Karlsruhe Institute of Technology

ETP-KA/2023-06

Anomaly Detection using Isolation Forests in Searches for Inelastic Dark Matter with a Dark Higgs at Belle II

Michael Binder

Bachelorthesis

9th June 2023

Institute of Experimental Particle Physics (ETP)

Advisor:	Prof. Dr. Torben Ferber
Coadvisor:	Dr. Giacomo De Pietro

Editing time: 1st March 2023 – 9th June 2023

Karlsruhe Institute of Technology

ETP-KA/2023-06

Anomalie Detektion mit Isolation Forests in Suchen nach inelastischer Dunkler Materie mit einem Dunklen Higgs bei Belle II

Michael Binder

Bachelorarbeit

9. Juni 2023

Institut für Experimentelle Teilchenphysik (ETP)

Referent:	Prof. Dr. Torben Ferber
Korreferent:	Dr. Giacomo De Pietro

Bearbeitungszeit: 1. März 2023 – 9. Juni 2023

Ich versichere wahrheitsgemäß, die Arbeit selbstständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde.

Karlsruhe, 9. Juni 2023

. . . .

(**Michael Binder**)

Disclaimer

This thesis builds on the work of Jonas Eppelt (KIT, ETP) on the search for Inelastic Dark Matter with a Dark Higgs. The production and reconstruction of the SM samples has been mainly done by the Belle II collaboration. Jonas Eppelt's reconstructions performed in his work are used. Parts of his program were also used and partly modified for own purposes. The simulation of the signals was performed by Patrick Ecker (KIT, ETP). The plots shown in this paper were also created by me using the plotting functions of Jonas Eppelt, unless otherwise noted. The Isolation Forest is based on the scikitlearn [1] implementation and partially modified or extended. All analyses in this thesis are performed by me.

Contents

1. Introduction

Particle physics is a dynamic field that seeks to unravel the puzzle of the universe and understand its fundamental nature. As knowledge expands, deviations from the well-established theoretical framework known as the Standard Model (SM) are emerging. These deviations, observed in the vast and intricate fabric of the universe, serve as clues to the existence of new particles and Physics Beyond the Standard Model (BSM). Among others, the Belle II experiment is used to search for Dark Matter (DM). A possible signature of the Dark Higgs is predicted by the theoretical Inelastic Dark Matter with a Dark Higgs (IDMDH) model. Using precise measurements of the SM and its processes, it is possible to study the collision process of an electron-positron pair effectively and to probe for evidence of unknown signals. However, the challenge is to detect these rare and anomalous events in the huge amount of data generated by such collisions. Anomaly Detection (AD) techniques have proven to be useful for the direct identification of unknown signals in complex data sets [2]. These techniques are widely used in various fields, including finance, where they are often used as powerful tools to detect fraud [3]. One particular method that has attracted considerable attention is the Isolation Forest (IForest) algorithm [4], which is known for its efficiency in detecting anomalies [5]. The search for the unknown in collision processes also offers a potential use of this approach. Motivated by the possibility of discovering new physical signatures, this work examines in detail the IForest algorithm within a search for the IDMDH model at Belle II.

This thesis focuses on training the IForest using Monte Carlo (MC) simulations of SM processes, exclusively on prompt decays of electron-positron collisions. By analyzing the behavior of IForest, including its hyperparameters and the effect of various input features, the study aims to understand its capabilities and limitations in identifying anomalies associated with the SM samples used and the IDMDH model. In addition, alternative techniques, such as averaging over an ensemble of Isolation Forests or iterative retraining, are explored. These techniques are investigated to assess the sensitivity of the IForest in Chapter 4.

In addition, this thesis includes a comparative analysis with Autoencoder (AE) [6], another AD approach, with an IForest model specifically derived from this study. The goal of this analysis is to gain valuable insight into the performance and efficiency of the IForest compared to AE. This comparative study is presented in Chapter 5 and highlights the relative sensitivity of these two AD techniques.

2. The Belle II Experiment and Physics Theory

The Belle II Experiment provides an opportunity to explore Physics Beyond the Standard Model (BSM). With its ability to produce precise Standard Model (SM) decays as output and due to its high luminosity, it provides an ideal platform to explore New Physics (NP). Searching for NP is of great interest, especially for Dark Matter (DM), since astrophysical observations have already suggested its existence. A specific model search for new physical phenomena exceeds the possibilities given by the nature of the unknown signals. Taking a more general approach allows for exploring a wider parameter space and increases the chances of discovering unexpected signs of unknown particles. Therefore, the consideration of a model-independent approach is justified in the search for anomalies specific to the theoretically assumed DM. To begin with, a concise technical introduction to the experiment is necessary for the scope of this work.

2.1. The Belle II Experiment

The Belle II detector operates at the SuperKEKB collider, located at KEK, Tsukuba, Japan. SuperKEKB [7] is an asymmetric electron-positron accelerator that operates at the energy of the $\Upsilon(4S)$ resonance at a center-of-mass energy of $\sqrt{\mathrm{s}} = 10.58\,\mathrm{GeV}$. Therefore it is also called a B-factory. Along the SuperKEKB accelerator, there are four experimental halls, namely Nikko, Fuji, Oho and Tsukuba. Figure 2.1 provides an overview of the accelerator's arrangement, with the Belle II detector located in the Tsukuba hall. The Main Ring (MR) consists of the Low Energy Ring (LER) for the positron beam at 4 GeV and the High Energy Ring (HER) for the electron beam at 7 GeV. Both electrons and positrons are injected into the 3 km long MR from a linear accelerator via beamlines. The collision of the two particles takes place at the Interaction Point (IP), around which the Belle II detector is placed. This particle accelerator, with its design for high luminosity, is advantageous for the generation of large amounts of data from the collision. At present, this is the accelerator with the highest instantaneous luminosity of $4.65 \times 10^34\mathrm{cm}^{-2}\mathrm{s}^{-1}$ in the world [8].

In the 4π-symmetric Belle II detector, several layers of subdetectors are responsible for detecting collision events, as illustrated in Fig. 2.2. The innermost detectors are the Pixel Detectors (PXD) and Silicon Vertex Detectors (SVD), used for particle trajectory detection. Charged particles are tracked by the Central Drift Chamber (CDC) in the Belle II detector

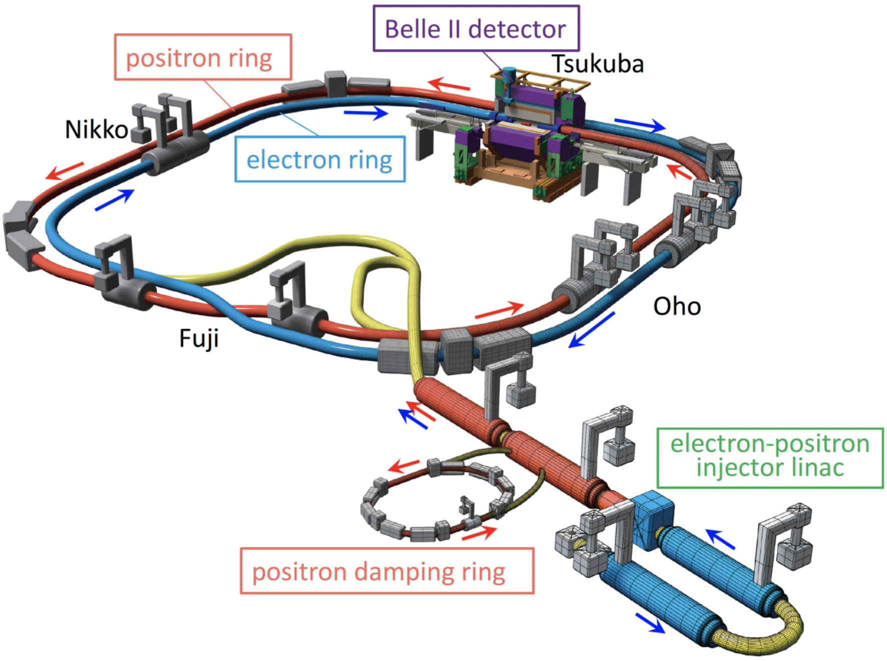


Figure 2.1.: Schematic of SuperKEKB with the Belle II detector at the particle Interaction Point (IP). The electron (positron) beams are marked in blue (red) with an additional linear accelerator for the particles and the positron damping ring [7].

using charged wires and helium-methane gas, providing essential charge and momentum information.

The Time-of-Propagation (TOP) counter measures the precise timing of particle interactions by detecting the emission of photons in a Cherenkov cone as charged particles traverse a radiator material. The Aerogel Ring-imaging Cherenkov (ARICH) measures the angular distribution of Cherenkov photons emitted in the silica aerogel source. Both subdetectors provide Particle Identification (PID) information, focusing on distinguishing kaons from pions throughout most of the momentum spectrum.

The next layer of the Belle II detector is the Electromagnetic Calorimeter (ECL), which is used to measure the energy of particles. For photon detection and electron identification, scintillation crystals with their strong light emission are used to convert absorbed high-energy gamma radiation by the emission of low-energy photons.

The K_L and Muon Detector (KLM) is located outside the superconducting solenoid as the outermost layer of the Belle II detector. Long-lived particles as K_L^0 and μ pass through the previous layers without being stopped due to their weaker interaction strength. It is composed of alternating layers of iron plate absorbers and resistive plate chambers to measure the energy loss of these particles. The superconducting solenoid provides a $1.5\,T$ magnetic field. More detailed information on the technical design of these detectors is given in [10].

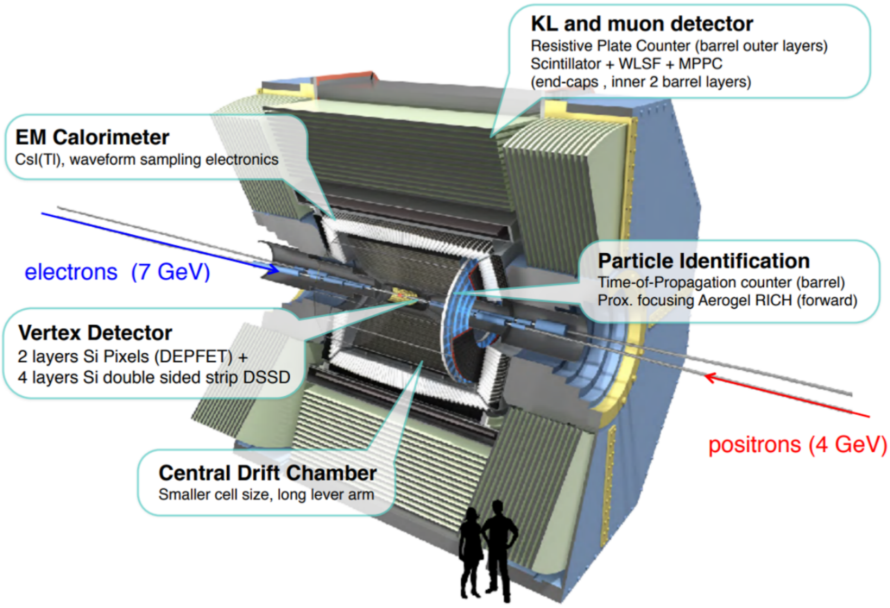

KL and muon detector
Resistive Plate Counter (barrel outer layers)
Scintillator + WLSF + MPPC
(end-caps , inner 2 barrel layers)
EM Calorimeter
CsI(Tl), waveform sampling electronics
electrons (7 GeV)
Particle Identification
Time-of-Propagation counter (barrel)
Prox. focusing Aerogel RICH (forward)
Vertex Detector
2 layers Si Pixels (DEPFET) +
4 layers Si double sided strip DSSD
positrons (4 GeV)
Central Drift Chamber
Smaller cell size, long lever arm

Figure 2.2.: Schematic of the Belle II detector layers. The asymmetric energy of the colliding particles and their direction, as well as the placements of the detectors, are displayed [9].

Through a combination of hardware and software algorithms, triggers evaluate various characteristics of detected particles to determine if an event has to be retained or discarded. The goal of triggers is to efficiently filter and reduce the amount of data while retaining the most relevant events for physics analysis. The exact operation of keeping events associated with triggers is described in [11]. For analysis, the Belle II Analysis Software Framework (`basf2`) [12] is provided by the collaboration. This software can be used to generate Monte Carlo (MC) particles as well as for combining the information from all the detectors introduced above and thus reconstructing the particle events.

2.1.1. Beam Background

The operation of the SuperKEKB accelerator leads to beam background events in addition to the collision process. Consequently, several processes that contribute to the generation of beam background events, studied during the operation of SuperKEKB [13], are presented:

- Touschek backgrounds: Coulomb interaction with particles in the same beam that scatter at an energy different from the nominal bunch energy.
- Beam-gas events: Particles that deviate from the nominal beam path and collide with the wall of the accelerator pipe.
- Synchroton radiation: Radiation in the energy range of several keV generated from emitted photons of moving charged particles.

- Luminosity background: Apart from the collision at the IP, other processes in the Belle II detector lead to an increase of the radiation dose and occupation of the hits in the detector by emitting photons, e.g., the radiative Bhabha scattering $e^+e^- \to e^+e^-\gamma$.
- Injection background: Because of the short lifetime of the beam, interactions between injected bunches to maintain a stable current can lead to the creation of additional particles through beam-beam collisions.

The evaluation of background is crucial due to the high luminosity, as it leads to an increased presence of background events. As a result, these background processes produce additional particles that do not originate from the collision itself but are still detected and, generally, worsen the detectors' performance. With future upgrades of SuperKEKB [14], the beam background becomes more relevant.

2.2. Physics Beyond the Standard Model

In the past, numerous observations supporting the existence of DM have been documented [15]. As a result, theories involving DM are being proposed for phenomena that cannot be explained by the SM. Some known phenomena in the field include the Comsmic Microwave Background (CMB), galaxy rotation, and the challenging strong CP problem. Tighter constraints on the mass of potential DM particles, based on the theoretical model presented in [16], prohibit them from being heavier than the gauge boson A' and thus tend to limit the DM to a lower mass spectrum.

An inelastic coupling to SM particles that can exist independently of the CMB bounds is very intriguing. This model is described in more detail in [16] and is called Inelastic Dark Matter with a Dark Higgs (IDMDH). In particular, the idea of introducing two DM particles, as well as bosonic particles interacting with the DM particles, is analogous to existing particles in the SM. In the context of the IDMDH model, a SM photon resulting from electron-positron collisions can kinetically mix with an induced Dark Photon A'. The lightest DM particle is introduced as χ_1 and can be excited to χ_2 by a Dark Photon. Similarly to the SM where the higgs particle gives mass to other particles, the dark particles get their mass from a dark Higgs h'. The process is illustrated in a simplified model shown in Fig. 2.3. In this particular case, as well as in other possible scenarios involving different final state particles, an SM photon mixes with a Dark Photon and emits a dark Higgs h' that decays into a muon-antimuon pair. The Dark Photon, in turn, decays into two DM particles, χ_1 and χ_2. The heavier χ_2 then decays into a χ_1 and a Dark Photon, which decays into an electron-positron pair. So in total there are four leptons (e^+, e^-), (μ^+, μ^-) and the missing energy. This results in seven free parameters [16] with:

- The mass of the A', $m_{A'}$
- The mixing angle of the SM photon to the A', ϵ
- The mass of the h', $m_{h'}$
- The mixing angle of the SM Higgs to the h', θ
- The mass of the χ_1, m_{χ_1}

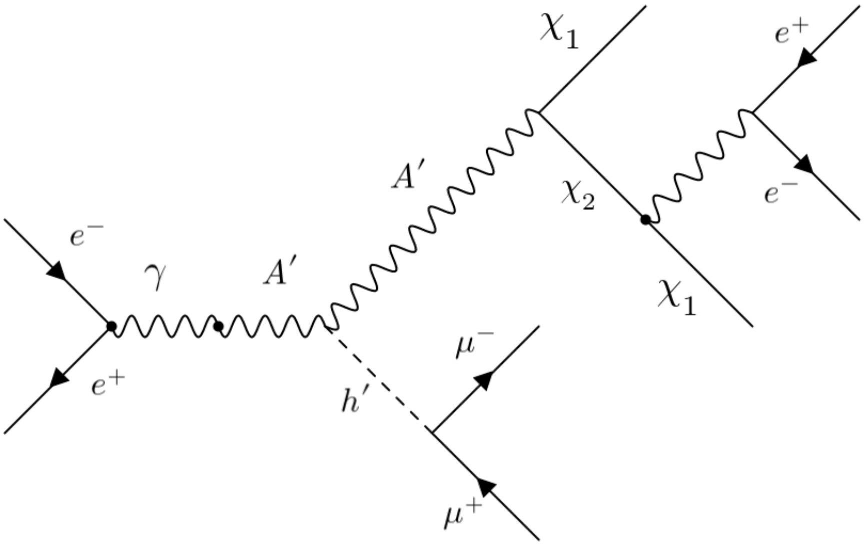


Figure 2.3.: Feynman diagram of the simplified IDMDH model adapted from [6].

- The coupling of the χ_1 and χ_2 to the h', f
- The coupling of the χ_1 and χ_2 to the A', g_X

The two dark χ_1 are expected to be undetectable, so there is a discrepancy with the accelerator energy when the remaining ECL clusters are summed up and missing energy is left over. In this work, based on the calculations of the master's thesis [6], the mass of the Dark Photon is set to $m_{A'} = 4 \cdot m_{h'}$ and the following hypotheses are assumed. The couplings are required to be less than $\sqrt{4\pi}$. Therefore, the h' must not be heavier than the Dark Photon A'

$$m_{h'}^2 \lesssim \frac{\sqrt{\pi}}{4g_X} m_{A'}^2. \tag{2.1}$$

With additional constraints for the masses due to the DM annihilation and CMB

$$\frac{f^4}{64\pi^2} m_{\chi_1} < m_{h'} \lesssim m_{\chi_1} < m_{A'}. \tag{2.2}$$

This results in a theoretical mass for the χ_2 of

$$m_{\chi_2} = m_{\chi_1} + \frac{f \cdot m_{A'}}{g_X} \tag{2.3}$$

In addition, the coupling constants are fixed to

$$f = \sqrt{4\pi\alpha_f} \approx 0.2476 \tag{2.4}$$

and

$$g_X = \sqrt{4\pi\alpha_D} \approx 1.12 \tag{2.5}$$

with $\alpha_f = 0.006$ and $\alpha_D = 0.1$. This gives the relation between masses $\chi_1$and χ_2

$$m_{\chi_1} + m_{\chi_2} + m_{h'} < 10.58\,\mathrm{GeV}. \tag{2.6}$$

The accelerator energy of 10.58 GeV results in:

$$\frac{f}{g_X} \cdot 4 \cdot m_{\chi_1} \approx m_{\chi_1} \tag{2.7}$$

additional limitation from [16] results in

$$m_{\chi_1} > m_{h'}\,, \tag{2.8}$$

and

$$\Delta m = m_{\chi_2} - m_{\chi_1} \geq 2 \cdot m_\mu. \tag{2.9}$$

The focus in this work is on the prompt decays of h' and χ_2. For this reason, the mixing angles ϵ and θ are set to a high value of 10^{-2}, controlling the lifetime of the Dark Photon A'.

2.2.1. Standard Model

The signature of the events studied in this work is characterized by four charged particles in the detector acceptance: a muon-antimuon pair resulting from the decay of the h' particle, and an electron-positron pair resulting from the decay of the χ_2 particle. Also, missing energy corresponds to the energy carried away by the χ_1 particle, which is not detected. A variety of known processes within the SM produce a similar signature. Therefore, such processes are detailed in this chapter.

2.2.2. Possible Collider Processes

The process $e^+e^- \to e^+e^-\mu^+\mu^-$ involves exactly the same particles that are expected in the final state. Therefore, the missing energy could be derived from reconstruction and measurement errors. In addition, tauon decays e.g. $\tau^- \to e^- + \bar{\nu}_e + \nu_\tau$ with undetectable neutrinos may mimic the final state of the dark process. Moreover, combinations of single-lepton-pair processes and the beam background, as well as other measuring and reconstructing errors, can imitate the final states that are sought. Furthermore, the similarity of the masses of the muons and pions implies a contribution from the $e^+e^- \to e^+e^-\pi^+\pi^-$ process. Due to the high production rate of the $B^\pm \to l^\pm\nu_l X$ or $B^0 \to K^0 l^+ l^-$ decays of the B factory accelerator, a background results from reconstruction errors. In addition to the production of b-quark and anti-b-quark pairs, the continuum background includes non-resonant decays $e^+e^- \to u\bar{u}$, $e^+e^- \to d\bar{d}$, $e^+e^- \to s\bar{s}$, and $e^+e^- \to c\bar{c}$, which contribute to the overall background. Additionally, within the context of the SM, there are various processes that exhibit high cross-sections. These processes involve the production of commonly observed particles, such as electrons, muons, quarks, and gauge bosons. As described in [6], the use of machine learning is motivated by the resulting large amount of data.

Table 2.1.: Summary of simulated process-based MC samples adapted from [6] with their corresponding luminosity and number of events.

process	simulated luminosity in fb^{-1}	number of events ($\cdot 10^6$)
$e^+e^- \to e^+e^-\mu^+\mu^-$	100	1883
$e^+e^- \to \tau^+\tau^-$	100	91.9
$e^+e^- \to e^+e^-\pi^+\pi^-$	100	189.5
$e^+e^- \to e^+e^-e^+e^-$	100	3955
$e^+e^- \to \mu^+\mu^-$	100	114.8
$e^+e^- \to e^+e^-$	10	2958
$e^+e^- \to B_0\bar{B}_0$	100	54
$e^+e^- \to B^+B^-$	100	51
$e^+e^- \to u\bar{u}$	100	160.5
$e^+e^- \to d\bar{d}$	100	40.1
$e^+e^- \to s\bar{s}$	100	38.3
$e^+e^- \to c\bar{c}$	100	132.9
$e^+e^- \to \mu^+\mu^-\mu^+\mu^-$	2000	0.35120
$e^+e^- \to K^0\overline{K}^0(\gamma)$	1000	0.886400

2.3. MC Simulations and Reconstruction

MC simulated samples are used for the comprehensive study of the Isolation Forest (IForest) in this thesis. These samples adapted from [6], correspond to each of the SM processes and are summarized in Table 2.1. Samples with a simulated luminosity other than $100 fb^{-1}$ are reweighted later. The dark Higgs signals were simulated for a few possible configurations for 25000 events adapted from [6]. All model parameters which are used for the simulation are listed in Tab. 2.2. Separate simulations are done for beam background events which are then overlayed in the event simulation process. In the further discussion, three exemplary signal samples are used to represent the three extreme cases:

Table 2.2.: Summary of the model parameters with their corresponding values, as adapted from [6].

model parameter	values
m_{χ_1} in $\mathrm{GeV}\,c^{-2}$	[0.25, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0]
$m_{h'}$ in $\mathrm{GeV}\,c^{-2}$	[0.25, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0]
$m_{A'}$ in $\mathrm{GeV}\,c^{-2}$	$4 \cdot m_{\chi_1}$
f	2.746×10^{-1}
g_X	1.12

- small masses DM: $m_{h'} = 0.5\,\mathrm{GeV}\,c^{-2}$, $m_{\chi_1} = 0.5\,\mathrm{GeV}\,c^{-2}$
- large masses DM: $m_{h'} = 2.5\,\mathrm{GeV}\,c^{-2}$, $m_{\chi_1} = 2.5\,\mathrm{GeV}\,c^{-2}$
- high mass splitting: $m_{h'} = 0.5\,\mathrm{GeV}\,c^{-2}$, $m_{\chi_1} = 3\,\mathrm{GeV}\,c^{-2}$

The reconstruction conditions for the particles taken from [6] are as follows:

The Final State Particle (FSP) were restricted in the following aspects for the reconstruction process:

- two pairs of opposite-charged tracks
- tracks originating from the IP
- tracks in the CDC acceptance of $17^\circ < \theta < 150^\circ$

To reduce the number of misidentified particles, the PID is used. It provides a probability of a detected particle being assigned to a specific particle type. In this regard, a variable known as binary PID is introduced

$$\mathrm{PID}(e, \mu) = \frac{\mathcal{L}_e}{\mathcal{L}_e + \mathcal{L}_\mu}, \tag{2.10}$$

where the value $\mathcal{L}_\ell$ represents the probability assigned to a specific particle ℓ. Thus, an electron (muon) is assumed if the PID value is 1 (0). Based on this assumption, a binary PID(e, μ) greater than 0.1 is assigned to electrons and smaller than 0.9 to muons. Intermediate particles h' and χ_2 are reconstructed by combining opposite-charged muons and electrons. In the case where both particles originate from the same parent particle, their tracks can be extrapolated to identify the decay vertex, allowing the selection of intermediate particles based on a vertex fit result. This results in the following vertex conditions for both particle pairs:

- decay vertex must originate from the IP
- all candidates with failed fits are rejected
- at least one of the vertex fits must fulfill $\chi_{prob} > 0.01$, where χ_{prob} is the p-value of the vertex fit

Further conditions for the reconstruction, referred to as *weakly selection samples* are:

- events with more tracks that fulfill the requirement are discarded
- π^0*veto*: The emission of photons in collision processes, such as $\pi^0 \to \gamma\gamma$, is considered due to the absence of photons in the signal state. This process occurs in tauon decays, specifically $\tau \to \pi + \pi^0 + \nu_\tau$. The selection conditions for photons are as follows:
  - The number of cluster hits in the ECL is greater than 1.5.
  - The cluster is located between 17° and 150° in the ECL
  - The reconstructed energy is smaller than 0.25 GeV
  - The absolute time difference between the collisions and measurement of the photon in the ECL must be smaller 200 ns
- The rest of the event of all remaining ECL clusters must be greater than 0.05 GeV; otherwise, they are discarded.

Additional selection on the samples are performed in this work referred to as *stricter selection samples.* A short overview of the selections is provided below:

- selection on the missing energy: Candidates with unphysical missing energy with $\mathrm{E_{miss}} < 0$ and $\mathrm{E_{miss}} > 10.58\,\mathrm{GeV}$ are removed
- vetoing π^0: events with $0 \text{ Gev}\,\mathrm{c}^{-2} < m_{\gamma\gamma} < 0.3\,\text{Gev}\,\mathrm{c}^{-2}$ are excluded

Detailed selection specifications can be found in [6].

3. Machine Learning and Anomaly Detection

In the field of Machine Learning (ML), Anomaly Detection (AD) refers to the process of identifying anomalous instances within a data set [17]. There are a number of definitions for the description of an anomaly [18], which are generally rare and few in number. Outliers are defined as distinct deviations from the normal behavior of a data distribution, providing a way to characterize and identify exceptional observations. This definition refers to the distance of the data points from most of the data set. Another definition is over-densities, which refers to regions in the data that show more events than expected. To measure abnormality, an Anomaly Score (AS) is established that provides an estimate of the anomaly's magnitude and quantifies the deviation of a data point or pattern from expected normality. A general definition of the score is not present and must be specified for each problem independently.

In the search for New Physics (NP), one tries to find unknown signals which deviate from known SM processes. To this end, the High Energy Physics (HEP) community has conducted research on anomaly detection methods through several challenges, such as the LHC Olympics 2020 [19], where different methods were developed using simulated collider events from two jets resulting from strong interaction. The goal is to test different approaches' potential and advance the search for new physics with anomaly detection algorithms. The Dark Machine Challenge [20] also includes studies of simulated proton-proton collisions for the Large Hadron Collider. The models are trained on pure SM samples, allowing the algorithms to learn the properties of the SM background. Many different methods, such as the presented autoencoder in [20], are considered and compared. These challenges are intended to drive the development of anomaly detection algorithms and to stimulate research into new physics.

In contrast to supervised learning algorithms, which require labeled training data, unsupervised algorithms are able to use the inherent patterns and structures in the data to detect deviations from normality [21]. This approach allows anomaly detection based solely on the intrinsic properties of the data set without the need for explicit classification. The tree-based unsupervised method known as Isolation Forest (IForest) offers a straightforward approach to direct anomaly detection. In this method, the sample is partitioned based on randomly chosen partition values, allowing for the identification of anomalies.

In order to exploit these advantages of the efficient algorithm, the IForest is discussed in this chapter.

3.1. Isolation Forest

Isolation Forests, an anomaly detection algorithm, uses binary trees as its basic structure. Binary trees allow for an efficient and iterative evaluation of the data by splitting instances into two child nodes based on decision values. Creating the IForest from unlabeled data involves recursively dividing subsamples from the data set, as illustrated in Fig. 3.1. Outliers are promptly identified as they exhibit a short path length in the tree, requiring fewer divisions to isolate them. This process is repeated multiple times to create a forest of trees. The average path length $E(h(x))$ of an instance x is then used to calculate the anomaly score by Eq.3.1,

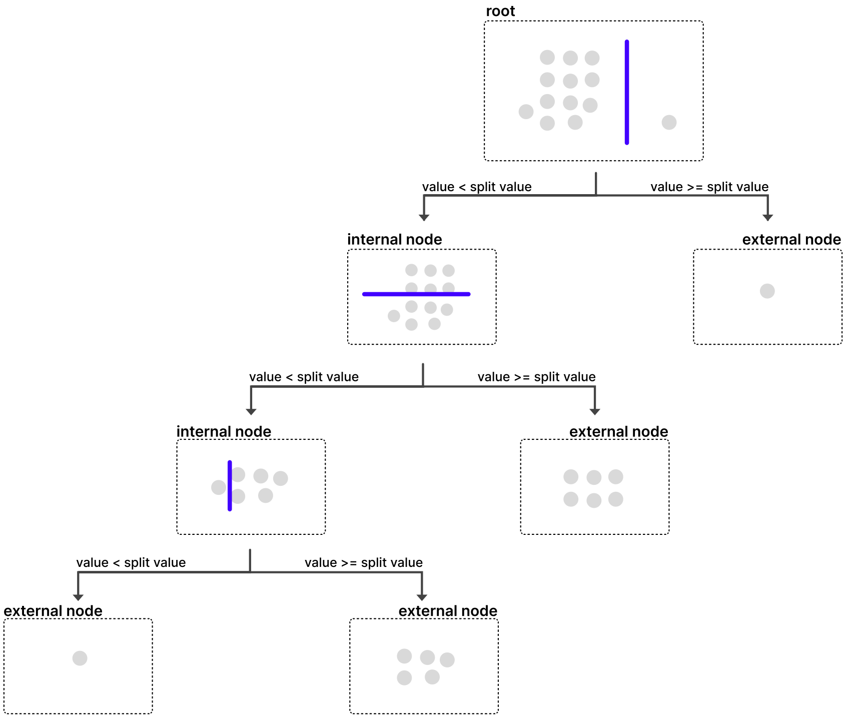


Figure 3.1.: Exemplary representation of the Binary Tree (BT) of the IForest. Simplified distribution starts from the root node and is partitioned into two daughter nodes by a random split value. External nodes represent the end nodes.

$$\mathrm{s(x,n)} = 2^{-\frac{\mathrm{E(h(x))}}{\mathrm{c(n)}}} \tag{3.1}$$

with c(n), where n represents the number of instances in the data set

$$\mathrm{c(n)} = 2\mathrm{H(n-1)} - (2(\mathrm{n}-1)/\mathrm{n}). \tag{3.2}$$

Serving as a normalization factor, with H(i) representing the harmonic number, derived from the average path length of terminating paths in external nodes, defined as an unsuccessful search in a binary search tree [4]. This value indicates the extent to which an instance is an outlier in the range between 0 and 1, where outlier values are closer to 1 and normal instances closer to 0. The growth of the binary trees terminates when it satisfies one of three conditions:

- only one instance in the external node
- instances in the external node having the same value
- the tree exceeds a height limit.

The model's training and testing stages are shown in Fig. 3.2, which represents the construction of the various binary trees in the training and testing stage, with the test data being passed to the trained forest by obtaining the path length for each instance. The height of the binary tree is limited by the condition $\log_2(k)$, where k represents the subsample size. This limitation is motivated in [4] to enhance the algorithm's efficiency by prioritizing shorter path lengths, which are indicative of anomalies. Subsequently, the average path length of the dataset instances is used to determine the anomaly score by Eq. 3.1. When using the IForest algorithm, it is important to consider the influence of swamping and masking effects, as their contribution is determined by the size of the subsample used for training. Swamping occurs when anomalies are misclassified as normal instances because they share characteristics with the majority of the data. Masking refers to the phenomenon of anomalies being detected as normal since they are surrounded by a dense region of normal data.

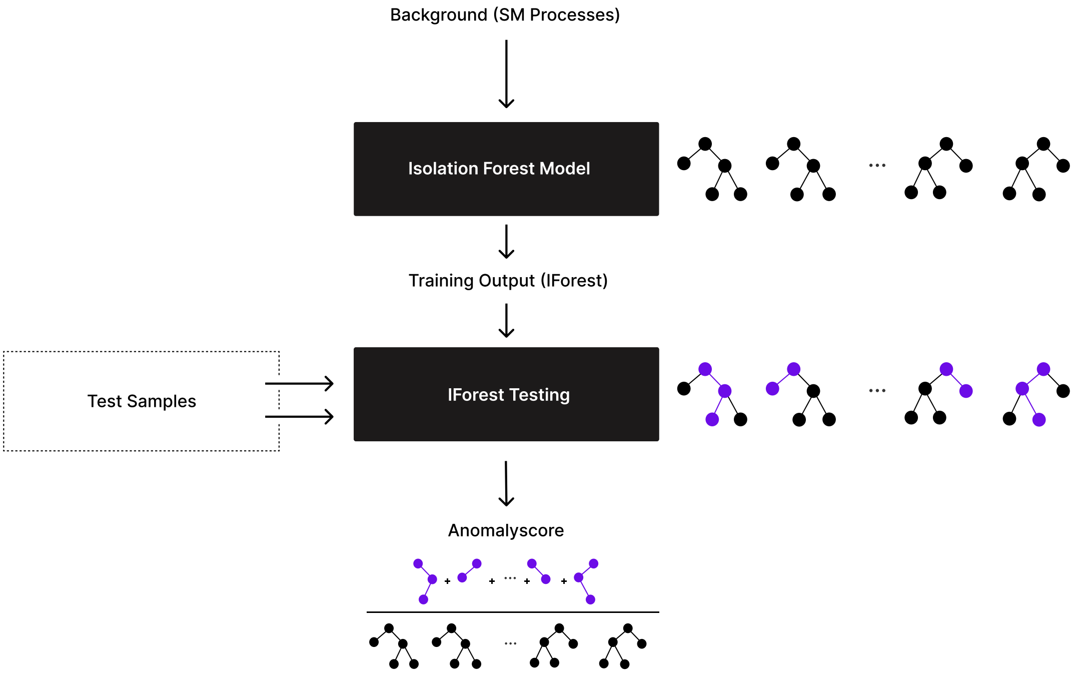


Figure 3.2.: The training phase represents the construction of binary trees on the training sample, and the test phase is the traversal of the test sample, resulting in an anomaly score corresponding to the average path length of the entire IForest.

3.1.1. Optimization and Challenges

Hyperparameter selection is a typical challenge for machine learning algorithms as well as for the IForest. By modifying the number of trees ($n\_estimators$), the chosen subsample size (*max_ samples*), and the number of randomly chosen input features (*max_features*), the IForest will show different outcomes after training. The parameters that can be changed depending on the forest are shown in Fig. 3.3. According to the original paper [4], the forest

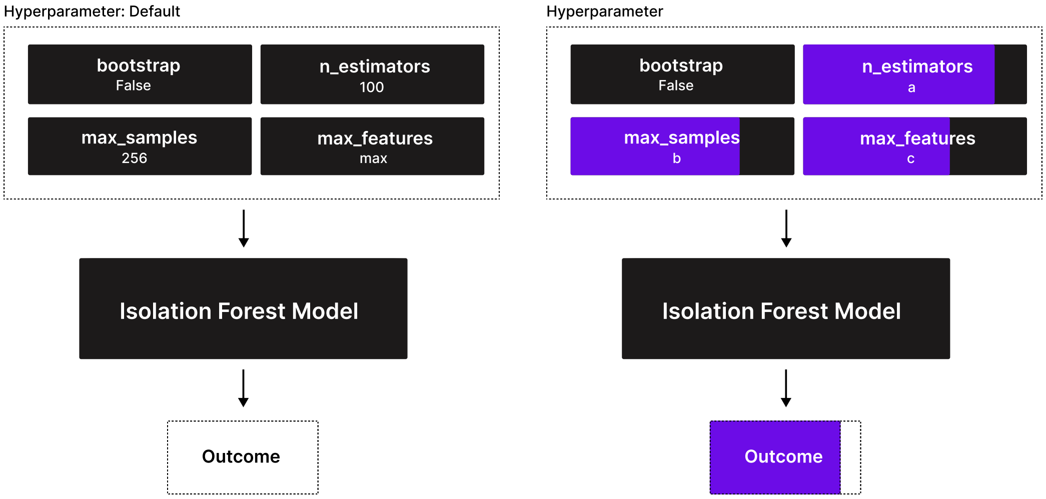


Figure 3.3.: Representation of the hyperparameters of the IForest algorithm. As usual in machine learning models, the hyperparameters are optimized to achieve improved results. The accompanying visualization represents the impact of various hyperparameters, denoted as values a, b, and c, on the performance of the model.

is set to its default values. However, optimizing these samples by changing the settings in the search for the Inelastic Dark Matter with a Dark Higgs (IDMDH) model is possible. The optimization can directly improve the performance of the IForest by more accurately categorizing the signals based on the trained SM background. With a small number of trees, a large dataset is trained on only a very small subset of it, which increases the underfitting of the IForest. This can be reduced by averaging over a larger number of trees since a larger data set is represented in the training. With a high-dimensional data set, it is essential to consider the number of random features for splitting within the binary tree. Relying on a subset of features does not adequately capture the distribution of the samples. To take the training data into account, a random subsample is selected for each tree repeatedly, which means that the *bootstrap* parameter must be set to true. The objective is to train a model on the background of Standard Model (SM) samples that can effectively isolate events as anomalies while reducing the problems of overfitting and underfitting. This becomes particularly relevant when considering different hyperparameter configurations.

3.2. Extended Isolation Forest

Based on the similar principle of a random selection of features and split values, the Extended Isolation Forest [22] is a different variant of the Standard Isolation Forest. Their difference is in the way they apply those selections to the distribution shown in 3.4. The selection criteria for the Standard Isolation Forest can be either horizontal or vertical along the selected features. This leads to a bias in isolating outliers, provided in the anomaly score

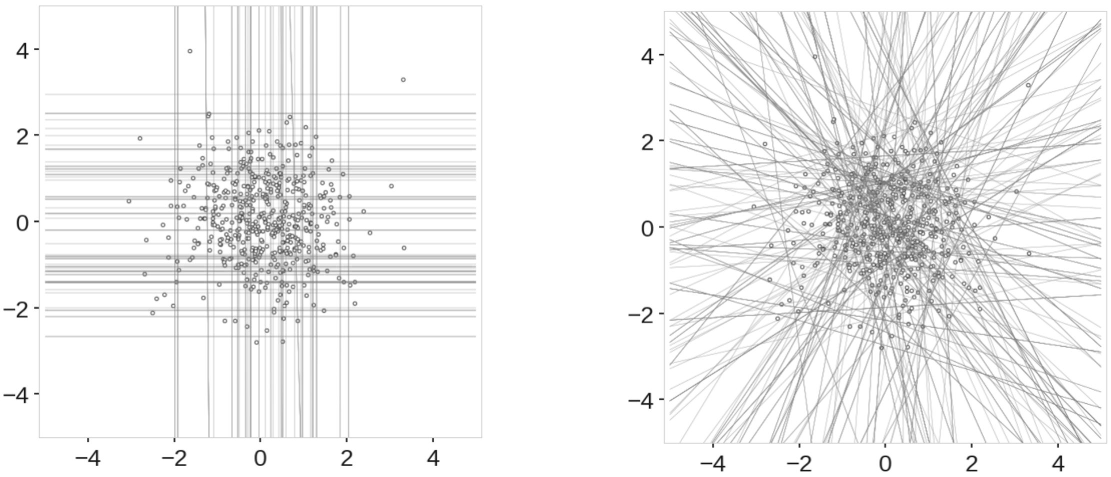


Figure 3.4.: Selection Criteria for the different IForest Methods: Standard Isolation Forest (left) and Extended Isolation Forest (right). Taken from [22].

map in Fig. 3.5 for the anomaly score. Because of the horizontal and vertical selections of the distribution, compared to selections with a random location in the distribution, some areas indicate a lower anomaly score for points that are out of the distribution. This problem is solved with the Extended Isolation Forest, where the anomaly score bias in the vertical and horizontal directions is resolved. The principle of the Extended Isolation Forest is based on the fact that the selections are applied with a randomly chosen slope in the space. As with the Standard Isolation Forest, the anomalous data points will stand out due to the small number of separations. For a high-dimensional data set, the selections are not straight lines. Instead, they are an N-1 dimensional Hyperplane of the N-dimensional feature space.

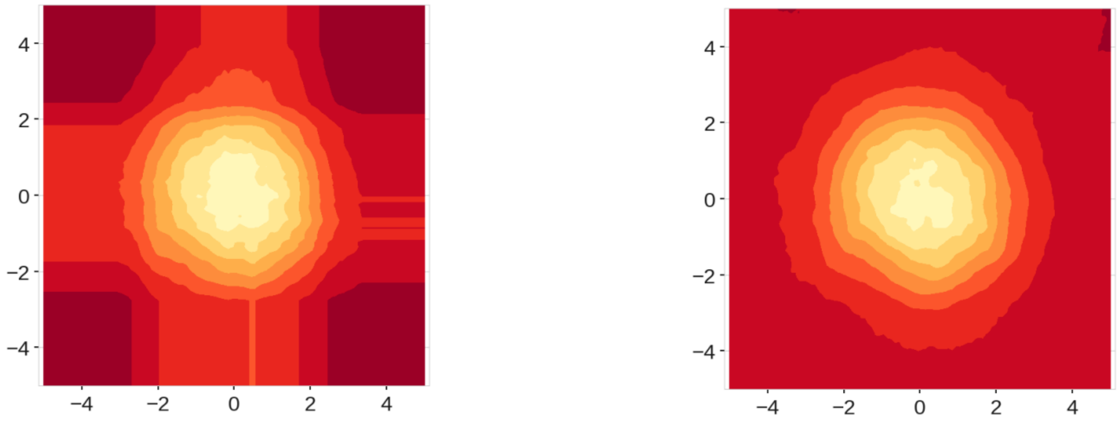


Figure 3.5.: Anomaly Score Map of the Standard Isolation Forest (left) and Extended Isolation Forest (right) performed on normally distributed clusters. Taken from [22].

These are constructed with a randomly chosen normal vector $\vec{n}$ and a randomly chosen intercept point p for any data $\vec{x}$: $(\vec{x} - \vec{p}) \cdot \vec{n} \leq 0$. Multiple levels of extension are used for high-dimensional data because of the N-1 dimensional hyperplanes. With hyperplanes that can cross any of the coordinate axes available on the fully extended level. The lowest level of extension is represented by the Standard Isolation Forest.

4. Isolation Forest Training Analysis

This work uses the *scikit-learn* [1] implementation of the Isolation Forest (IForest) algorithm for training on Standard Model (SM) samples, introduced in section 2.3. Bootstrapping is enabled during training to ensure a broader representation in the construction of binary trees and to avoid under-sampling on a subsample. Since hyperparameters are values that modify the training of IForest, hyperparameter optimization is performed to obtain a matched configuration for categorizing the SM and signal samples. By considering input features such as particle kinematics, Particle Identification (PID), and energies, it becomes possible to investigate their contribution to the distinction between background and signal peaks. The entire background and signal samples are used to compute the Anomaly Score (AS) resulting from passing through instances of the given sample. To evaluate the performance of the IForest, the Punzi Figure of Merit (PFOM) as a metric to assess sensitivity is used. Furthermore, the influence of different features and the individual contribution of SM processes are investigated.

Finally, the optimized IForest is compared to the autoencoder approach in Chapter 5.

4.1. Plain Isolation Forest

The default settings provided in Table 4.1 are initially utilized to configure and evaluate the IForest model. These default settings are based on the implementation provided by the *scikit-learn* library [1]. A total of 20 input features are used, with the choice of

- Four-vectors
- Missing momentum
- Missing energy

being provided by the events of the Final State Particle (FSP). A second set of features provides additional information about the detected particles. This increase is summarized in Tab. 4.2 .

Table 4.1.: Default hyperparameter settings are used for IForest training, except for the bootstrap value.

n_estimators	max_samples	max_features	bootstrap
100	256	all	True

Table 4.2.: Summary of the two input feature sets utilized in the training and testing stages of the IForest.

Input Features	Count
Four-vector	16
Missing momentum	3
Missing energy	1
Transverse momentum	4
binaryPID(e, μ)	4
ϕ	4
θ	4
$\sum$	36

Input Features	Count
Four-vector	16
Missing momentum	3
Missing energy	1
$\sum$	20

Punzi Figure of Merit

To evaluate the performance of the IForest with different modifications made, the PFOM is used [23]:

$$\text{PFOM} = \frac{\epsilon}{\frac{a}{2} + \sqrt{B}} \tag{4.1}$$

with the signal efficiency $\epsilon = \text{N}_{\text{after selection}}/\text{N}_{\text{produced events}}$, where N is the count of events and B the number of remaining background events after the selection. The signal efficiency is an indication of the ability to accurately detect signal events, while the background count is related to the presence of false positives. The significance level described in terms of relating to one-sided Gaussian tests at a given significance is denoted by the parameter a. In this case, its chosen value is $a = 1$ in order to be consistent with the autoencoder study in [6].

Evaluation of Isolation Forest

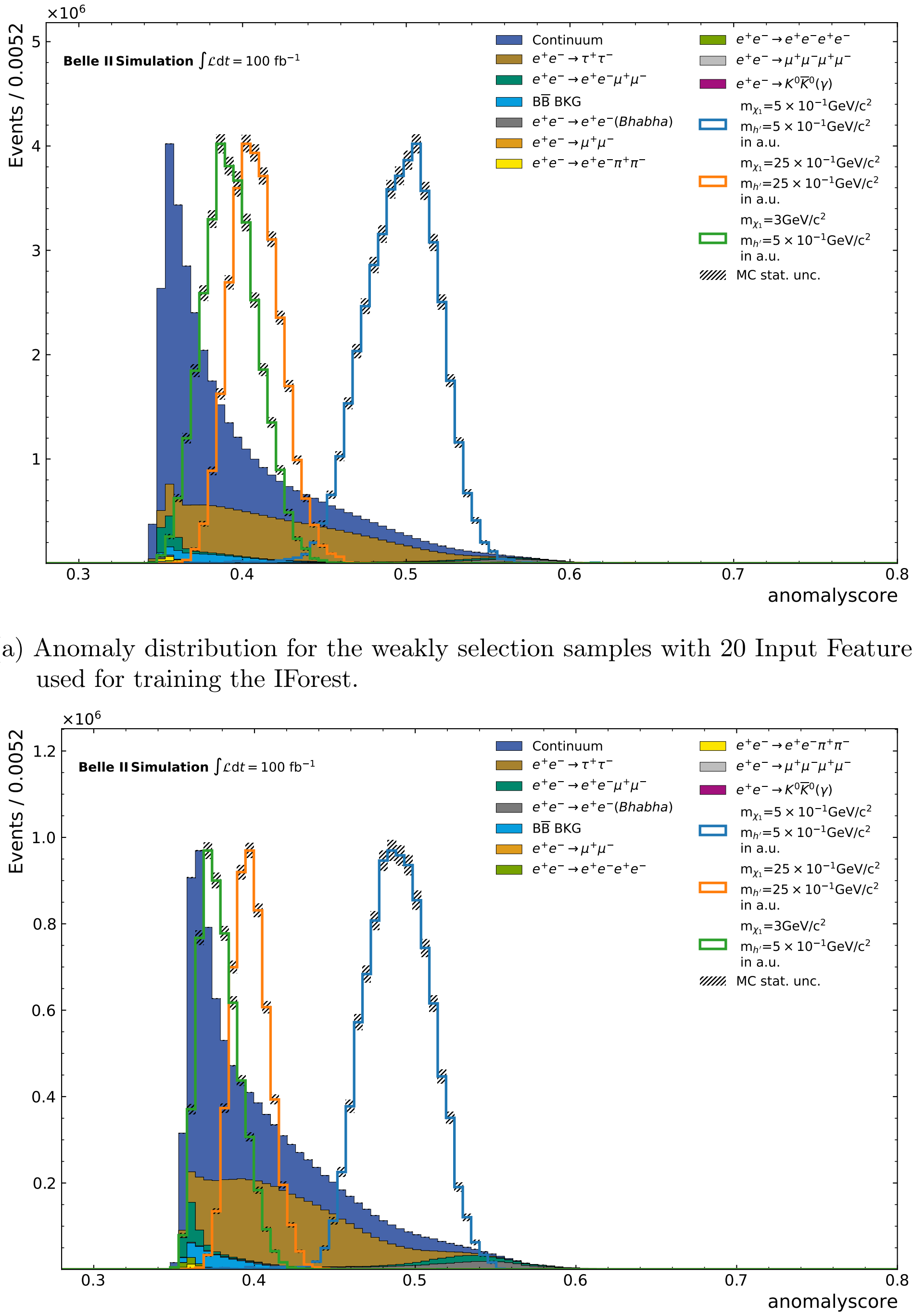
In the evaluation of IForest, the distribution of the anomaly score is examined. Therefore, the deviation of the anomaly scores between the events observed in the SM samples and the example signals described in Section 2.3 is investigated. For that, only the SM samples are trained, and then both these and the signal events are passed through the forest. In Fig. 4.1a, the background distribution of SM samples is depicted using filled bins for each process. The background events in the distribution are weighted by the simulated luminosity to an integrated luminosity of 100 fb^{-1}. The resulting *anomalyscore* is a dimensionless quantity whose distribution is given by the number of events in each bin. The signals are also represented for the different model parameters as an outline of the distribution in an arbitrary unit. A distinction arises in the calculation of background and signal events, evident from the peaks that correspond to a higher density of these events for an anomaly score. Thus, the analysis reveals that the high-density SM sample region exhibits a lower anomaly score compared to the three signals. A difference in scoring is also apparent for these signals in particular. While the peaks of the heavy and high mass difference signals show a less distinct detection, the signal for the light mass shows a much higher value shift in the anomaly score compared to the high-density background events. In the case of the stricter selection samples used for training (Fig. 4.1b), the detection performance for the heavy mass and large mass difference signals is worse. This can be attributed to significant overlap with the background distribution, making it more challenging to categorize them as anomalies.

In order to estimate the performance between the two results of the IForest with different training samples, the PFOM (Eq.4.1) is calculated. Additionally, the signal efficiency is determined by quantifying the number of events that remain after applying selections to the sample discussed in Section 2.3. The selection criteria for the anomaly score are set at the 1% and 99% percentile, resulting in 36 equally distributed steps. Based on the simulated luminosity, the background is weighted. Fig. 4.2 illustrates the PFOM values for the weakly and stricter selection samples. The differences are summarized in Tab. 4.3 for the maximal PFOM with the relative difference between these samples in percentage. The light mass signals have the highest PFOM for the different model parameters. This means that the signal is most sensitive to the light mass, as expected from the distribution of the anomaly scores. The maximum PFOM also confirms the lower separation between the background and the low-sensitivity signals. Thus, comparing the two low and high-sensitivity samples shows that an improvement in sensitivity is possible.

Table 4.3.: Summary of the maximal PFOM for the two IForests trained on weakly selection samples and stricter selection samples. The relative difference in the selection sensitivity (max. PFROM) of the stricter with respect to the weakly selection samples is given in percentage.

	$\mathrm{PFOM}_{\mathrm{max,\,weak}}$	$\mathrm{PFOM}_{\mathrm{max,\,strict}}$	relative difference
$\mathrm{m}_{\chi_1} = 5 \times 10^{-1}\mathrm{GeV/c^2}$ $\mathrm{m}_{\mathrm{h}'} = 5 \times 10^{-1}\mathrm{GeV/c^2}$	0.000427	0.000758	72.8%
$\mathrm{m}_{\chi_1} = 25 \times 10^{-1}\mathrm{GeV/c^2}$ $\mathrm{m}_{\mathrm{h}'} = 25 \times 10^{-1}\mathrm{GeV/c^2}$	0.000221	0.000375	69.7%
$\mathrm{m}_{\chi_1} = 3\mathrm{GeV/c^2}$ $\mathrm{m}_{\mathrm{h}'} = 5 \times 10^{-1}\mathrm{GeV/c^2}$	0.000189	0.000313	65.6%

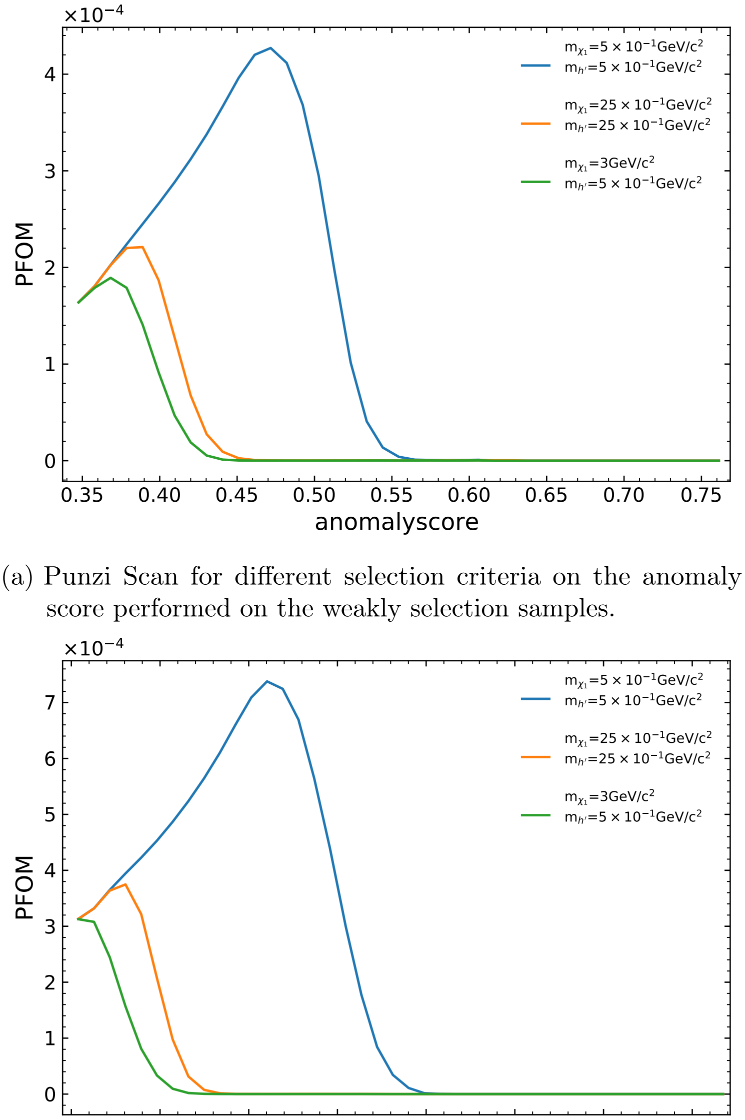
The distribution of the anomaly score for additional input features with information on particle identification, transverse momentum, and angular coordinates (ϕ, θ) used to describe the direction of the emitted particles is shown in Appendix. A.1. A widening of the sharp background peak is observed. The signals of the heavy and large mass differences overlap even more with the background, making them indistinct from the background. Comparing the maximal PFOM between the respective sample restriction types and considering the contribution of additional input features results in an improvement of the light mass signal in the weakly selected samples. However, the sensitivity is reduced in the more strictly selected samples. In addition, the categorization of the two weakly sensitive signals leads to a more "normal" categorization in terms of the anomaly score compared to the majority of the background. This does not contribute to a favorable detection of Dark Higgs signals.



(a) Anomaly distribution for the weakly selection samples with 20 Input Features used for training the IForest.

(b) Anomaly distribution for the stricter selection samples with 20 Input Features used for training the IForest.

Figure 4.1.: The IForest algorithm performed on the weakly and stricter selection samples in a direct comparison of the anomaly score distribution. Showing the effects on the Dark Higgs signals with higher anomaly scores compared to the peak of the SM sample background.



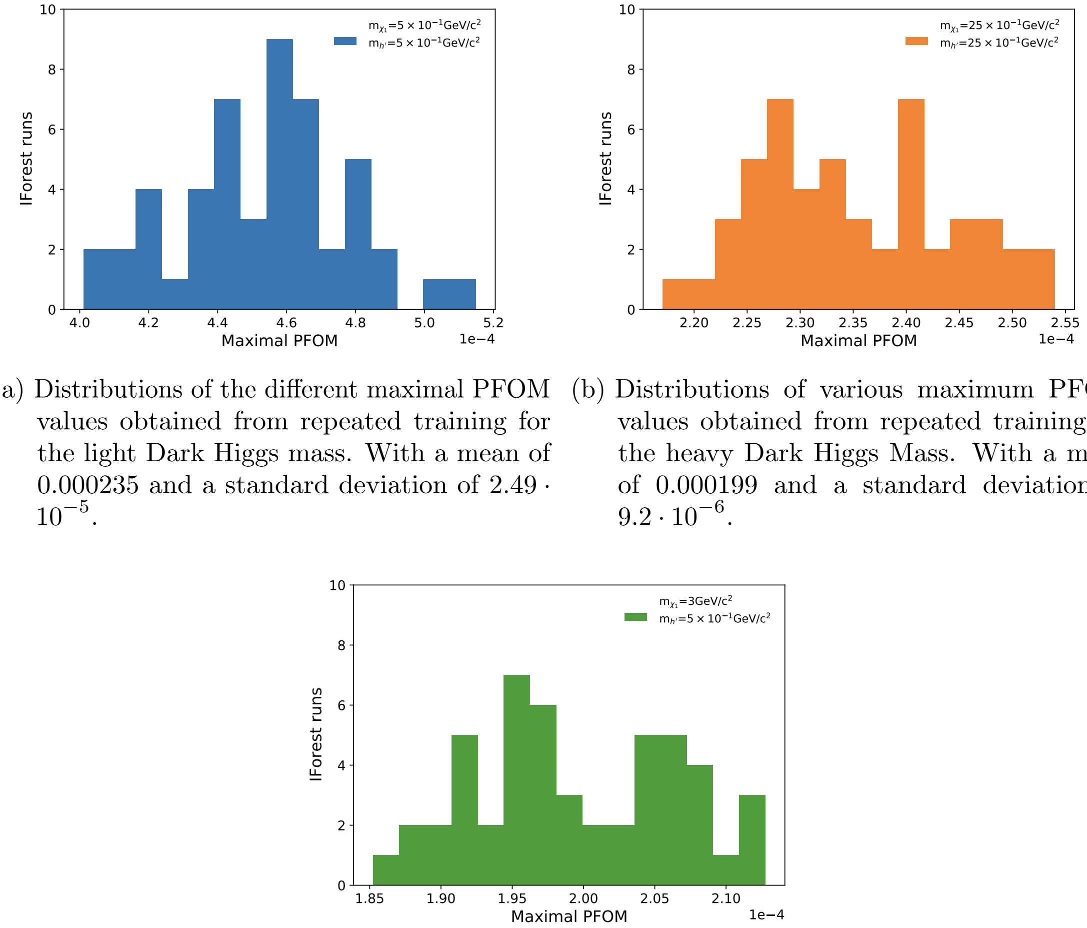
(a) Punzi Scan for different selection criteria on the anomaly score performed on the weakly selection samples.

(b) A Punzi Scan performed on the stricter selection samples, considering different selection criteria based on the anomaly score.

Figure 4.2.: The PFOM for different selection criteria applied to the anomaly score. In which the separation between background and signal is evaluated for the weakly (a) and stricter (b) selection samples.

Statistical fluctuations of training

The randomness involved in training the IForest results in varying outcomes for the same forest. This occurs because the entire forest observes a different distribution of the high-dimensional sample when the subsample is randomly selected. Therefore, an examination is conducted to assess the magnitude of fluctuations when training the IForest multiple times. For this purpose, the same IForest is trained 50 times using default values.



(a) Distributions of the different maximal PFOM values obtained from repeated training for the light Dark Higgs mass. With a mean of 0.000235 and a standard deviation of $2.49 \cdot 10^{-5}$.

(b) Distributions of various maximum PFOM values obtained from repeated training for the heavy Dark Higgs Mass. With a mean of 0.000199 and a standard deviation of $9.2 \cdot 10^{-6}$.

(c) Maximal PFOM for the strong splitting between Dark Higgs and Dark χ_1 Masses. With a mean of 0.000453 and a standard deviation of $2.49 \cdot 10^{-6}$.

Figure 4.3.: Fluctuations of the maximal PFOM for exemplary chosen Signals trained 50 times on the same IForest. The distribution of the maximal PFOM shows the unstable sensitivity, revealing a greater fluctuation of the same IForest.

In Fig. 4.3, the distributions of the maximal PFOM for the different signals

- light Dark Higgs Signal (a)

- heavy Dark Higgs Signal (b)
- strong splitting between Dark Higgs and Dark χ_1 masses (c)

are obtained after 50 runs of the same IForest. For light masses in Fig. 4.3a, a stronger variation of the values is given, while for heavy masses in Fig. 4.3b and mass differences in Fig. 4.3c, a lower scattering is present. This represents the dispersion of the sensitivity deviations between the runs. Accordingly, the light mass signal has a maximum deviation of 30% between minimum and maximum values in this distribution. For the heavy mass signal, the largest deviating values of the maximal PFOM account for 1.4%, which increases to 13.5% for large mass differences. This indicates that two of the tested signals have a high spread of results in terms of sensitivity. While the remaining background count in Appendix A.2 shows similarities to the maximum PFOM, the distribution of signal efficiencies in Appendix A.3 appears to be evenly distributed. Consequently, the fluctuations in the results reveal the instability of the IForest. Although these variations are inherent to randomness, this investigation demonstrates a significant variance in the results. Therefore, it is important to minimize these fluctuations. A simple suggestion is to increase the number of trees in the IForest so that more subsamples are used for training across multiple trees, thus representing a larger fraction of the input sample. But as the computation time is highly dependent on the number of trees, increasing this parameter significantly increases the overall computation time. However, this behavior is not investigated further in the subsequent analysis.

4.2. Hyperparameter Optimization

4.2.1. Model Impact

The influence of different hyperparameters is investigated by examining the distribution of the anomaly score. In this investigation, two parameters are fixed to their default values, while the other free parameter is systematically changed. To evaluate the impact of hyperparameter on the IForest, different configurations are considered for use on the stricter selection samples provided in Tab. 4.4.

Table 4.4.: Summary of the Hyperparameter configurations for the study on the influence on the model

Hyperparameter	Configurations
n_estimators	[10, 80, 100, 125, 200, 256, 500, 1000]
max_samples	[10, 80, 100, 125, 200, 256, 500, 1000]
max_features	[1, 2, 4, 10, 11, 15, 26, 36]
bootstrap	True

Therefore, extreme ranges of values as the *n_ estimators* of 10 and 1000 for the number of trees are added to the range of parameters for construction in the training stage. Accordingly, for a number of trees of 10, the Fig. 4.4

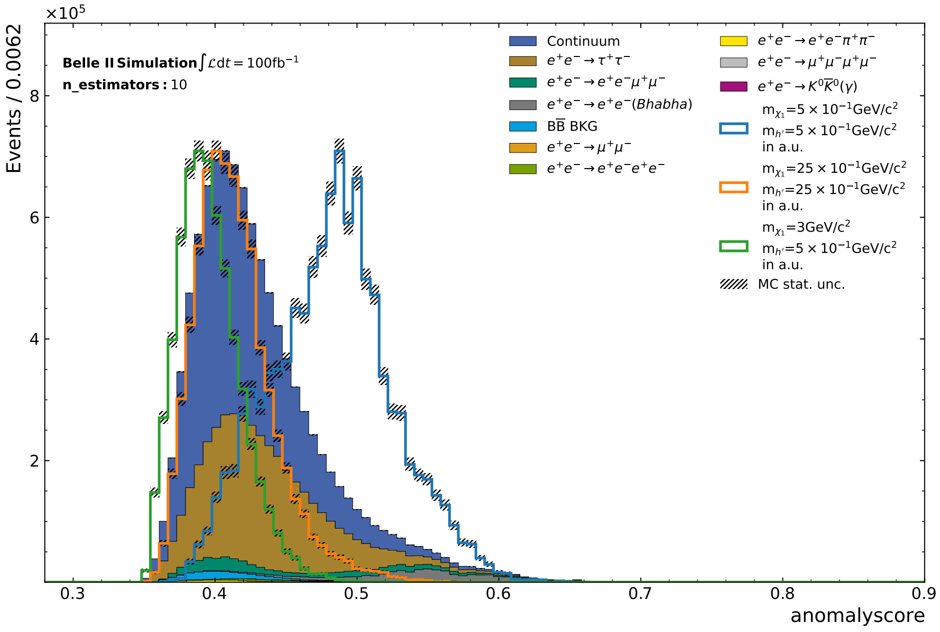


Figure 4.4.: Training of the IForest on the Hyperparameter: *n_ estimators*=100, *max_ samples*=256 and *max_ features*=36.

is obtained. Along the anomaly score axis, the signals exhibit a widening trend, indicating an increase in the range of the anomaly score distribution. Resulting in an almost complete overlapping of the light mass signal with the SM samples. Also, the other signals with heavy mass and large mass differences strongly overlap with the high-density events of the SM samples. This leads to a worse classification of the signal events because the samples trained on the ten trees have a very similar distribution to the background. Increasing the number of trees with the parameter *n_ estimator* as shown in Appendix A.5.1 results in a compression of the anomaly score distribution and a reduction of the width of the Dark Higgs peaks for all signal model configurations. This is most evident for the light mass.

A forest trained with *max_ samples*=10 in Fig. 4.5 shows a strong similarity to a Gaussian distribution since the events of the background are densely distributed around an anomaly score. In addition, the distribution of the signal clearly has an overlap with the background. Therefore, no interpretation of signal and background with respect to anomalies is possible. Greater separation between the light mass signal and the high-density SM events of the low anomaly score is observed as the subsample size increases, as displayed in Appendix A.5.2. This separation is particularly apparent when the IForest is trained on larger subsample sizes but is not as distinct for other signals.

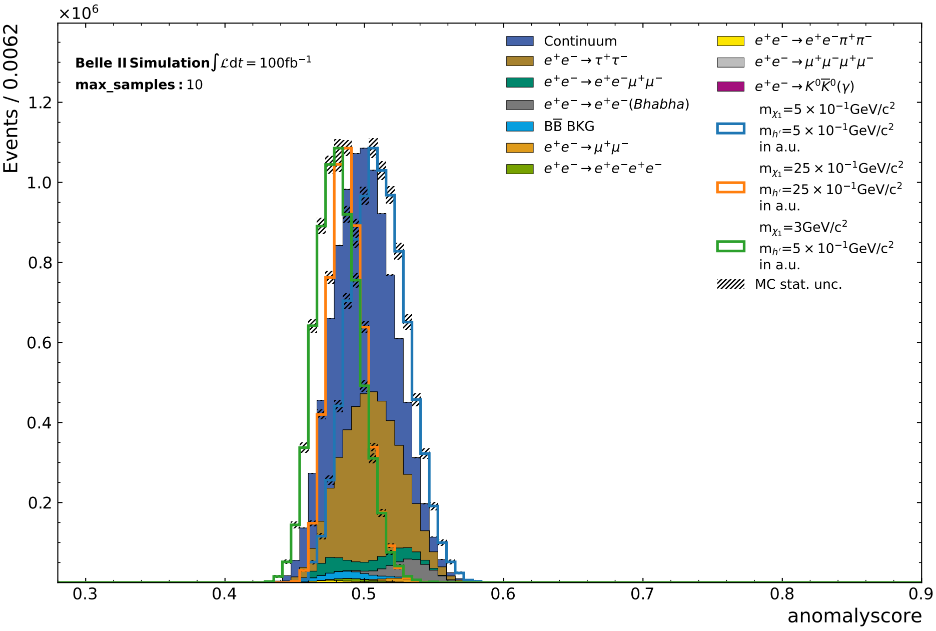


Figure 4.5.: Training of the IForest on the Hyperparameter: $n\_estimators = 100$, $max\_samples = 10$ and $max\_features = 36$.

Furthermore, the parameter *max_features*, which specifies the number of randomly chosen input features, with the value $max\_features = 1$ shown in Fig. 4.6, exhibits similar behavior as the subsample size since also, in this case, only a very small part of the input sample is used for training. The range of the anomaly scores increases without significant change in the peak width of the background and signal, as shown in Appendix A.5.3.

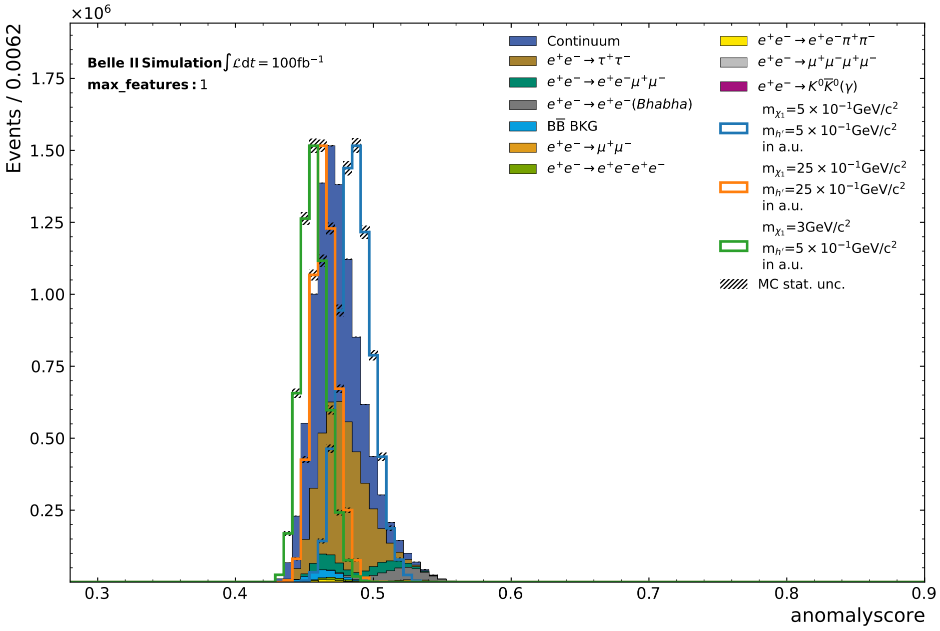


Figure 4.6.: Training of the IForest on the Hyperparameter: n_estimators=100, $max\_subsamples = 256$ and $max\_features = 1$.

As the parameters increase, the distribution of the anomaly score shifts, stretches, or widens. Given this behavior of the hyperparameters, low values are not advantageous for detection because distributions concentrated around a particular anomaly score have no meaning in terms of signal detection. Thus, the hyperparameter choice is quite relevant for anomaly detection. To improve performance, optimizing the hyperparameters can help to improve the detection of these Dark Higgs signals, studied in the following section, using a restrictive range of hyperparameters.

4.2.2. PFOM Grid Search

Based on the findings of the hyperparameter evaluation, the search range for the number of trees and the subsample size is narrowed. The acceptable range for these parameters is now limited to 100-500, as specified in Tab. 4.5. Despite the expected reduced statistical variation in sensitivity for a higher number of trees, the value is limited because of the high computational cost. With this smaller range of parameters, a grid search is performed for the weakly selection samples. Thus, the PFOM (Eq. 4.1) is used as a metric of sensitivity to

Table 4.5.: Summary of the restricted Hyperparameter configurations for the Grid Search

Hyperparameter	Configurations
n_estimators	[100, 150, 200, 250, 500]
max_samples	[100, 150, 200, 256, 500]
max_features	[16, 20, 24, 36]
bootstrap	True

Table 4.6.: Summary of the resulting best hyperparameters from the Grid Search

n_estimators	max_samples	max_features	bootstrap
500	256	all	True

background and signal behavior. Fig. 4.7, which displays all hyperparameter configurations and shows a negligible variation of the sensitivity for the heavy Dark Higgs and large mass differences of the Dark Higgs and Dark χ_1 signals. Additionally, the light mass signal demonstrates a more prominent sensitivity fluctuation. As the number of trees increases, the sensitivity increases gradually with the number of trees. The value of *max_features* has a pronounced impact on the sensitivity, resulting in a substantial decrease for lower values and a continuous rise with an increasing number of *max_features*. This is investigated by fixing two parameters and iterating over the free parameter to find the best possible hyperparameters. With 500 trees and a sub-sample size of 256, the number of features selected distinguishes a slight change in sensitivity, as seen in Fig. 4.8. The resulting best hyperparameters are listed in Tab. 4.6, leading to an increase in sensitivity of 4.2 % for the light Dark Higgs signal. Thus, a strong influence of the hyperparameters is not given. The reason behind this phenomenon is the overall larger inclusion of more subsamples represented in the training process as the number of trees increases. When the number of trees exceeds 150, the deviation between the maximal PFOM values is reduced, and the stability of IForest is consolidated. However, the best hyperparameters are not used due

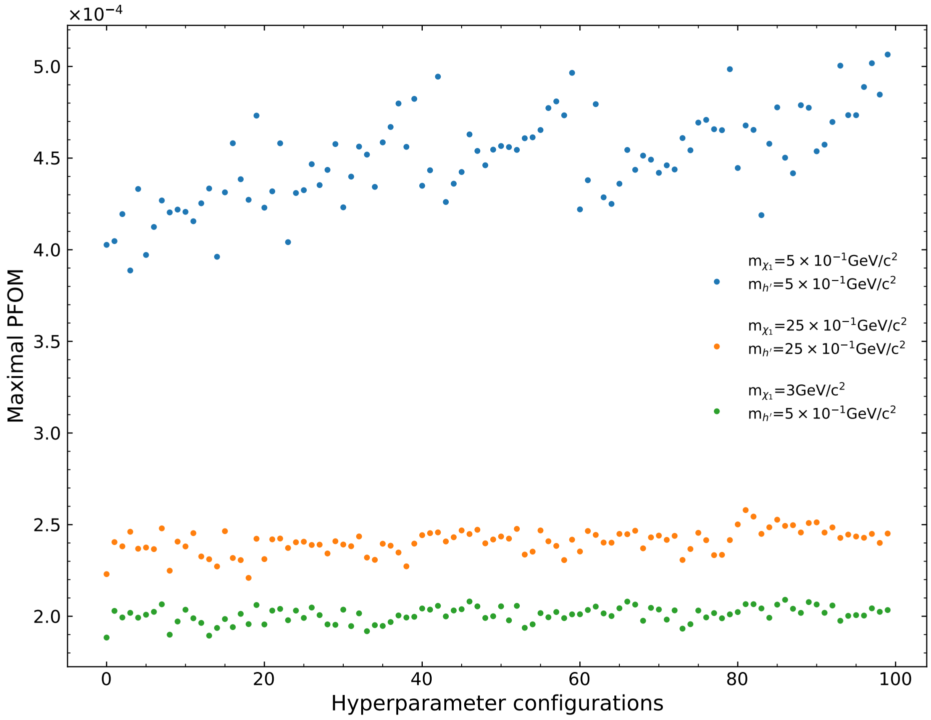


Figure 4.7.: Grid search result for the maximum PFOM for all hyperparameter configurations provided in 4.5 individually for the selected signals.

to the increased computation time of almost 3 hours. Instead, the parameters shown in Tab. 4.7 are chosen. Considering that the decrease in sensitivity is only 1.6%, the reduced computational time significantly benefits the subsequent analysis.

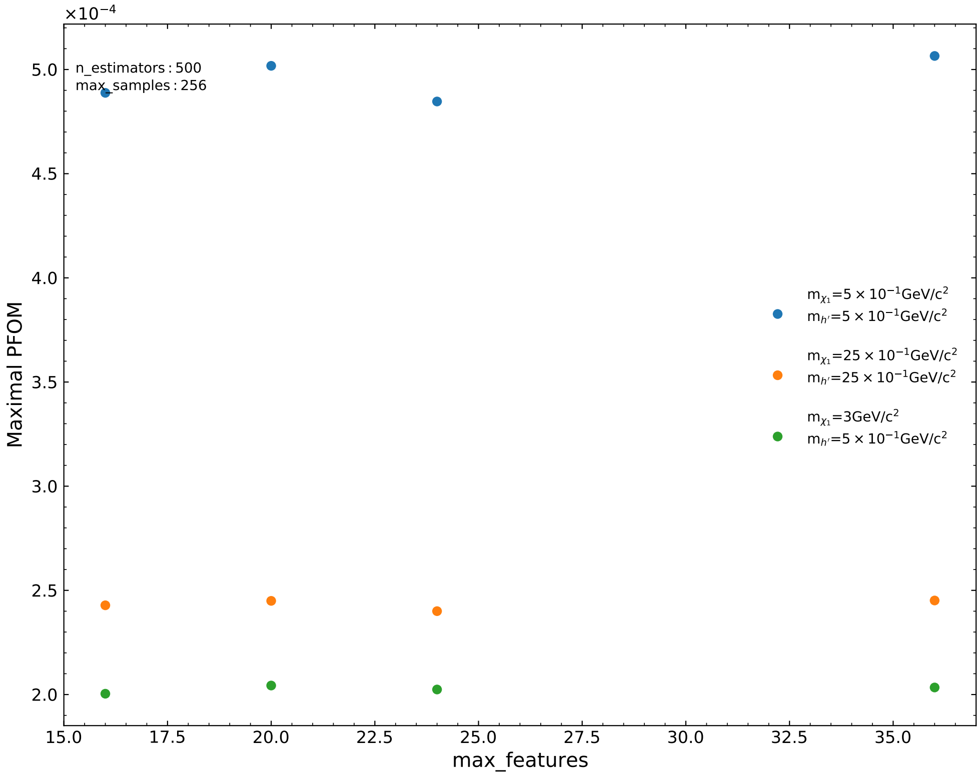


Figure 4.8.: Fixed parameters of n_estimators=500 and max_sample=256 used for iteration in a specific region of parameters. No significant difference is evident for these values.

Table 4.7.: Summary of the hyperparameters used for the following analysis.

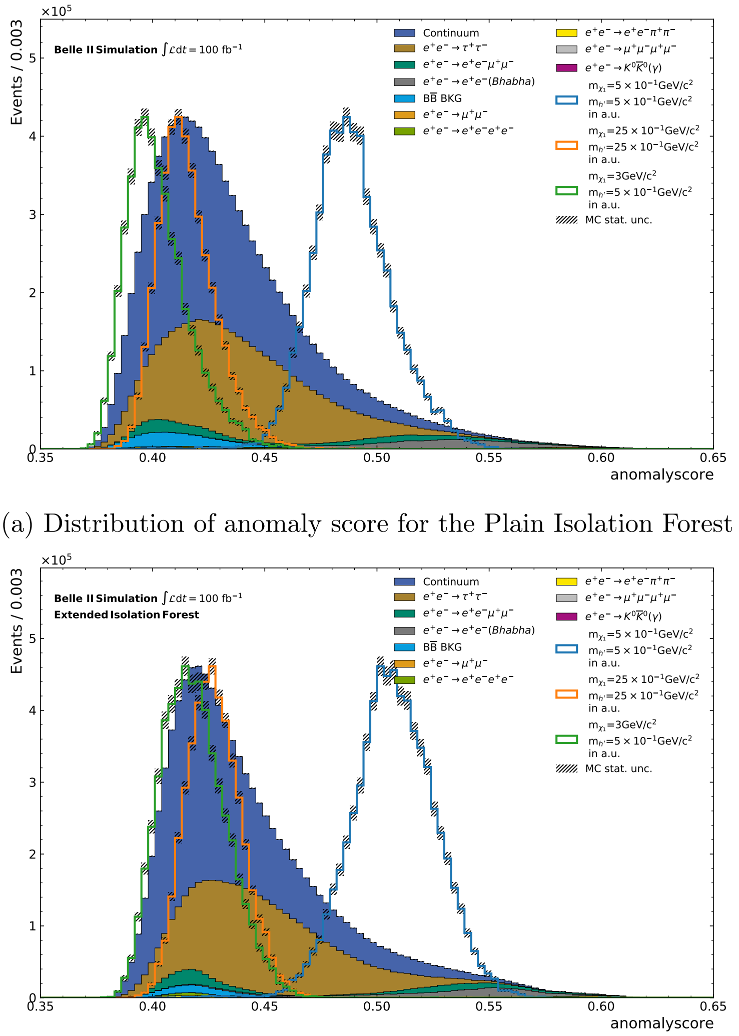
n_estimators	max_samples	max_features	bootstrap
250	256	all	True

4.3. Extended Isolation Forest

As introduced in section 3.2, the Extended Isolation Forest is studied because it explores a different approach to partitioning the distribution with randomly selected slopes. The hyperparameters used in the analysis of the Plain Isolation Forest remain unchanged. However, an additional extension level is introduced for this model. The extension level determines the maximum number of possible intersections of the hyperplanes used for the high-dimensional samples, and it is set to 35.

In contrast to the Plain Isolation Forest, the study of the Extended Isolation Forest is severely limited by runtime costs. In this model, the duration for training and calculating the anomaly score for the same data measures 8 hours. This computation time far exceeds the time required for the Plain Isolation Forest. Consequently, a more rigorous selection of samples for analysis is required to justify using the Extended Isolation Forest.

Fig. 4.9 shows a clear difference in the distribution of the anomaly values, especially for the Dark Higgs signals. These peaks in the distribution shift towards higher anomaly values. Looking at the PFOM curve in Fig. 4.10 it provides a negligible rise in sensitivity with this extended method. More precisely, the maximal PFOM value increased from 0.000681 to 0.000729, improving the sensitivity by 7.2% for the light Dark Higgs signal. Furthermore, the study with different configurations of Dark h' and Dark χ_1 masses in Fig. 4.11 shows that the Extended Isolation Forest is more sensitive to a few light masses

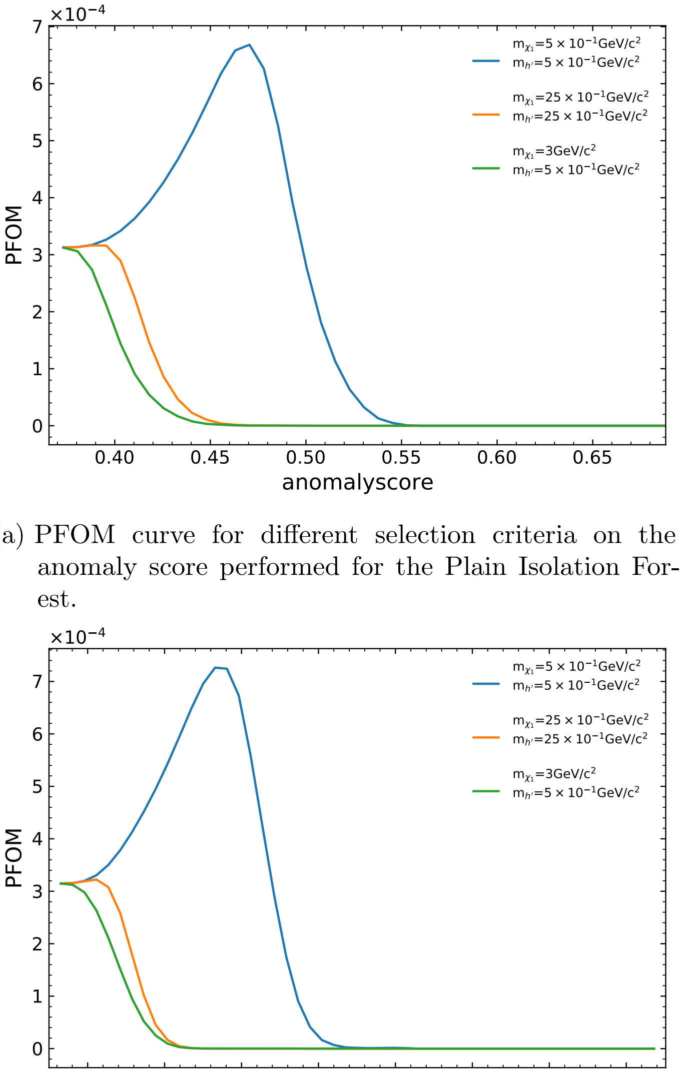


(a) Distribution of anomaly score for the Plain Isolation Forest

(b) Distribution of anomaly score for the Extended Isolation Forest

Figure 4.9.: Different distributions of the anomaly score for the Plain Isolation Forest (a) and the Extended Isolation Forest (b) are shown.

in ranges of $m_{h'} < 1.0$ and $m_{\chi_1} < 1.0$. In addition, the Extended Isolation Forest rejects similar amounts of background as the Plain Isolation Forest, with a minor change in signal efficiency.

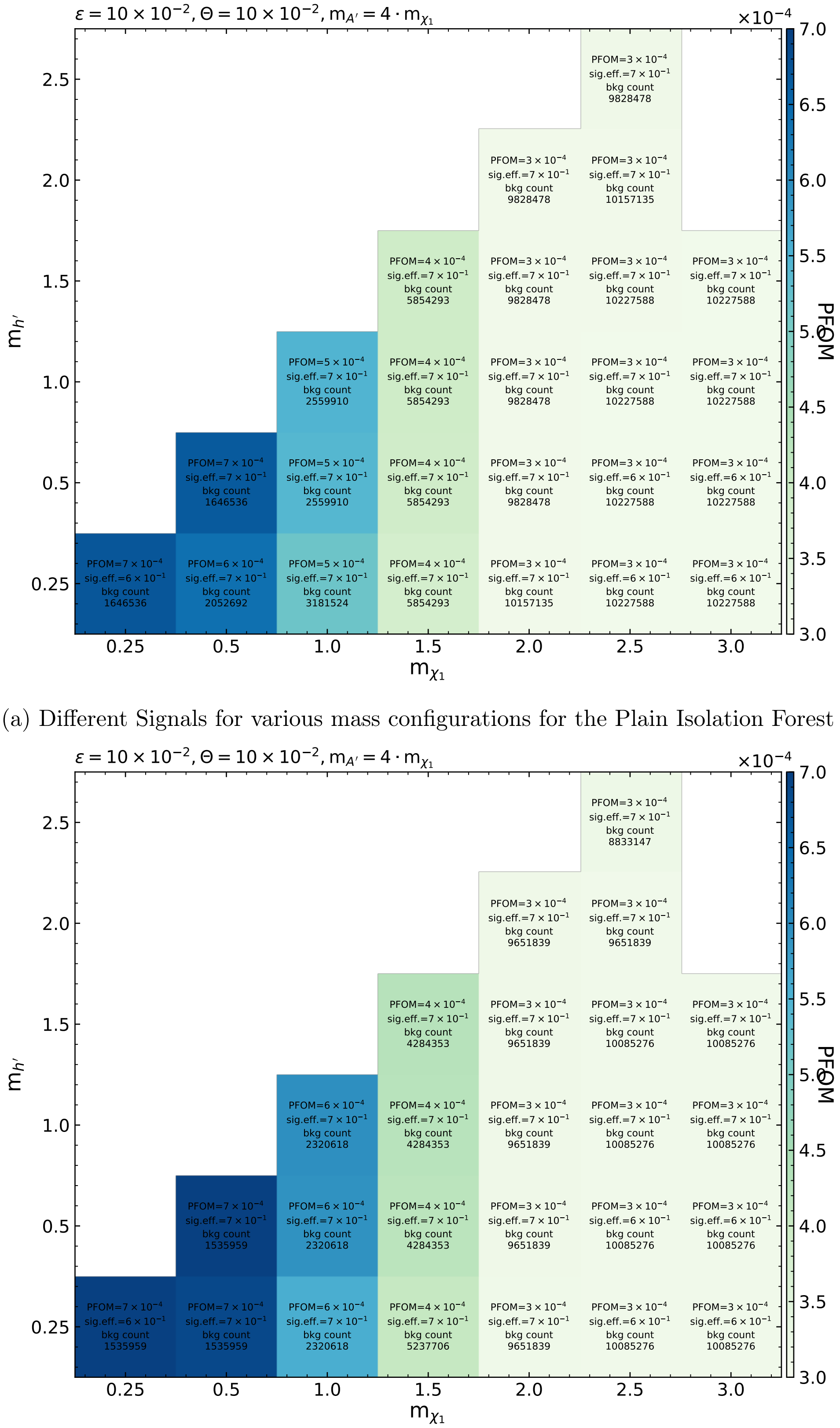


(a) PFOM curve for different selection criteria on the anomaly score performed for the Plain Isolation Forest.

(b) PFOM curve for different selection criteria on the anomaly score performed for the Extended Isolation Forest.

Figure 4.10.: The PFOM for the representative selected signals in comparison of the Plain Isolation Forest (a) and Extended Isolation Forest (b).

Due to the increased computational time and marginal improvement in sensitivity for the signals, the Extended Isolation Forest approach is not examined in further analysis.



(a) Different Signals for various mass configurations for the Plain Isolation Forest

(b) Different signals for various mass configurations for the Extended Isolation Forest

Figure 4.11.: The maximal PFOM for different configurations of the Dark h' and Dark χ_1 for the Plain Isolation Forest (a) and the Extended Isolation Forest (b) as well as the signal efficiency and remaining background after selection.

4.4. Contribution to Model

The following study investigates the impact of different training runs on separately trained IForests. This involves training on each SM process and a subset of the features grouped into the different information of the particles. Specifically, the effectiveness of the IForest trained on individual background and input features regarding the contribution of isolating signals in the anomaly score distribution. By considering these factors, insights are gained into the capability of the IForest to contribute to the advancement of this method.

4.4.1. Input Feature

For the examination of the contribution of the features, the focus is on the different kinematics and other information from the collision events, which are listed in Tab. 4.8. The different information content of the input features is utilized to understand the individual influence on the model. This is accomplished by training the model on each SM process and separately for each input feature group. Due to the small sample size of the $e^+e^- \to K^0\overline{K}^0(\gamma)$ process, all events in the subsample are used for training. For the other processes, 256 events for each tree are taken according to the *max_ sample* hyperparameter. The resulting distribution of the anomaly score for one SM process and feature group is represented in Appendix A.6. Therefore, the performance of the separately trained IForest is evaluated in terms of its ability to differentiate between background and signal peaks with respect to the input features. The separation strength between these peaks is examined as a measure of performance. A clear distinction between these distributions indicates that the features have good discriminative power while overlapping distributions indicate that the features may not be contributing effectively to the distinction. Certain features, such as the missign Four-vector and transverse momentum, stand out in all ten SM processes, showing a robust categorization of signal peaks with high anomaly scores compared to the SM sample. For example, in the $e^+e^- \to e^+e^-\mu^+\mu^-$ process for the missing momentum and missing energy or in the $e^+e^- \to e^+e^-e^+e^-$ process for the transversal momentum. In addition, the input features that have shared information content show a very different degree of isolation of the outliers of these distributions. This can be seen for the Four-vector and the transversal momentum, where the latter contains less information. Nevertheless, it shows a better separation between the background and the signal. This suggests that the way the information is presented to the IForest is important for this method of anomaly

Table 4.8.: Summary of different Input Feature groups used for individual training on the IForest.

Input Feature	Count
Missing momentum/energy	4
binary PID (e, μ)	4
Transverse momentum	4
Four-vector	16
θ	4
ϕ	4
Invariant mass	2

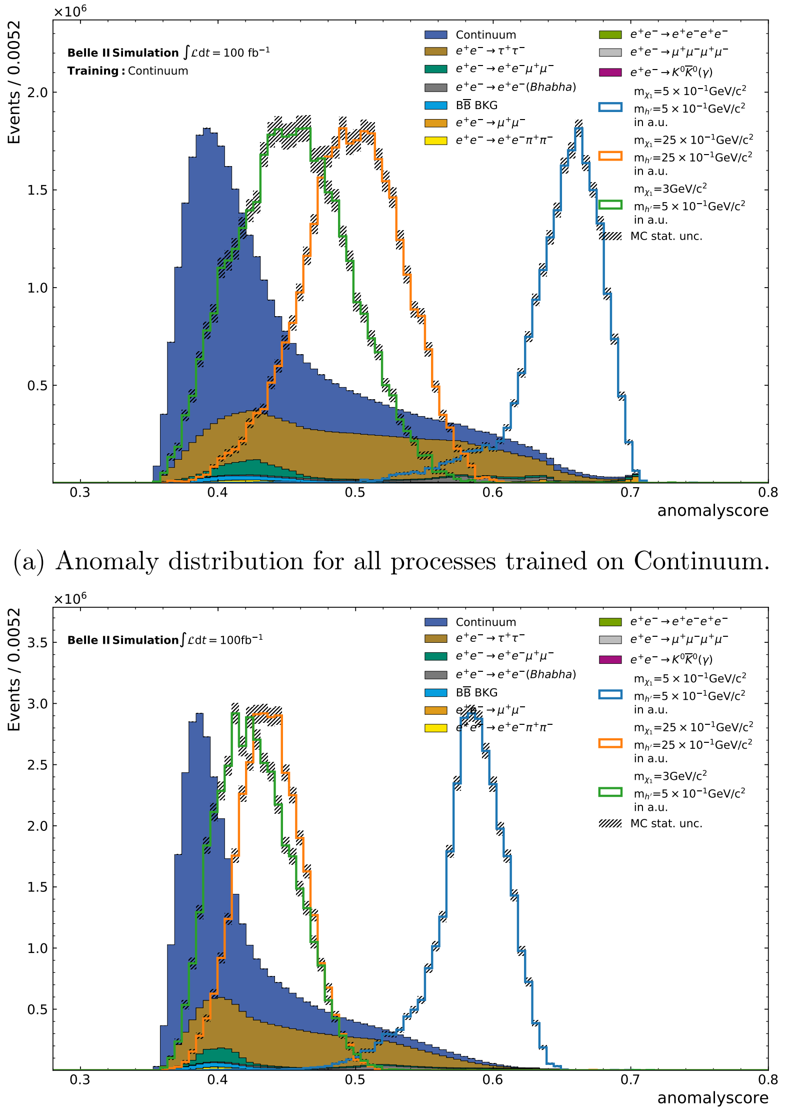
detection. Furthermore, binaryPID(e, μ) information also delivers poor characteristics for detecting anomalies, as they mainly lead to overlap with the SM sample distributions. For these reasons, missing momentum, missing energy, and transverse momentum are used as input features for the following study, as they are generally beneficial in all given processes.

4.4.2. Background Samples

To investigate the individual SM processes, since these are not learned specifically in the combined training, the IForest is set up separately on the distribution of the SM sample processes in Tab. 4.9. This is used to infer the behavior of the trained forest on all processes and their contribution to signal detection as a distinction from the SM background. The separate training is performed using the same well-suited input features, namely the missing Four-vector and transverse momentum, as explored in Section 4.4.1. The IForest trained on a single SM sample learns more about a process distribution because only that distribution is used to construct the forest, and more samples are used in total per process. According to the dominant share of Continuum and $e^+e^- \to \tau^+\tau^-$ processes, the weighting of these processes within the training is questioned. The assumption is that in order to perform more effective partitioning on these distributions, training on a particular process's distribution, as discussed earlier, must be performed. When determining the anomaly score in the test phase, all SM samples are passed, which allows the observation of how the IForest calculates the anomaly score for not seen events. The IForest trained on the Continuum (Fig. 4.12a) contributes significantly to the model training due to its distribution's similarity to the forest trained on all SM samples (Fig. 4.12b). This similarity is also reflected in the calculation of the anomaly score for the SM sample events in both distributions.

Table 4.9.: Summary of the SM processes used for the individual training of the IForest with rounded percentage values.

SM Background Process	Events	Percentage
Continuum	23802002	62.4%
$e^+e^- \to \tau^+\tau^-$	10807984	28.3%
$e^+e^- \to \mu^+\mu^-\mu^+\mu^-$	1132608	3.0%
$e^+e^- \to e^+e^-\mu^+\mu^-$	1085954	2.8%
$B\overline{B}\,BKG$	940570	2.5%
$e^+e^- \to \mu^+\mu^-$	158074	0.4%
$e^+e^- \to e^+e^-\pi^+\pi^-$	106766	0.3%
$e^+e^- \to e^+e^-e^+e^-$	67066	0.18%
$e^+e^- \to e^+e^-(Bhabha)$	58225	0.15%
$e^+e^- \to K^0\overline{K}^0(\gamma)$	175	$4.6 \cdot 10^{-4}$%
$\sum$	38159424	100%



(a) Anomaly distribution for all processes trained on Continuum.

(b) Anomaly distribution for IForest trained on all SM samples.

Figure 4.12.: Distributions of the anomaly score for the training on Continuum with the entire background (a) and the training on all SM samples (b) are shown. The uncertainties in the background events are not visible in the plot due to their small magnitude.

The results in Appendix A.7 illustrate that the anomaly score calculation of other process events is strongly affected by the separate training, as reflected in the extreme classification for the anomaly score of the not trained processes. This is to be expected since the distributions partly overlap. Nevertheless, there is a clear tendency indicating a strong deviation of the anomaly score distribution with respect to the trained SM process. Particularly evident in

- $e^+e^- \to e^+e^-e^+e^-$
- $e^+e^- \to e^+e^-\pi^+\pi^-$
- $e^+e^- \to \mu^+\mu^-\mu^+\mu^-$

processes that have the peculiarity of containing at least two of the four expected particles. The analysis suggests that the majority of the data set does not overlap with these particular events, leading to a classification of the remaining background as abnormal being prevalent. Furthermore Fig. 4.13 presents a significant bias in favor of samples with a larger proportion,

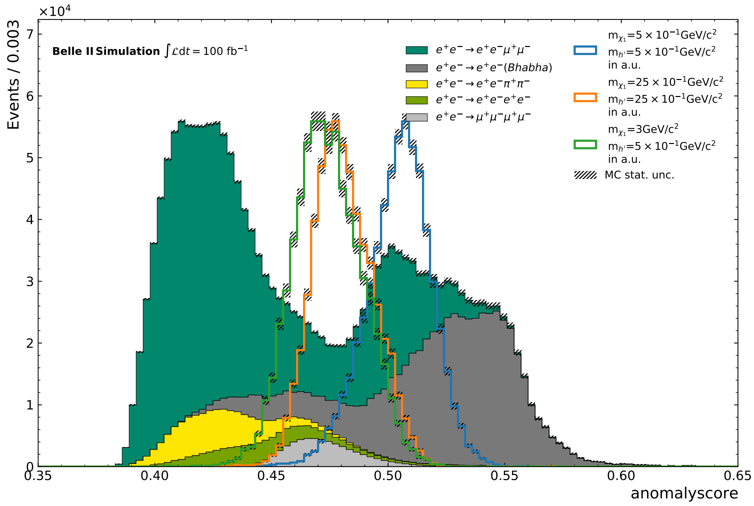


Figure 4.13.: Distribution of the anomaly score for an IForest trained on a subset of given processes.

such as the $e^+e^- \to e^+e^-\mu^+\mu^-$ process when calculating the anomaly score for Bhabha events. The denser fraction of Bhabha events shows a higher value of the anomaly score compared to the distribution trained only on the Bhabha process in Appendix A.7.2. This may be a consequence of a type of overtraining on instances in the $e^+e^- \to e^+e^-\mu^+\mu^-$ process that overlap with those in the Bhabha process.

These insights serve as inspiration for the ensemble approach of IForests, which is presented in Chapter 4.6.

4.4.3. Impact of global PIDs

Recognizing the significance of particle information, it is now incorporated into the model. The binary PID is used to indicate whether a particle of interest is classified as a muon or an electron. However, as discussed in Section 4.4.1, these features tend to result in a less precise distinction of background and signal. Several particle identification features used for model training are listed in Tab. 4.10. The lepton PID used gives the probability that a detected particle is an electron or a muon, with all particles considered for the PID. For example, if the probability of an electron is 0.5, then the probability of the other five particles combined is 0.5.

Table 4.10.: Overview of the additional Input Features regarding the Particle Identification.

Input Feature	Count
Missing momentum/Energy	4
Transverse momentum	4
binary PID(e,μ)	4
electron ID noTOP	4
kaon ID	4
proton ID	4
pion ID	4
muon ID	4
deuteron ID	4
lepton PID	4
$\sum$	40

An examination of the distribution presented in Fig. 4.14 indicates a major improvement

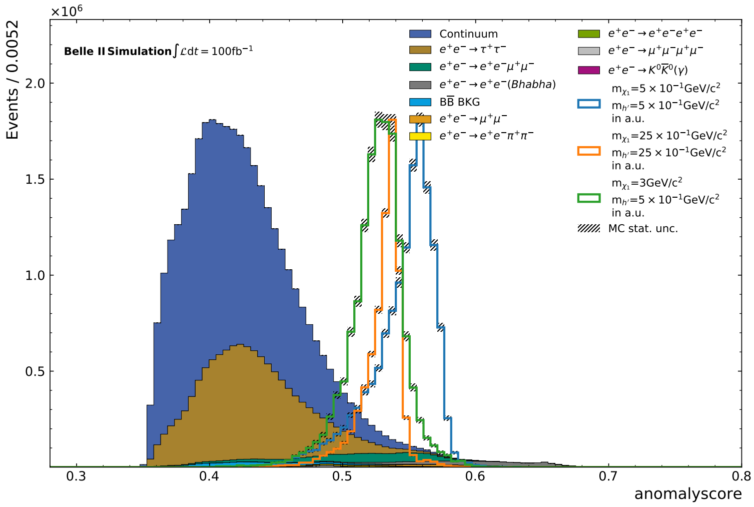
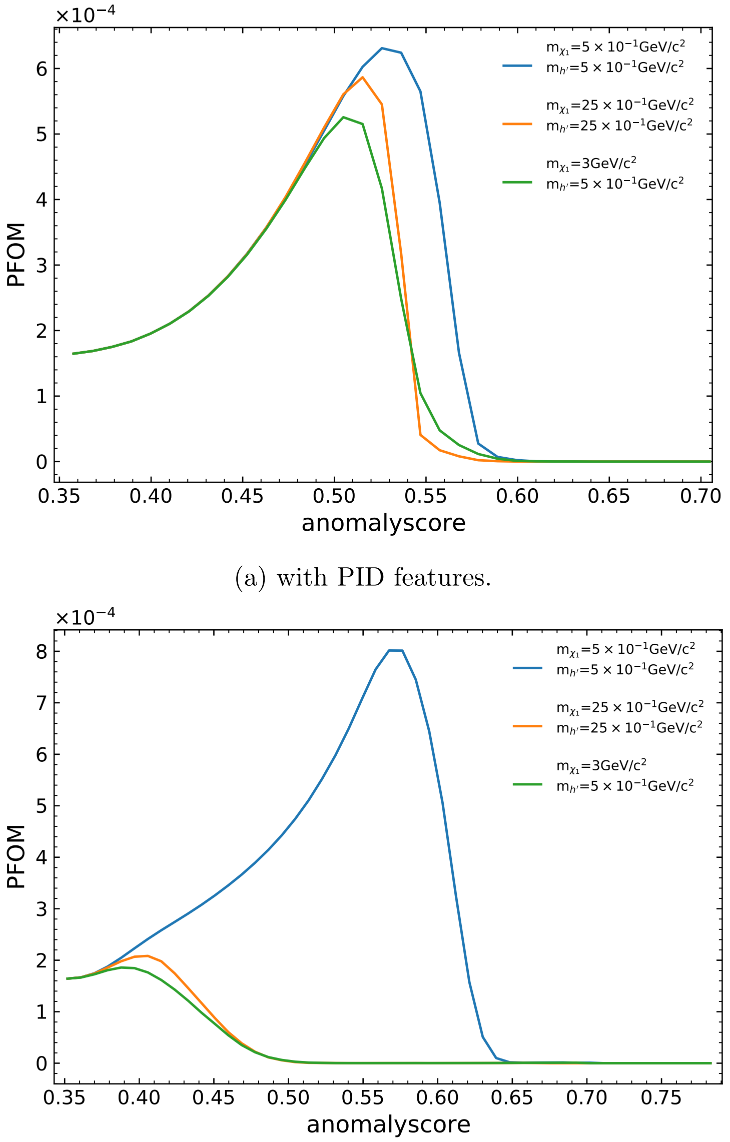


Figure 4.14.: Distribution of the anomaly score for the IForest with PID information used for training.

in model-independent signal detection. This improvement is attributed to the fact that the signals are closely grouped within a similar anomaly score range.

The PFOM curve (Fig. 4.15) presents an improved sensitivity for all signals with the addition of PID information. These signals, which performed worse in the previous analysis, now compete with the sensitivity of the light mass. Compared to the IForest with missing Four-vector and transverse momentum as input, there is a drop of about 30%, particularly concerning the light mass. Nevertheless, the result with the PIDs is an improvement due to the independence of the model parameters of the signals tested for anomaly detection.



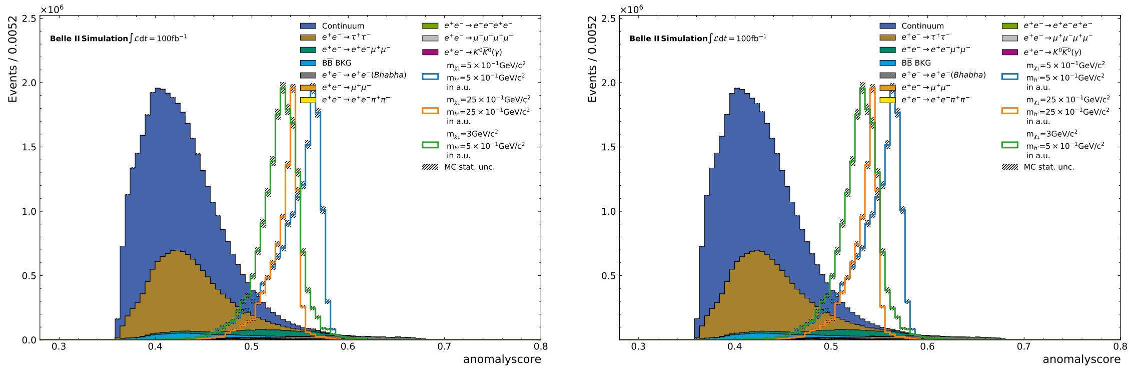
(a) with PID features.

(b) without PID features.

Figure 4.15.: PFOM curve for the IForest trained with PID features (a) and without PID features (b).

4.5. Evaluating Overfitting

The assessment of model overfitting is related to the classification accuracy of trained and unknown samples. If the model's performance remains consistent when applied to unseen data during testing, this suggests minimal overfitting and high accuracy. To investigate this in the context of the IForest model using the studied input features (missing four-vector, lateral momentum, global PIDs), a large data set is randomly divided into an 80% training set and a 20% test set. Both sets are then passed through the IForest algorithm. The anomaly score distributions (Fig. 4.16) are obtained by scaling the events based on the sample size ratio to an integrated luminosity of 100 fb^{-1}. This scaling takes into account the difference in event counts. The distributions reveal no significant variation in the event distribution for the anomaly score. By calculating the PFOM, scaling it to the sample size, and then analyzing the curves, only minor differences between the curves for the light mass are apparent (Fig. 4.17). This observation of curve overlap for the PFOM confirms that the IForest model, when applied to unseen data, has comparable classification performance to the training data, which comprises 80% of the data set. These results indicate the absence of overfitting. Due to time constraints in this study, it was not possible to conduct further testing with different ratios of test and training samples.



(a) Distribution of 20% samples.

(b) Distribution of 80% samples.

Figure 4.16.: Distribution of the anomalyscore for the 80% random samples (b) used for training and 20% for testing (a) on the trained IForest. The background events are scaled according to the proportion of samples to 100 fb^{-1}.

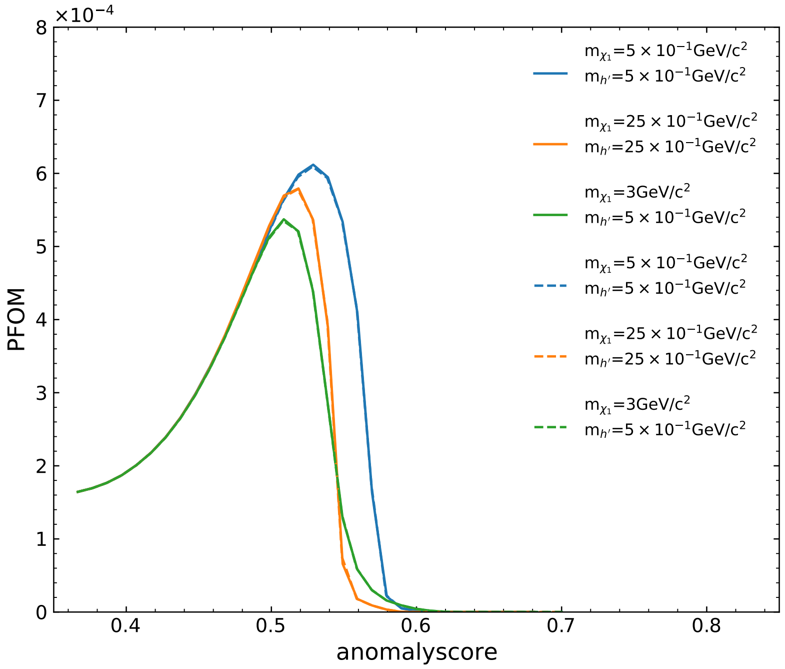


Figure 4.17.: PFOM curve for the Train/Test Split IForest. The dashed line represents the sensitivity for the 80% used for training, while the straight line represents the 20% that the IForest did not see during training.

4.6. Ensemble of Isolation Forests

The IForest is modified by performing an ensemble of forests, each is trained on one process of the SM background, as shown in Fig. 4.18. To train the forests, the missing Four-vector and the transverse momentum are used as input features. Then the resulting anomaly scores are averaged over the entire ensemble, yielding the *average anomalyscore*. The distribution of the individual forests trained on each process in Appendix A.8 exhibits a tendency towards the center of the anomaly score range, which is primarily influenced by the extreme classification behavior of each separate IForest towards an anomaly score of 0 or 1. Fig. 4.19, illustrates the compression of the distribution along the anomaly score, showing the concentration towards the center. The sensitivity improvements (Fig. 4.20) are due to the fact that for each process, the resulting anomaly score is given equal weight in the evaluation. Accordingly, the application of additional techniques for selecting the anomaly score, such as minimum, maximum, weighting based on luminosity or weighting based on the sample size is displayed in Appendix A.4, A.9. The results presented in Appendix A.9.1 show that using the minimum and maximum scores shifts the distributions towards either low or very large anomaly scores. Compared to the Plain Isolation Forest with default settings, this causes the model to be less sensitive. Therefore, using extremes in the anomaly score distribution can negatively impact the model's ability to effectively discriminate between normal and anomalous sample events. A weighting based on sample size or luminosity results in an anomaly score with a dominant contribution from the Continuum and $e^+e^- \to \tau^+\tau^-$ due to their large proportion of the total sample. These provide a sensitivity that is comparable to that of the IForest trained on all SM samples.

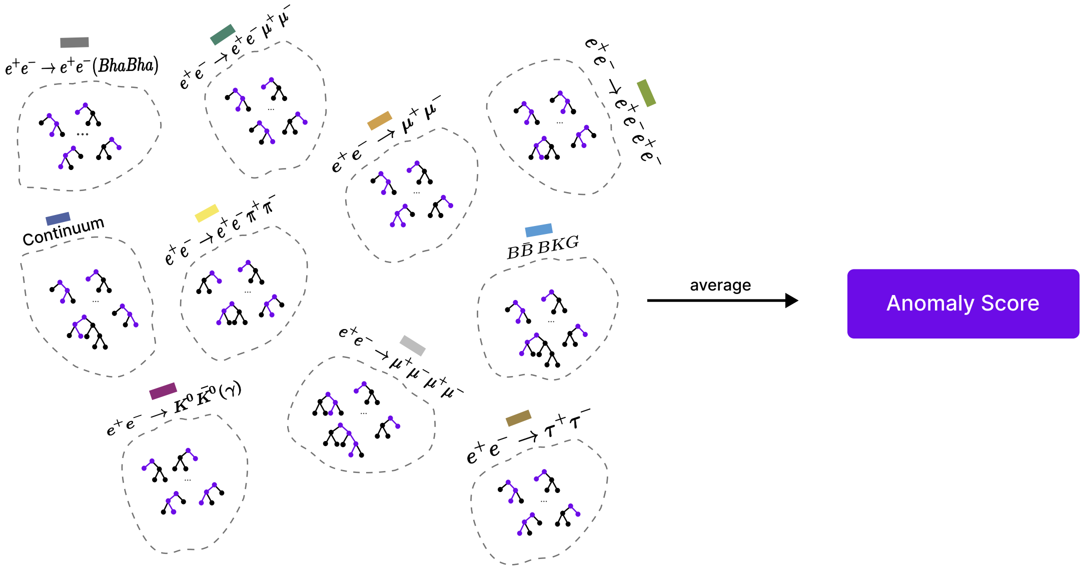

$e^+e^- \to e^+e^-(BhaBha)$
$e^+e^- \to e^+e^-\mu^+\mu^-$
$e^+e^- \to \mu^+\mu^-$
$e^+e^- \to e^+e^-e^+e^-$
Continuum
$e^+e^- \to e^+e^-\pi^+\pi^-$
$B\bar{B}\,BKG$
average
Anomaly Score
$e^+e^- \to \mu^+\mu^-\mu^+\mu^-$
$e^+e^- \to K^0\bar{K}^0(\gamma)$
$e^+e^- \to \tau^+\tau^-$

Figure 4.18.: Schematic of the Ensemble Isolation Forest, which consists of several IForests trained on separate SM processes. The new anomaly score is the average of the individual IForests in the ensemble.

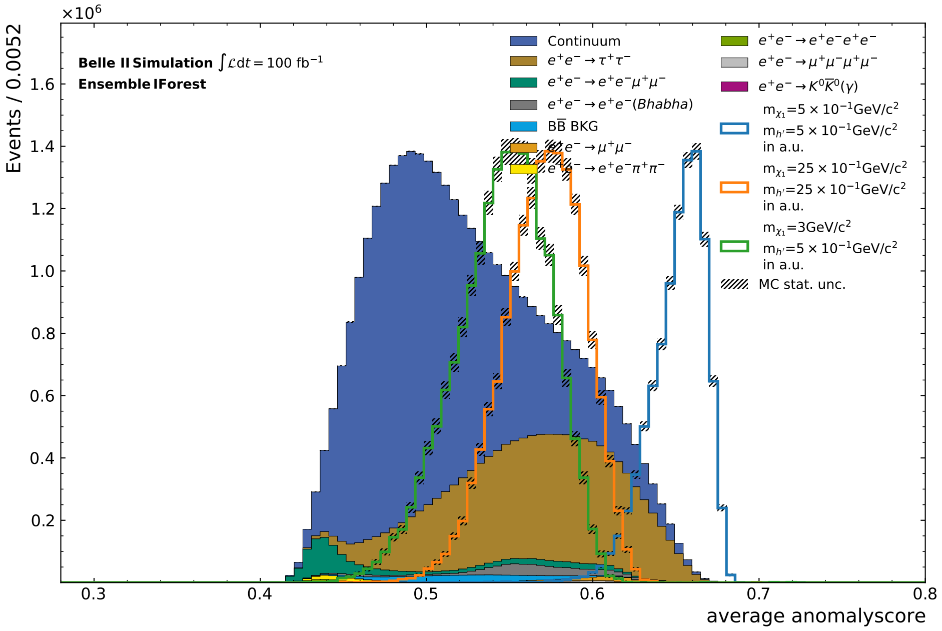


Figure 4.19.: Distribution of the averaged anomaly scores for the Ensemble of Isolation Forests.

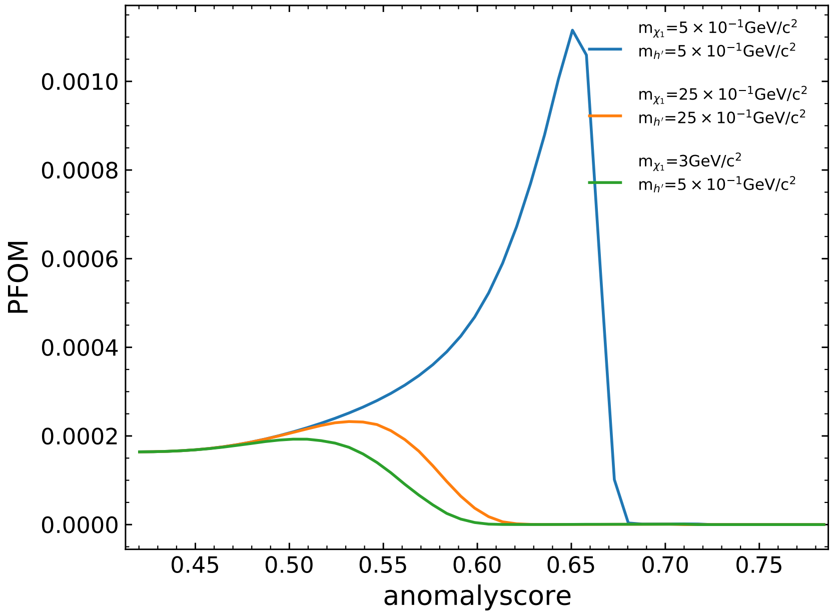


Figure 4.20.: PFOM curve of the averaged anomaly scores for the Ensemble of Isolation Forests.

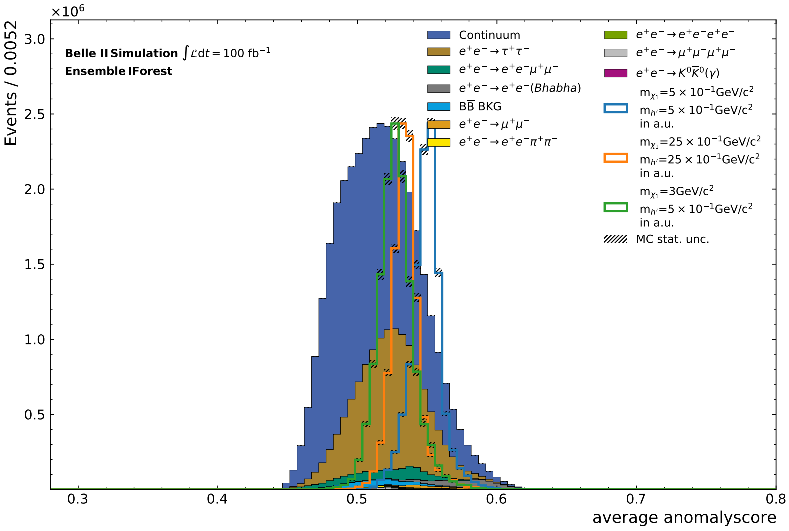


Figure 4.21.: Distribution of the averaged anomaly scores for the Ensemble of Isolation Forests with PID information used for training.

Ensemble of Isolation Forest with PIDs

In order to determine whether applying the ensemble of Isolation Forests increases sensitivity only for certain features, the global PIDs are now included. The training procedure with the 40 input features and the subsequent analysis of the anomaly score distributions for the processes in Appendix A.8.1 display a distinct behavior in the distributions of the individually trained IForests, especially concerning the signals. As a result, the distribution in Fig. 4.21 shows a strong overlap between background and signals, which is also reflected in the PFOM in Fig. 4.22 for the poor performance of the averaged forests. Furthermore, due to the extremely low sensitivity for signals with high mass and large mass difference for the suitable features missing Four-vector and transverse momentum, there is no general advantage of this method despite the higher sensitivity for signals with low mass. This reveals a strong dependency on the IForest and its input in the Ensemble of Isolation Forests, making it not recommended for general use.

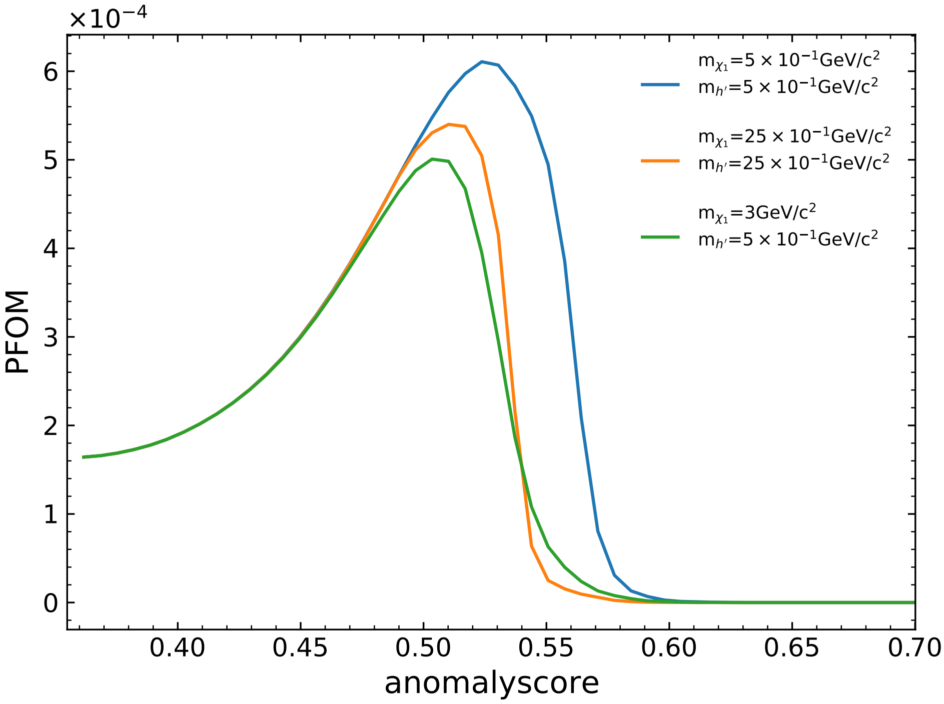


Figure 4.22.: PFOM curve for the Ensemble of Isolation Forests with PID information.

4.7. Iterative Training

In order to efficiently reduce the input sample through a beneficial selection based on a separation between background and signals, an iterative method is used. By using the IForest trained on separate SM processes in the previous Section 4.4.2, suitable selections are determined. An advantageous selection can be made with the distribution of the anomaly score trained separately on the $e^+e^- \to e^+e^-\mu^+\mu^-$ process illustrated in Fig. 4.23, which

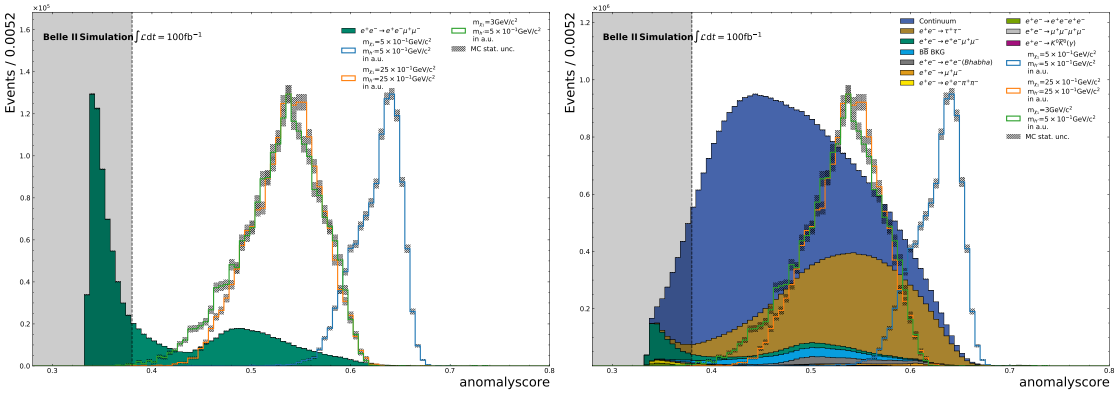


Figure 4.23.: Removed background events on the anomaly score distribution (right) for the IForest trained on the $e^+e^- \to e^+e^-\mu^+\mu^-$ process separately. Individual distribution of the $e^+e^- \to e^+e^-\mu^+\mu^-$ process (left), on which the criterion for selection without signal efficiency loss for an $anomalyscore > 0.38$ is based.

allows selection almost without efficiency loss on the example signals with an anomaly value greater than 0.38. The remaining background event counts after the selection are listed in Tab. 4.11. Only a small portion of the background, about 13.4%, is removed. Subsequently, training is performed on the entire background using only one IForest. As a result, the distribution for the IForest trained on the entire reduced data set with the selection criterion of the anomaly score of the separately trained $e^+e^- \to e^+e^-\mu^+\mu^-$ process is obtained. This distribution is shown in Appendix A.17. The width of the SM sample peak is tighter compared to the signal peaks. In addition, the signal peaks for heavy masses and strong mass differences are closer to the SM sample peak, indicating some overlap in their distributions. Light mass signals portray a distinct distribution that is separate from the SM peak. To evaluate the performance of the selection, the PFOM metric is chosen.

As depicted in Fig. 4.24, the reduction of the background using the selected process results in increased sensitivity for the light Dark Higgs signal. However, it also leads to a deterioration in performance for the other two signals. This indicates a greater overlap of signals with the background for signals with low sensitivity, as already observed in the distribution of the anomaly score. In the following analysis, additional tests are performed on the iterative training approach using additional information from PID.

Table 4.11.: Summary of the SM processes remaining after the performed selection on $e^+e^- \to e^+e^-\mu^+\mu^-$. The event count of the processes is not weighted on their luminosity.

SM Background Process	Events	rejected background
Continuum	19333337	18.77%
$e^+e^- \to \tau^+\tau^-$	10285312	4.84%
$e^+e^- \to \mu^+\mu^-\mu^+\mu^-$	1132607	8.83%
$e^+e^- \to e^+e^-\mu^+\mu^-$	1052735	3.06%
$B\overline{B}\,BKG$	846062	10.05%
$e^+e^- \to \mu^+\mu^-$	158053	0.013%
$e^+e^- \to e^+e^-\pi^+\pi^-$	103564	3.00%
$e^+e^- \to e^+e^-e^+e^-$	66548	0.77%
$e^+e^- \to e^+e^-(Bhabha)$	57828	0.68%
$e^+e^- \to K^0\bar{K}^0(\gamma)$	166	5.14%
$\sum$	33036212	13.43%

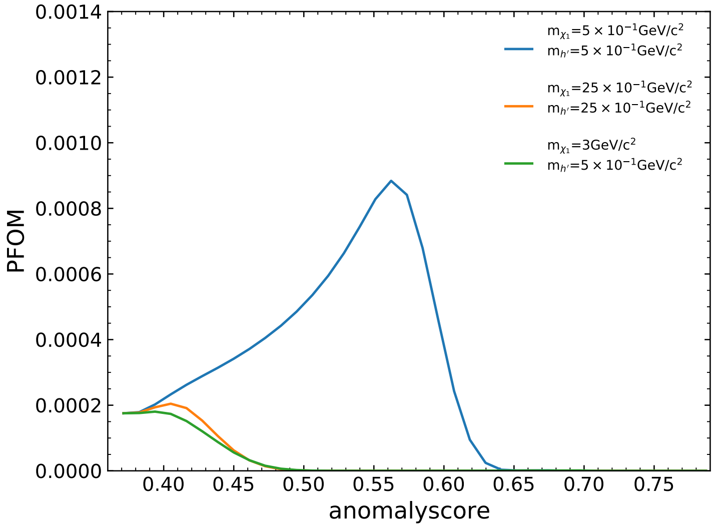


Figure 4.24.: PFOM showing the sensitivity on several selections along the anomaly score for the IForest trained on the reduced samples according to the selection criterion of the *anomaly score* > 0.38.

Iterative training method with PIDs

Another test is performed using the PID characteristics for the iterative method. The selection criteria based on the individually trained SM processes are shown in Tab. 4.12. The selection in the IForest with all the SM samples to be trained is referred to as 'collective'. Here, the selection criterion $e^+e^- \to e^+e^-\mu^+\mu^-$ is taken from the IForest trained without PIDs. A suitable selection on the continuum-trained IForest is identified at a selection criterion of 0.48 for the separate trained forest on the SM processes. This selection effectively reduces a significant part of the background, as summarized in Tab. 4.13, in particular

Table 4.12.: Overview of the individual selection criteria for the iterative training of the IForest.

Training	$e^+e^- \to e^+e^-\mu^+\mu^-$	Continuum	Collective
Selection Criterion	0.38	0.48	0.45

removing the Continuum and $e^+e^- \to \tau^+\tau^-$ events. Subsequently, the IForest with collective SM samples are considered and trained on all SM samples. Again, with a selection criterion of 0.45, much of the background can be removed as described in Tab. 4.13b.

The retrained IForest on $e^+e^- \to e^+e^-\mu^+\mu^-$ and collective sample selections are the only ones that contain a clear separation between background and signal peaks, shown in Appendix A.17, A.20. For the continuum selection, only the events that exhibit similar characteristics are retained due to the significant reduction in the sample size. As a result, an overlap of signal and background in the distribution is given. Therefore, this selection using the continuum leads to quality loss as the IForest reaches its limit of detecting anomalies in overlapping samples. With the sensitivity given by the PFOM in Appendix A.10, it is clear that with the $e^+e^- \to e^+e^-\mu^+\mu^-$, there is an improvement for all signals, but for the IForest without PIDs, the light mass signal loses some sensitivity. As expected, the sensitivity for the substantial background reduction with the continuum method is very poor. Selection on the entire trained background shows similar behavior to the selection on the $e^+e^- \to e^+e^-\mu^+\mu^-$ process. Therefore, these two methods are particularly well suited for increasing the overall sensitivity of all signals.

The additional analysis of the invariant mass of h' reflects the reduction of background events while preserving the remaining signal events (Fig. 4.25). This selection proves advantageous for the detection of the signals since it results in the removal of a significant number of background events while preserving the signal characteristics. Although iterative training does not provide a large increase in sensitivity, the global PIDs and these effective selections can be applied to train a model that provides relatively high sensitivity for

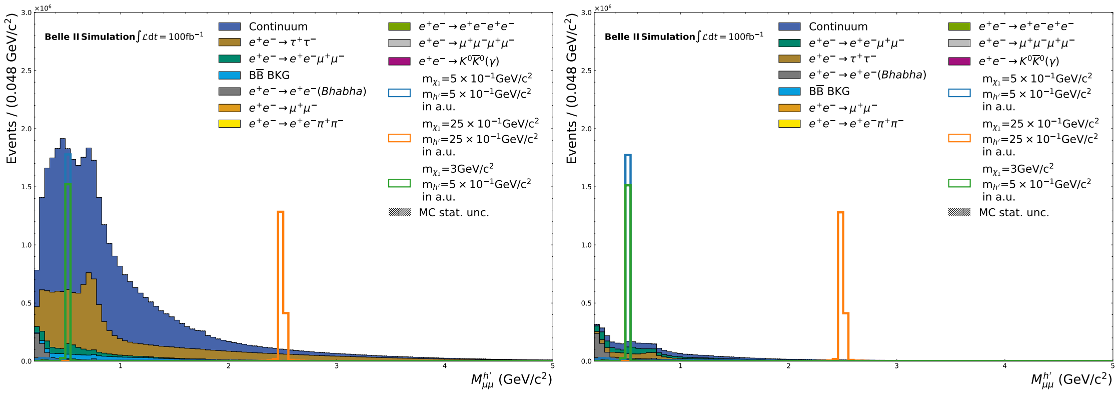


Figure 4.25.: Distribution of the invariant mass of the h' for the selection on the IForest trained on all SM samples. By selecting without losing many events, the number of signal events before and after selection remains almost unchanged.

independent model parameters.

Table 4.13.: Summary of the SM process events remaining after the performed selections presented in Tab. 4.12 .

(a) Overview of the SM processes remaining after the performed selection on Continuum. The event count of the processes is not weighted on their luminosity.

SM Background Process	Events	rejected background
Continuum	1939787	91.85%
$e^+e^- \to \tau^+\tau^-$	1158132	89.28%
$e^+e^- \to \mu^+\mu^-\mu^+\mu^-$	708041	37.48%
$e^+e^- \to e^+e^-\mu^+\mu^-$	799687	26.36%
$B\overline{B}\,BKG$	172496	81.66%
$e^+e^- \to \mu^+\mu^-$	154479	2.27%
$e^+e^- \to e^+e^-\pi^+\pi^-$	67802	36.49%
$e^+e^- \to e^+e^-e^+e^-$	59946	10.62%
$e^+e^- \to e^+e^-(Bhabha)$	56749	2.54%
$e^+e^- \to K^0\bar{K}^0(\gamma)$	104	40.57%
$\sum$	5117223	86.59%

(b) Overview of the SM processes remaining after the performed selection on Collective Background. The event count of the processes is not weighted on their luminosity.

SM Background Process	Events	rejected background
Continuum	5137734	78.41%
$e^+e^- \to \tau^+\tau^-$	3119249	71.14%
$e^+e^- \to \mu^+\mu^-\mu^+\mu^-$	1122954	0.85%
$e^+e^- \to e^+e^-\mu^+\mu^-$	926682	14.67%
$B\overline{B}\,BKG$	307249	67.33%
$e^+e^- \to \mu^+\mu^-$	157631	0.28%
$e^+e^- \to e^+e^-\pi^+\pi^-$	89430	16.24%
$e^+e^- \to e^+e^-e^+e^-$	64642	3.61%
$e^+e^- \to e^+e^-(Bhabha)$	57924	0.52%
$e^+e^- \to K^0\bar{K}^0(\gamma)$	132	24.57%
$\sum$	10983627	71.22%

Signals for various Masses

Additionally, to ensure that the selections do not unintentionally also remove other signal instances when discarding events based on a given anomaly score criterion, different configurations of the Dark Higgs and Dark χ_1 masses are now taken into account. This serves to verify that an undetected signal is not affected by the selections and is not completely removed. The overview provided in Tab. 4.14 illustrates the signal efficiencies obtained through the selection on the distribution of the IForest trained on the collective background at an anomaly score value of 0.45. It indicates that the removal of different signals based on the variation of their masses is not observed due to the overall high signal efficiency.

Table 4.14.: Overview of the signal efficiency for various masses regarding the strong selection on the collective SM sample distribution with the selection criterion of an $anomalyscore > 0.45$.

$m_{h'}$	m_{χ_1}	signal. eff.
0.25	0.25	0.998
0.25	0.5	0.999
0.5	0.5	0.999
0.25	1.0	0.997
0.5	1.0	0.997
1.0	1.0	0.996
0.25	1.5	0.995
0.5	1.5	0.996
1.0	1.5	0.997
1.5	1.5	0.995
0.25	2.0	0.994
0.5	2.0	0.993
1.0	2.0	0.993
1.5	2.0	0.993
2.0	2.0	0.993
0.25	2.5	0.988
0.5	2.5	0.990
1.0	2.5	0.993
1.5	2.5	0.996
2.0	2.5	0.997
2.5	2.5	0.998
0.25	3.0	0.994
0.5	3.0	0.995
1.0	3.0	0.995
1.5	3.0	0.997

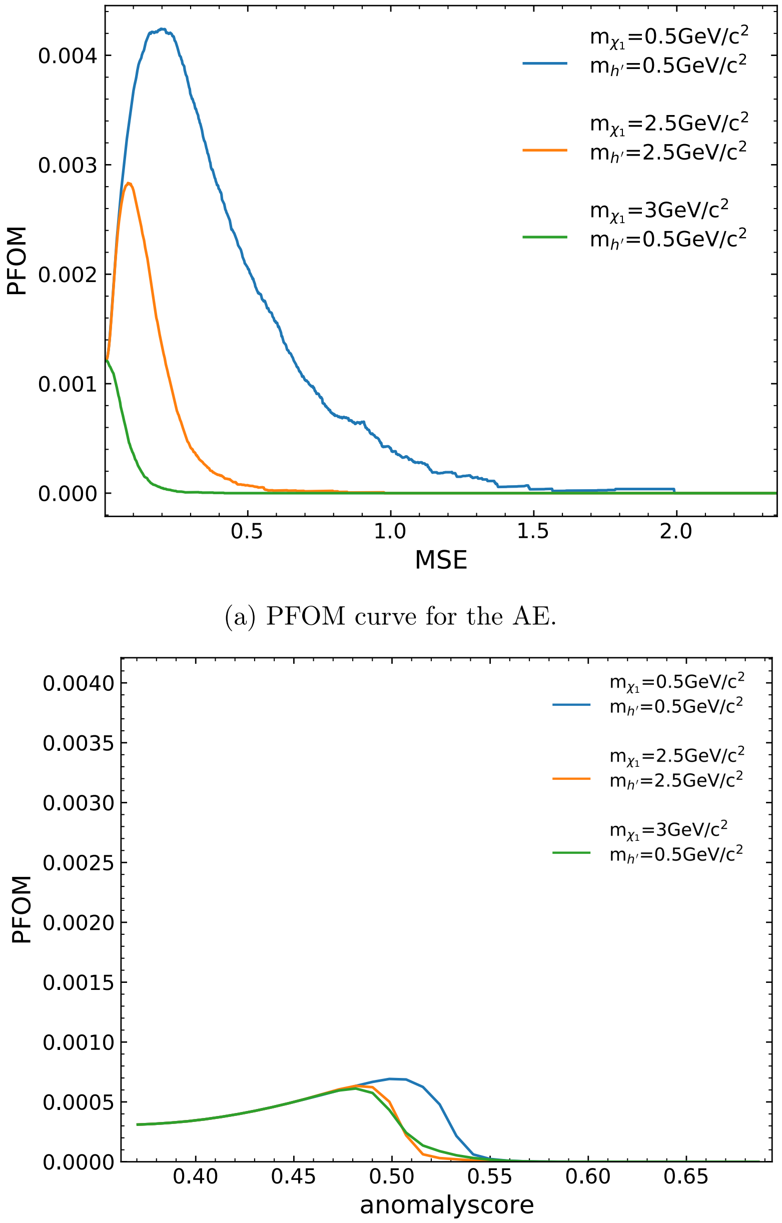
5. Comparison with Autoencoder

In this chapter, a comparison is drawn between Autoencoder (AE) and Isolation Forest (IForest), two methods used for Anomaly Detection (AD). The AE approach uses deep neural networks in an encoder-decoder architecture to encode samples into a reduced parameter space and decode them back to their original form. The discrepancy between the input and output is measured using the mean square error (MSE) metric, which serves as the Anomaly Score (AS). In the following, a comparison of the performance of two models for anomaly detection is made, focusing on the 8-dimensional Basic Autoencoder as the best model in the study [6].

There is a significant difference in sensitivity between the IForest and AE approaches. The AE shows much higher sensitivity in detecting the light mass signal compared to the best IForest model, which achieves only a quarter of the sensitivity. Despite the relatively lower sensitivity, the IForest result, which is characterized by its independence from signal parameters, is selected for further analysis. It is important to note that direct comparison between the two methods in this study is limited due to the use of different input formats and samples.

Apparent differences at the initial level are that the IForest algorithm is a more practical approach to anomaly detection due to the simplicity and efficiency of binary trees. Unlike AE, IForest does not require any preparation for sample reduction since the method works well on samples with only a few selections. A direct comparison of the PFOM curve (Fig. 5.1) with the corresponding anomaly scores, yield an order of magnitude difference in sensitivity between the two methods, further strengthening the case for the AE approach. However, the sensitivity differences between different signals are less pronounced for IForest than for AE. Furthermore, the sensitivity remains relatively constant for mass configurations (Fig. 5.2) in the range of $m_{h'} < 1.5$ and $m_{\chi_1} < 1.5$, indicating the independence from various model parameters. Remarkably, IForest exhibits significantly higher signal efficiency compared to AE, with a twofold difference for light masses and more than a threefold difference for heavy masses.

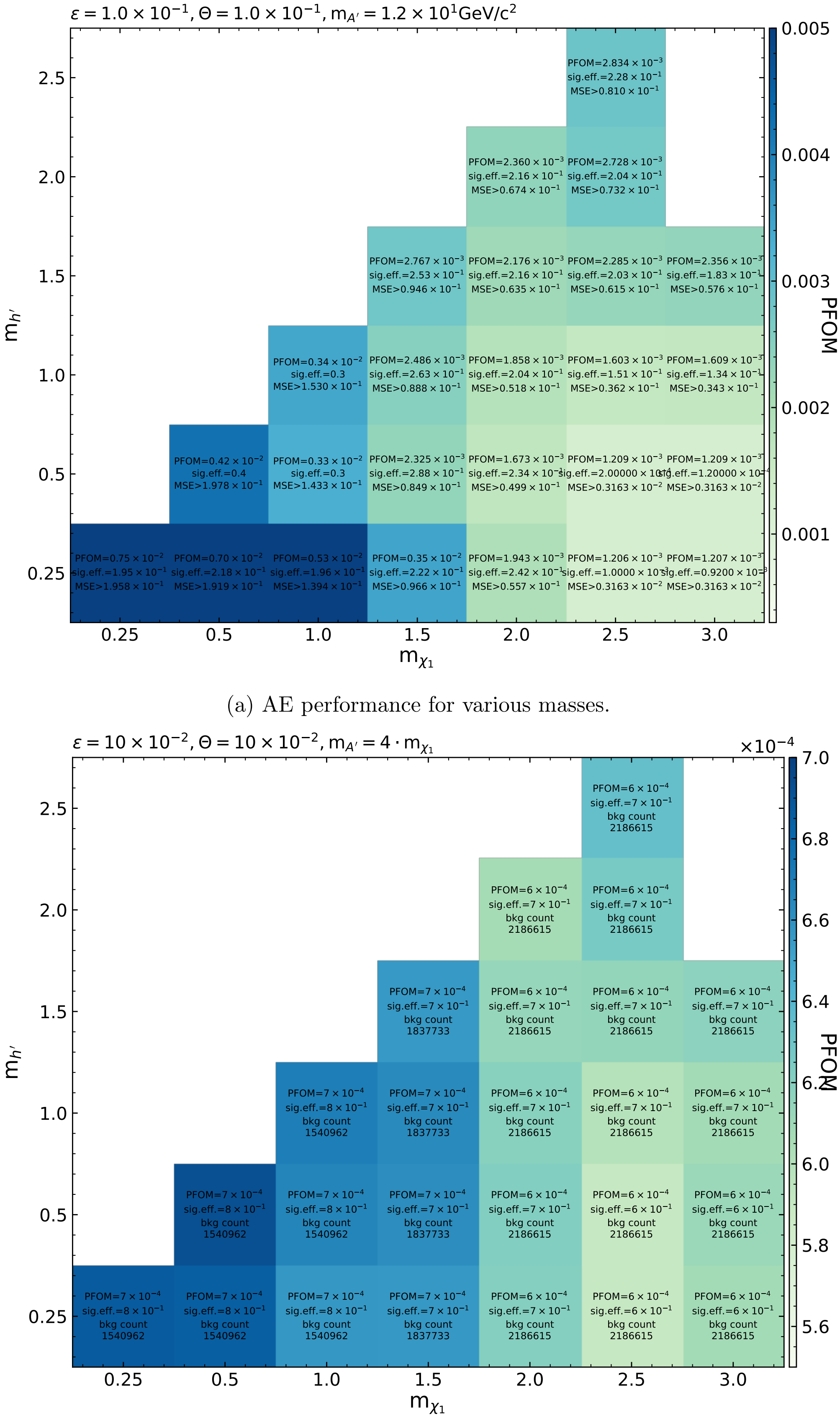
This comparison underlines the potential of IForest as a model-independent approach for the detection of Dark Higgs signals. In addition, the ability of IForest to process particle information by particle identification (PID) offers further advantages since the inherent limitations of the PID distribution are problematic for AE application. The AE exhibits higher sensitivity, as indicated by the maximal PFOM, making it a more suitable AD method compared to the IForest. Therefore, in the context of searching for unknown signals, the AE is considered a more appropriate approach due to its enhanced sensitivity.



(a) PFOM curve for the AE.

(b) PFOM curve for the IForest.

Figure 5.1.: Direct comparison of the PFOM for different selections applied using the corresponding AS for the AE (a) and the IForest (b).



(a) AE performance for various masses.

(b) IForest performance for various masses.

Figure 5.2.: Summary of the sensitivity and signal efficiency for the AE (a) and the IForest (b) considering the respective range of the maximum PFOM for variations of the masses $m_{h'}$ and m_{χ_1}.

6. Conclusions and Outlook

In this thesis, a comprehensive investigation of the Isolation Forest (IForest) method for Anomaly Detection (AD) is carried out in the context of searching for the Inelastic Dark Matter with a Dark Higgs (IDMDH) model at Belle II. The study focuses on training the IForest on Monte Carlo (MC) simulations of Standard Model (SM) processes that arise from prompt decays of electron-positron collisions. The IForest exhibits a certain level of variance in the anomaly score distribution due to its inherent randomness. An analysis of the hyperparameters reveals an underfitting behavior for small hyperparameter values while optimizing the hyperparameters improves the sensitivity of the computed Punzi Figure of Merit (PFOM). However, the model shows limited parameter dependence when the number of trees exceeds 100, and higher numbers of trees lead to stabilized sensitivity results.

The influence of different input features on anomaly detection with the IForest is explored, highlighting the effectiveness of the missing four-vector and transversal momentum in classifying signals as abnormal. Incorporating global PIDs with particle type information further enhances sensitivity for signals with heavy mass and significant mass differences, providing a model with signal parameter-independent sensitivity. To assess the overfitting, the training data is split into separate training and testing sets, revealing the stability of the IForest in capturing the behavior of anomaly score distributions for both seen and unseen samples, indicating the absence of overfitting.

An alternative approach, the Ensemble of Isolation Forests, where separate processes of the Standard Model are trained individually, does not generally improve the sensitivity in terms of the averaged anomaly score. The sensitivity enhancement highly depends on the input features used. Additional methods, such as iterative re-training of the IForest with selections for background efficiency reduction based on anomaly score distributions, either lead to reduced sensitivities with lower maximum PFOM due to similar instances of background and signal or yield similar maximum PFOM values for all tested signals.

In the final comparison with the Autoencoder (AE), the IForest achieves a level of sensitivity below that of the AE. This is evident from the significantly lower maximum PFOM achieved by the IForest, indicating a sensitivity of approximately one order of magnitude lower than that of the AE. However, the IForest demonstrates significantly better signal efficiency for various mass configurations than the AE.

Outlook

Overall, there are several ways to improve the IForest algorithm. By reviewing techniques and modifications as presented in [24] to improve anomaly detection. In addition, hybrid versions [25] of the IForest and nested IForest [5] as further developed methods can be implemented to address the challenges of over-density anomalies. These advances indicate that the potential of IForest as a method for discovering new physical phenomena has not yet been exhausted.

Bibliography

[1] L. Buitinck et al., "API design for machine learning software: experiences from the scikit-learn project," Sep, 2013. https://arxiv.org/abs/1309.0238.

[2] M. Crispim Romão, N. F. Castro, and R. Pedro, "Finding new physics without learning about it: anomaly detection as a tool for searches at colliders," *The European Physical Journal C* **81** no. 1, (2021) 27. https://doi.org/10.1140/epjc/s10052-020-08807-w.

[3] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: A review of anomaly detection techniques and recent advances," *Expert Systems with Applications* **193** (2022) 116429. https://www.sciencedirect.com/science/article/pii/S0957417421017164.

[4] Fei Tony Liu, Kai Ming Ting, Zhi-Hua Zhou, " Isolation Forest," p. 413–422. Eighth IEEE International Conference on Data Mining, 2008. doi:10.1109/ICDM.2008.17.

[5] J. Yanga, Y. Guoa, L. Cai, "Using a nested anomaly detection machine learning algorithm to study the neutral triple gauge couplings at an e^+e^- collider," Mar, 2022. https://arxiv.org/abs/2111.10543.

[6] J. Eppelt, "Anomaly Detection in Searches for Inelastic Dark Matter with a Dark Higgs," Master's thesis, Karlsruhe Institute of Technology (KIT), 2022.

[7] **SuperKEKB accelerator team**, K. Akai, K. Furukawa, and H. Koiso, "SuperKEKB collider," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **907** (Nov, 2018) 188–199. https://doi.org/10.1016%2Fj.nima.2018.08.017.

[8] Y. F. et al., "The superkekb has broken the world record of the luminosity," in *Proc. IPAC'22*, no. 13 in International Particle Accelerator Conference, pp. 1–5. JACoW Publishing, Geneva, Switzerland, 07, 2022. https://jacow.org/ipac2022/papers/moplxgd1.pdf.

[9] Z. Liptak et al., "Measurements of beam backgrounds in SuperKEKB Phase 2," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1040** (Oct, 2022) 167168. https://doi.org/10.1016%2Fj.nima.2022.167168.

[10] **The Belle II Collaboration**, T. Abe et al., "Belle II Technical Design Report," Nov, 2010. https://arxiv.org/abs/1011.0352.

[11] Y. Iwasaki, B. Cheon, E. Won, X. Gao, L. Macchiarulo, K. Nishimura, and G. Varner, "Level 1 trigger system for the Belle II experiment," *IEEE Trans. Nucl. Sci.* **58** (2011) 1807–1815.

[12] T. Kuhr, C. Pulvermacher, M. Ritter, and T. H. N. Braun, "The Belle II Core Software," *Computing and Software for Big Science* **3** no. 1, (Nov, 2018) 1. https://doi.org/10.1007/s41781-018-0017-9.

[13] A. Natochiia et al., "Beam background expectations for Belle II at SuperKEKB," Aug, 2022. https://arxiv.org/abs/2203.05731.

[14] P. Križan (for the Belle II Collaboration), "The Belle II Upgrade Program," Nov, 2022. https://arxiv.org/abs/2203.05731.

[15] G. Bertone, D. Hooper, "A History of Dark Matter," May, 2016. https://arxiv.org/abs/1605.04909.

[16] M. Duerr, T. Ferber, C. Garcia-Cely, C. Hearty, and K. Schmidt-Hoberg, "Long-lived dark Higgs and inelastic dark matter at Belle II," *Journal of High Energy Physics* **2021** no. 4, (Apr, 2021) . https://doi.org/10.1007%2Fjhep04%282021%29146.

[17] M. Goldstein, S. Uchida, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data,". https://doi.org/10.1371/journal.pone.0152173.

[18] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.* **41** no. 3, (Jul, 2009) . https://doi.org/10.1145/1541880.1541882.

[19] G. Kasieczka et al., "The LHC Olympics 2020: A community challenge for anomaly detection in high energy physics," *Reports on Progress in Physics* **84** no. 12, (Dec, 2021) 124201. https://doi.org/10.1088%2F1361-6633%2Fac36b9.

[20] T. Aarrestad et al., "The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider," Dec, 2021. https://arxiv.org/abs/2105.14027.

[21] Z. Ghahramani, "Unsupervised Learning," *Bousquet, O., von Luxburg, U., Rätsch, G. (eds) Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science(), vol 3176. Springer, Berlin, Heidelberg.* **3176** (2004) 1–2, 80.

[22] Sahand Hariri, Matias Carrasco Kind, Robert J. Brunner, "Extended Isolation Forest," Nov, 2018. https://arxiv.org/abs/1811.02141v3.

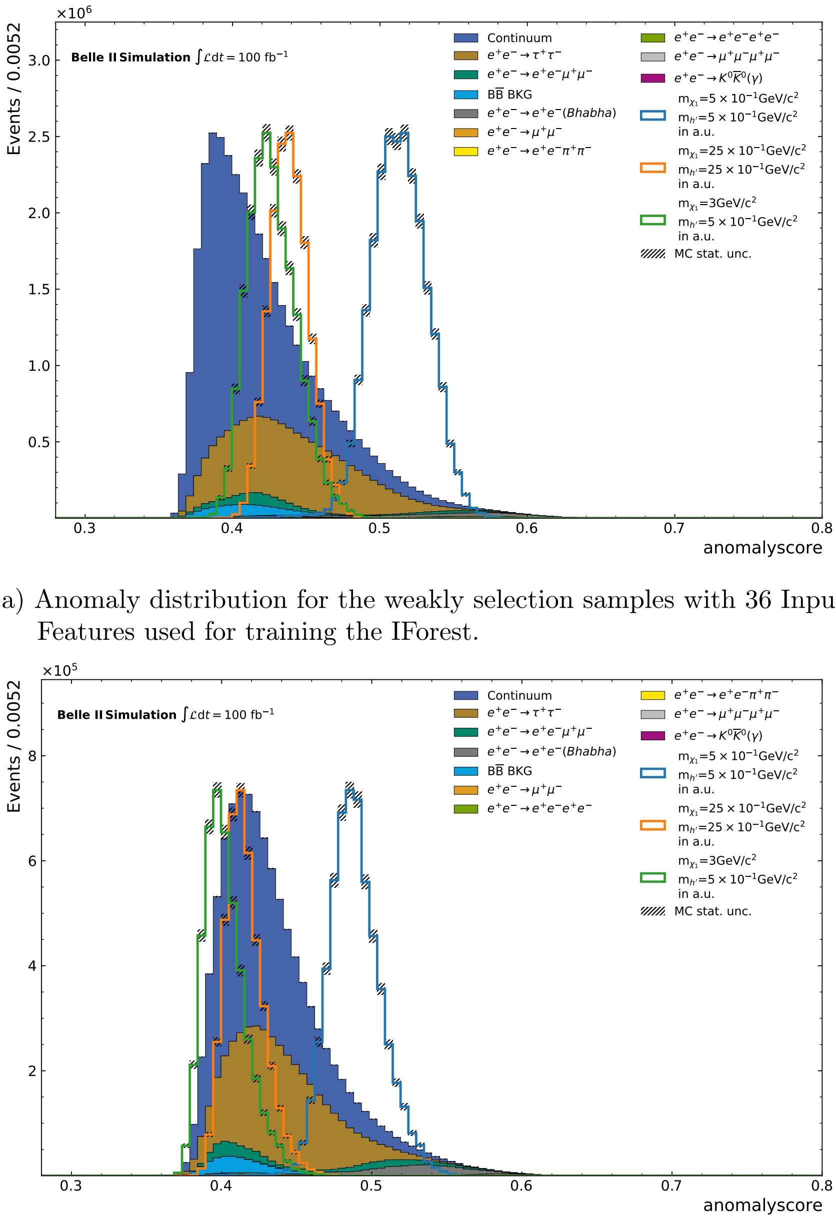
[23] G. Punzi, "Sensitivity of searches for new signals and its optimization," 2003. https://arxiv.org/abs/physics/0308063.

[24] Y. Chabchoub, M. U. Togbe, A. Boly, and R. Chiky, "An in-depth study and improvement of Isolation Forest," *IEEE Access* **10** (2022) 10219 – 10237. https://hal.science/hal-03537102.

[25] P. Marteau, S. Soheily-Khah, N. Béchet, "Hybrid Isolation Forest - Application to Intrusion Detection," May, 2017. https://arxiv.org/abs/1705.03800.

A. Appendix

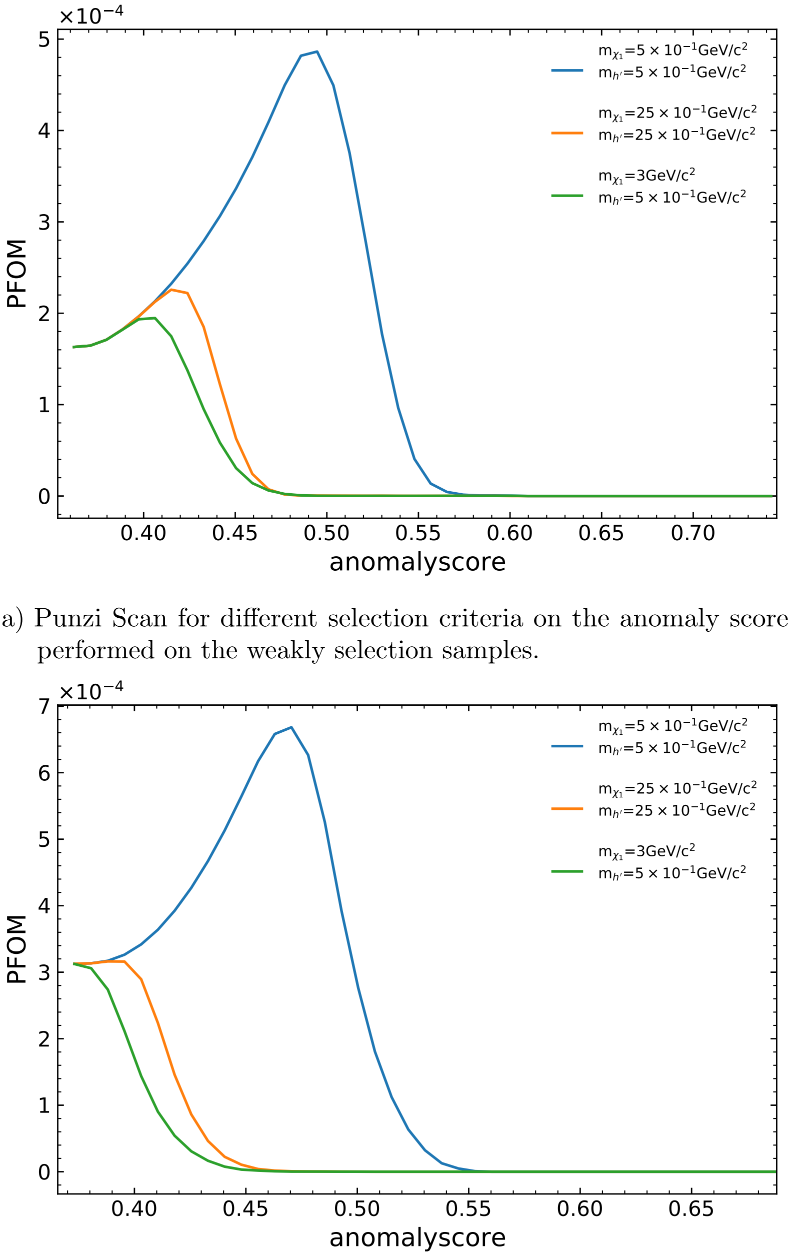
A.1. Plain Isolation Forest trained on additional features: $binaryPID(e, \mu)$, p_t, θ and ϕ



(a) Anomaly distribution for the weakly selection samples with 36 Input Features used for training the IForest.

(b) Anomaly distribution for the stricter selection samples with 36 Input Features used for training the IForest.

Figure A.1.: Usage of a different subset of input features on the weakly (a) and stricter (b) selection samples in a direct comparison of the anomaly score distribution. The effects on the Dark Higgs signals detected in regard to the anomaly score distribution are shown.

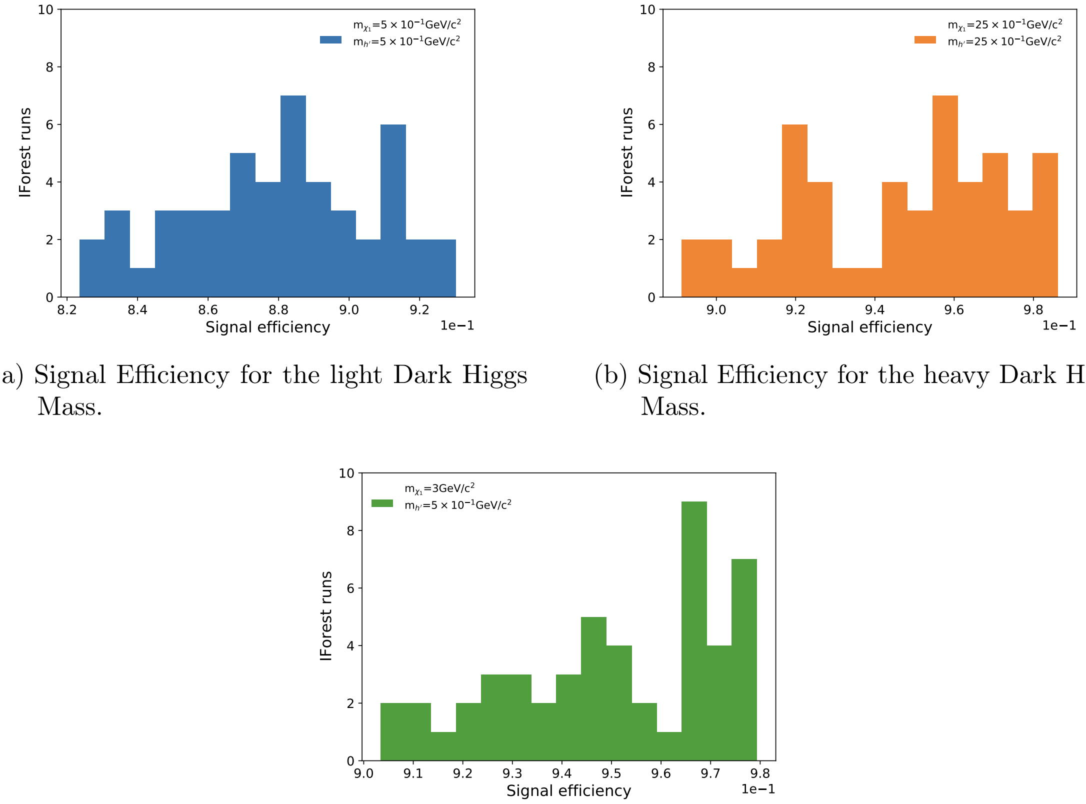


(a) Punzi Scan for different selection criteria on the anomaly score performed on the weakly selection samples.

(b) Punzi Scan for different selection criteria on the anomaly score performed on the stricter selection samples.

Figure A.2.: The PFOM for different selection criteria applied to the anomaly score. In which the separation between background and signal is evaluated for the weakly (a) and stricter (b) selection samples trained on more information given by the 36 input features.

A.2. Statistical fluctuations for the signal efficiency of the IForest



(a) Signal Efficiency for the light Dark Higgs Mass.

(b) Signal Efficiency for the heavy Dark Higgs Mass.

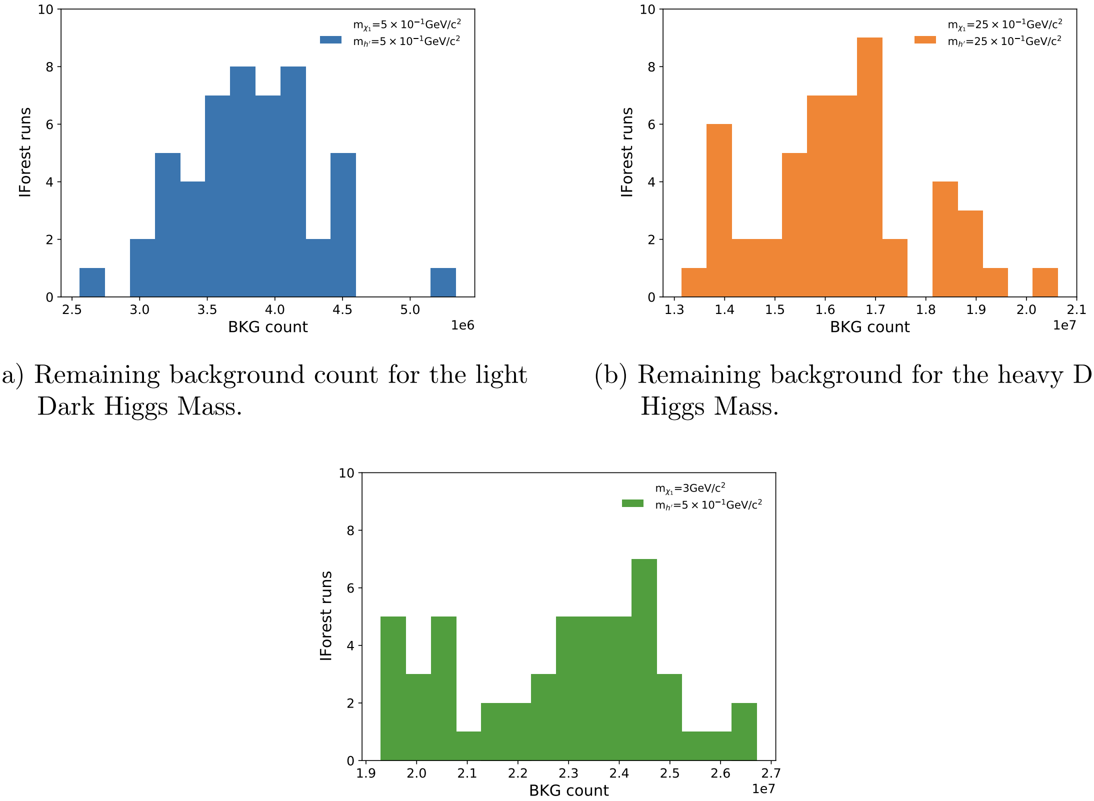
(c) Signal Efficiency for the strong splitting between Dark Higgs and Dark χ_1 Masses.

Figure A.3.: Fluctuations of the signal efficiency for exemplary chosen Signals trained 50 times on the same Isolation Forest. The distribution shows the unstable sensitivity given a larger fluctuation of the same Isolation Forest.

Table A.1.: Overview of the mean and standard deviation for the signal efficiency.

Signal Efficiency	mean	standard deviation
$m_{\chi_1} = 5 \times 10^{-1}\mathrm{GeV/c^2}$ $m_{h'} = 5 \times 10^{-1}\mathrm{GeV/c^2}$	0.880	0.027
$m_{\chi_1} = 25 \times 10^{-1}\mathrm{GeV/c^2}$ $m_{h'} = 25 \times 10^{-1}\mathrm{GeV/c^2}$	0.946	0.026
$m_{\chi_1} = 3\mathrm{GeV/c^2}$ $m_{h'} = 5 \times 10^{-1}\mathrm{GeV/c^2}$	0.950	0.022

A.3. Statistical fluctuations for the remaining background count of the IForest



(a) Remaining background count for the light Dark Higgs Mass.

(b) Remaining background for the heavy Dark Higgs Mass.

(c) Remaining background for the strong splitting between Dark Higgs and Dark χ_1 Masses.

Figure A.4.: Fluctuations of the remaining background count for exemplary chosen Signals trained 50 times on the same Isolation Forest. The distribution shows the unstable sensitivity given a larger fluctuation of the same Isolation Forest.

Table A.2.: Overview of the mean and standard deviation for the remaining background count.

Background Count	mean	standard deviation
$m_{\chi_1} = 5 \times 10^{-1}\mathrm{GeV/c^2}$ $m_{h'} = 5 \times 10^{-1}\mathrm{GeV/c^2}$	3812482	480163
$m_{\chi_1} = 25 \times 10^{-1}\mathrm{GeV/c^2}$ $m_{h'} = 25 \times 10^{-1}\mathrm{GeV/c^2}$	16283915	1621311
$m_{\chi_1} = 3\mathrm{GeV/c^2}$ $m_{h'} = 5 \times 10^{-1}\mathrm{GeV/c^2}$	22783374	1985134

A.4. Anomaly Score distribution for IForest

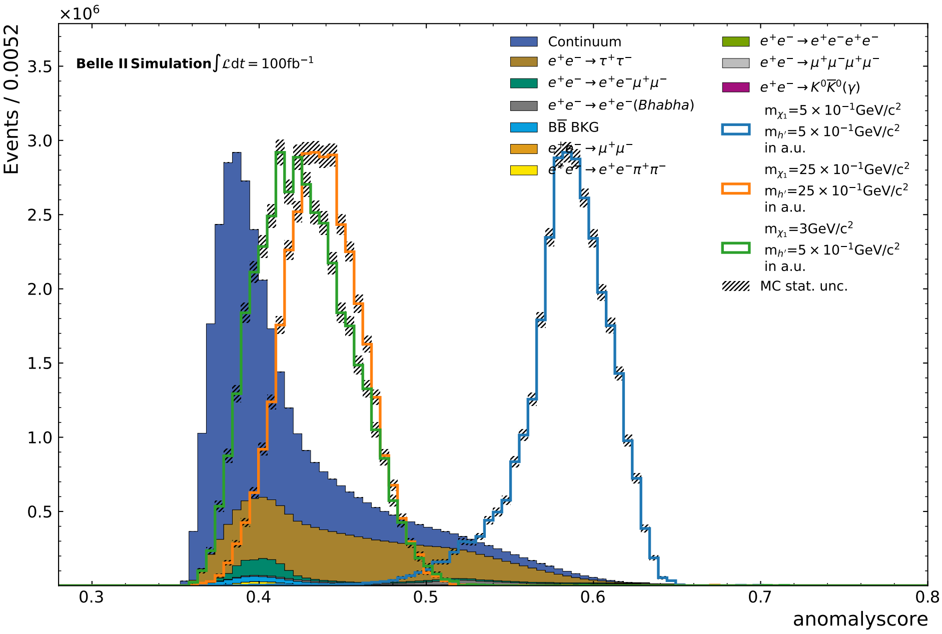


Figure A.5.: Distribution of the AS for the optimized Isolation Forest on the favorable Input Features (p_t, missing Four-vector).

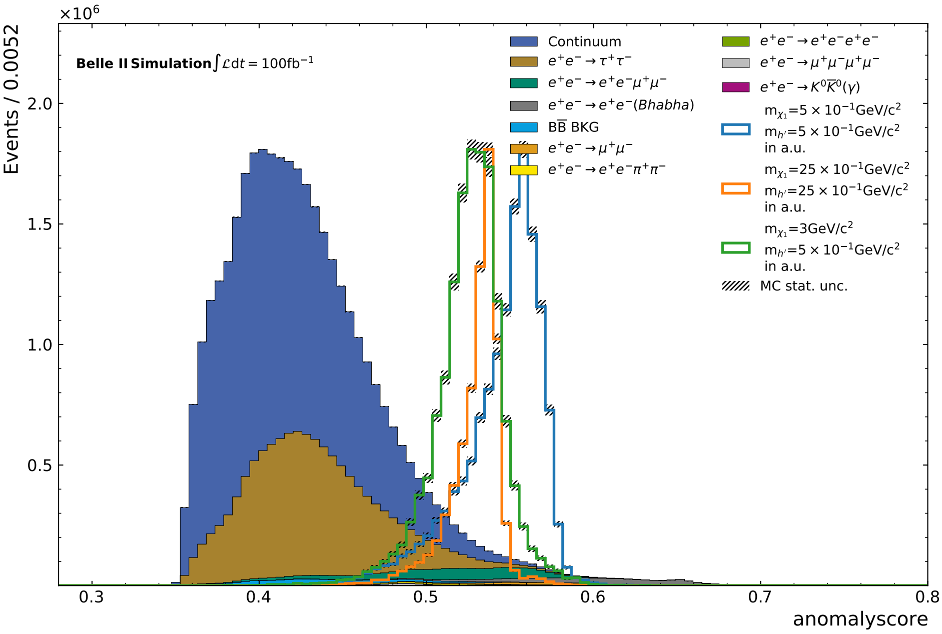


Figure A.6.: Distribution of the AS for the optimized Isolation Forest on the favorable Input Features (p_t, missing Four-vector) with PID information.

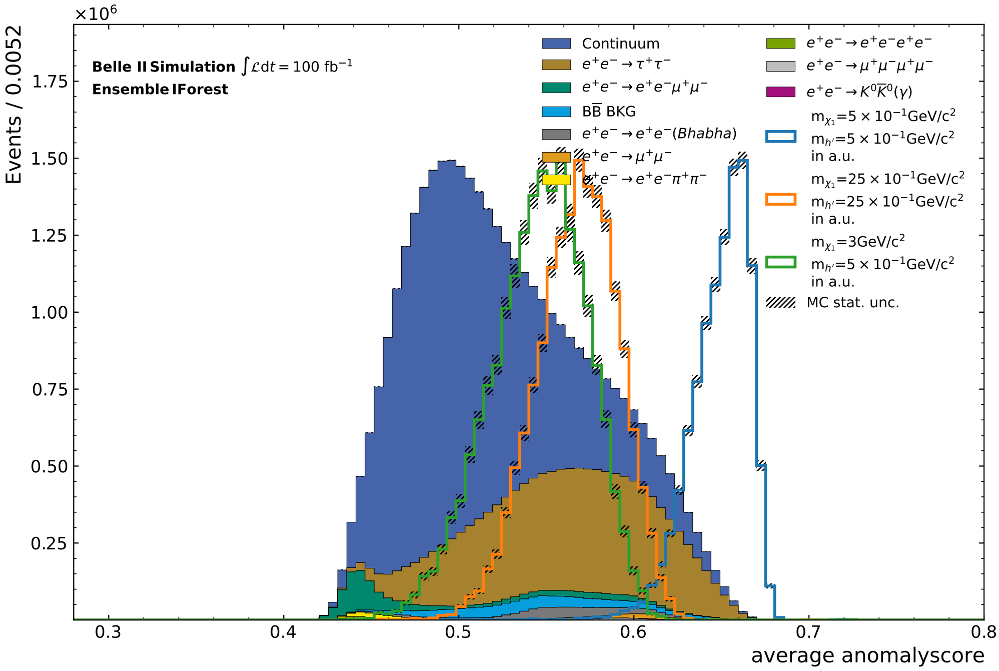


Figure A.7.: Distribution of the averaged AS for the Ensemble of Isolation Forests on the favorable Input Features (p_t, missing Four-vector).

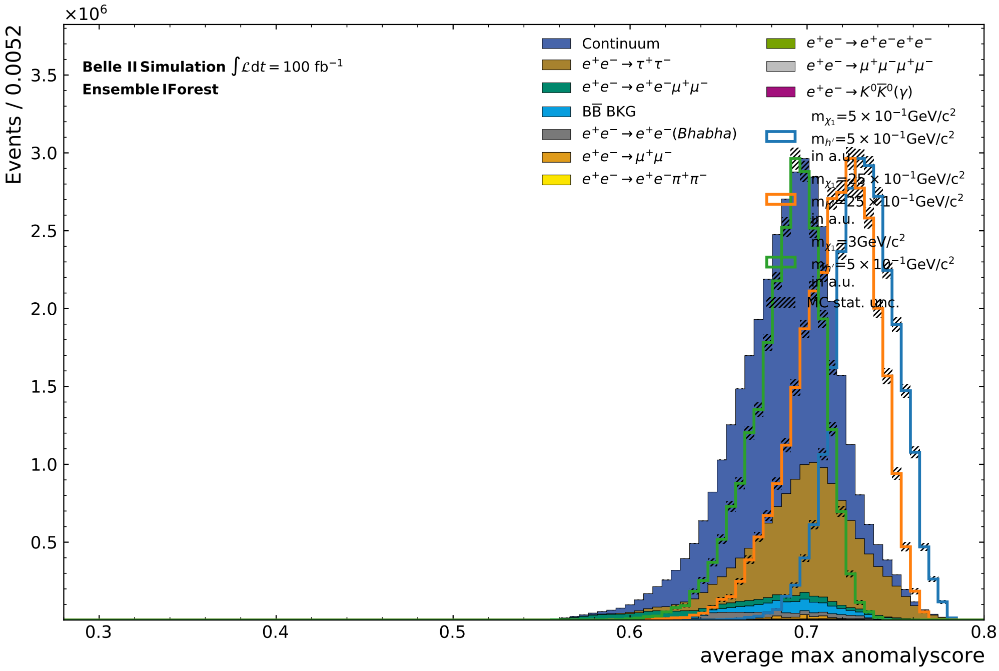


Figure A.8.: Distribution of the averaged maximal AS for the Ensemble of Isolation Forests on the favorable Input Features (p_t, missing Four-vector).

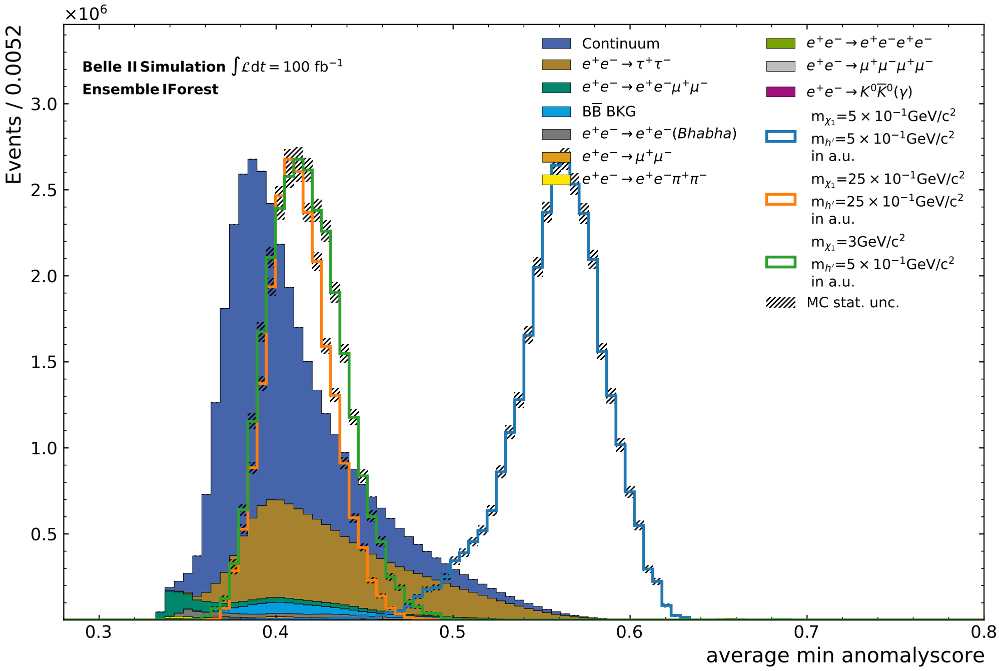


Figure A.9.: Distribution of the averaged minimal AS for the Ensemble of Isolation Forests on the favorable Input Features (p_t, missing Four-vector).

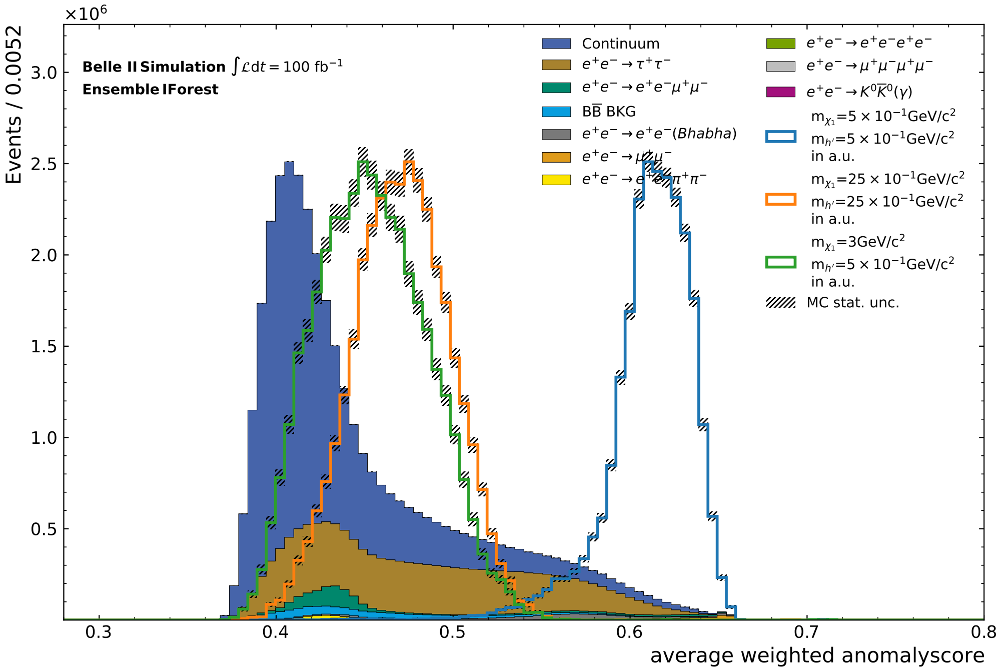


Figure A.10.: Distribution of the averaged AS weighted by sample size for the Ensemble of Isolation Forests on the favorable Input Features (p_t, missing Four-vector).

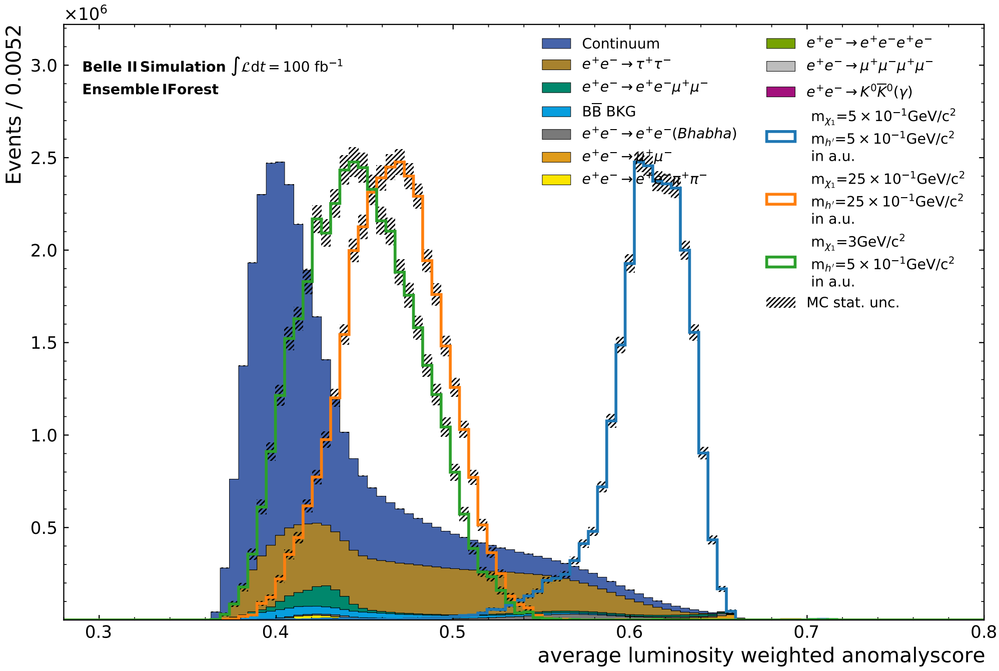


Figure A.11.: Distribution of the averaged AS weighted by luminosity for the Ensemble of Isolation Forests on the favorable Input Features (p_t, missing Four-vector).

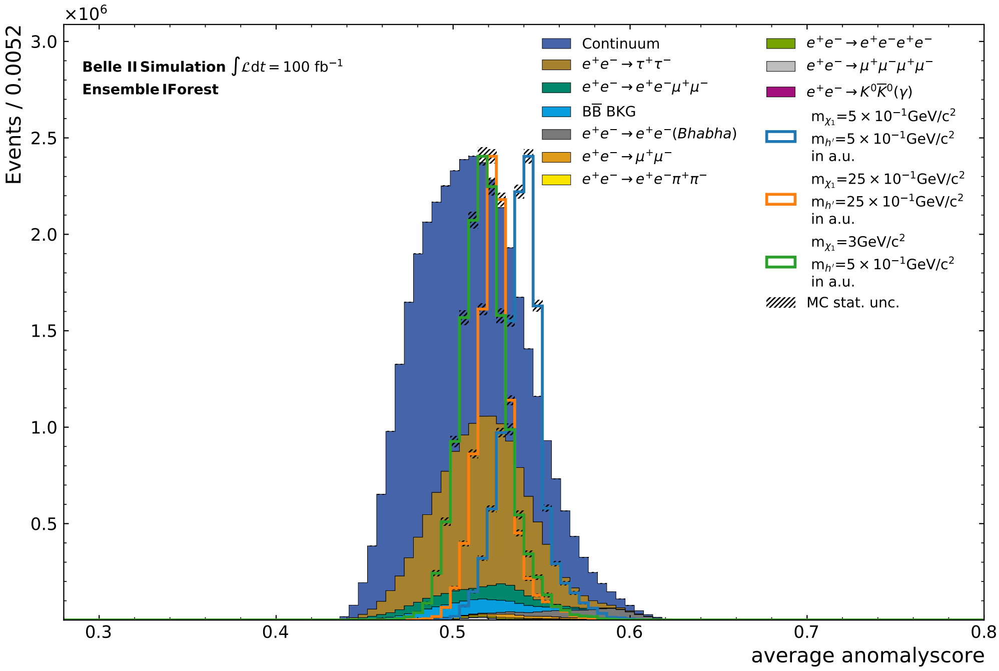


Figure A.12.: Distribution of the averaged AS for the Ensemble of Isolation Forests on the favorable Input Features (p_t, missing Four-vector) with PID information.

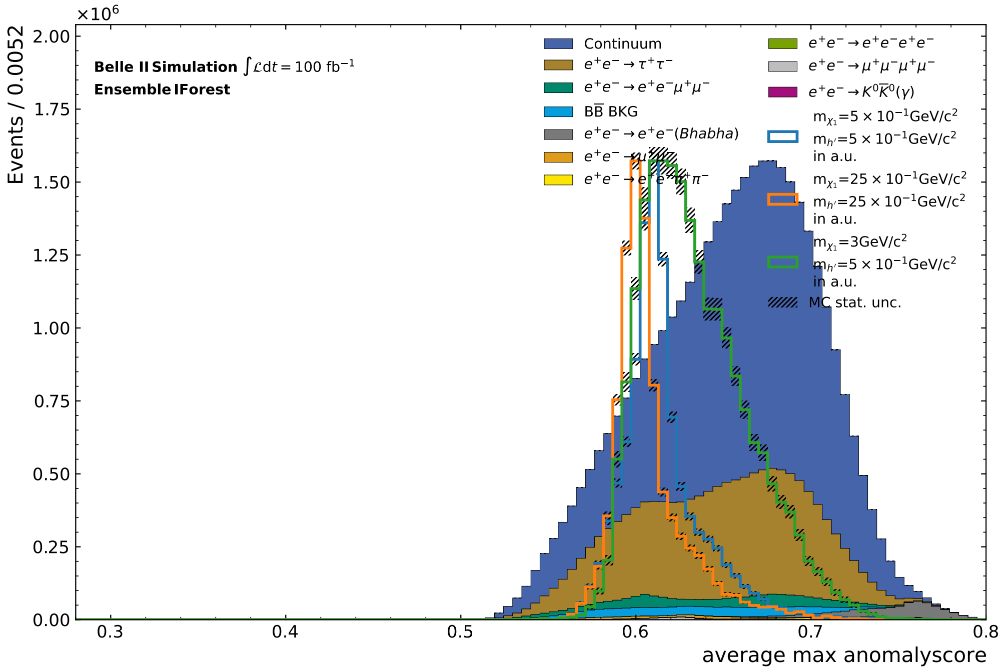


Figure A.13.: Distribution of the averaged maximal AS for the Ensemble of Isolation Forests on the favorable Input Features (p_t, missing Four-vector) with PID information.

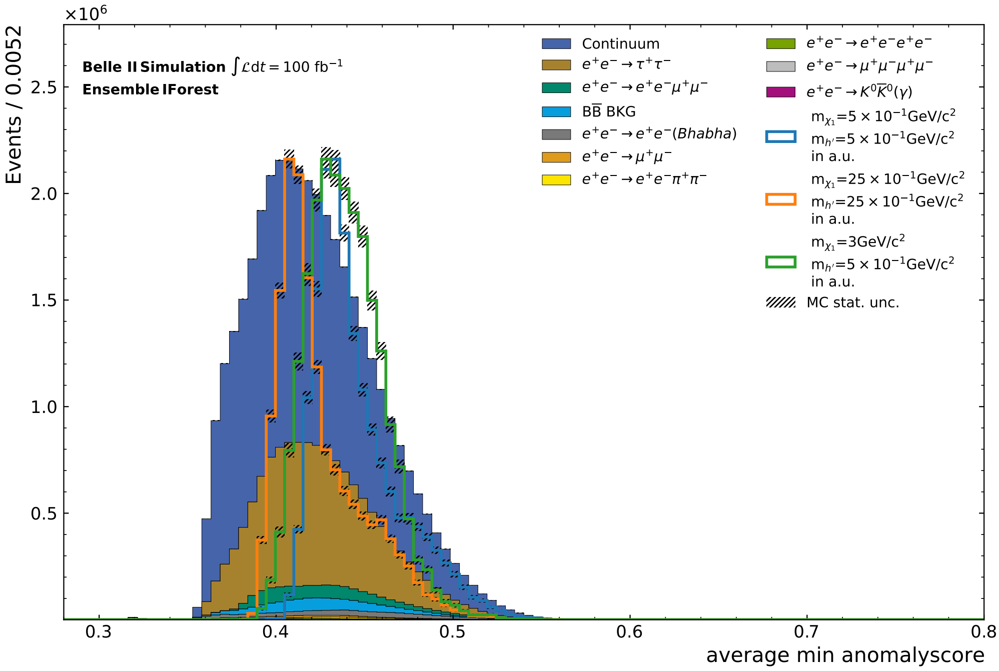


Figure A.14.: Distribution of the averaged minimal AS for the Ensemble of Isolation Forests on the favorable Input Features (p_t, missing Four-vector) with PID information.

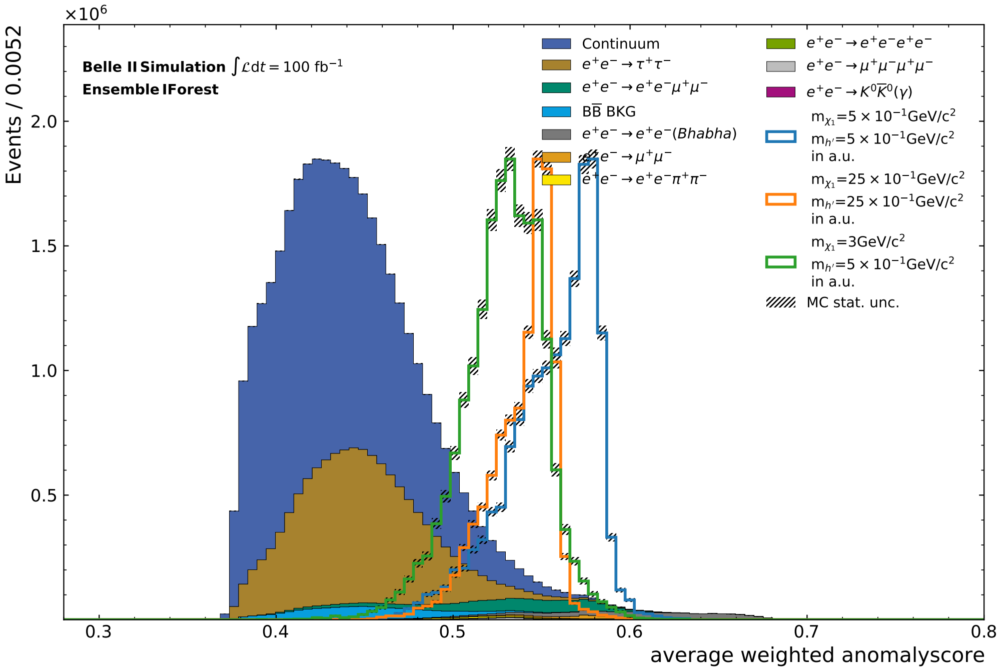


Figure A.15.: Distribution of the averaged AS weighted by sample size for the Ensemble of Isolation Forests on the favorable Input Features (p_t, missing Four-vector) with PID information.

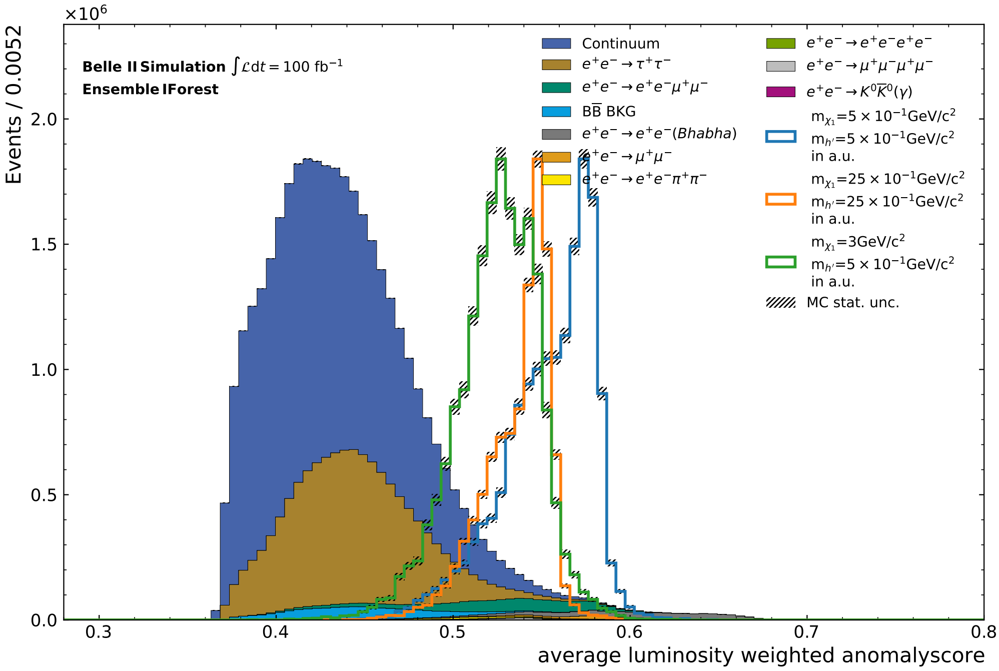


Figure A.16.: Distribution of the averaged AS weighted by luminosity for the Ensemble of Isolation Forests on the favorable Input Features (p_t, missing Four-vector) with PID information.

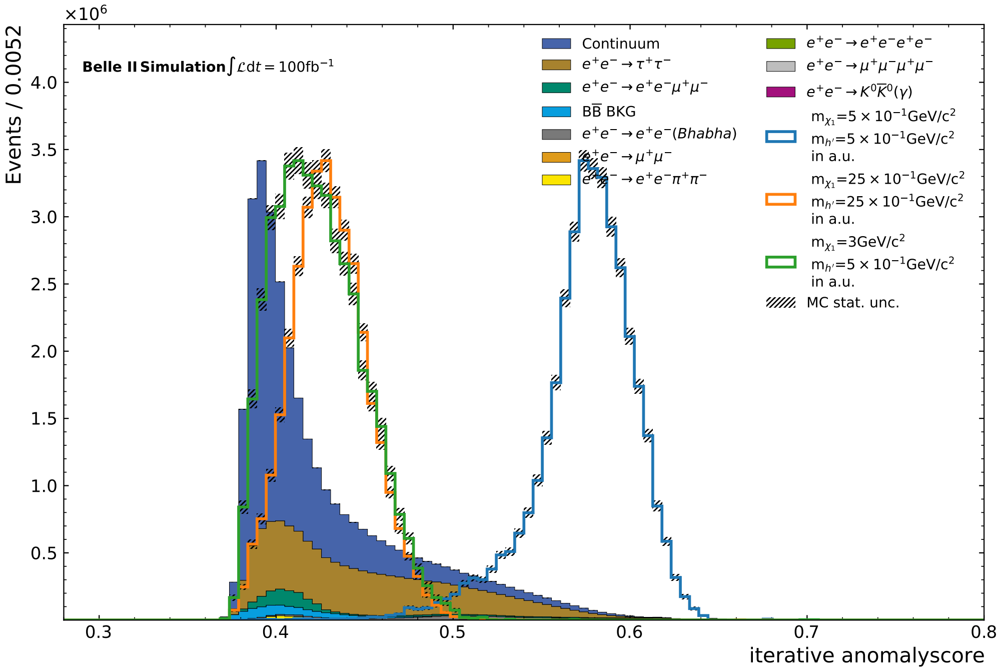


Figure A.17.: Iterative retrained IForest on the selection criterion of an anomaly score of 0.38 for the $e^+e^- \to e^+e^-\mu^+\mu^-$ process.

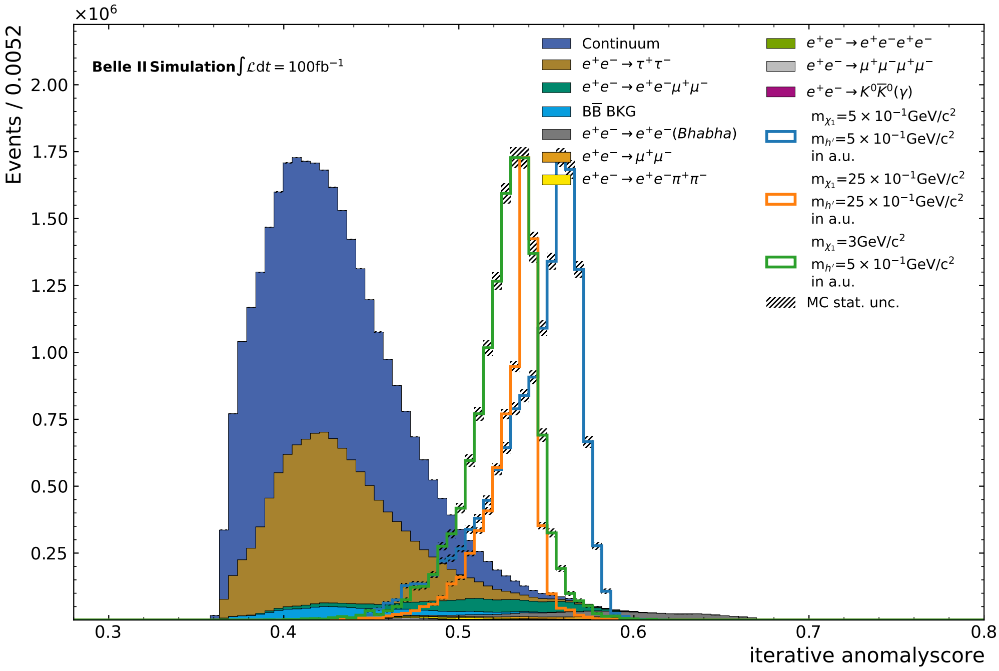


Figure A.18.: Iterative retrained IForest on the selection criterion of an anomaly score of 0.38 for the $e^+e^- \rightarrow e^+e^-\mu^+\mu^-$ process with PID information.

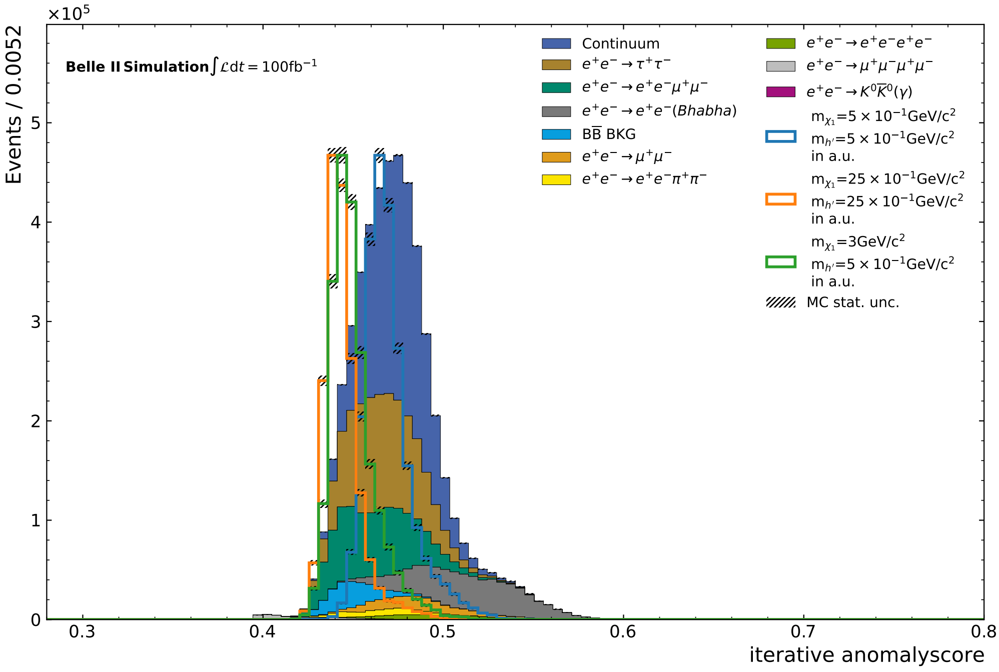


Figure A.19.: Iterative retrained IForest on the selection criterion of an anomaly score of 0.48 for the Continuum with PID information.

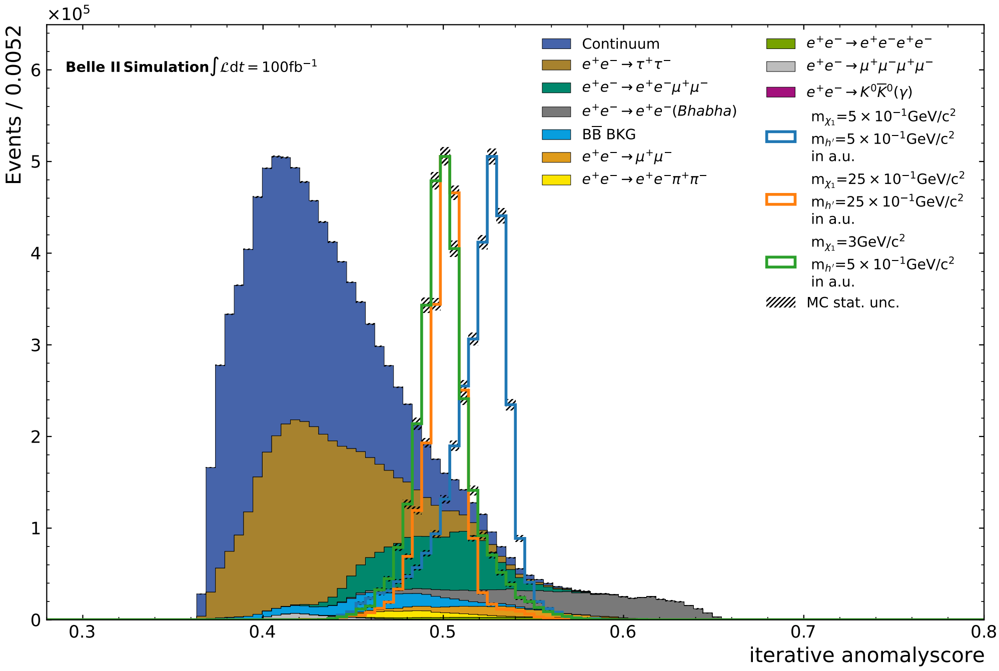


Figure A.20.: Iterative retrained IForest on the selection criterion of an anomaly score of 0.45 for all processes with PID information.

A.5. Model Impact Details of Hyperparameters

A.5.1. Model Impact Details of n_estimators parameter

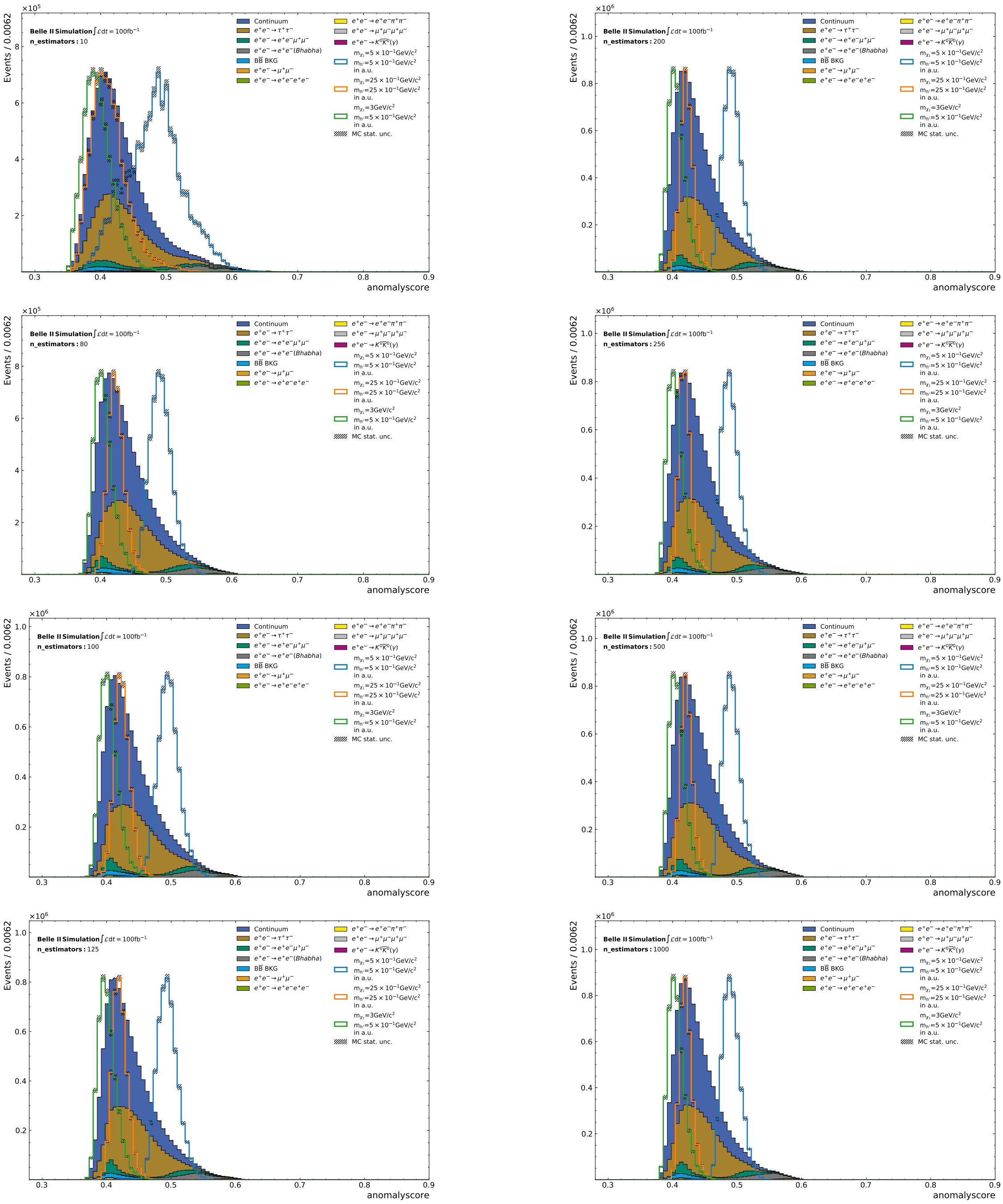


Figure A.21.: Evaluation of the Model Impact regarding the Hyperparameter $n\_estimators = 10 - 1000$ which corresponds to the number of trees of the Isolation Forest.

A.5.2. Model Impact Details of max_samples parameter

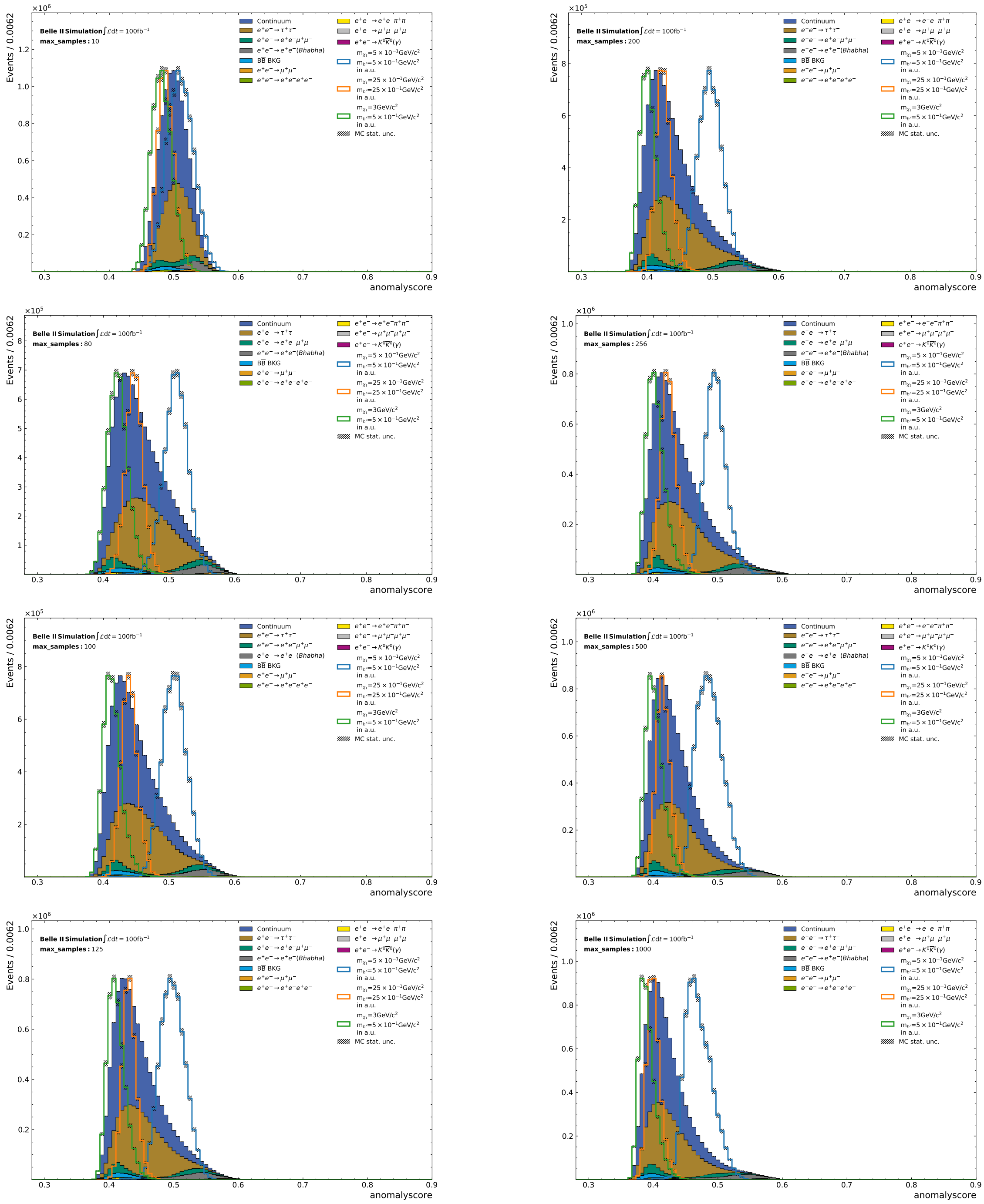


Figure A.22.: Evaluation of the Model Impact regarding the Hyperparameter *max_samples* = 10 − 1000 which corresponds to subsample size of the instances used for training each tree of the Isolation Forest

A.5.3. Model Impact Details of max_features parameter

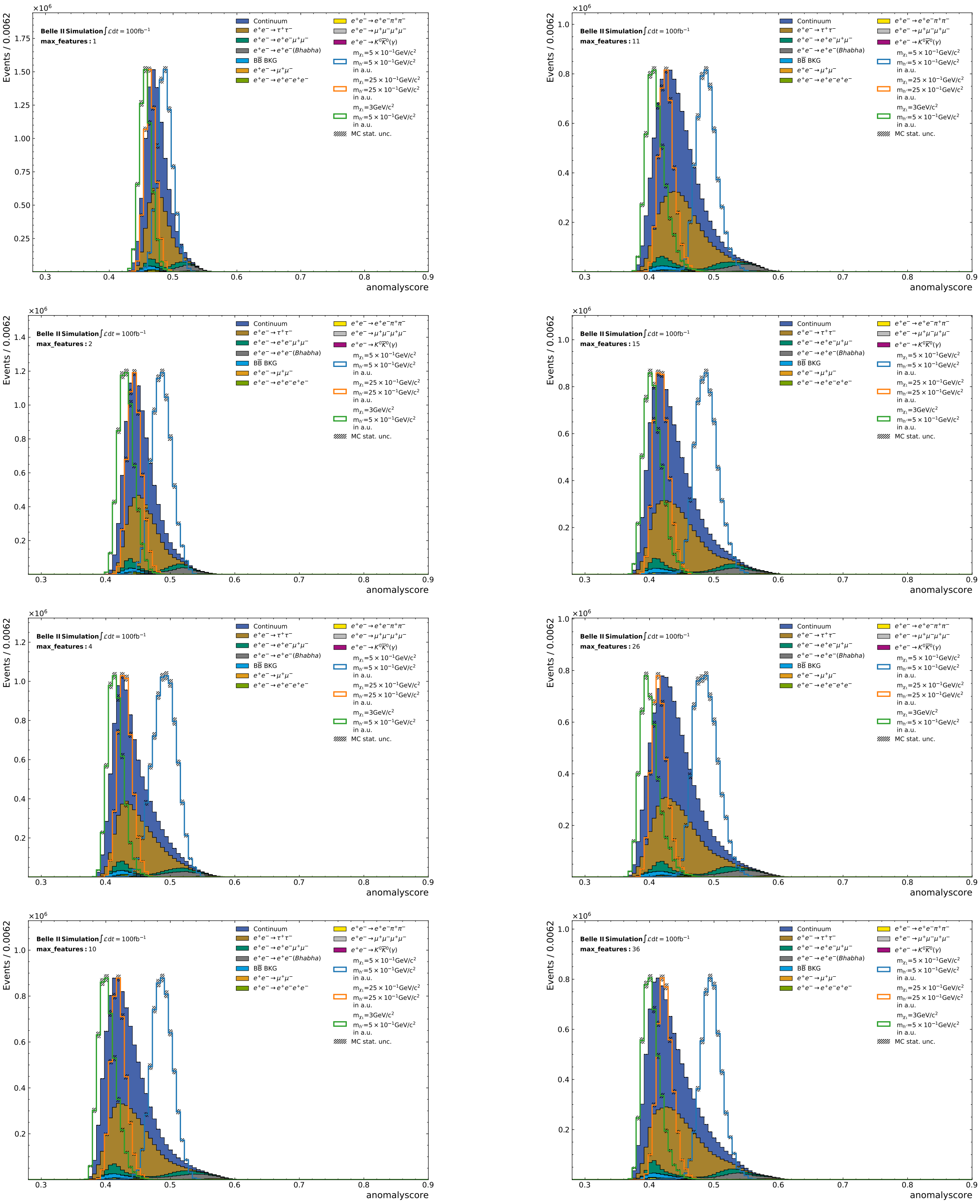


Figure A.23.: Evaluation of the Model Impact regarding the Hyperparameter $max\_features = 1-36$ which corresponds to the subset of features randomly chosen to use for the partitioning in each tree of the IForest.

A.6. Input Feature contribution to the distribution of the AS for the IForest

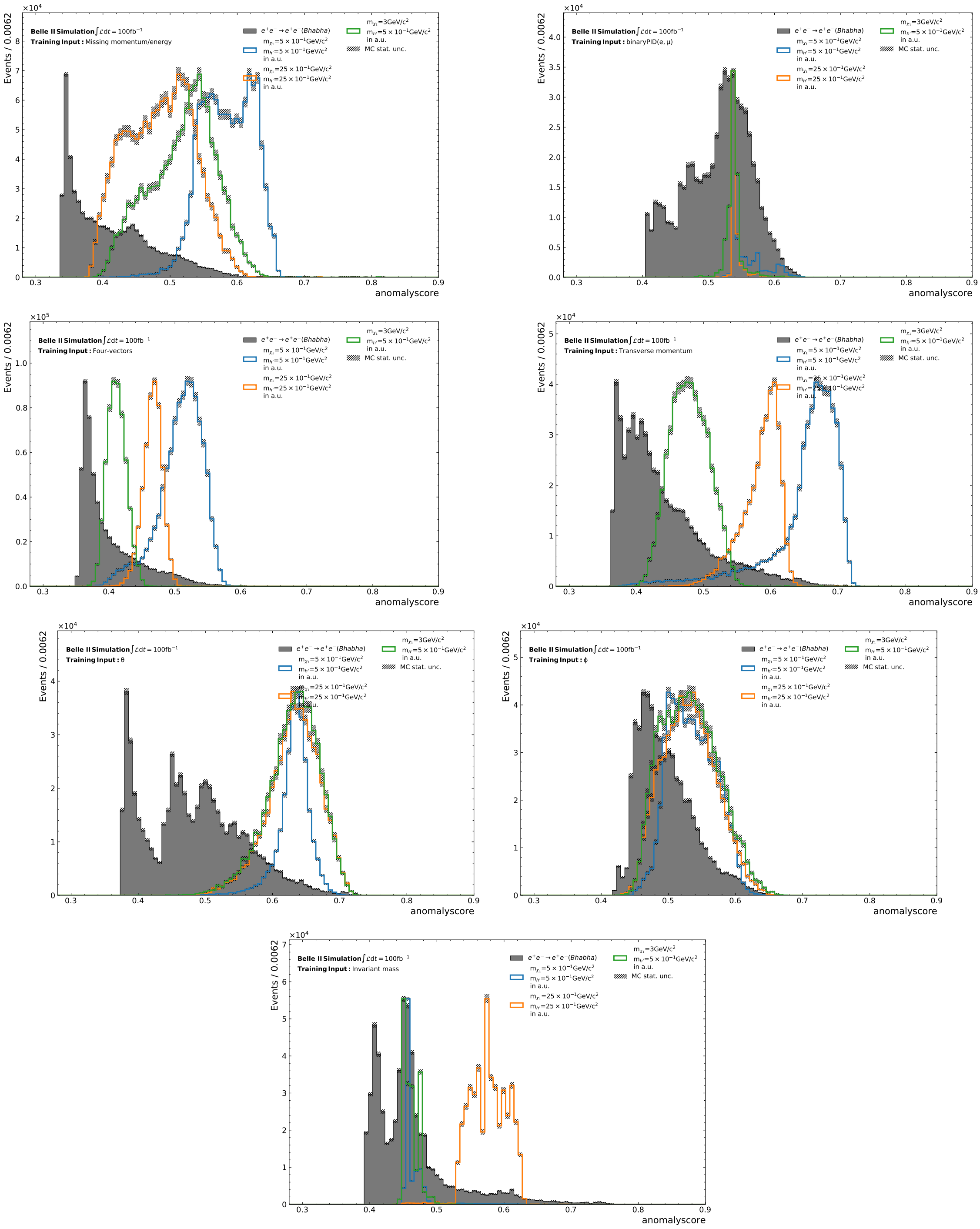


Figure A.24.: Contribution of each Input Feature group shown for the Bhabha process trained separately.

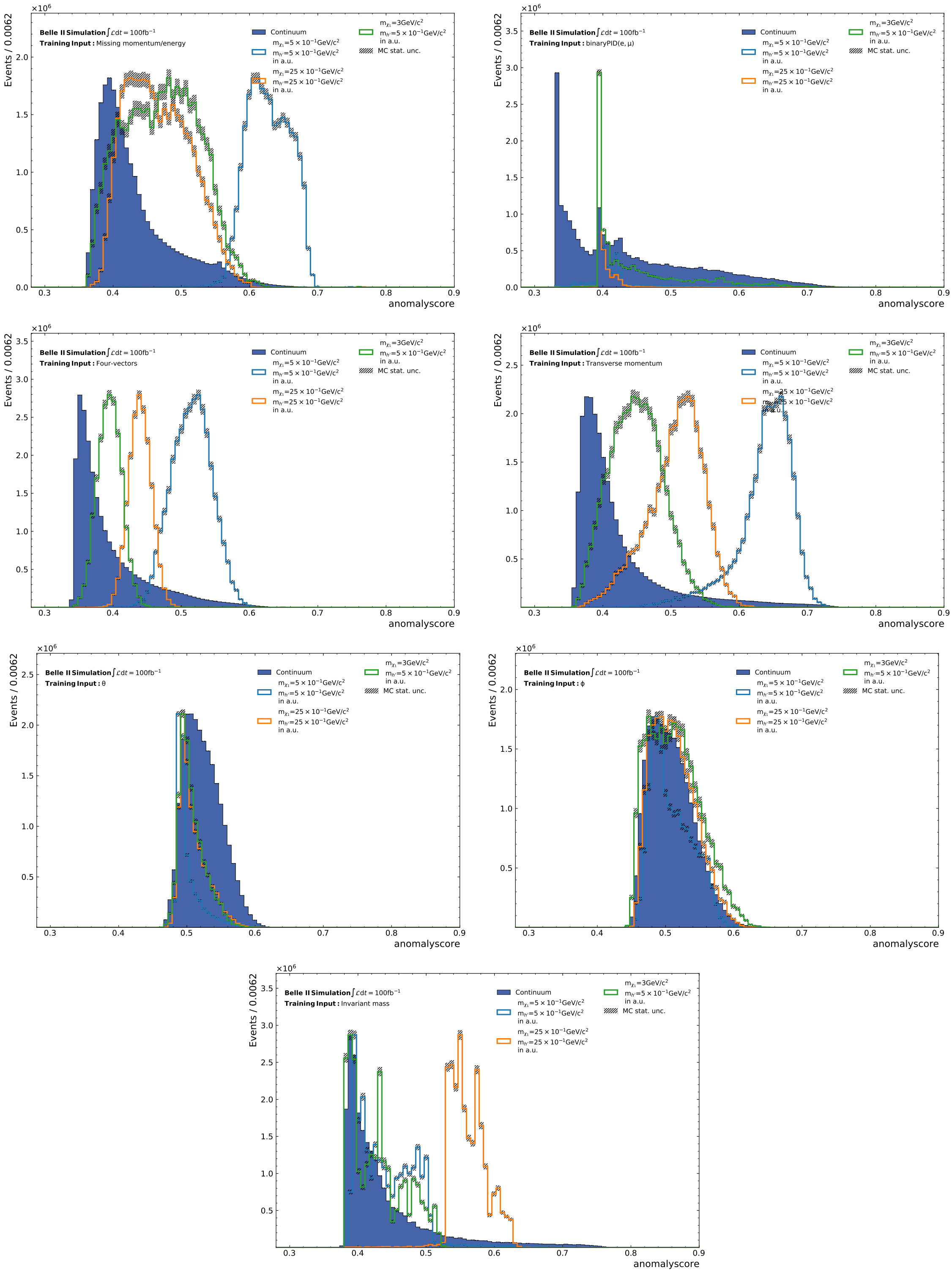


Figure A.25.: Contribution of each Input Feature group shown for the Continuum trained separately.

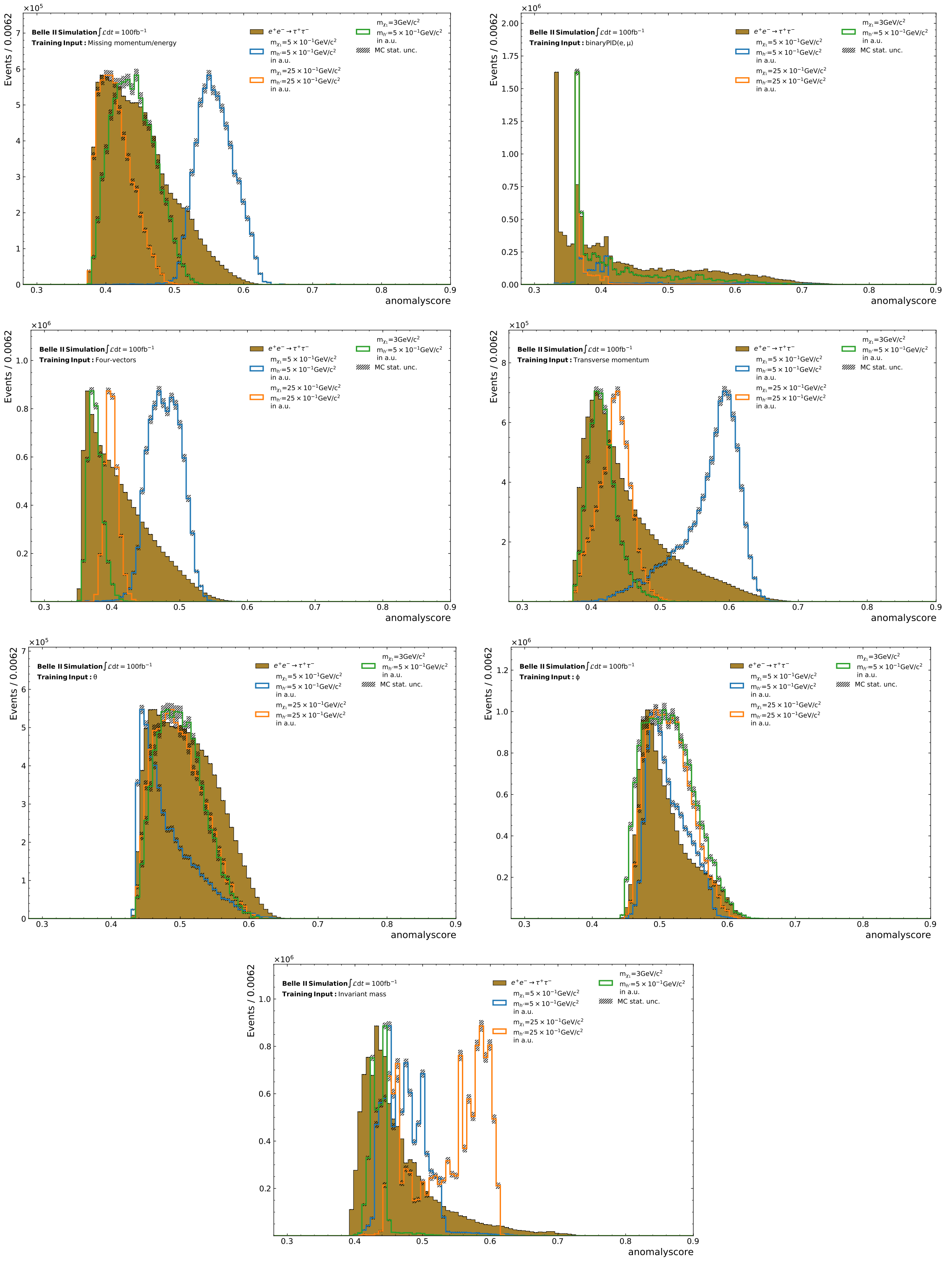


Figure A.26.: Contribution of each Input Feature group shown for the $e^+e^- \rightarrow \tau^+\tau^-$ process trained separately.

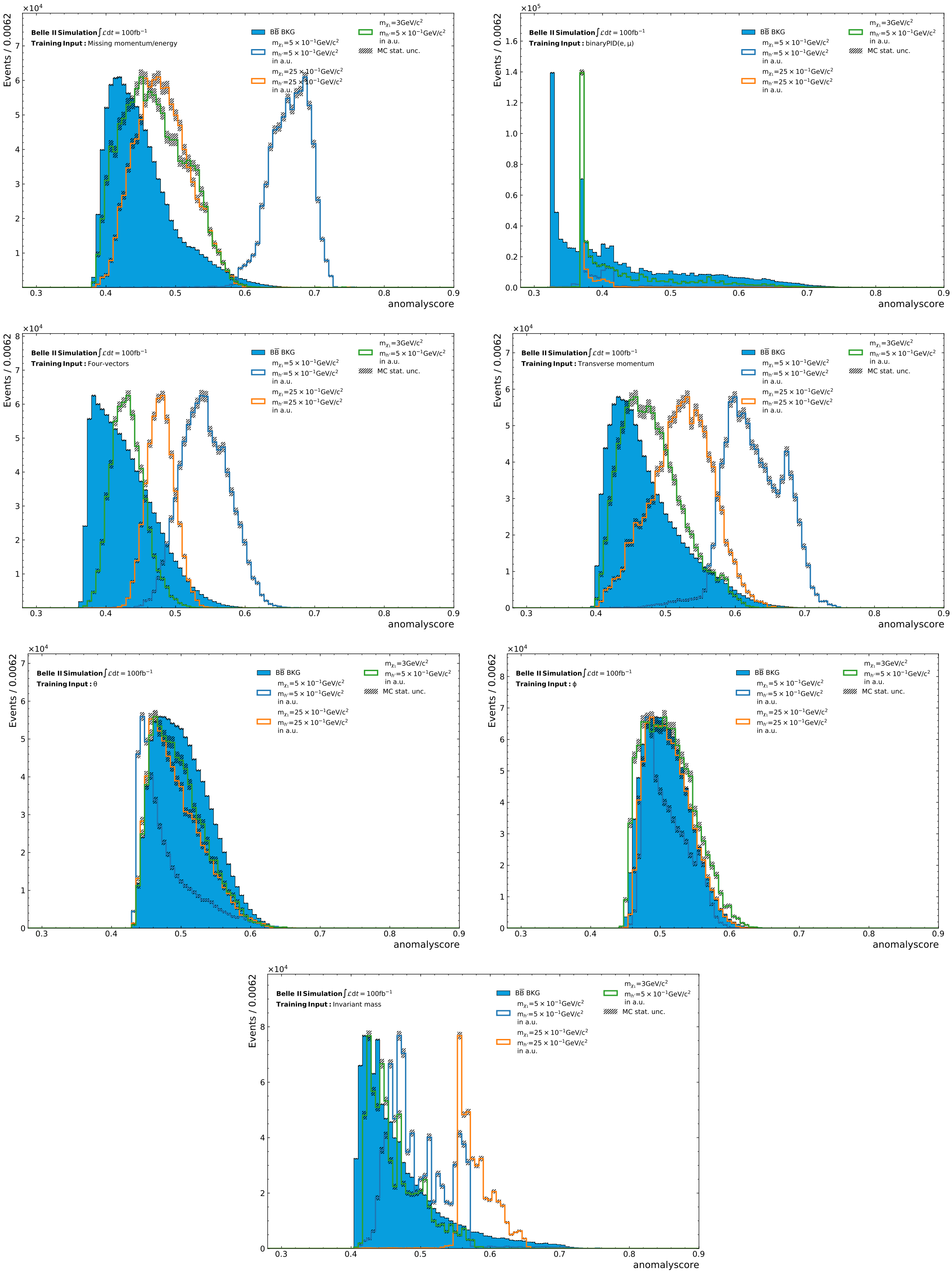


Figure A.27.: Contribution of each Input Feature group shown for the $B\overline{B}\,BKG$ process trained separately.

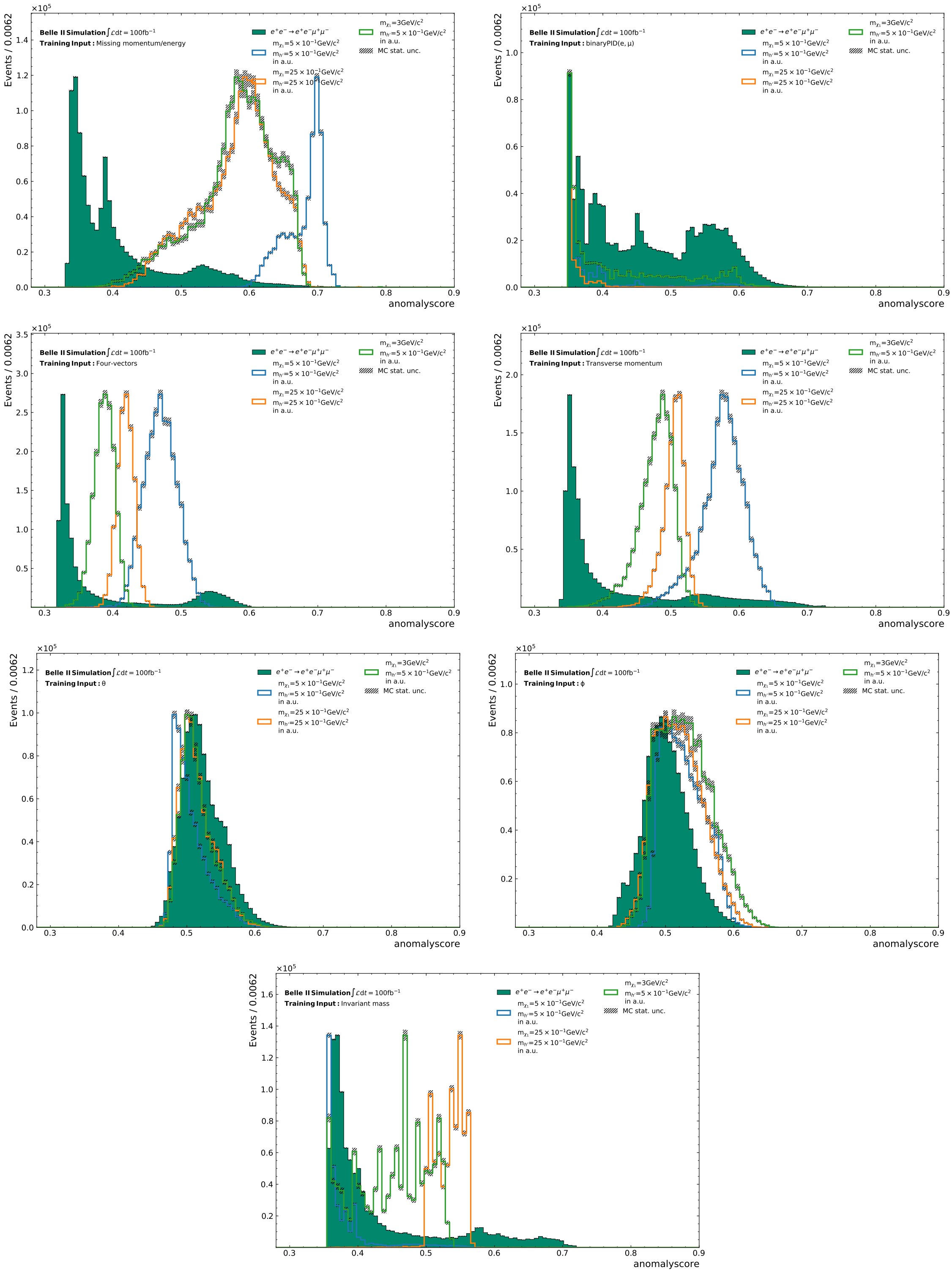


Figure A.28.: Contribution of each Input Feature group shown for the $e^+e^- \to e^+e^-\mu^+\mu^-$ process trained separately.

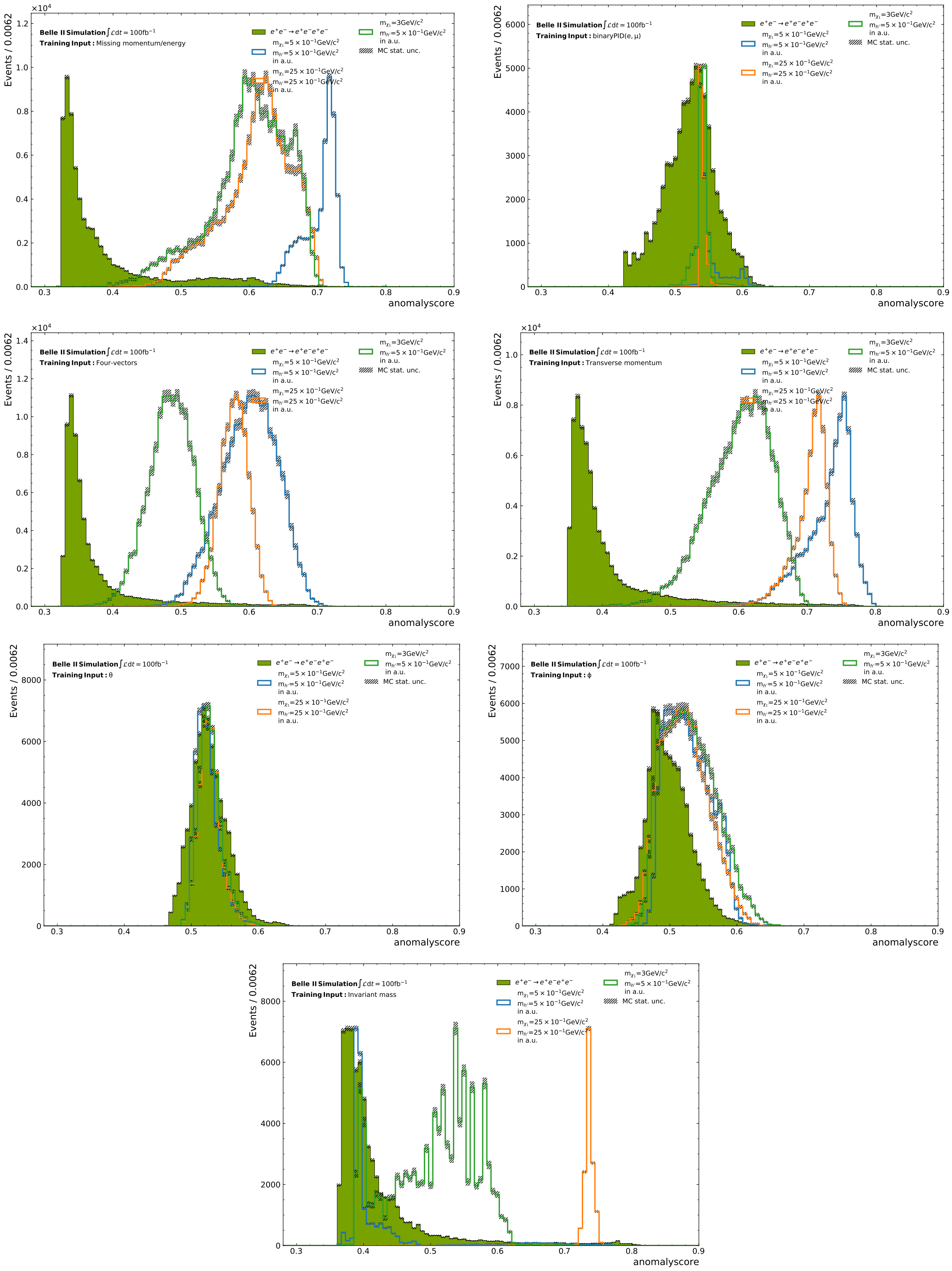


Figure A.29.: Contribution of each Input Feature group shown for the $e^+e^- \rightarrow e^+e^-e^+e^-$ process trained separately.

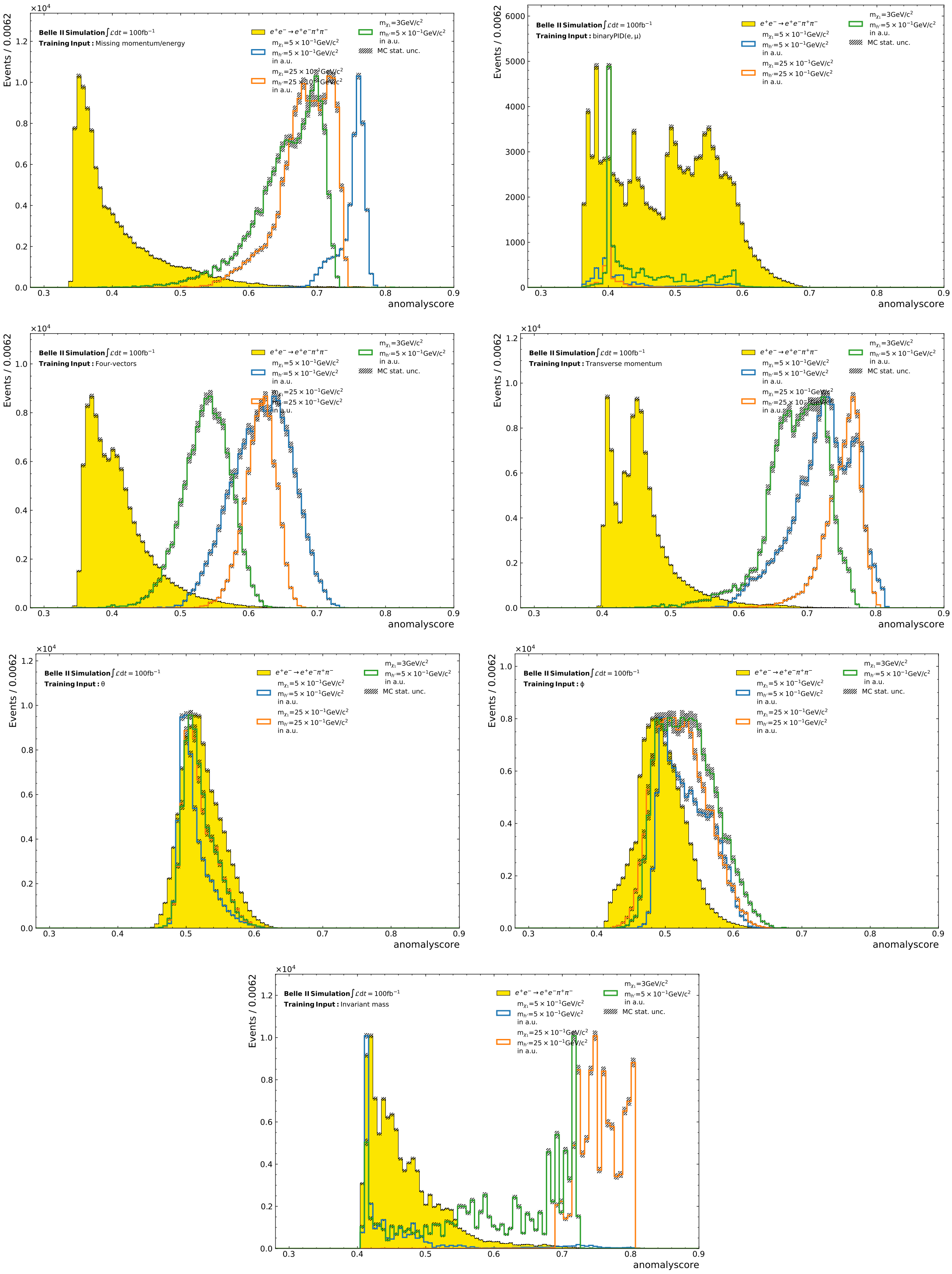


Figure A.30.: Contribution of each Input Feature group shown for the $e^+e^- \to e^+e^-\pi^+\pi^-$ process trained separately.

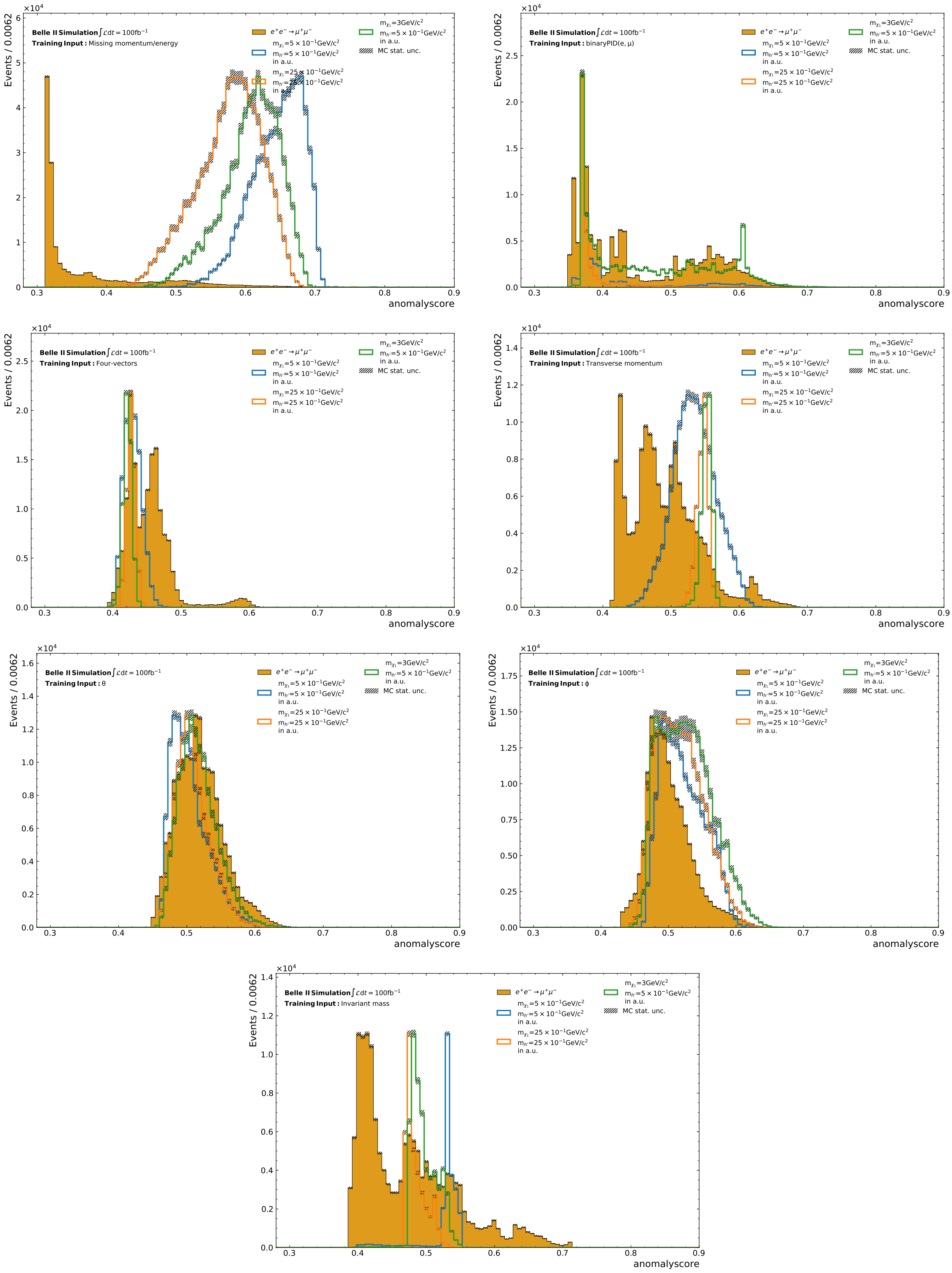


Figure A.31.: Contribution of each Input Feature group shown for the $e^+e^- \to \mu^+\mu^-$ process trained separately.

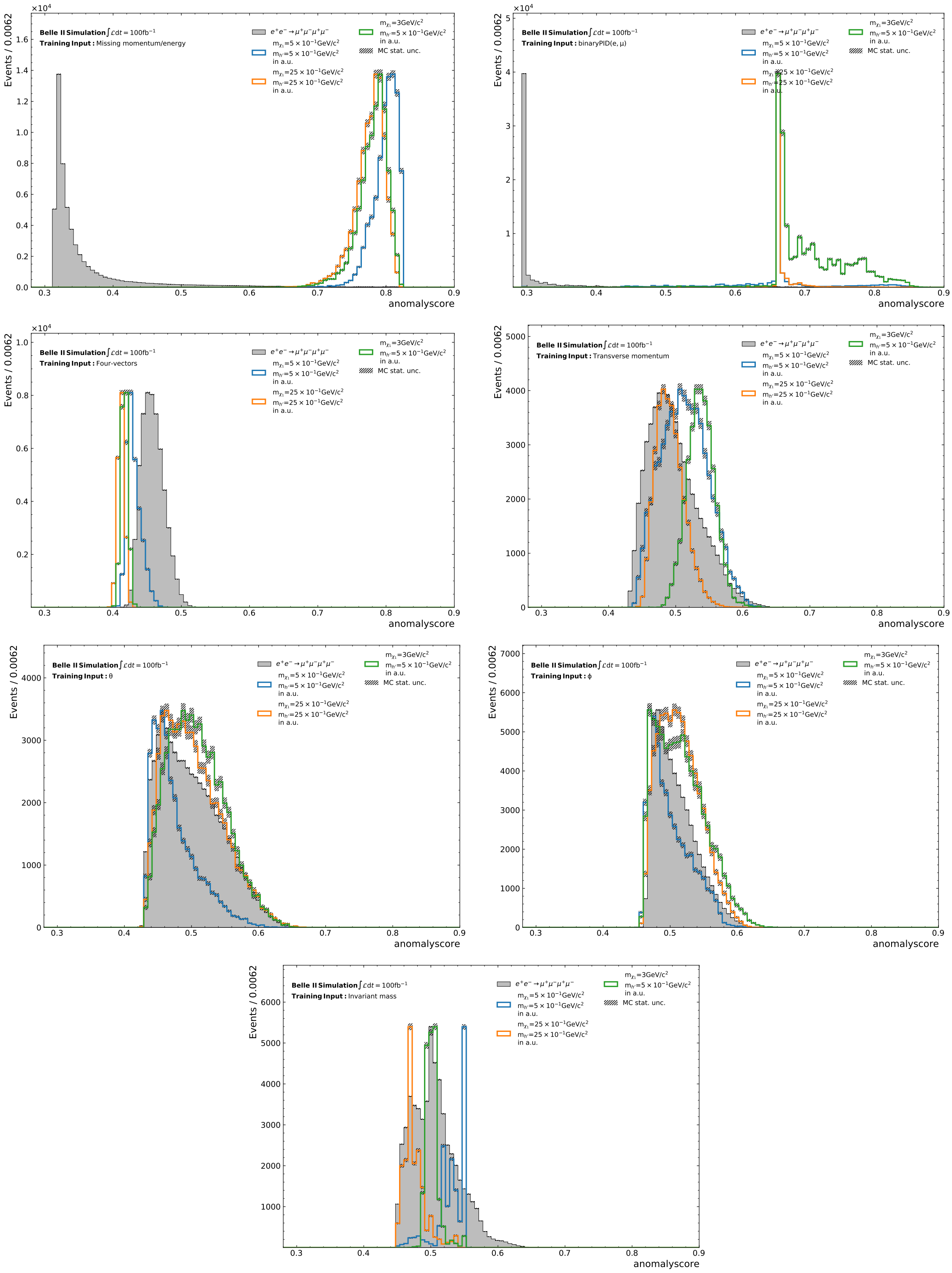


Figure A.32.: Contribution of each Input Feature group shown for the $e^+e^- \to \mu^+\mu^-\mu^+\mu^-$ process trained separately.

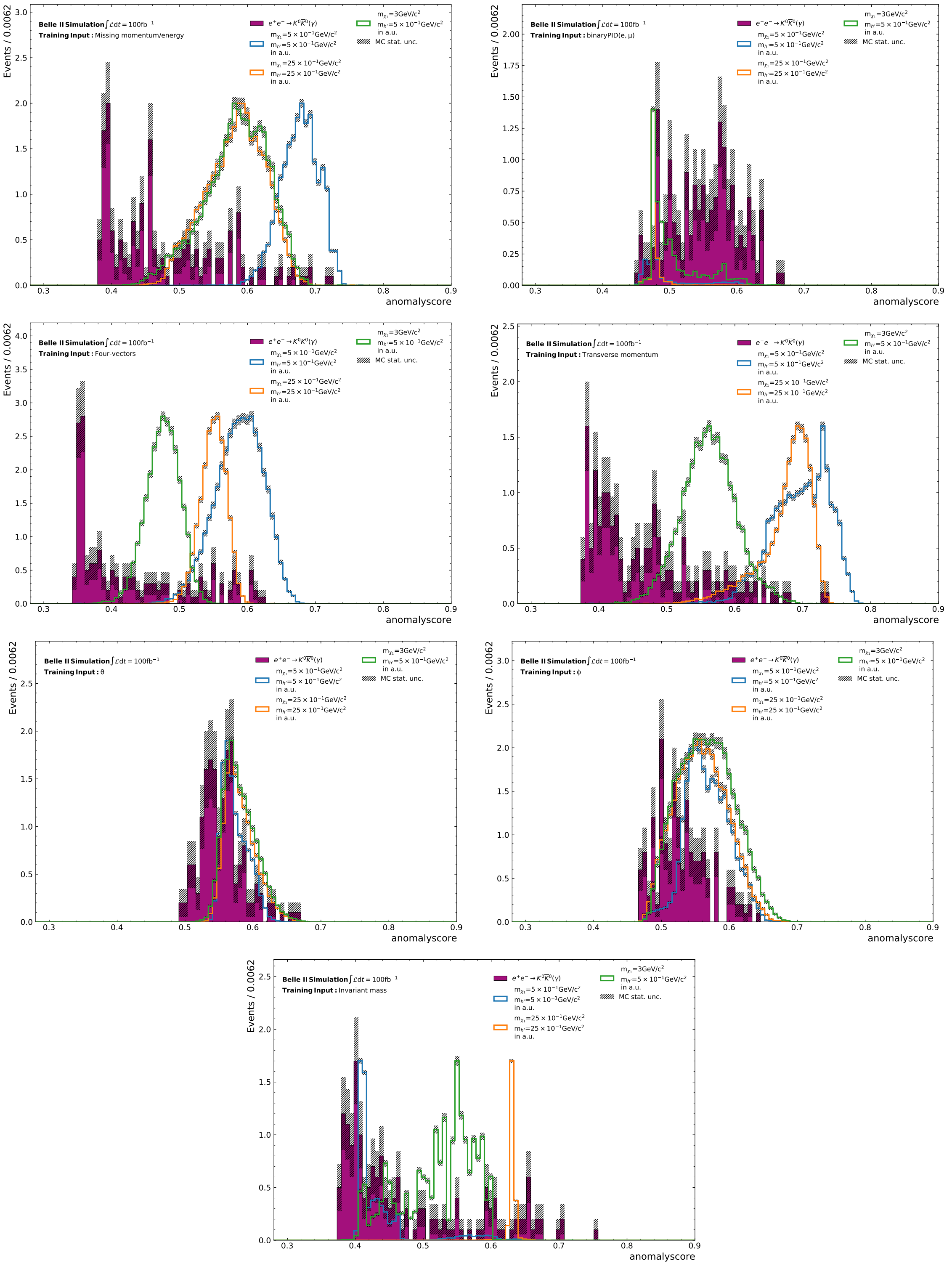
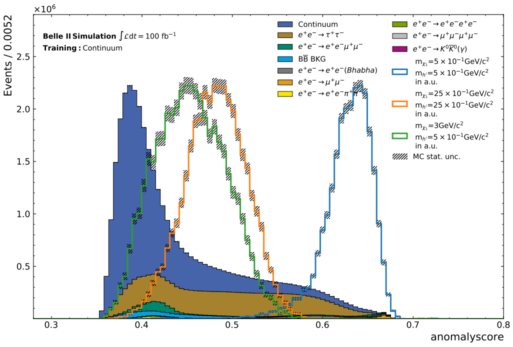
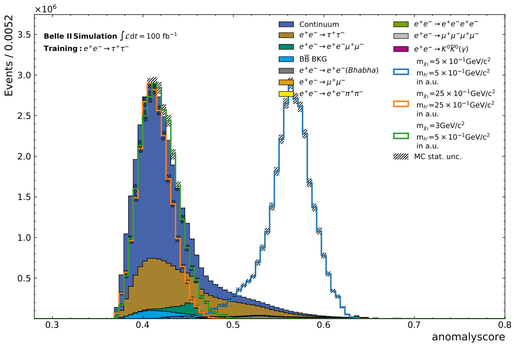


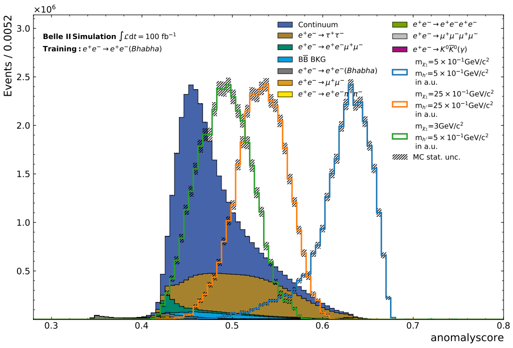
Figure A.33.: Contribution of each Input Feature group shown for the $e^+e^- \to K^0\overline{K}^0(\gamma)$ process trained separately.

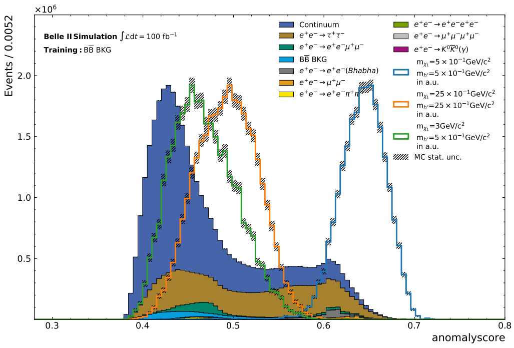
A.7. SM sample contribution to the distribution of the AS for the IForest

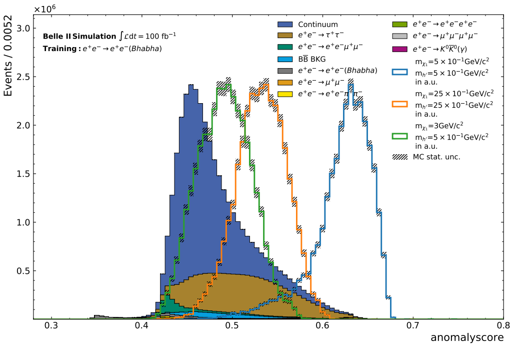
A.7.1. Distribution of the AS of separate trained SM process for favorable Input Features (p_t, missing Four-vector)

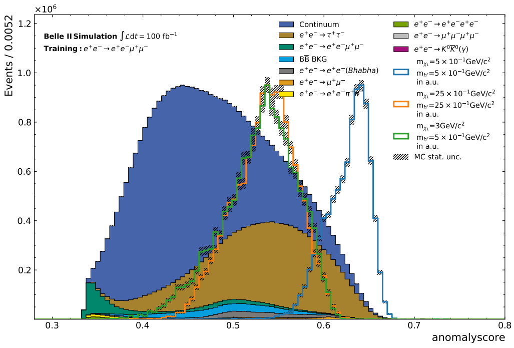


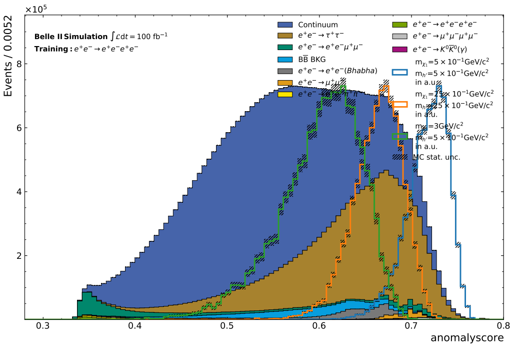


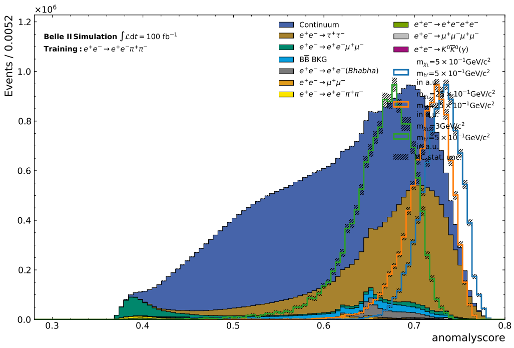












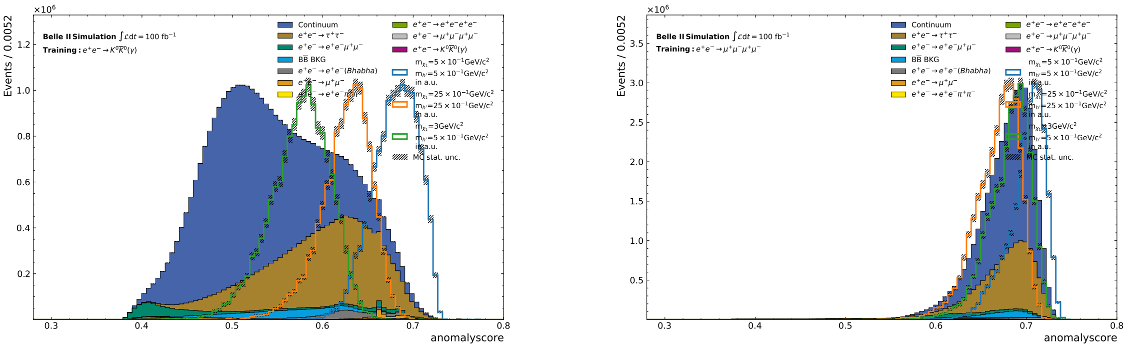
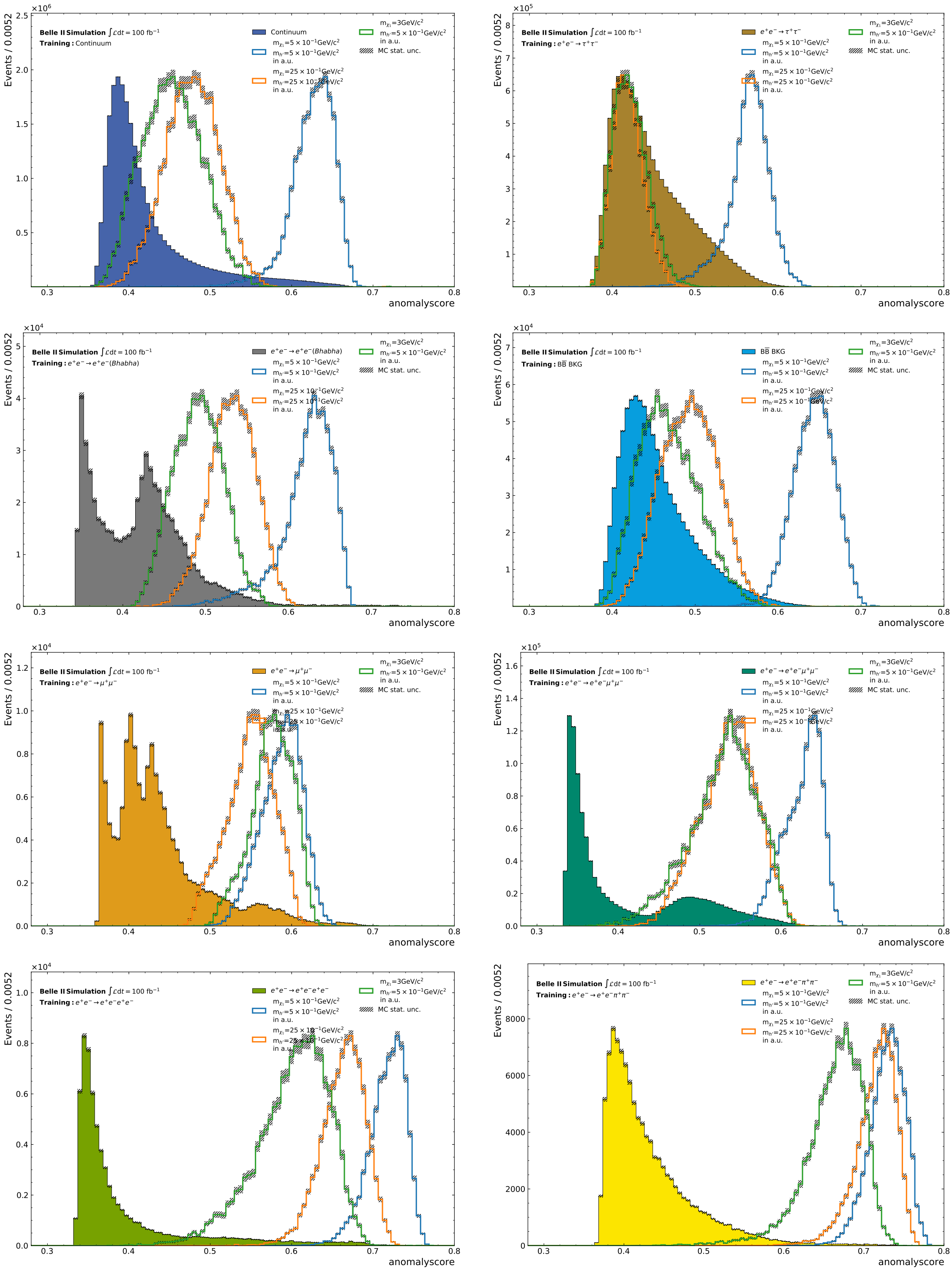


Figure A.34.: Contribution of each SM Background process shown for each process trained separately given the anomaly score distribution for the entire background.

A.7.2. Distribution of the AS of individual SM process for favorable Input Features



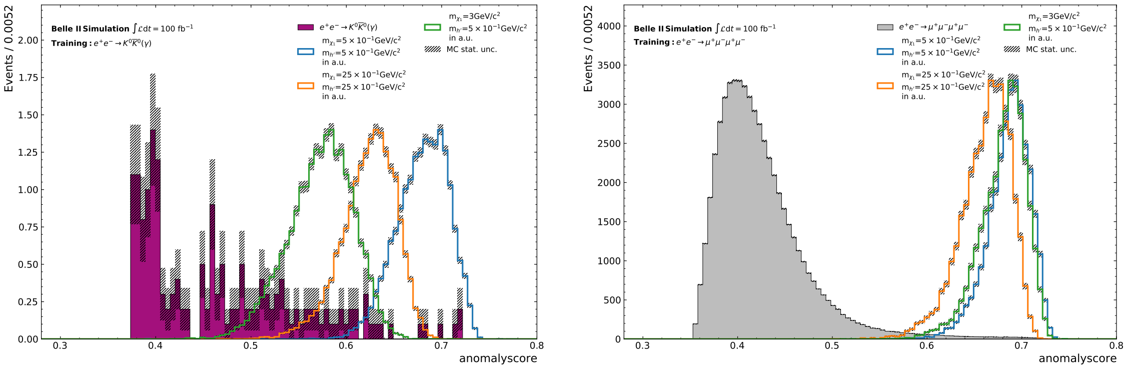
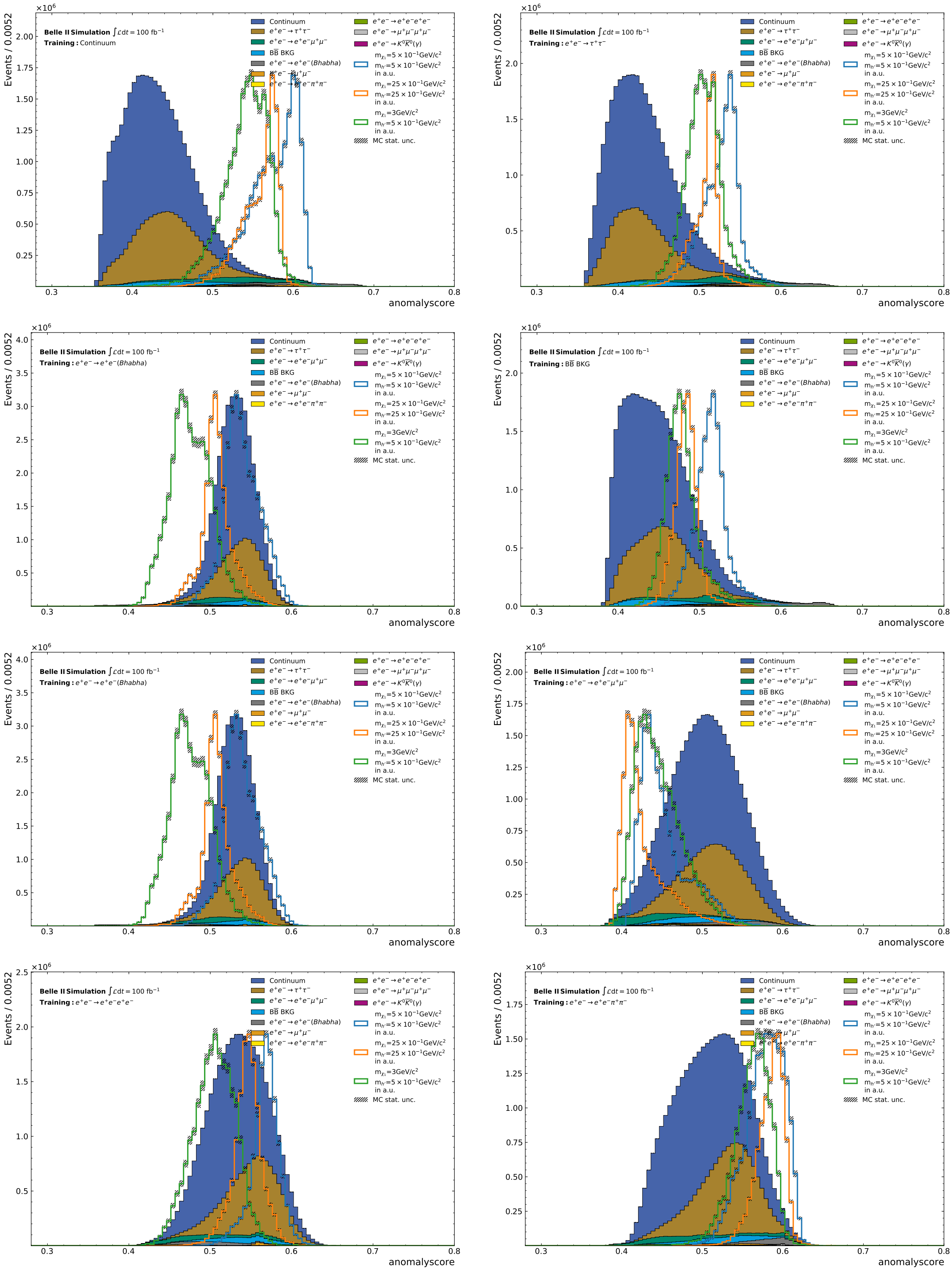


Figure A.35.: Contribution of each SM Background process shown for each process trained separately, given only the anomaly score distribution for the certain process.

A.8. Distribution of the AS of separate trained SM process for favorable Input Features with PID information.



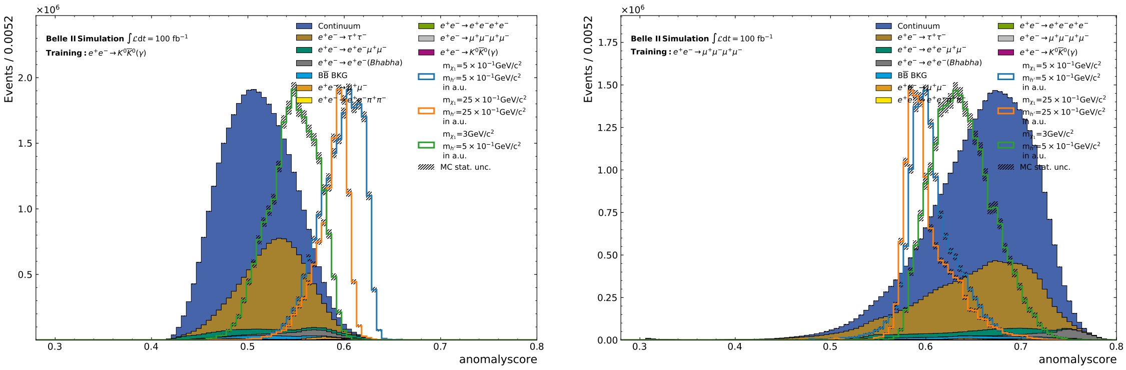
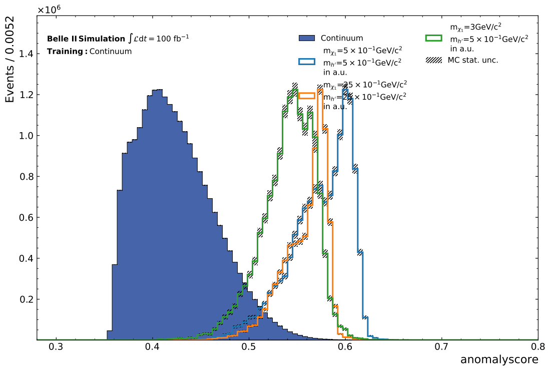
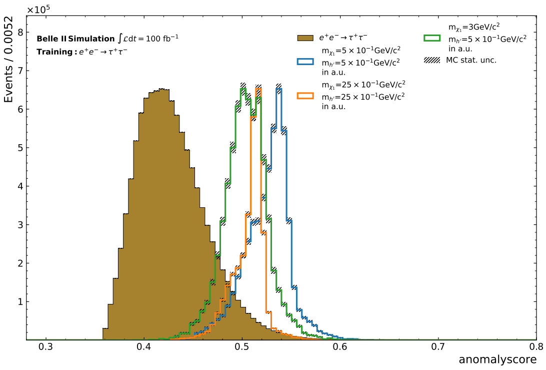
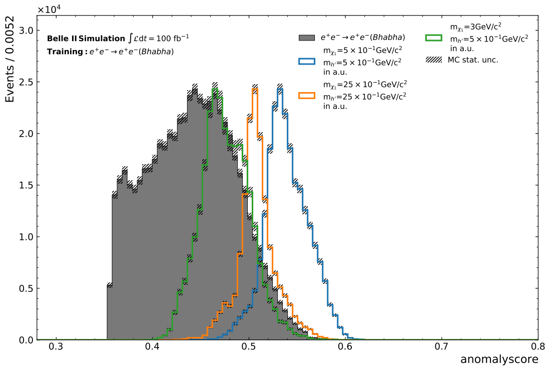


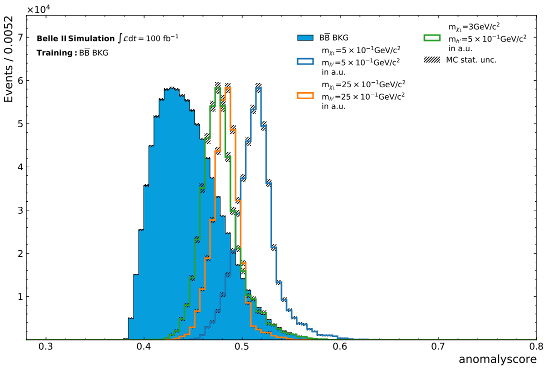
Figure A.36.: Contribution of each SM process shown for each process trained separately with PID information given distribution of the anomaly score for the entire background.

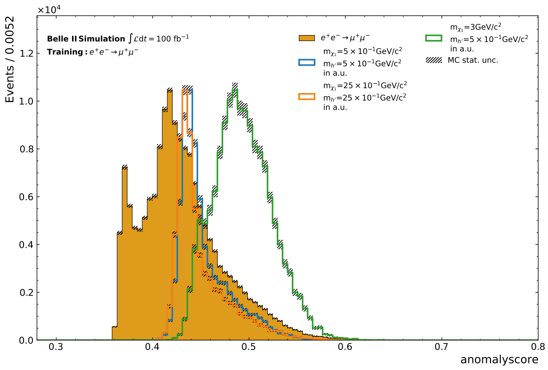
A.8.1. Distribution of the AS of individual SM process for favorable Input Features with PID information.

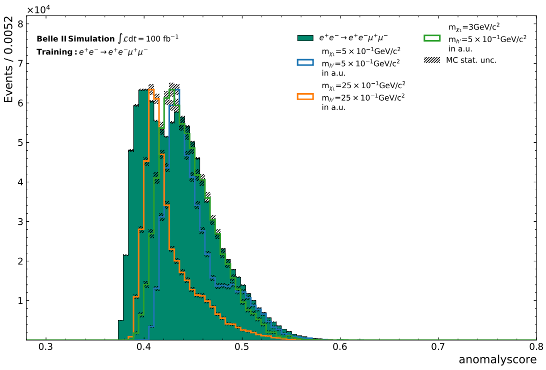


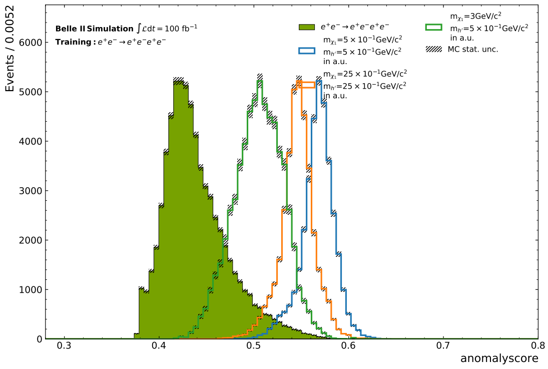


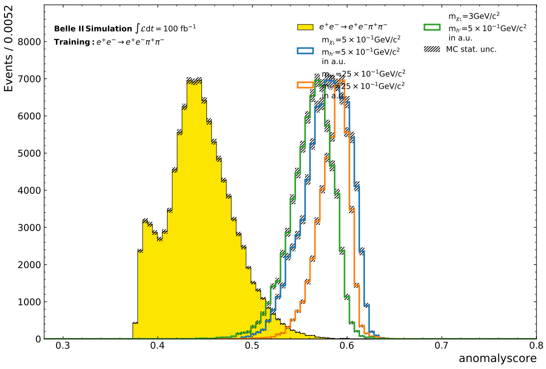












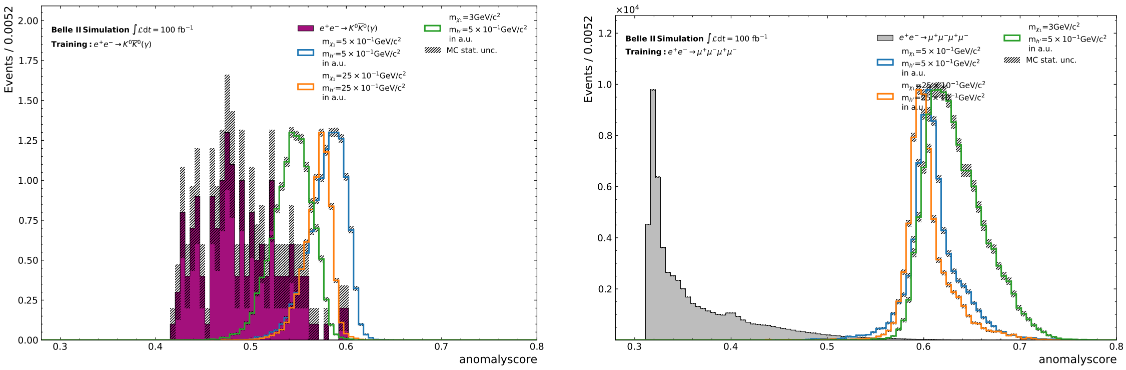


Figure A.37.: Contribution of each SM process shown for each process trained separately with the usage of PID information given the distribution of the anomaly score for the certain process.

A.9. PFOM curve for different AS of the Ensemble of Isolation Forests

A.9.1. PFOM curve using the favorable Input Features (p_t, missing Four-vector).

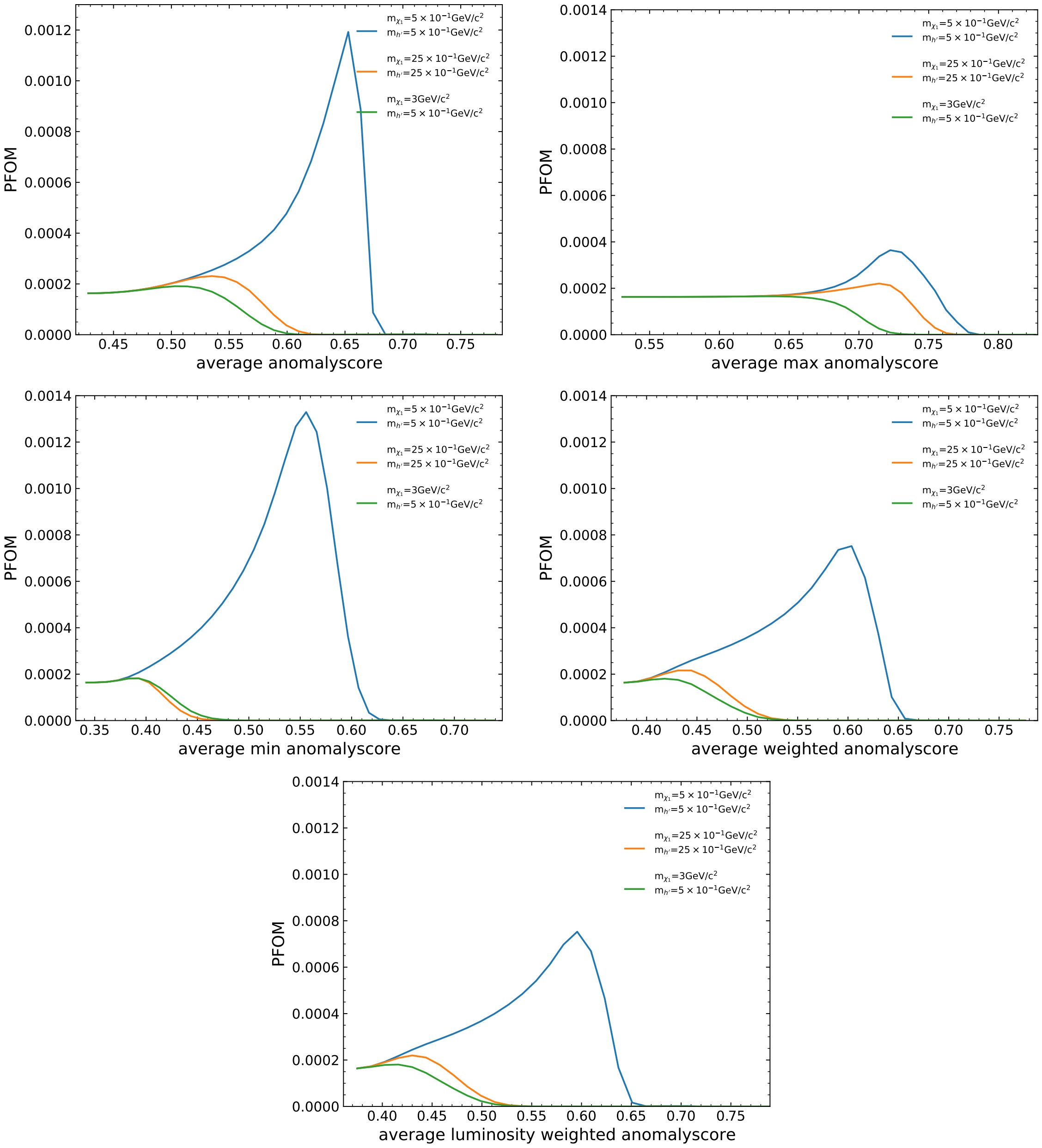


Figure A.38.: PFOM curve for each method of processing the different anomaly score distributions for favorable Input Features (p_t, missing Four-vector).

A.9.2. PFOM curve using the favorable Input Features with additional PID information.

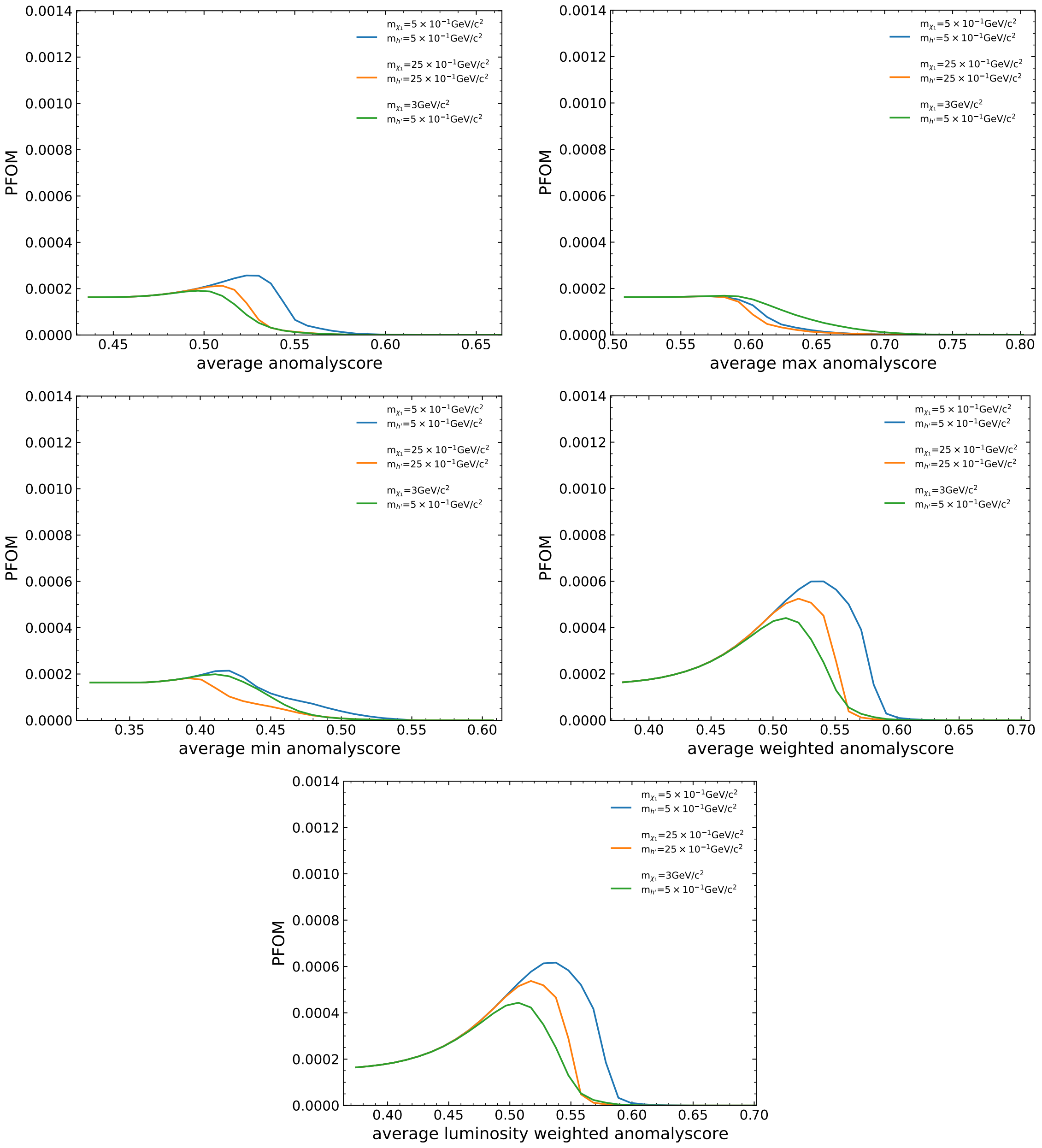
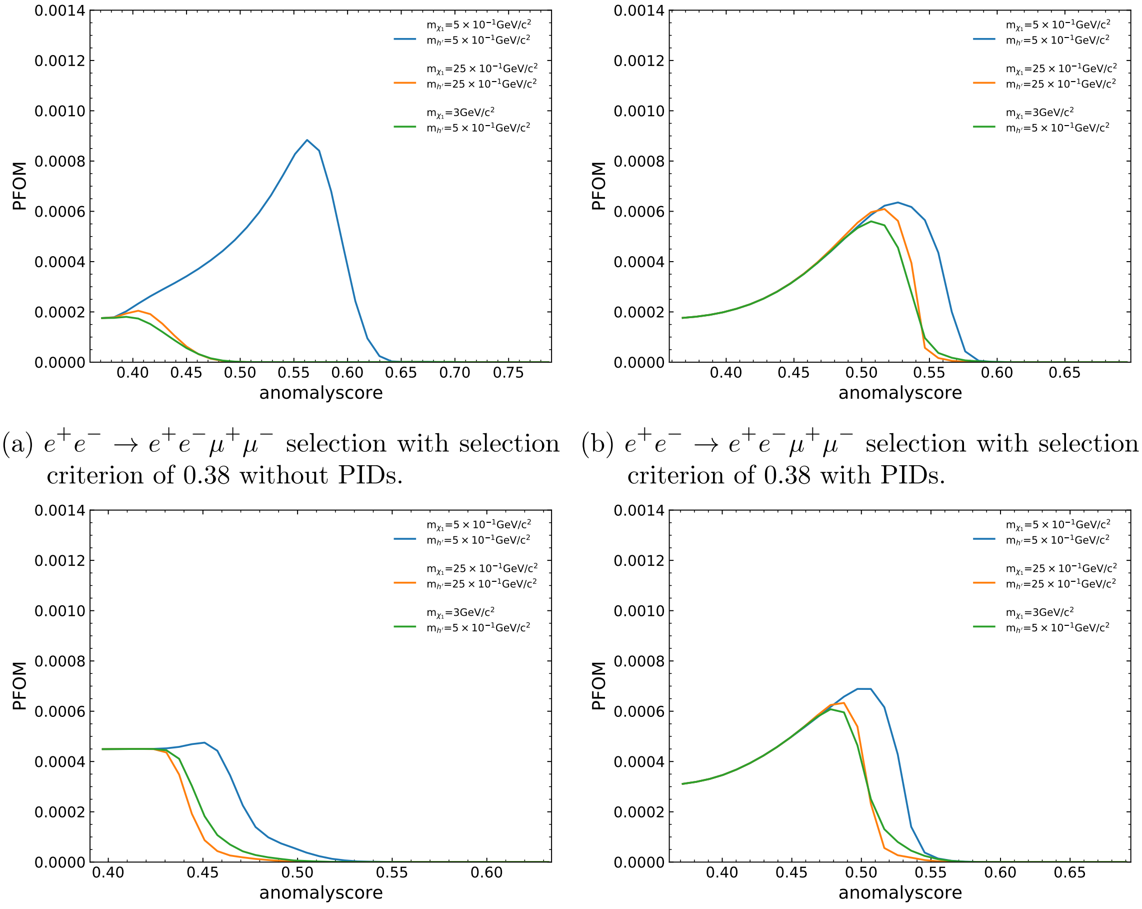


Figure A.39.: PFOM curve for each method of processing the different anomaly score distributions for favorable Input Features (p_t, missing Four-vector) with additional PID information.

A.10. PFOM curve for different selections on the AS of the iterative trained IForest.



(a) $e^+e^- \to e^+e^-\mu^+\mu^-$ selection with selection criterion of 0.38 without PIDs.

(b) $e^+e^- \to e^+e^-\mu^+\mu^-$ selection with selection criterion of 0.38 with PIDs.

(c) Continuum selection with selection criterion of 0.48.

(d) Collective selection with selection criterion of 0.45.

Figure A.40.: Selections on different selection criteria on distributions of separately trained IForests and one forest trained on all SM samples.

B. List of Figures