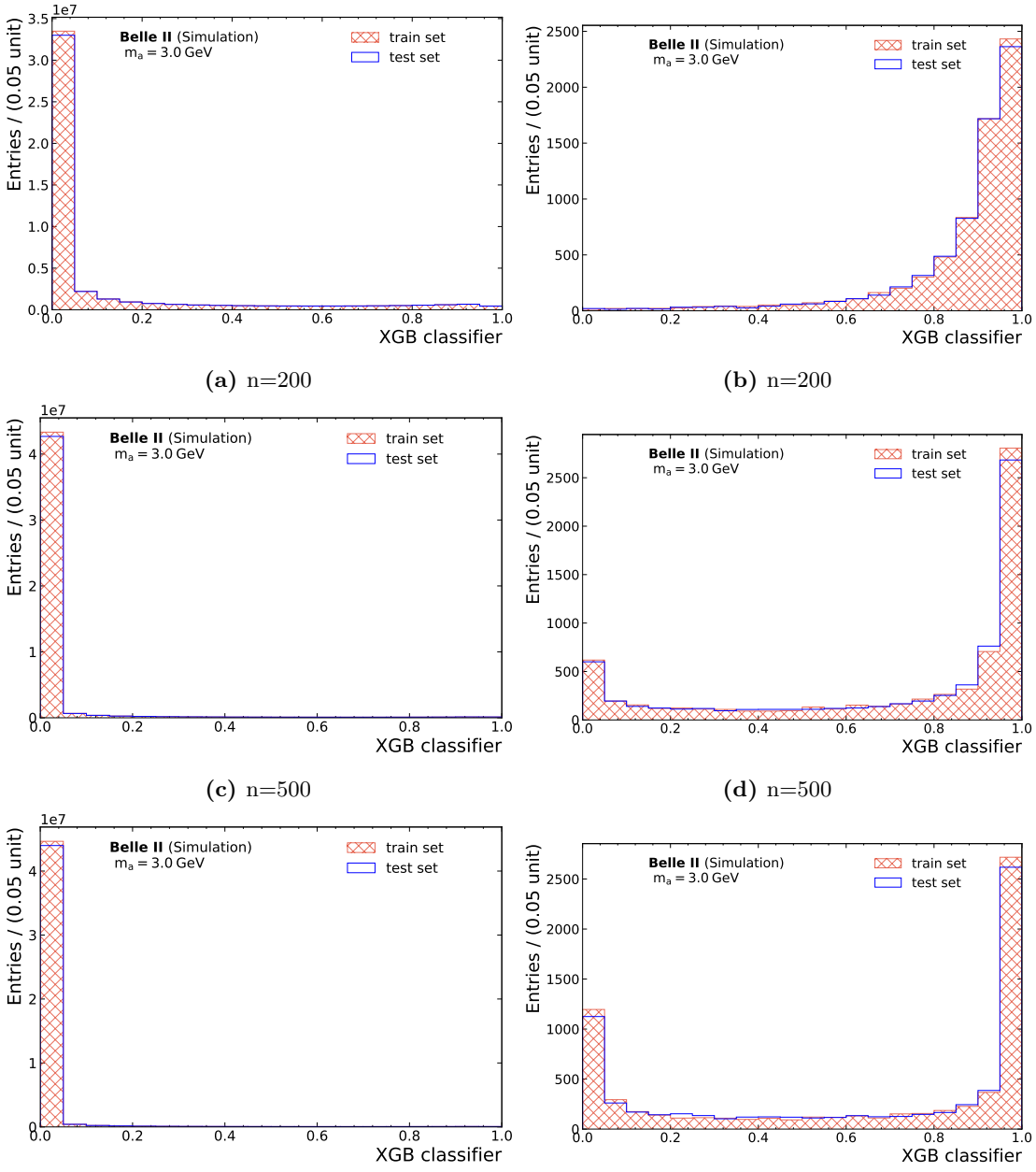


Figure 3.2: Train set and test set comparison of the signal samples with n = 500 and different m_a.

According to table 3.2 $n = 500$ is the best value for an ALP mass of $m_a = 3.0\,\text{GeV}$. For comparison purpose, $n = 500$ will also be used for $m_a = 5.0\,\text{MeV}$ and $m_a = 4.6\,\text{GeV}$ even though the shape of the signal histogram for $m_a = 5.0\,\text{MeV}$ in Figure 3.2 clearly indicates a sub-optimal n value. At this point, it should be stated, that the background histograms for all three hypotheses look nearly identical like the (**b**) plot from Figure 3.3.



(a) n=200 **(b)** n=200

(c) n=500 **(d)** n=500

(e) n=800 **(f)** n=800

Figure 3.3: Train set and test set histogram comparison for $m_a = 3.0\,\text{GeV}$. The graphs on the left-hand side show the background classification while the right-hand side shows the signal classification for different numbers of trees n.

CHAPTER 4

Results

In this chapter, we will evaluate the sensitivity S for $B \to K\alpha$ presuming that the branching fraction is the same as $B \to K\nu\bar{\nu}$ due to its experimental similarities. Another big aspect is to evaluate whether training multiple models with different ALP hypotheses yields a better result in classifying the signal for an unknown m_a or one model trained on multiple m_a. Using the branching fraction for a given decay and the signal efficiency of the detectors, a prediction is made as to how many ALPs are created in the process. Comparing that to the number of background particles gives us a good estimation if the detected particle is an ALP or just a background fluctuation. The threshold for a discovery to be considered significant is 5 σ or a one-in-a-million probability. The Punzi figure of merit (PFOM) [16] is used to evaluate the best selection criterion (SC) to maximize the sensitivity.

4.1 Signal efficiency

The signal efficiency ϵ_s is defined as the ratio between the number of events detected N_{signal} divided by the total number of generated MC signal samples N_{MC}

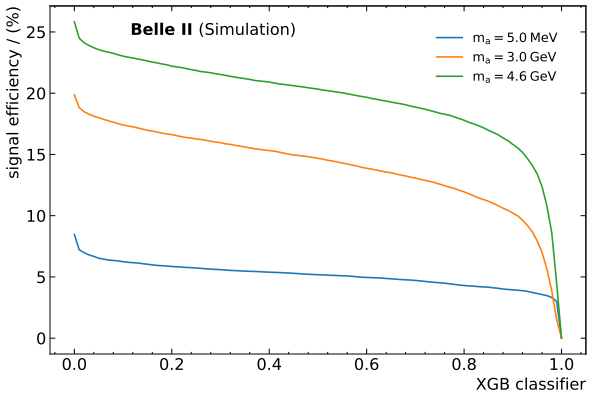
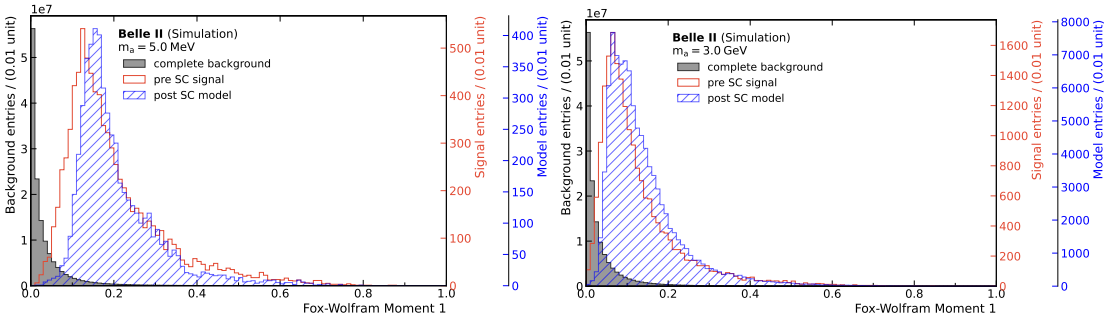


Figure 4.1: Signal efficiency for the prediction models with the different hypotheses.

$$\epsilon_s = \frac{N_{signal}}{N_{MC}} \tag{4.1}$$

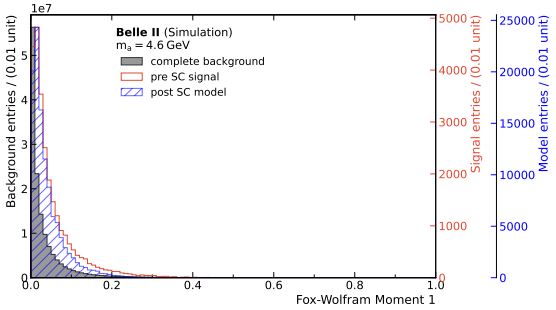
The shape of the plot is a good indicator of how well the XGBoost classifier performs. Figure 4.1 shows the efficiency for the three ALP hypotheses. Larger masses have a higher signal efficiency caused by signal generation but the relative drop off between 0.1 and 0.95 is lower for $m_a = 5.0\,\text{MeV}$ indicating an overall better plot.

4.2 PFOM



(a) Signal sample m_a=5.0 MeV

(b) Signal sample m_a=3.0 GeV



(c) Signal sample m_a=4.6 GeV

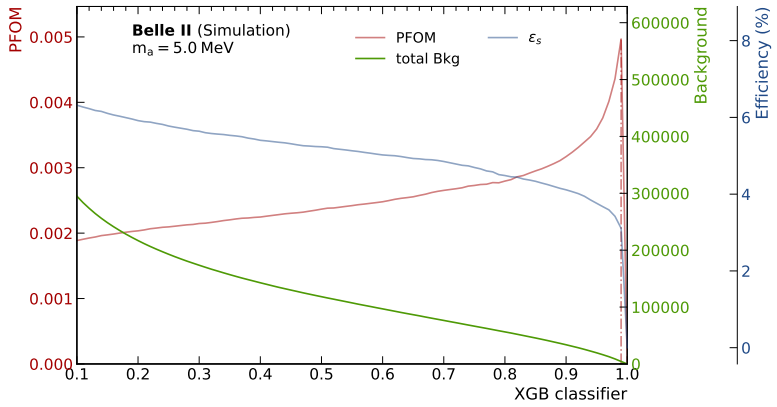
Figure 4.2: Histogram comparison for background sample, signal sample, and the sample set left after the selection cut applied at $PFOM_{max}$ for the models $M_{0.005}$, $M_{3.0}$, and $M_{4.6}$

The PFOM is used for particles with unknown branching fractions to determine an SC with the least amount of background while keeping the signal efficiency as high as possible to maximize the possibility of discovery for a given significance $\gamma\,\sigma$. It is defined as

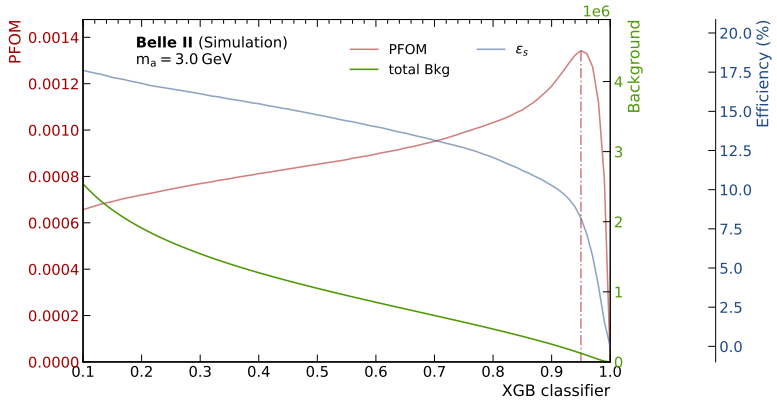
$$PFOM = \frac{\epsilon_s}{\sqrt{B} + \frac{\gamma}{2}} \tag{4.2}$$

where B is the number of background events in SC. In this case, the significance $\gamma = 5$ is applied at the maximum $PFOM_{max}$ to get the best sensitivity S. According to figure 4.3 m_a=5.0 MeV has the highest $PFOM_{max}$ despite having the lowest overall efficiency.

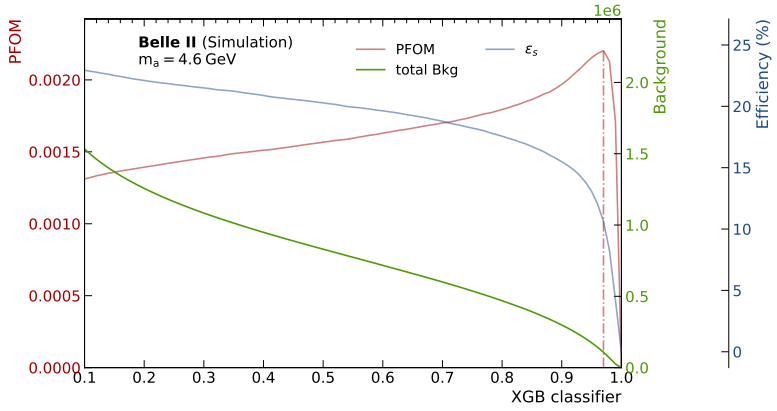
However, its shape displays a volatile working point caused by the number of estimator trees. Volatile working points are too reliant on the given samples and often yield sub-optimal results for samples with a slight deviation caused by natural uncertainties. Model $M_{0.005}$ fits the given simulation sample too well. Figure 4.2 displays a comparison between the background histogram with an integrated luminosity of 100 fb^{-1}, the signal samples with 33 333 generated events, and the histogram of the SC for the first Fox-Wolfram moment H_1. The trained models picked up the difference between signal and background. According to the feature importance function of the sklearn module, the first Fox-Wolfram moment and the harmonic moment of the thrust T_0 are the most significant value for distinguishing between signal and background for the different m_a.



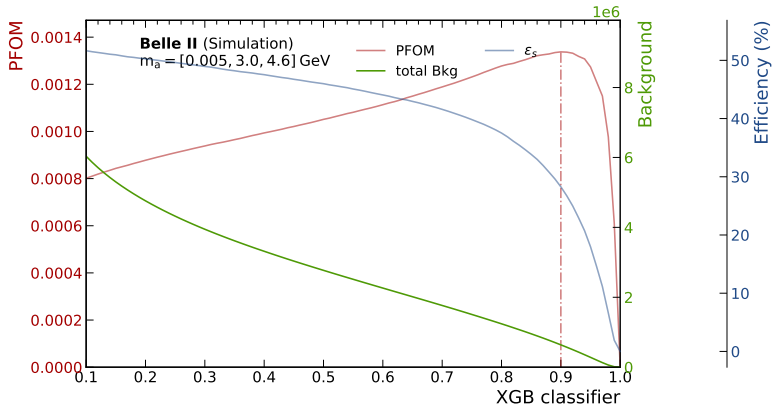
(a) Model $\mathrm{M}_{0.005}$



(b) Model $\mathrm{M}_{3.0}$



(c) Model $\mathrm{M}_{4.6}$



(d) Model $\mathrm{M}_{\mathrm{all}}$

Figure 4.3: PFOM for the different hypotheses $m_a = \{0.005, 3.0, 4.6\}$ GeV and their corresponding XGBoost models $\mathrm{M}_{0.005}$, $\mathrm{M}_{3.0}$, $\mathrm{M}_{4.6}$ and $\mathrm{M}_{\mathrm{all}}$. The dashed line indicates the best selection cut for the highest sensitivity projection.

4.3 Sensitivity

Table 4.1: Sensitivity values of different models M_i applied to the signal samples m_a.

Models M_i	**Sensitivity in σ for m_a**		
	5.0 MeV	3.0 GeV	4.6 GeV
$M_{0.005}$	$(9.3 \pm 0.5) \times 10^{-3}$	-	-
$M_{3.0}$	-	$(12.7 \pm 0.3) \times 10^{-3}$	-
$M_{4.6}$	-	-	$(16.9 \pm 0.4) \times 10^{-3}$
M_{all}	$(3.1 \pm 0.4) \times 10^{-3}$	$(5.9 \pm 0.3) \times 10^{-3}$	$(10.7 \pm 0.3) \times 10^{-3}$

The selection cut performed with the PFOM gets us an estimation for S for α using the XGBoost algorithm as a searching tool. Because of the unknown branching fraction (BF for $B^+ \to K^+\alpha$), we assume this decay is as likely as the ($B^+ \to K^+\nu\bar{\nu}$) decay [9]. S is defined as

$$S = \frac{N_s}{\sqrt{B_{SC}}} = \frac{\epsilon_s(PFOM_{max}) \cdot BR \cdot L \cdot \sigma(e^+e^- \to b\bar{b})_{\Upsilon(4S)}}{\sqrt{B_{SC}}} \tag{4.3}$$

where L is the integrated luminosity, and $\sigma(e^+e^- \to b\bar{b})_{\Upsilon(4S)}$ is the cross-section of the ($e^+e^- \to b\bar{b}$) decay at the $\Upsilon(4S)$ resonance. B_{SC} is the number of background events in the selection cut. Furthermore, BR denotes the combined BF for both decays

$$BR = 2 \cdot BF(\Upsilon(4S) \to B^+B^-) \cdot BF(B^+ \to K^+\nu\bar{\nu}) = 2 \cdot 0.514 \cdot 4.6 \times 10^{-6} \tag{4.4}$$

Due to time constraints, the calculations of upper limits for BF were beyond the scope of this thesis. Instead, the significance for a fixed BF was calculated to provide a benchmark for the sensitivity of the measurement. To determine the standard deviation of S, assuming the branching fraction is the same as the $B \to K\nu\bar{\nu}$ event of the SM cross-validation was performed in 10 smaller simulation samples with $L = 10\,\text{fb}^{-1}$. Table 4.1 reveals that the statistical significance is less than $5\,\sigma$ for all hypotheses, resulting in the rejection of the ALP hypothesis. Even though the $M_{0.005}$ has the best performance in terms of sensitivity, it still is way below the $5\,\sigma$ threshold, despite operating at an extremely unstable working point. By comparison, the $M_{3.0}$ would need n = 2000 estimators for a similar plot and $PFOM_{max}$ yielding a sensitivity of $S = (0.17 \pm 0.5)\,\sigma$. While increasing the number of estimators may get us fewer background events for the selection cut, it also increases the false negative rate as seen in figure 3.3 resulting in a smaller $\epsilon_s(PFOM_{max})$. The engineered variables used in the model training aren't sufficient enough to counteract the imbalanced simulation sample.
Additionally, another model M_{all} has been trained for the whole mass spectrum $m_a < m_B - m_K$ using all three signal samples for the training. According to figure 4.4 M_{all} has an easier time identifying the signal samples for larger m_a. This is caused by the different feature importance rankings on both ends. For $m_a < 2\,\text{GeV}$ the most important feature is

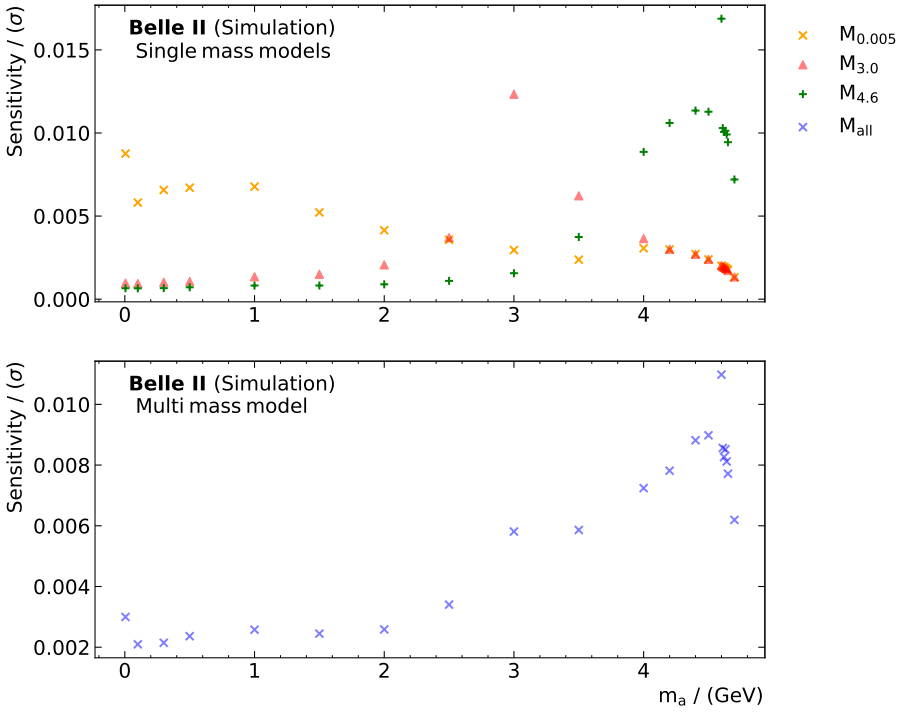


Figure 4.4: Trained models M_i applied on the whole m_a spectrum. Local peaks at 0.005, 3.0, and 4.6 GeV indicate the signal samples used in the training process.

the first Fox-Wolfram moment H_1 while for $m_a > 2\,\text{GeV}$ it is the harmonic moment of the thrust T_0. If one of those variables is not the dominant feature, XGBoost has difficulty classifying the signal sample indicated by the shape of the signal mass models. Due to the low signal efficiency rate for $m_a = 0.005\,\text{GeV}$ the multi-mass model prioritizes T_0 as the classification variable resulting in a similar plot as $M_{4.6}$. Despite its worse performance compared to the individual models on their respective signal samples the shape of the PFOM suggests improvement possibilities for M_{all}. This shows the difficulties of using ML algorithms in the search for new Physics.
With equation 4.3 we can reevaluate the branching fraction $BF(m_a)$ for the two body decay $(B^+ \to K^+\alpha)$ resulting in

$$BF(m_a = \{0.005,\ 3.0,\ 4.6\}\,\text{GeV}) = \{1.7,\ 3.5,\ 2.4\} \times 10^{-3}. \tag{4.5}$$

CHAPTER 5

Conclusion

This thesis evaluates the use of the XGBoost classifier in search of new physics beyond the standard model. In particular, the invisible particle α in the two-body decay of $B^+ \to K^+\alpha$ at the $\Upsilon(4S)$ resonance with $m_a < m_B$ - m_K. The model differentiates between the signal sample m_a consisting of 100 000 Monte Carlo generated events and the background sample with an integrated luminosity of $100\,\mathrm{fb}^{-1}$ by using the different engineered variables from the Belle II experiment to discriminate the background. Due to in-build regularization functions, a new metric had to be introduced for hyperparameter evaluation. Using $m_a = 3.0\,\mathrm{GeV}$ as a benchmark model $\mathrm{M}_{0.005}$ and $\mathrm{M}_{4.6}$ were trained on the extreme end of the hypotheses $m_a = \{0.005, 4.6\}\,\mathrm{GeV}$ by using the same hyperparameters for comparison purpose. All models' sensitivity calculation was performed by maximizing the PFOM and using the branching fraction of the experimentally similar $B \to K\nu\bar{\nu}$ event. In addition, all three models were applied to the whole m_a spectrum to see if a single mass model is sufficient for classifying an unknown m_a. Furthermore, $\mathrm{M}_{\mathrm{all}}$ was specifically trained on all three signal samples 5.0 MeV, 3.0 GeV, and 4.6 GeV for this purpose. To conclude the findings, single mass models are not suited for classifying outside their area because of the steep drop-off in sensitivity. This is due to the change in the feature importance ranking. Combining all three single mass models yields a better result but the overall sensitivity is still $S < 0.1\,\sigma$. Comparing the sensitivity of $\mathrm{M}_{\mathrm{all}}$ for small m_a with $\mathrm{M}_{0.005}$ we can see a huge difference. Although S for $\mathrm{M}_{0.005}$ is lower than the single mass models, the difference is not too big for $m_a > 2.5\,\mathrm{GeV}$.

5.1 Outlook

This section will discuss potential improvements for using the XGBoost classifier. First, the integrated regularization tools in the XGBoost classifier make it difficult to determine the best models for highly imbalanced simulation samples. The default regularization method is the Ridge method (L2). It forces the weights of each variable to be small but not zero. The Lasso regression (L1) on the other hand uses feature selection and eliminates the least significant features. Both have their own hyperparameters for adjusting the impact ranging from 0 to infinity. Another regularization parameter for gradient boosting is γ. Determining the suitable method must be the first step for model training. This

might improve the single mass model on both extremes of the m_a spectrum.
By combining the first Fox-Wolfram moment H_1 and T_0 the multi-mass model could be improved for all ALP hypotheses. A straightforward proposal would be

$$V_{new} = \frac{1}{m_a} \cdot H_1 + m_a \cdot T_0 \tag{5.1}$$

For small m_a F_1 prevails while for large m_a T_0 dominates. In addition, we have to consider the signal efficiency for the different mass hypotheses. The scale_pos_weight hyperparameter has to be first applied to the signal samples before combining them for training.

CHAPTER 6

List of Figures

CHAPTER 7

List of Tables

CHAPTER 8

Bibliography

[1] Torben Ferber et al. "Displaced or invisible? ALPs from B decays at Belle II" (Jan. 2022). arXiv: 2201.06580 [hep-ph].

[2] Particle Data Group Collaboration. "Review of Particle Physics". *PTEP* 2020.8 (2020), p. 083C01.
DOI: 10.1093/ptep/ptaa104.

[3] Florian Bernlochner et al. "Online Data Reduction for the Belle II Experiment using DATCON". *EPJ Web Conf.* 150 (2017). Ed. by C. Germain et al., p. 00014.
DOI: 10.1051/epjconf/201715000014. arXiv: 1709.00612 [hep-ex].

[4] C. S. Wu et al. "Experimental Test of Parity Conservation in β Decay". *Phys. Rev.* 105 (1957), pp. 1413–1414.
DOI: 10.1103/PhysRev.105.1413.

[5] J. H. Christenson et al. "Evidence for the 2π Decay of the K_2^0 Meson". *Phys. Rev. Lett.* 13 (1964), pp. 138–140.
DOI: 10.1103/PhysRevLett.13.138.

[6] Howard Georgi, David B. Kaplan, and Lisa Randall. "Manifesting the Invisible Axion at Low-energies". *Phys. Lett. B* 169 (1986), pp. 73–78.
DOI: 10.1016/0370-2693(86)90688-X.

[7] BaBar Collaboration. "Search for Long-Lived Particles in e^+e^- Collisions". *Phys. Rev. Lett.* 114.17 (2015), p. 171801.
DOI: 10.1103/PhysRevLett.114.171801. arXiv: 1502.02580 [hep-ex].

[8] BaBar Collaboration. "Search for a Dark Leptophilic Scalar in e^+e^- Collisions". *Phys. Rev. Lett.* 125.18 (2020), p. 181801.
DOI: 10.1103/PhysRevLett.125.181801. arXiv: 2005.01885 [hep-ex].

[9] BaBar Collaboration. "Search for $B \to K^{(*)}\nu\overline{\nu}$ and invisible quarkonium decays". *Phys. Rev. D* 87.11 (2013), p. 112005.
DOI: 10.1103/PhysRevD.87.112005. arXiv: 1303.7465 [hep-ex].

[10] D. J. Lange. "The EvtGen particle decay simulation package". *Nucl. Instrum. Meth. A* 462 (2001). Ed. by S. Erhan, P. Schlein, and Y. Rozen, pp. 152–155.
DOI: 10.1016/S0168-9002(01)00089-4.

[11] S. Jadach, B. F. L. Ward, and Z. Was. "The Precision Monte Carlo event generator K K for two fermion final states in e+ e- collisions". *Comput. Phys. Commun.* 130 (2000), pp. 260–325.
DOI: 10.1016/S0010-4655(00)00048-5. arXiv: hep-ph/9912214.

[12] BaBar Collaboration. "The Physics of the B Factories". Vol. 74. 2014, p. 3026.
DOI: 10.1140/epjc/s10052-014-3026-9. arXiv: 1406.6311 [hep-ex].

[13] CLEO Collaboration. "Search for exclusive charmless hadronic B decays". *Phys. Rev. D* 53 (1996), pp. 1039–1050.
DOI: 10.1103/PhysRevD.53.1039. arXiv: hep-ex/9508004.

[14] Geoffrey C. Fox and Stephen Wolfram. "Observables for the Analysis of Event Shapes in e^+e^- Annihilation and Other Processes". *Phys. Rev. Lett.* 41 (1978), p. 1581.
DOI: 10.1103/PhysRevLett.41.1581.

[15] Tao Zhang et al. "Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning". *Journal of Advances in Modeling Earth Systems* 13 (May 2021).
DOI: 10.1029/2020MS002365.

[16] Giovanni Punzi. "Sensitivity of searches for new signals and its optimization". *eConf* C030908 (2003). Ed. by L. Lyons, R. P. Mount, and R. Reitmeyer, MODT002. arXiv: physics/0308063.