GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Fakultät für
Physik

Bachelor's Thesis

Optimierung der Suche von Kopplungsvariationen im VBF $HH \to b\bar{b}WW^*$ Kanal am ATLAS

Optimisation of the search for coupling variations in the VBF $HH \to b\bar{b}WW^*$ channel at ATLAS

prepared by

Friedrich Konrad Hoppe

from Berlin

at the II. Physikalisches Institut

Thesis number:	II.Physik-UniGö-BSc-2026/02
Thesis period:	7th November 2025 until 27th February 2026
First referee:	Prof. Dr. Stanley Lai
Second referee:	Prof. Dr. Steffen Schumann

Abstract

Higgs boson pair production via vector boson fusion (VBF) in the $HH \to b\bar{b}WW^*$ decay channel allows a better understanding of the Higgs potential, which ultimately leads to a better understanding of the Standard Model (SM) itself.
In this thesis, the decay mode of Higgs boson pairs in the $HH \to b\bar{b}WW^*$ channel with one leptonically decaying W boson is investigated using the VBF production mode. Precise cross-section measurements of this process allow for constraints on the coupling strength mediators between the Higgs boson and vector bosons.
Within the ATLAS detector at CERN, a large quantity of data is produced during measurements. To separate the signal processes from the background, different machine learning models, namely a feed forward neural network and boosted decision trees, are proposed in this thesis. Both models are trained on simulated data from the ATLAS experiment at CERN. The clean classification of signal and background processes enables increased sensitivity in the search for the VBF $HH \to b\bar{b}WW^*$ process.

Keywords: Particle physics, Higgs boson pair production, Vector boson fusion, Machine Learning, Neural network, Boosted decision trees

Contents

1 Introduction

In 2012 scientists at the ATLAS and CMS experiment at CERN discovered the Higgs boson [1][2], that was being theoretically predicted by Peter Higgs, François Englert and Robert Brout in 1964 [3][4]. The Higgs boson has a mass of around 125 GeV[5]. It was the last undiscovered particle, of the Standard Model (SM), which is a theory that describes the weak, the electromagnetic and the strong force all at once.
To experimentally test the SM, particle colliders have been built at large scales. The largest example for this is the proton-proton collider LHC at CERN. When two constituents of the proton collide, they can produce other particles, including the Higgs boson.
Currently, measurements of parameters such as the cross-sections, and with that the coupling strength mediators of the Higgs potential that determine the shape of the Higgs potential, are being conducted. For this, the Higgs boson pair production is of particular interest. Here, the decay channel of $HH \to b\bar{b}WW^*$ with one W boson decaying leptonically is investigated. The vector boson fusion (VBF) production for Higgs boson pairs is in the focus of this thesis, as it allows the studying of the coupling between two vector bosons and two Higgs bosons (HHVV). The leading order VBF Feynman diagrams are shown in Figure 1.1.

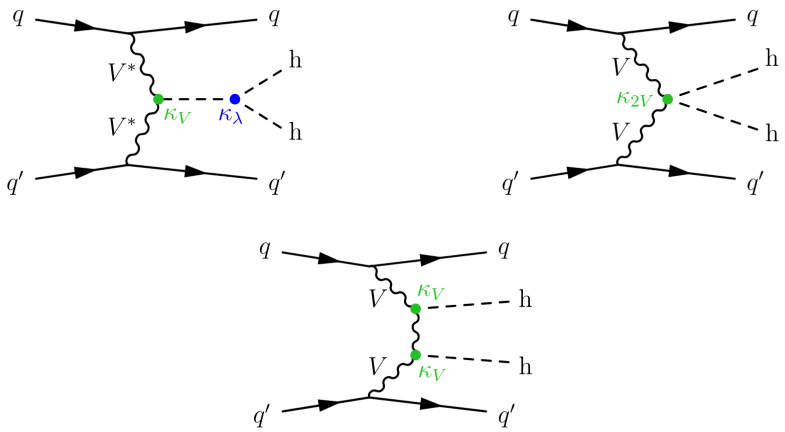


Figure 1.1: Leading order Feynman diagrams for Higgs boson pair production via VBF[6].

In this thesis, a neural network and boosted decision trees are trained on Monte Carlo simulated data for the signal events and several background events, mainly $t\bar{t}$ production. Both machine learning models tackle the problem of separating the given events into signal and background.
The structure of this thesis is as follows. First, the Standard Model is presented in Chapter 2, then the applied machine learning models are introduced in Chapter 3. After that, the LHC and the ATLAS detector is described in Chapter 4. The specific decay channel, which is observed in this thesis is explained in Chapter 5. Finally, the optimisation of the machine learning models is further implemented in Chapter 6. The thesis ends with a conclusion in Chapter 7.

2 The Standard Model and beyond

In this chapter, the theoretical framework of the Standard Model (SM) of particle physics is introduced. The SM, as formulated by Glashow [7], Salam [8], and Weinberg [9], provides a description of the strong, weak, and electromagnetic interactions of elementary particles.
The model postulates the existence of fundamental particles, categorised into fermions and vector bosons, along with the Higgs boson. A schematic representation of these particles and some of their key properties is provided in Figure 2.1.
Furthermore, the SM predicts essential parameters governing particle interactions, including transition amplitudes, cross-sections of various processes, and the strength of these interactions, as determined by the coupling constants of the bosons with other particles.

2.1 Framework and content

The SM predicts certain particles, that are presented in this section and shown in Figure 2.1. The fermions, listed in the centre of Figure 2.1, all possess spin-$\frac{1}{2}$. The six quarks are grouped into three generational doublets: the up and down quarks, the charm and strange quarks, and the top and beauty quarks. All quarks carry a colour charge, with the up, charm, and top quarks having an electric charge of $Q = \frac{2}{3}$ and the down, strange, and beauty quarks having $Q = -\frac{1}{3}$.
Quarks cannot exist alone, as colour charge must be neutral for any directly measurable particle. Thus, quarks combine to form heavier composite particles, such as pions or protons, or decay into lighter particles through one of the three fundamental interactions.
The leptons consist of the electron (e), muon (μ), and tau-lepton (τ), each paired with its corresponding neutrino (ν_e, ν_μ, and ν_τ). Unlike quarks, leptons do not carry colour charge. The electron, muon, and tau-lepton each have an electric charge of $Q = -1$, while the neutrinos are electrically neutral.
The gluon (g), the mediator of the strong interaction, is massless and electrically neutral but carries two colour charges: one colour and one anti-colour. With three colours and three anti-colours, there are eight possible charge combinations, corresponding to the

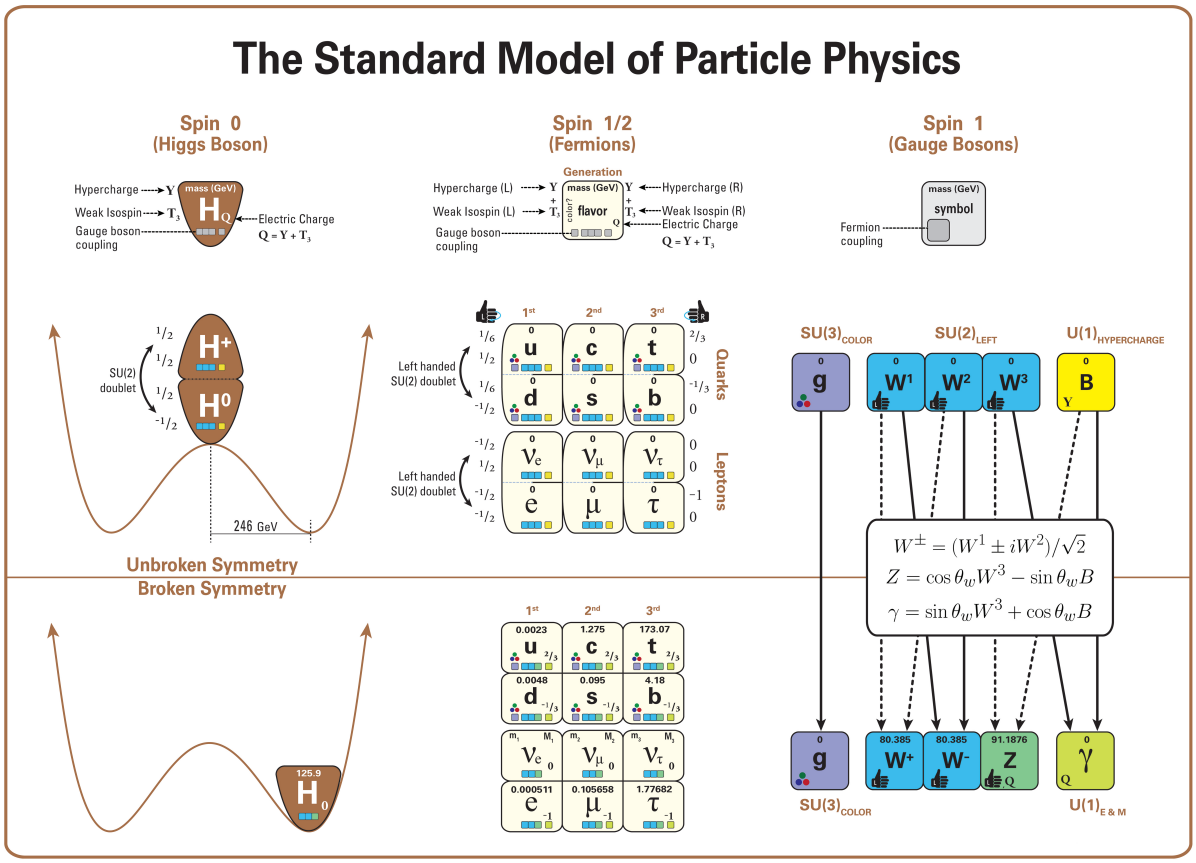


Figure 2.1: The image shows all particles predicted by the SM. They are all assumed to be fundamental particles. In the upper half, all particles are shown before the electroweak symmetry breaking, in the lower half after the symmetry breaking. From left to right the particles are ordered due to their spin: first the spin 0 Higgs boson, then the spin $\frac{1}{2}$ fermions and thirdly the spin 1 gauge bosons.

generators of the $\mathrm{SU}_c(3)$ group. The gluon couples to all particles with non-vanishing colour charge, including quarks and itself.

The photon (γ) is massless and carries neither electric nor colour charge. It mediates the electromagnetic force by coupling to the electromagnetic field and interacts with all quarks, charged leptons, and the $W^{\pm}$ bosons, but does not self-couple.

The W^+ boson carries a positive electric charge, while the W^- boson carries a negative electric charge, and the Z^0 boson is electrically neutral. They all mediate the weak force. None of these bosons possess colour charge, so they do not participate in the strong interaction.

The Higgs boson (H), is a neutral scalar particle associated with the Higgs field. Unlike the gauge bosons, it does not mediate a fundamental force but arises as a result of

the mechanism of electroweak symmetry breaking. It carries neither electric nor colour charge and therefore does not participate directly in the electromagnetic or strong interactions. The Higgs boson couples to particles proportionally to their mass, interacting with quarks, charged leptons, and the $W^{\pm}$ and Z^0 bosons. Through its non-vanishing vacuum expectation value, the Higgs field generates the masses of the weak gauge bosons and fermions, while leaving the photon and gluon massless.

Spontaneous symmetry breaking

In order to understand the implications of spontaneous symmetry breaking, it is useful to recall the gauge structure of the Standard Model and its electroweak sector. The pattern of symmetry breaking and the resulting particle spectrum are determined by the underlying gauge symmetry. The symmetry group of the SM is [10]

$$G_{SM} \quad = \quad \mathrm{SU}_c(3) \times \mathrm{SU}_L(2) \times \mathrm{U}_Y(1) \tag{2.1}$$

Before the spontaneous symmetry breaking, there are four massless gauge bosons for the electro-weak unification: W^1, W^2, W^3 only coupling to left-handed particles, and B coupling to the hypercharge. After the symmetry breaking of the electro-weak symmetry group $G_{ew} = \mathrm{SU}(2)_{\mathrm{Left}} \times \mathrm{U}(1)_Y$, the mass carrying W^+, W^- and Z^0, as well as the massless photon occur. The translation between these two states is described by

$$W^{\pm} = \frac{1}{\sqrt{2}} \left(W^1 \mp i W^2 \right), \tag{2.2}$$

$$\begin{pmatrix} Z \\ A \end{pmatrix} = \begin{pmatrix} \cos\theta_W & -\sin\theta_W \\ \sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} W^3 \\ B \end{pmatrix} \tag{2.3}$$

where θ_W is the weak mixing angle.

The electroweak interaction is described by the gauge symmetry SU(2) × U(1), which unifies the weak and electromagnetic interactions. The corresponding quantum numbers are the weak isospin T, specifically its third component T_3, the weak hypercharge Y, and the electric charge

$$Q = T_3 + \frac{Y}{2}. \tag{2.4}$$

Before electroweak symmetry breaking, weak isospin and weak hypercharge are conserved quantum numbers, reflecting the invariance of the theory under $\mathrm{SU}(2)_L$ and $\mathrm{U}(1)_Y$ gauge transformations. The symmetry is spontaneously broken by the Higgs mechanism,

$$\mathrm{SU}_L(2) \times \mathrm{U}_Y(1) \longrightarrow \mathrm{U}_e(1), \tag{2.5}$$

leaving electric charge Q as the conserved quantity after symmetry breaking.
As a consequence of electroweak symmetry breaking, three of the four gauge bosons acquire mass. These are the charged weak bosons $W^{\pm}$ and the neutral Z^0 boson, while the remaining massless gauge boson is identified with the photon γ.

2.2 Higgs mechanism and Higgs boson pair production

In this section, the theoretical foundations of the Higgs mechanism within the SM are explored. For that, the Higgs mechanism[10], as well as the Higgs boson pair production is presented.
The shape of the Higgs potential [11]

$$V(\phi) \quad = \quad \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2 \tag{2.6}$$

is pictured in Figure 2.2. The Higgs boson is included in the SM particles on the left of Figure 2.1.

2.2.1 Higgs mechanism

In the Higgs mechanism, the Higgs filed develops a non vanishing vacuum expectation value (VEV) and the electroweak symmetry breaks spontaneously.
The Higgs term of the SM Lagrangian $\mathcal{L}_{Higgs}$ with the Higgs potential V can be written as

$$\mathcal{L}_{Higgs} \quad = \quad (D^\mu\phi)^\dagger(D_\mu\phi) \; - V(\phi) \tag{2.7}$$

$$\quad = \quad (D^\mu\phi)^\dagger(D_\mu\phi) \; - \mu^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2, \tag{2.8}$$

where ϕ is a complex scalar field in the representation of the $\mathrm{SU}_L(2)$ group and D_μ is the covariant derivative. Note that, if $\mu^2 < 0$ the ground state will not be in the centre any more, and the Higgs field obtains a non-vanishing vacuum expectation value (VEV) $\phi^\dagger\phi = \frac{\nu^2}{2} = \frac{-\mu^2}{2\lambda}$. Since the direction in which the symmetry is broken is physically equivalent, the following ground state is chosen

$$\langle \phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu \end{pmatrix}. \tag{2.9}$$

The electroweak symmetry group $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ breaks into the exact symmetry of the electro-magnetic interaction, which possesses $\mathrm{U}(1)_e$ symmetry. Considering a small fluctuation around the ground state with a scalar function $h : x \mapsto h(x)$. The Higgs field becomes

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu + h \end{pmatrix} \tag{2.10}$$

in the unitary gauge. Inserting this into the Lagrangian $\mathcal{L}$, expanding all terms and diagonalising the matrices, one obtains Equation 2.11 [12].

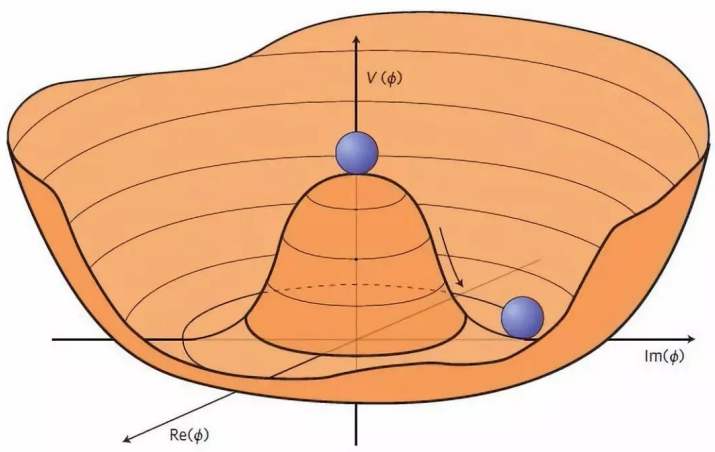


Figure 2.2: The image shows the Higgs Potential $V(\phi)$, where ϕ is complex. In the middle, the Higgs boson in its initial state is shown. After the electro-weak symmetry breaking, the Higgs fades to its second position, the ground state. There, the Higgs boson obtains a VEV.

$$\begin{aligned} \mathcal{L}(h) \quad = \quad & -\frac{\lambda v^4}{4} + \frac{1}{2}(-2\mu^2)h^2 + \lambda v h^3 + \frac{\lambda}{4}h^4 + \frac{g_2^2 v}{2} W_\mu^- W^{+\mu} h \\ & + \frac{g_2^2}{4} W_\mu^- W^{+\mu} h^2 + \frac{v}{4(g_2^2 + g_1^2)} Z_\mu Z^\mu h + \frac{1}{8(g_2^2 + g_1^2)} Z_\mu Z^\mu h^2 + \cdots \end{aligned} \tag{2.11}$$

Here the first term is constant and therefore does not contribute to the dynamics and $m_H = -2\mu^2$. The second term represents the mass of the Higgs boson, while as the third term contains the tri-Higgs coupling constant, which is proportional to λv. This coupling constant can be measured in Higgs boson pair production and is therefore extremely

valuable for the further experimental approach to the Higgs boson. The other terms shown in Equation 2.11 describe the Higgs couplings to the W and Z bosons, with coefficients set by the electroweak gauge couplings g_2, g_1, and v. The obtained masses of the H, W and Z bosons have been measured by the ATLAS experiment to be $m_H = 125.09 \pm 0.24$ GeV [5] $m_W = 80.370 \pm 0.019$ GeV [13] and $m_Z = 91.1876 \pm 0.0021$ GeV [14].

2.2.2 Higgs boson pair production via VBF and gluon fusion

Higgs boson pair production gives great insight into the properties of the Higgs boson. Certain parameters of interest are the self coupling constants of the Higgs boson, since they contribute directly to the potential. Therefore, measurements of Higgs boson pair production are essential to a better understanding of the Higgs potential. In general, there are two quantities that can be measured: The cross-section of Higgs boson pair production, and coupling modifiers. The coupling modifiers are defined as κ_V for H coupling with one vector boson, κ_{2V} for H coupling with two vector bosons, κ_λ Higgs boson self coupling and κ_t for H coupling to the top quark. All modifiers are equal to 1 in the Standard Model. The vertices of the corresponding couplings are shown in Figure 2.3.

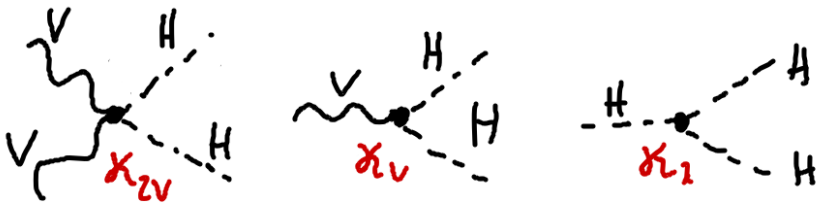


Figure 2.3: Here, the Feynman vertices of the processes that the coupling modifiers refer to are shown.

The Higgs boson can decay via different channels. The branching ratios for the most likely decay channels are listed in Table 2.1

There are many other processes related to the Higgs boson pair production, but the gluon-gluon-Fusion (ggF), and Vector Boson Fusion (VBF) are the production processes, that are the most accessible experimentally. The SM prediction for cross-sections at 13 TeV for $m_H = 125$ GeV are [16].

Decay channel	Branching ratio
$H \to b\bar{b}$	$58.09\% \pm 0.52\%$
$H \to \tau^+\tau^-$	$6.271\% \pm 0.067\%$
$H \to \mu^+\mu^-$	$0.0218\% \pm 0.0003\%$
$H \to gg$	$8.57\% \pm 0.22\%$
$H \to \gamma\gamma$	$0.227\% \pm 0.005\%$
$H \to ZZ^*$	$2.637\% \pm 0.028\%$
$H \to WW^*$	$21.52\% \pm 0.26\%$
$H \to ZZ$	$2.64\% \pm 0.03\%$

Table 2.1: Standard Model Higgs boson branching ratios for $m_H = 125$ GeV [15].

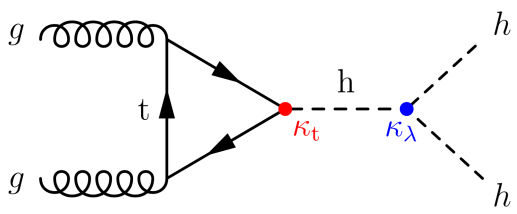
$$\sigma_{\text{ggF}}(pp \to HH) = 31.0^{+2.2}_{-2.4}\,\text{fb}, \tag{2.12}$$

$$\sigma_{\text{VBF}}(pp \to HH) = 1.73^{+0.04}_{-0.04}\,\text{fb}. \tag{2.13}$$

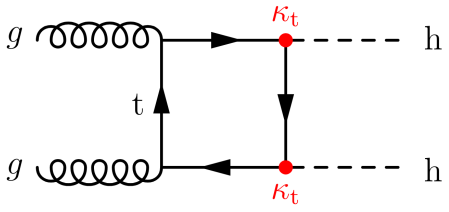
The total Higgs boson pair production cross-section in the Standard Model at $\sqrt{s} = 13\,\text{TeV}$ is approximately

$$\sigma(pp \to HH) \approx 33\,\text{fb}, \tag{2.14}$$

with ggF contributing more than 90% of the total rate and VBF representing the second-largest production mechanism [16].
For the ggF, there are two leading order Feynman diagrams are shown in Figure 2.4.



(a) Most probable decay of two gluons, fusing via a top quark to a single Higgs boson off shell, which then decays to two other Higgs bosons.



(b) Two gluons fusing via exchanging a top quark to two other top quarks. Again, they exchange a top quark and thereby produce two separate Higgs bosons

Figure 2.4: The two leading order events for ggF producing two Higgs bosons. Both processes interfere destructively. The important coupling constants for this process are κ_λ and κ_t [6]

The process on the left is often referred to as the triangle ggF process, and on the right as the box ggF process. The two interfere with each other destructively, which leads to a lower cross section.

The Feynman diagrams of the three main processes of VBF production are shown in Figure 2.5.

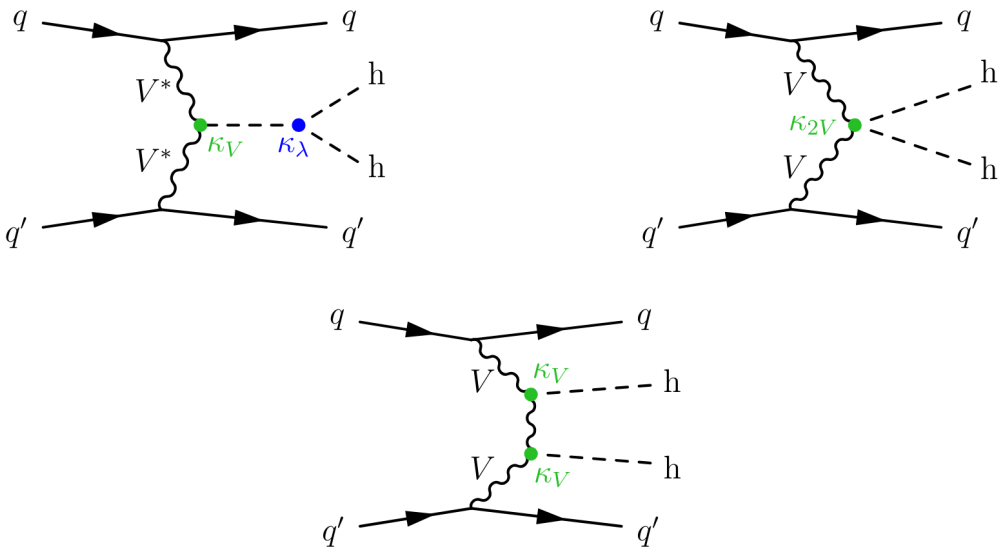


Figure 2.5: In all three VBF cases of leading order shown here, two quarks interact with each other and the included gauge bosons produce one or two Higgs bosons afterwards. The important coupling constants are the ones for VVH, $VVHH$ and HHH couplings, where V stands for the massive gauge boson. [6]

This process involves two quarks interacting via vector bosons. Consequently, the VBF production is sensitive to the coupling constant for vector bosons to the Higgs boson, which are linked to the modifiers κ_V and κ_{2V}. In particular, κ_{2V} can only be investigated in this production mode.

2.3 Limitations of the standard model and extensions to the Higgs sector

Although SM is extremely successful in describing the electromagnetic, strong and weak interaction, it does not include gravity [17], which remains unreconciled with quantum field theory at high energies. It also fails to explain dark matter [18], despite strong astrophysical evidence of its existence.

Extensions to the Higgs sector are proposed in many beyond-the-SM (BSM) theories, suggesting additional Higgs bosons or modified potentials [19].

In Higgs boson pair production, setting the VBF coupling constant $\kappa_{2V} \neq 1$ increases the

$HHVV$ coupling and enhances the cross-section. This process favours boosted topologies and higher-energy events [20]. This is also shown for the simulated samples of this thesis in chapter 9.
Currently, there are no experimental signs of any violation of the SM.

3 Data science methods

In this chapter, the two machine learning models, namely the neural network (NN) and the boosted decision trees (BDT)[21] are presented. To distinguish between signal and background processes, Monte-Carlo simulated samples are used to train the machine learning models.

3.1 Neural networks

In this section, the general architecture of a neural network is explained. A feed forward neural network [22] can be described as a composition of maps $\mathbb{R}^{n_0} \to \mathbb{R}^{n_1} \to \cdots \to \mathbb{R}^{n_p} \to \mathbb{R}^m$, where n_0 is the number of input features, n_1 to n_p the number of neurons per hidden layer and m the number of output classes. In the case of $m = 2$, one single output neuron is sufficient, since the outputs of the two classes are linearly dependant on each other. Each input vector of one event is mapped to an output vector with entries that later correspond to the probabilities of the event being in a certain class.

One single neuron

The building blocks of a neural network are the neurons. Each neuron is fed a n_i sized vector x and acts on it with

$$f\left(\sum_{j=1}^{n_i} x_j\theta_j + \theta_0\right) \quad = \quad a \tag{3.1}$$

where f is the activation function, θ_j are the weights of the network, θ_0 is the bias and a the 1-dimensional output of the neuron.

Activation functions

If only the biases and weights were to be considered, the neural network would not do much more, than a matrix multiplication. Since one wants the neural network to learn nonlinear relations, one must add a nonlinear function to it, which is the activation function. These

can differ for each layer of neurons. The two activation functions that are used in this research are the leaky RELU function [23]

$$f(x) = \begin{cases} x & \text{for } x \geq 0, \\ \alpha x & \text{for } x < 0, \end{cases} \tag{3.2}$$

with default $\alpha = 0.01$ and the SELU function [24]

$$f(x) = \begin{cases} \lambda x & \text{for } x \geq 0, \\ \lambda\beta(e^x - 1) & \text{for } x < 0, \end{cases} \tag{3.3}$$

with default $\beta = 1.67326$ and $\lambda = 1.0507$. Both functions are plotted in Figure 3.1 (left: RELU, right: SELU). The SELU function is used after the input layer, whereas after each hidden layer, the LeakyReLU function is applied.

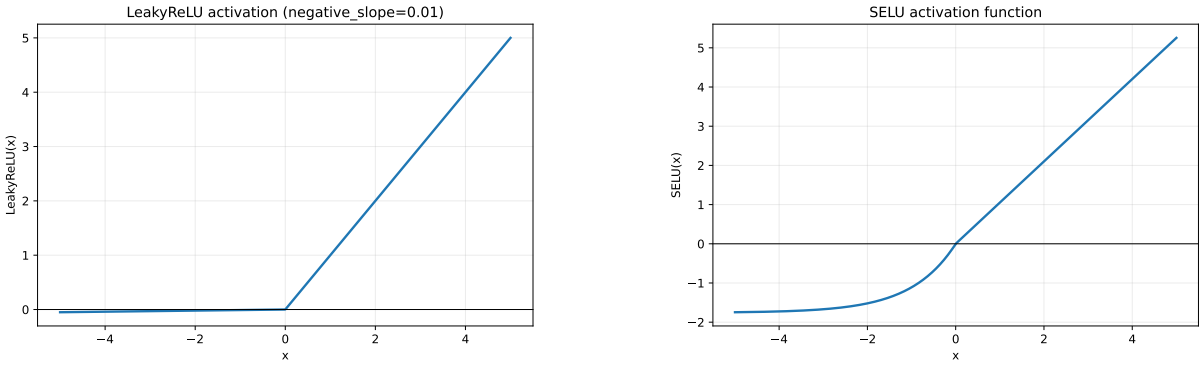


Figure 3.1: The leaky RELU function (left) and the SELU function (right), both being continuous at 0

Both functions were chosen because they are easy to compute. Leaky RELU was chosen above RELU ($f(x) = \max(0, x)$ to avoid producing dead neurons by multiplying with 0. The input layer consists of all input features, after that, several hidden layers of neurons are stacked. The last layer with the number of output classes equal to the number of neurons is called output layer. Figure 3.2 shows an example of a neural network structure. In case of $m = 2$, the NN maps every event to a specific value in [0,1] in a continuous way.

Loss function and Adam optimizer

The training process needs a clear numerical measure to show how different the network's predictions are from the correct target values. This is defined as the Loss function L. The

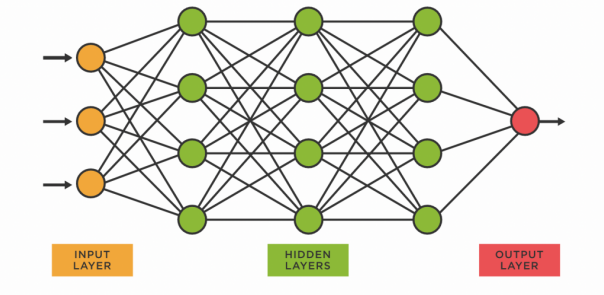


Figure 3.2: Schematic display of a feed forward neural network. The network is separated into an input layer, several hidden layers and an output layer. Each layer has a specific number of neurons, that are linked to the layers before and after them.[25]

target of the neural network is to minimise the loss function L. In this research, the cross entropy loss function [22]

$$L(q,p) \quad = \quad -\sum_{i=1}^{s} \omega_i \left(p_i \log(q_i) + (1-p_i)\log(1-q_i)\right) \tag{3.4}$$

is used for two output classes (signal and background), where p_i are the true classes and q_i are the predictions of the network after applying the sigmoid function. s is the number of samples and ω_i are the event weights. To minimize the loss function, the Adam optimizer was used[26]. This optimizer determines how to update the weights and biases:

$$\theta_{t+1} \quad = \quad \theta_t - \frac{m_t}{\sqrt{v_t}+\varepsilon} \cdot l, \tag{3.5}$$

where l is the learning rate, ε is set to 10^{-8} to avoid dividing by 0 and m_t and v_t are defined as

$$m_t \quad = \quad \frac{\beta_1}{1-\beta 1} m_{t-1} + \frac{\partial L}{\partial \theta_t} \tag{3.6}$$

$$v_t \quad = \quad \frac{\beta_2}{1-\beta_2} v_{t-1} + \left(\frac{\partial L}{\partial \theta_t}\right)^2. \tag{3.7}$$

The initial values at the start of the training are $m_0 = 0$ and $v_0 = 0$. The momentum parameters are $\beta_1 = 0.9$ and $\beta_2 = 0.999$. β_1 ensures a stable update direction by weighting past gradients, while β_2 adjusts the step size according to the historical variance of the gradients, as this helps to stabilise and accelerate the optimisation process.

Hyperparameters

The parameters that have direct impact on the structure or the calculation of the network are called hyperparameters. All of these can be optimized based on the task and the sample size.

Number of epochs gives the number of times the network trains with all the training data.

Number of layers gives the number of (hidden) layers of neurons the network has.

Batch size gives the number of events per batch. The network trains on each batch separately.

Learning rate is the factor that regulates how fast the weights are being altered.

Hidden size corresponds to n_i, the number of layers in a hidden layer.

3.2 Boosted decision trees

In this section, the second machine learning model that is applied in this research is presented. A boosted decision tree (BDT) [21] is also a supervised machine learning method, based on the idea of decision trees. The BDT model used in this research is XGBoost [27]. A decision tree f maps an event described by a vector $x \in \Omega = \Omega_1 \times \cdots \times \Omega_M$ (with $x_m \in \Omega_m$ being the input features) to a vector $\hat{x}$ of the same size, that is called a leaf. After the training of the tree, for every j leaves, the ratio of the two class labels defines the weight θ_j of the leaf. The weights give the prediction value of a certain event to be in a certain class, so $f(x^i) = \theta_j(x^i)$ where x^i denotes one particular event. Now f can be described as a composition of several indicator functions

$$1_{A_m}(x_m) = \begin{cases} 1 & \text{for} \quad x_m \in A_m \\ 0 & \text{for} \quad x_m \notin A_m \end{cases} , \tag{3.8}$$

where A_m are subspaces of the input features spaces Ω_m, splitting them into two real intervals A_m and $\Omega_m \backslash A_m$. A schematic display for a decision tree is shown in Figure 3.3. The machine learning part comes in, when at a certain branch, the A_m are chosen, in a way that maximizes the gain

$$G = \frac{1}{2} \left(\frac{G_{A_m}^2}{H_{A_m} + \lambda} + \frac{G_{\Omega_m \backslash A_m}^2}{H_{\Omega_m \backslash A_m} + \lambda} - \frac{(G_{A_m} + G_{\Omega_m \backslash A_m})^2}{H_{A_m} + H_{\Omega_m \backslash A_m} + \lambda} \right) , \tag{3.9}$$

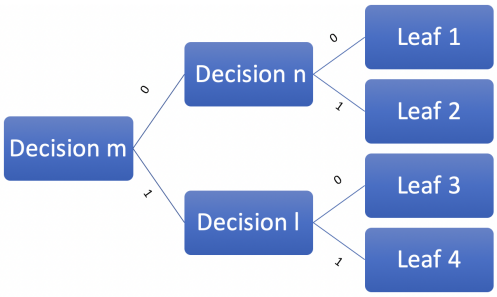


Figure 3.3: The image shows the schematic display of a simple decision tree with three different dicisions to put the input vector into one of the four leafs. In the decision trees used for this research the trees are more complex but work in the same way.

where G_{A_m} and $G_{\Omega_m \setminus A_m}$ are the sums of the gradients, and H_{A_m} and $H_{\Omega_m \setminus A_m}$ are the sums of the Hessians of the loss function. The used data and features and the order in which they occur are randomly chosen for each tree of the system of boosted decision trees. In a system of boosted decision trees, several decision trees are stacked in a row. The boosted decision trees measure the loss of the results via an objective function $\text{obj}(\theta_j, f_t) = L(\theta_j) + \Theta(f_t)$ where in case of the used XGBoost model the L denotes the logarithmic loss function and Θ denotes a regularization function that reduces the complexity of the system. Namely:

$$\begin{aligned} \text{obj}(\theta_j, f_t) &= L(\theta_j) + \Theta(f_t) && (3.10)\\ &= \sum_{i=1}^{n} l(y^i, \hat{y}^i) + \frac{1}{2}\lambda \sum_{j=1}^{T} \theta_j^2 && (3.11)\\ &= \sum_{i=1}^{n} l\left(y^i, \hat{y}^{i,(t-1)} + f_t(x^i)\right) + \frac{1}{2}\lambda \sum_{j=1}^{T} \theta_j^2 + \text{const} && (3.12) \end{aligned}$$

where y_i denotes the labels of an event, n the number of events and K the number of trees. The loss function can then be expanded via

$$l\left(y^i, \hat{y}^{i,(t-1)} + f_t(x^i)\right) \approx l\left(y^i, \hat{y}^{i,(t-1)}\right) + g^i f_t(x^i) + \frac{1}{2} h^i f_t(x^i)^2 \quad (3.13)$$

$$\text{with gradient} \qquad g^i = \frac{\partial l\left(y^i, \hat{y}^{i,(t-1)}\right)}{\partial \hat{y}^{i,(t-1)}} \quad (3.14)$$

$$\text{and Hessian} \qquad h^i = \frac{\partial^2 l\left(y^i, \hat{y}^{i,(t-1)}\right)}{\partial (\hat{y}^{i,(t-1)})^2}. \quad (3.15)$$

The objective function can now be written as

$$\mathbf{obj}(f_t) \approx \sum_{i=1}^{n} \left[g^i f_t(x^i) + \frac{1}{2} h^i f_t(x^i)^2 \right] + \frac{1}{2} \lambda \sum_{j=1}^{T} \theta_j^2 + \text{const.} \tag{3.16}$$

with T being the number of leaves, and λ being a regularization parameter. These help to repress overfitting via repressing high weights, and in the default case λ is set to 1. Firstly, the best choice of weights that minimizes the objective function is

$$\theta_j^* = -\frac{\sum_{i \in I_j} g^i}{\sum_{i \in I_j} h^i + \lambda}, \tag{3.17}$$

where I_j is the subset of events being in leaf j. The weights are then updated correspondingly. Then the prediction of the model is updated with

$$\hat{y}^{i,(t)} = \hat{y}^{i,(t-1)} + \eta \cdot f_t(x^i) \tag{3.18}$$

with the learning rate η. And lastly, the next tree f_{t+1} is trained on the pseudoresiduals $\tilde{y}^i = -g^i$, which are derived from the negative gradients of the loss function with respect to the updated predictions $\hat{y}^{i,(t)} = \hat{y}^{i,(t-1)} + \eta \cdot f_t(x^i)$. These pseudoresiduals serve as the new "labels" to correct the errors of the current ensemble.
In the end, the BDT maps every event to a value in [0,1] in a continuous way.

4 LHC and the ATLAS detector

4.1 LHC

To explore the fundamental constituents of matter and the forces governing their interactions, the Large Hadron Collider (LHC) [28] serves as a pivotal experimental facility.

The Large Hadron Collider (LHC) [28], operated by the European Organization for Nuclear Research (CERN) and situated on the border of France and Switzerland, is the world's largest particle collider. With a circumference of 27 km, the LHC is designed to collide protons and heavy ions at a centre-of-mass energy of up to $\sqrt{s} = 14\,\text{TeV}$ and a luminosity of $L = 10^{34}\,\text{cm}^{-2}\text{s}^{-1}$. Its primary purpose is to probe particle collisions, enabling both the evaluation of Standard Model (SM) predictions and the search for beyond-Standard-Model (BSM) phenomena. The distinct operational phases of the LHC are referred to as runs: during Run 1, protons achieved centre-of-mass energies of $\sqrt{s} = 7$ to 8 TeV; in Run 2, this increased to $\sqrt{s} = 13\,\text{TeV}$; and in Run 3, a centre-of-mass energy of $\sqrt{s} = 13.6\,\text{TeV}$ was reached.

The acceleration of protons at the LHC follows a multi-stage process. Initially, protons are accelerated in the Linear Accelerator 4 (LINAC 4) to an energy of approximately 160 MeV. They are then transferred to the Proton Synchrotron Booster (PSB), where their energy is increased to 2 GeV. Subsequently, the protons enter the Proton Synchrotron (PS), reaching an energy of 26 GeV, before being further accelerated to 450 GeV in the Super Proton Synchrotron (SPS). Finally, the protons are injected into the LHC, where they attain their ultimate collision energy. This staged acceleration is essential to achieve the high energies required for the experiments.

The protons are confined to a circular trajectory by a magnetic field of up to 8 T, produced by superconducting electromagnets operating at a temperature of 2 K. Protons are put into bunches with a bunch crossing interval of 25 ns and an event rate of 40 MHz. In Run 3, an average of 46.5 interactions occurred per bunch crossing.

4.2 ATLAS detector

To investigate the outcomes of high-energy proton collisions, the LHC relies on sophisticated detector systems, primarily ATLAS[29], CMS[30], ALICE[31], and LHCb[32]. In this thesis, the focus lies on the analysis of simulated events from the ATLAS detector. For this, a detailed description of its design and functionality is presented here.
The ATLAS detector is cylindrical, with dimensions of 44 m in length, 25 m in height, and a total weight of approximately 7000 t. It uses a right-handed coordinate system, where the x-axis points toward the centre of the LHC ring and the y-axis points upward. The kinematic information of a highly relativistic particle is put in a four-vector:

$$p \in \mathbb{R}^{1,3}, \qquad p = \begin{pmatrix} E \\ \vec{p} \end{pmatrix}. \tag{4.1}$$

Due to the cylindrical symmetry of the ATLAS detector, it is advantageous to use the polar angle θ and the azimuthal angle φ instead of Cartesian coordinates. The pseudorapidity

$$\eta = -\log\left(\tan\left(\frac{\theta}{2}\right)\right), \tag{4.2}$$

is conventionally used in particle physics because, for highly relativistic particles, the difference of η remains Lorentz-invariant under longitudinal boosts. With this, the four-vector can also be expressed in terms of the particle's mass m, transverse momentum p_T, azimuthal angle φ, and pseudorapidity η. The transformation between the two coordinate systems is given by:

$$E = \sqrt{m^2 + p_T^2 \cosh^2(\eta)}, \tag{4.3}$$
$$p_x = p_T \cos(\varphi), \tag{4.4}$$
$$p_y = p_T \sin(\varphi), \tag{4.5}$$
$$p_z = p_T \sinh(\eta). \tag{4.6}$$

The transverse momentum p_T is particularly useful because the observed processes are invariant under rotations in φ for any given event.
To characterise and distinguish particles produced in high-energy collisions, the ATLAS detector employs a combination of spatial and energy measurements. A key observable in this context is the angular distance in the η-φ plane, defined as:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\varphi)^2}. \tag{4.7}$$

The ATLAS detector provides coverage up to $|\eta| < 4.9$ and comprises multiple subsystems, each designed to measure different aspects of particle interactions. A transverse-plane sector of the ATLAS detector is illustrated in Figure 4.1.

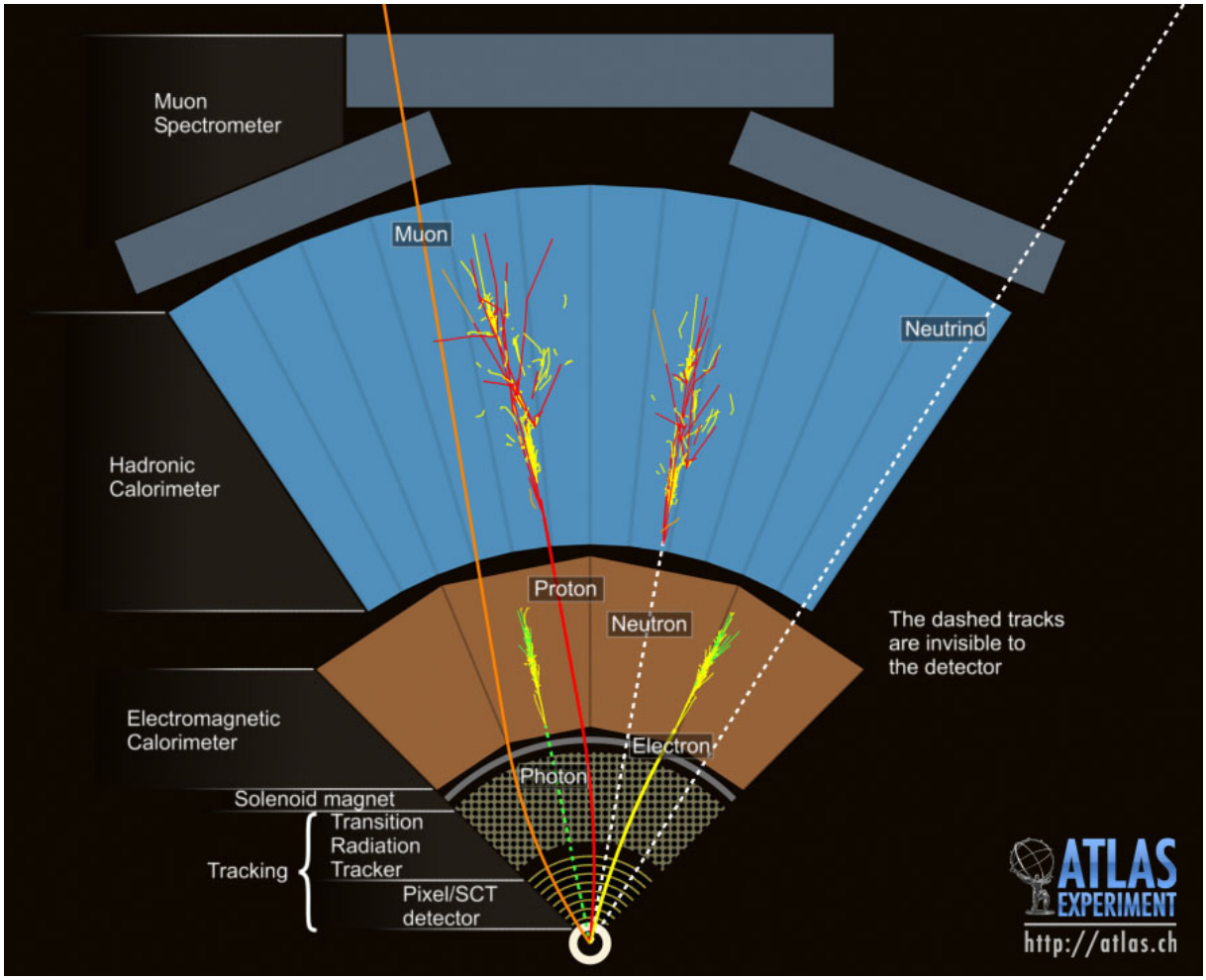


Figure 4.1: Sector of ATLAS in the transverse plane, showing the detector subsystems and example particle interactions. The layout includes the inner tracking detector, followed by an electromagnetic calorimeter, a hadronic calorimeter, and finally the muon spectrometer.

The Inner Detector (ID) serves as a tracking system, immersed in a 2 T magnetic field, enabling the measurement of the transverse momentum p_T and charge of particles within $|\eta| < 2.5$. This pseudorapidity limit arises from the ID's design, which restricts efficient track reconstruction beyond this range due to reduced spatial resolution and coverage at shallow angles relative to the beam axis. The ID integrates four components: the Insertable B-layer (IBL), the Pixel detector, the Semiconductor Tracker (SCT), and the Transition Radiation Tracker (TRT). The transverse momentum resolution of the ID is

$$\frac{\sigma_{p_T}}{p_T} = 0.05\%\ \text{GeV}^{-1} \cdot p_T \oplus 1\%. \tag{4.8}$$

The Electromagnetic Calorimeter (ECAL) and Hadronic Calorimeter (HCAL) measure particle energies up to $|\eta| < 3.2$. Their design alternates between active layers, which measure energy deposits, and passive layers, which induce particle showers. In the ECAL, passive layers interact electromagnetically, primarily through bremsstrahlung with electrons and positrons, creating cascades of particles until a critical energy is reached. The ECAL has a total thickness of less than $22X_0$ with X_0 being the radiation length. This is sufficient to contain most electromagnetic showers. The energy resolution of the ECAL is

$$\frac{\sigma_E}{E} = 10\%\ \text{GeV}^{\frac{1}{2}} \cdot \frac{1}{\sqrt{E}} \oplus 0.7\% \tag{4.9}$$

for EM showers. In the HCAL, hadrons interact via the strong force, producing showers dominated by hadronic particles and gluons. The HCAL has a thickness of $7.4\lambda_{\text{int}}$, with λ_{int} being the hadronic interaction length. For the central and end-cap regions ($|\eta| < 3.2$), the HCAL energy resolution is

$$\frac{\sigma_E}{E} = 50\%\ \text{GeV}^{\frac{1}{2}} \cdot \frac{1}{\sqrt{E}} \oplus 3\% \tag{4.10}$$

for pion showers. In the forward region ($3.1 < |\eta| < 4.9$), the resolution goes down to

$$\frac{\sigma_E}{E} = 100\%\ \text{GeV}^{\frac{1}{2}} \cdot \frac{1}{\sqrt{E}} \oplus 10\%. \tag{4.11}$$

Together with the end-cap calorimeters, the ATLAS calorimeter system provides coverage up to $|\eta| < 4.9$.

To detect and measure muons that are highly penetrating particles that interact minimally with matter, the ATLAS detector incorporates a muon spectrometer. Muons, being leptons, do not participate in strong interactions and, due to their high mass, only rarely undergo electromagnetic interactions, primarily losing energy through minimal ionisation. The muon spectrometer utilises gas-filled drift chambers immersed in a magnetic field of approximately 1 T in the end-cap regions and 0.5 T in the barrel, providing coverage for $|\eta| < 2.7$.

Since muons are electrically charged, the spectrometer measures their trajectories and transverse momenta with high precision. For muons with high transverse momenta, the momentum resolution of the spectrometer is approximately

$$\frac{\sigma_{p_T}}{p_T} = 10\% \quad \text{at} \quad p_T = 1\,\text{TeV}. \tag{4.12}$$

Given the LHC's bunch crossing interval of 25 ns, the ATLAS detector generates roughly 60 TB/s of raw data. To manage this vast data stream, a sophisticated trigger system [33] is employed to select events of high physical interest while discarding less relevant collisions. The system comprises two main stages: a hardware-based Level 1 (L1) trigger and a software-based High-Level Trigger (HLT). These triggers drastically reduce the data rate, decreasing the event rate from 40 MHz at the collision level to 100 kHz after L1 and further to about 3 kHz following the HLT stage.

5 VBF $HH \to b\bar{b}WW^*$ production in the 1-lepton final state

After having presented the Higgs boson pair production in Chapter 2, this chapter focusses on the exact decay channel that is the subject to this thesis.
To reconstruct and analyse the complex final states produced in high-energy collisions, particles measured within a shower are grouped into composite objects known as jets. At ATLAS, the clustering of particles into jets is performed using the anti$-k_t$ algorithm [34], which sequentially combines particles based on their relative transverse momentum (p_T) and angular separation, defined by $R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$.
In scenarios where a decaying particle possesses a high transverse momentum p_T, its decay products are spatially close within the detector. In such cases, it is advantageous to merge all decay products into a large-radius (LR) jet, typically with a cone size of $R = 1.0$ (in contrast to $R = 0.4$ for standard small radius jets). This configuration is referred to as a boosted topology.
However, this approach introduces challenges, as hadronic decay products from unrelated particles or even pileup events may be incorrectly added into the jet even more than it is the case for small radius jets.

5.1 Decay topology

To investigate the production and decay of Higgs boson pairs, this thesis focuses on the $HH \to b\bar{b}WW^*$ decay channel, specifically the scenario where one W boson decays hadronically and the other leptonically, as illustrated in Figure 5.1. In this process, two quarks interact via vector boson fusion (VBF) to produce a pair of Higgs bosons. One Higgs boson subsequently decays into a beauty quark and its antiparticle ($b\bar{b}$). When the Higgs boson carries a large transverse momentum p_T, its decay products are Lorentz-boosted in the transverse plane, resulting in highly collimated quarks. For values of $\kappa_{2V} \neq 1$, the decay topology is more likely to exhibit a boosted configuration[20].
Both beauty quarks hadronise and are detected as jets within the ATLAS detector. Due

to their collimated nature, the two jets overlap significantly, making it impractical to resolve their individual constituents. Thus, they are reconstructed as a single large-radius jet with $\Delta R = 1$. Given the Higgs boson mass of approximately $m_H = 125\,\text{GeV}$, the invariant mass of the two quarks is expected to peak at this value.

In the $b\bar{b}WW^*$ decay channel, the second Higgs boson decays into a pair of W bosons. Since the mass of the W boson is about $m_W = 80\,\text{GeV}$, and the Higgs boson mass is about 125 GeV, at least one W boson must be off-shell. In this analysis, the hadronically decaying W boson is favoured to be on-shell during preselection (as said in section 5.2), so its decay products, a quark and an antiquark are also expected to form a jet with an invariant mass close to m_W. For this reason, the two quarks from the hadronically decaying W_{had} are combined into a single large-radius jet.

This study restricts the $b\bar{b}WW^*$ channel to events with one lepton in the final state. Consequently, the other W boson decays leptonically, producing either a lepton l^- and a right-handed antineutrino $\bar{\nu}_l$, or an antilepton l^+ and a left-handed neutrino ν_l. While the lepton is directly measurable in the detector, the neutrino kinematics can only be deduced through the missing transverse energy E_T^{miss} and its azimuthal angle φ_T^{miss}. The two quarks from the VBF process are emitted in opposite directions with polar angles close to the beam pipe, resulting in two forward jets with high values of $|\eta|$ and a large pseudorapidity separation $\Delta\eta$.

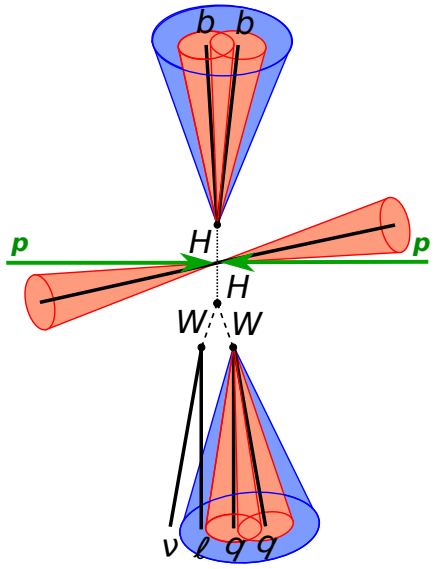


Figure 5.1: Decay topology of the $HH \to b\bar{b}WW^*$ channel with one lepton in the final state.

5.2 Data preparation

To explore both Standard Model (SM) and beyond-Standard-Model (BSM) scenarios, this thesis utilises Monte Carlo simulated data from the ATLAS experiment. The Monte Carlo simulation models the physical processes of proton-proton collisions at the LHC, as well as the response of the ATLAS detector. The dataset comprises simulated events for both signal and background processes. The signal data includes vector-boson fusion (VBF) Higgs pair production in the final state described in section 5.1, generated for different values of the coupling modifier $\kappa_{2V} = 0$, $\kappa_{2V} = 0.5$, $\kappa_{2V} = 1$, $\kappa_{2V} = 1.5$, $\kappa_{2V} = 2$, and $\kappa_{2V} = 3$. For the training of machine learning models, only signal events with $\kappa_{2V} = 2$ and $\kappa_{2V} = 3$ are used. This selection is motivated by the fact that $\kappa_{2V} = 1$ corresponds to the SM prediction, while $\kappa_{2V} = 2$ and $\kappa_{2V} = 3$ represent BSM scenarios of one single leading order VBF process (Figure 2.5 on the right). All available backgrounds and signals are listed in Table 5.1 with the corresponding Monte-Carlo generators used in this research, as well as the number of data points produced.

Event	Monte-Carlo generator	Number of simulated evens
signal	Madgraph (ME) + Pythia 8 (PS)	10944 ($\kappa_{2V} = 2, 3$)
W jets	Sherpa 2.2.11	143692
Z jets	Sherpa 2.2.11	74310
single top	Powheg and Pythia 8 (PS)	7513
$t\bar{t}$	Powheg and Pythia 8 (PS)	95570
dijet	Pythia 8 (PS)	1166
diboson	Sherpa2.2.14	68266
single H	Powheg and Pythia 8 (PS)	69437

Table 5.1: Different physical processes, their corresponding Monte-Carlo generators and the number of simulated events.

To effectively distinguish signal from background, this analysis focuses on the dominant background process for VBF Higgs boson pair production: the $t\bar{t}$ production channel. While other background channels such as single Higgs boson decay, single top quark decay, Z-jets, W-jets, and diboson decay contribute to the overall background, the $t\bar{t}$ process is the most significant, as can be seen in the feature distributions in Chapter 8. Thus, the training of the machine learning models is restricted to $t\bar{t}$ events. A Feynman diagram illustrating $t\bar{t}$ production is shown in Figure 5.2.

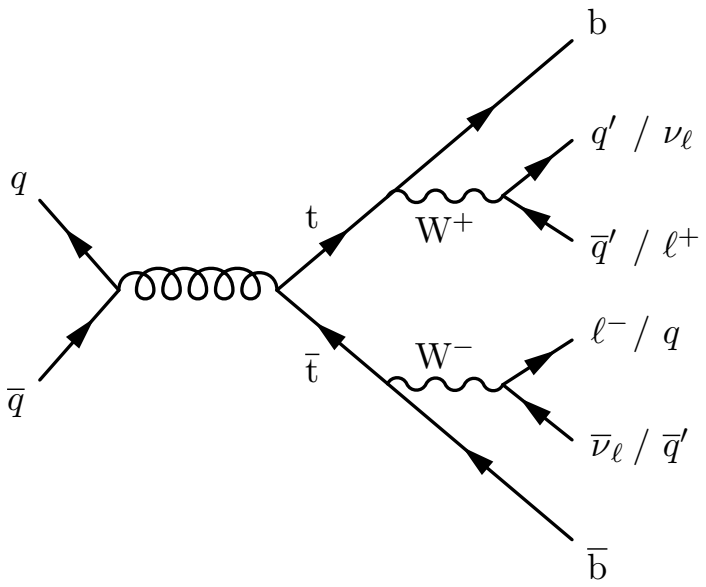


Figure 5.2: Feynman diagram of $t\bar{t}$ decaying into $b\bar{b}WW^*$ with one lepton in the final state [6].

For the specific $t\bar{t}$ decay topology that mirrors the final state of the $b\bar{b}WW^*$ channel in VBF Higgs pair production, two top quarks with a mass of $m_t = (172.95 \pm 0.53)$ GeV [35] each decay into a beauty quark b and a W boson, respecting charge conservation. In this analysis, the final state requires one W boson to decay hadronically and the other leptonically, analogous to the signal process. This $t\bar{t}$ decay channel thus reproduces the same final-state particles as $HH \to b\bar{b}WW^*$, except for the two forward jets characteristic of VBF production. However, such jets can still arise in the detector from other processes, such as pileup events, soft QCD radiation, or underlying events. To optimise the identification of signal events, the preselection process assigns the highest transverse momentum (p_T) b-tagged large-radius (LR) jet as the candidate for H_{bb}, while the highest p_T non-b-tagged LR jet is assigned as the hadronically decaying W boson (W_{had}). The terms leading and subleading refer to the ordering of jets by their transverse momentum p_T. With prior knowledge of these decay topologies, a preselection of events can be applied to reduce the number of background events before the machine learning algorithms begin training. The preselection criteria are as follows:

1. Exactly two large radius jets (2 LR jets) with $p_T > 200\,\text{GeV}$ and $m > 40\,\text{GeV}$
2. Transverse momentum of the charged lepton (electron or muon) is sufficiently large ($p_{T,\text{lep}} > 10$ GeV)
3. The lepton candidate passes identification criteria. For the electron tight ID and

tight track only-varrad isolation is used, while for the muons tight ID and p flow loose-varrad isolation is used.

4. Confirmed $H_{b\bar{b}}$ LR jet. This happens via a cut of 0.94 on the b-tagging score. The b-tagging algorithm combines impact parameter and secondary-vertex information in a multivariate discriminant to separate b-jets from light-flavour jets, with 0.94 corresponding to a high-purity working point.

5. A pair of forward jets is required, satisfying the condition that both jets are standard sized, not overlapping with the LR jets (ΔR(jet, LRjet) > 1.4) and are well separated in pseudorapidity, with $|\Delta\eta| > 3.5$.

6. H_{bb} and W_{had} masses are above 40 GeV.

Given the mass selection criteria applied to W_{had}, signal events are more likely to feature an on-shell decay of the hadronically decaying W boson.
With these preselections applied and using only LHC Run 2 simulated data, the dataset comprises 95570 $t\bar{t}$ background events and 10944 HH signal events for the coupling modifiers $\kappa_{2V} = 2$ and $\kappa_{2V} = 3$.
The properties stored in the event are the 4-vector kinematics of the

- leading large radius jet
- subleading large radius jet
- leading forward jet
- subleading forward jet
- leading small radius jet
- subleading small radius jet
- lepton jet
- hadronically decaying Higgs boson candidate
- hadronically decaying W boson candidate.

The data also includes information about the missing transverse energy E_T^{miss}, the angle, where this energy is missing φ_T^{miss} (which provides information about the neutrino), the number of small radius jets N_{SR} and the number of forward jet pairs N_{2f}. In the decay channel considered in this thesis, the charged lepton is restricted to either a muon or

an electron, since the τ lepton would decay further. Each simulated event is labelled by lepton flavour in order to distinguish between electrons and muons in the final state.
To ensure that simulated events accurately reflect both the expected physics and the experimental conditions, each event is assigned an event weight ω. This weight is calculated by the Monte Carlo simulation, taking into account the production cross-section σ, the integrated luminosity $\mathcal{L}$, the branching ratios of the relevant decays, Monte Carlo weights derived from perturbative theory corrections, and scale factors that correct for discrepancies between simulation and recorded data.
The Boosted Decision Tree (BDT) algorithm minimises a weighted loss function, assuming non-negative event weights ($\omega_i \geq 0$). When there are some negative weights, for instance due to NLO corrections of the cross section, the empirical misclassification error,

$$\varepsilon = \frac{\sum_i \omega_i \mathbb{I}(y_i \neq h(x_i))}{\sum_i \omega_i}, \tag{5.1}$$

can fall outside the expected range $[0, 1]$. In gradient boosting, negative weights invert the sign of the gradient contributions, causing the algorithm to maximise rather than minimise the loss for these events. This reversal destabilises the boosting procedure, leading to inconsistent and unreliable model performance, as the algorithm is no longer optimising the intended objective.
The neural network is trained to minimise the weighted cross-entropy loss, defined as:

$$\mathcal{L}(q, p) = -\sum_{i=1}^{s} \omega_i \left(p_i \log(q_i) + (1 - p_i) \log(1 - q_i)\right), \tag{5.2}$$

where $p_i \in \{0, 1\}$ are the true binary labels, $q_i \in [0, 1]$ the predicted probabilities from the sigmoid output, s the number of samples, and ω_i the event weights.
The contribution of each event to the loss is proportional to ω_i. If there is only a small number of negatively weighted events (especially in regions of low predicted scores, since the ROC curve is plotted logarithmically), their statistical impact on the sums used in the ROC calculation can cause fluctuations or sideways deviations in the ROC curve.
To avoid these problems, all negative event weights have been replaced by 10^{-6}.
To ensure balanced learning between signal and background events, the event weight plays a crucial role in guiding the machine learning model by indicating the relative importance of each event during training. Given the inherent class imbalance where background events significantly outnumber signal events the neural network (NN) tends to learn more about the background than the signal, potentially biasing the model.
To mitigate this imbalance, the event weights are adjusted to incorporate the signal class weight. Each event is assigned the updated event weight defined as

$$w_i = w_{\text{event}} \cdot \left(1 + \frac{\sum_{\text{events}} w_{\text{bkg}}}{\sum_{\text{events}} w_{\text{sig}}} \cdot \delta_{i,\text{sig}}\right). \tag{5.3}$$

This reweighting ensures that the NN gives appropriate importance to signal events, balancing the learning process. For the Boosted Decision Tree (BDT) implemented via XGBoost, this adjustment is unnecessary, as the algorithm internally computes and applies class weights to handle imbalances.

For both models, the dataset is separated into training (80%), validation (10%), and test (10%) samples. All performance evaluations and final assessments are conducted exclusively on the test sample to ensure unbiased results.

6 Separation of Signal and Background

This chapter discusses the kinematic variables used to distinguish between signal and background events in the analysed samples. The following sections outline the baseline variables and their distributions, highlighting the differences between signal and background processes. After that, the machine learning models, namely the BDT and NN are optimised to the different feature sets.

6.1 Kinematic Variables

The kinematic variables are essential for the discrimination of signal and background events. The distributions of these variables are analysed to identify features that can effectively separate the two classes.

Baseline Variables

Although the simulated dataset contains the mass m, transverse momentum p_T, azimuthal angle φ, and pseudorapidity η of the jets and charged lepton (item 5.2), the φ information is effectively identical for all signals and backgrounds, thus training on this variable would not be beneficial to separate signal from background.
The baseline set of feature variables is defined as the (m, η, p_T) information of all jets, the lepton, the H_{bb} candidate, and the W_{had} candidate, as well as the missing transverse energy E_T^{miss}. Additionally, the number of small-radius jets N_{SR} and the number of forward jet pairs N_{2f} are included. Example distributions for these variables are shown in Figures 6.1 to 6.6, while the remaining distributions are provided in Chapter 8.
First, the mass of the leading large-radius jet, m_{lead}^{LR}, is shown in Figure 6.1. The signal events are divided into $\kappa_{2V} = 2$ and $\kappa_{2V} = 3$ samples, represented in dark blue and red respectively, while the background events originate from $t\bar{t}$ (blue), diboson, single H, dijet, single t, W, and Z production. The $t\bar{t}$ background dominates the other background processes, and there is no significant difference between the two signal categories. For

clarity, the signal and background distributions are normalised to unity. The $t\bar{t}$ distribution peaks at approximately 80 GeV, suggesting that the leading large-radius jet in $t\bar{t}$ events likely originates from the hadronically decaying W_{had}. In contrast, the signal distributions exhibit a strong maximum at around 125 GeV, corresponding to the Higgs boson mass. This indicates that the H_{bb} is more likely to possess higher p_T than the decaying W boson, which is expected given that the topology becomes less boosted with each subsequent decay. All distributions drop below 40 GeV due to the mass cut applied to the large-radius jets. The significant differences between the signal and background distributions in both Figure 6.1 and Figure 6.2 suggest that these feature variables will perform well in the machine learning models.

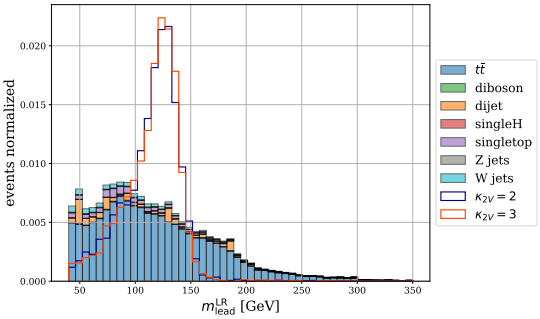


Figure 6.1: Number of normalised events as a function of the mass of the leading large-radius jet in GeV.

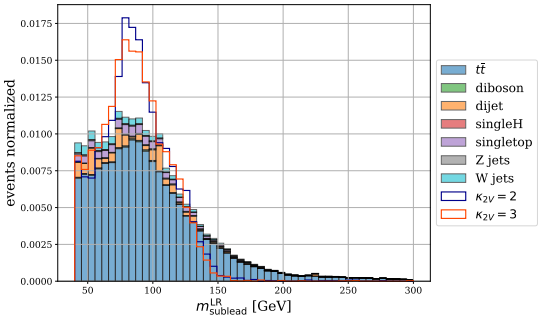


Figure 6.2: Number of normalised events as a function of the mass of the subleading large-radius jet in GeV.

Secondly, the mass of the subleading large-radius jet is shown in Figure 6.2. For the signal distributions, a clear maximum is observed at the W boson mass of 80 GeV, with a smaller local maximum at the H boson mass of 125 GeV. This is consistent with the features in Figure 6.1. For the background distributions, the distributions again peaks at the W boson mass, indicating little preference for which large-radius jet is leading or subleading. There are no events with $m^{\text{LR}}_{\text{sublead}} < 40\,\text{GeV}$ due to the mass cut on the large-radius jets. Next, the transverse momentum of the lepton is considered. The distributions for the signal and background processes are shown in Figure 6.3. Due to the cut $p_T^{\text{lep}} > 10\,\text{GeV}$, the distributions start at values of 10 GeV. In Figure 6.3, the distributions peak at around 60 GeV and then fade out. The signal distributions have a longer tail of events with high transverse momenta, reflecting the boosted nature of the signal topology.

The transverse momentum of the tagged H_{bb} (Higgs boson decaying via $b\bar{b}$) is shown in Figure 6.4. The signal distributions show much higher values than the backgrounds, indicating that HH production is significantly more boosted than the selected backgrounds.

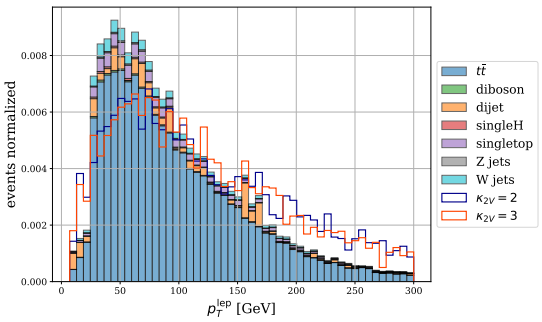


Figure 6.3: Number of normalised events as a function of the transverse momentum of the lepton.

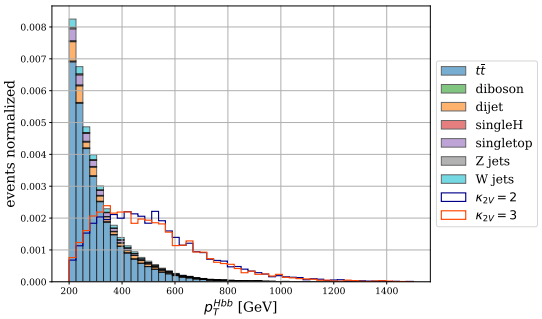


Figure 6.4: Number of normalised events as a function of the transverse momentum of the Higgs boson candidate decaying via $b\bar{b}$ in GeV.

Above 700 GeV, the events are predominantly signal, making this variable highly discriminative. The $p_T^{H_{bb}}$ distributions fall off at 200 GeV due to the applied cut.

The pseudorapidity η of the lepton is shown in Figure 6.5. The ATLAS detector is centred around the interaction point. In the very central region, defined by $|\eta| < 0.1$, the inner detector and the calorimeters have limited acceptance[29]. Therefore, leptons produced in this region are not reconstructed by these subdetectors. Only the muon spectrometer contributes in this region. In addition, in the region $|\eta| \in [1.37, 1.52]$, which corresponds to the transition between the barrel and end-cap calorimeters, the detector performance is similarly reduced, resulting in another minimum for the distributions in Figure 6.5. For high p_T, the reconstructed objects tend to have low η, especially for the signal distributions, due to the boosted topology of HH production.

Within the ATLAS detector at CERN, lepton reconstructions are considered to be reliably reconstructed within the tracking acceptance of $|\eta| < 2.5$, as defined by the coverage of the ID [36]. Thus, the distributions are shown in that region of η.

Finally, the pseudorapidity of the subleading forward jet is presented in Figure 6.6. Jets in the forward region are reconstructed with reliable performance within the calorimeter acceptance up to $|\eta| < 4.5$, consistent with standard ATLAS jet reconstruction criteria [37]. Therefore, the distributions of the pseudorapidities of the subleading forward jet $\eta^{\text{fwd}}_{\text{sublead}}$ for signal and background are shown in that region. The leading forward jet, by definition, has higher p_T than the subleading forward jet, resulting in η peaking at 0 for the leading forward jet. Due to the preselection, the two forward jets must satisfy $\Delta\eta > 3.5$, causing the $\eta^{\text{fwd}}_{\text{sublead}}$ to peak at $|\eta| \approx 3.5$.

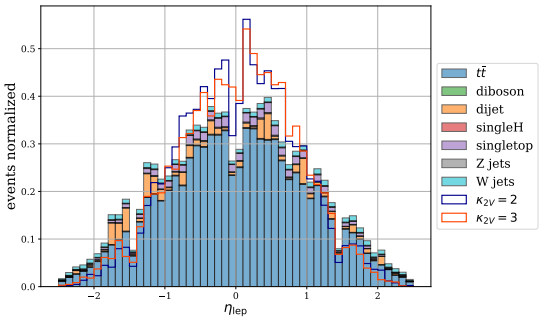


Figure 6.5: Number of normalised events as a function of the pseudorapidity η for the lepton.

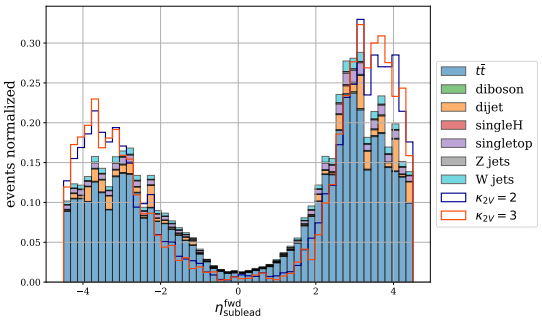


Figure 6.6: Number of normalised events as a function of the pseudorapidity for the subleading forward jet.

6.1.1 Reasonable Features

To improve model interpretability and avoid redundancy, the next step is to restrict the feature set to variables with clear physical significance. This not only aids in validating the model's predictions but also eliminates variables that are either redundant or irrelevant to the underlying physics.

A physically reasonable feature set is thus defined as the (m, η, p_T) of the following:

1. leading and subleading forward jets,
2. leading and subleading large-radius jets,
3. lepton (including information about the lepton type).

Additionally, the missing transverse energy E_T^{miss} and the number of forward jet pairs $N_{2\text{f}}$ are included. Information about the H_{bb} and W_{had} candidates is excluded, as it is largely redundant with the information contained in the large-radius jets. To illustrate this, Figure 6.7 shows the mass of the W_{had} candidate $m^{W_{\text{had}}}$ and Figure 6.8 shows the mass of the subleading large-radius jet $m^{\text{LR}}_{\text{sublead}}$.

Information about the small-radius jets is also excluded, as their relevance is limited. In the VBF HH decay channel, two forward jets with small radius are produced. If the leading or subleading small-radius jets correspond to these forward jets, the information is redundant. Otherwise, these jets are random and irrelevant to the signal event.

To enhance the discriminative power of the feature set, six new composite feature variables are introduced. The first new feature is ΔR_{2f}, the angular distance between the two forward jets shown in Figure 6.9 and defined as

$$\Delta R_{2f} = \sqrt{(\Delta\eta)^2_{2f} + (\Delta\varphi)^2_{2f}}. \tag{6.1}$$

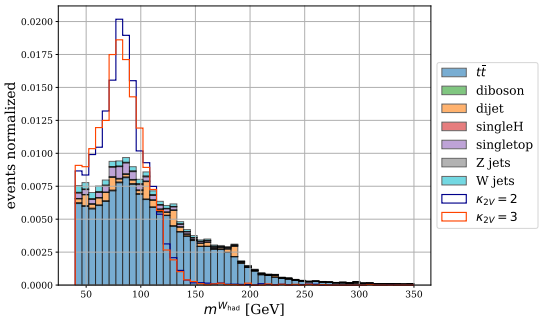


Figure 6.7: Number of normalised events as a function of the mass of the W_{had} candidate.

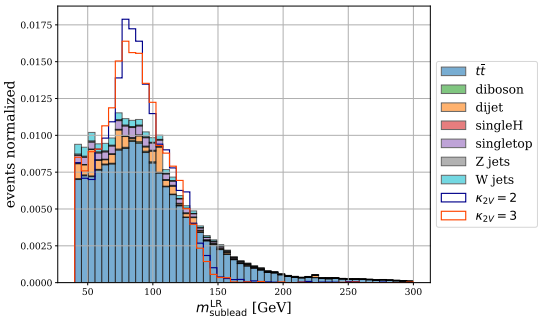


Figure 6.8: Number of normalised events as a function of the mass of the subleading large-radius jet.

Due to the preselection criterion $\Delta\eta_{2f} > 3.5$, the distributions begin at 3.5. For background events, the forward jets are randomly selected within the detector from underlying events or pileup, leading to lower $\Delta\eta_{2f}$ values compared to the signal events, where two distinct forward jets are measured.

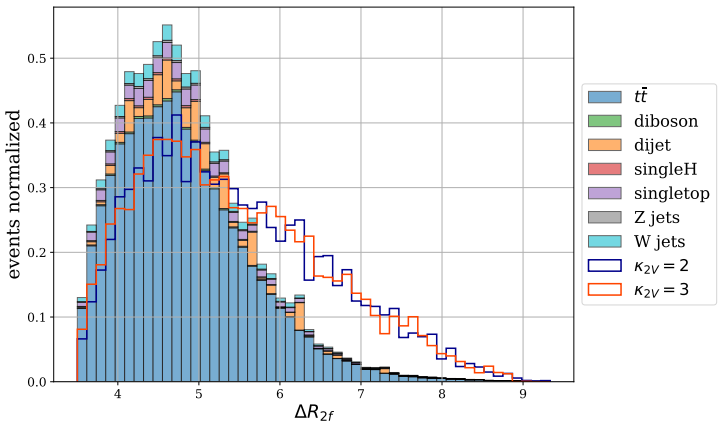


Figure 6.9: Number of normalised events as a function of the angular distance ΔR_{2f} of the leading and subleading forward jets.

The next feature is the angular distance between the lepton and the subleading large-radius jet $\Delta R_{\text{lep, subleadLR}}$ shown in Figure 6.10. For the $HH \to b\bar{b}WW^*$ decay, the leptonically decaying W boson and thus the lepton should be near the hadronically decaying W boson due to the boosted topology of the channel. The signal distributions peak at low $\Delta R_{\text{lep, subleadLR}}$ values, with an additional peak at π due to local maxima at $-\pi$ and π in the $\Delta\varphi$ distribution. This occurs when the subleading large-radius jet does not correspond to the W_{had}, and $\Delta\varphi$ measures the angle between the lepton and the H_{bb} candidate.

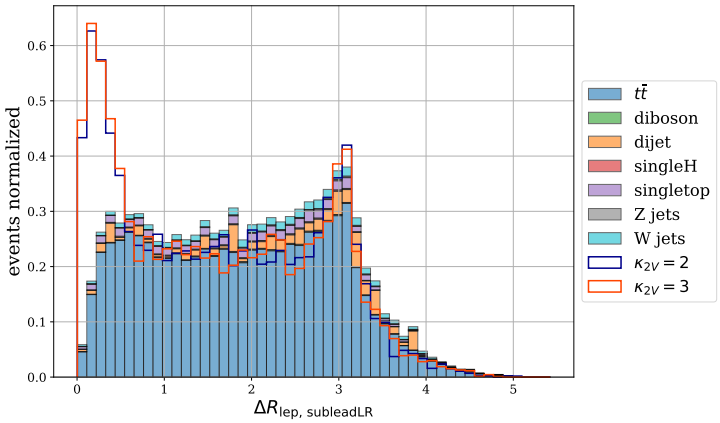


Figure 6.10: Number of normalised events as a function of the angular distance between the lepton and the subleading large-radius jet, $\Delta R_{\text{lep, subleadLR}}$.

The invariant mass of the two leading forward jets m_{forward},

$$\begin{aligned} m_{\text{forward}} &= \sqrt{(p_{f1} + p_{f2})^{\mu}(p_{f1} + p_{f2})_{\mu}} \\ &= \sqrt{(E_{f1} + E_{f2})^2 - (\vec{p}_{f1} + \vec{p}_{f2})^2}. \end{aligned} \tag{6.2}$$

where $f1$ and $f2$ denote the leading and subleading forward jets, is introduced as the next feature. The corresponding distributions are shown in Figure 6.11. The signal distributions exhibit higher invariant masses, as the Higgs boson pair production requires substantial energy from the VBF quarks, that then further decay into forward jets. Thus, the signal distributions indicate higher energies for the forward jets compared to random jets from underlying events or pileup in case of the background samples. This is evident in the distribution, where the signal peaks at higher values and exhibits a longer tail compared to the background.

Next, the ratio of the leading and subleading large-radius jet masses, ζ_{LRjet}, is defined as:

$$\zeta_{\text{LRjet}} = \frac{m_{\text{leading LR}}}{m_{\text{subleading LR}}}. \tag{6.3}$$

The corresponding distributions for signal and background are shown in Figure 6.12. For the signal samples, $m_{\text{leading LR}}$ is typically the Higgs boson mass, and $m_{\text{subleading LR}}$ is the W boson mass, resulting in a maximum at $125/80 \approx 3/2$, as can be seen in the plot. For $t\bar{t}$ events with a final state similar to the signal events, the one LR jet can contain the jets of either a beauty quark or a beauty quark and a charged lepton. The other LR jet then contains any combination of the jets of the second beauty quark and the decay products from the hadronically decaying W boson. For many of these options, this leads to a mass

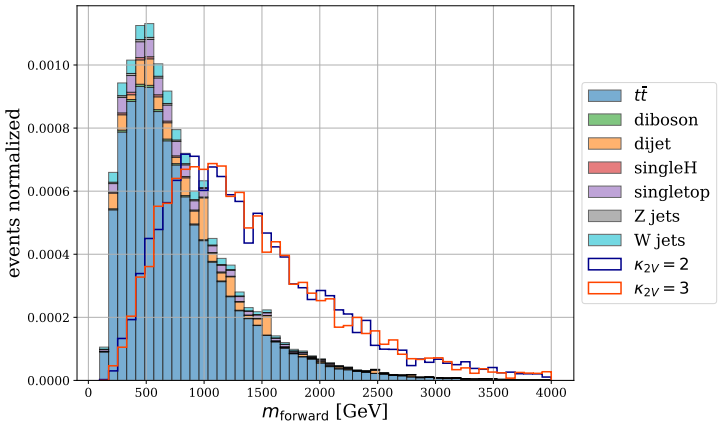


Figure 6.11: Number of normalised events as a function of the invariant mass of the two VBF quarks in GeV.

ratio of about $\zeta_{\text{LRjet}} \approx 1$ for the considered $t\bar{t}$ events. Thus, a maximum can be seen in the $t\bar{t}$ distribution in Figure 6.12.

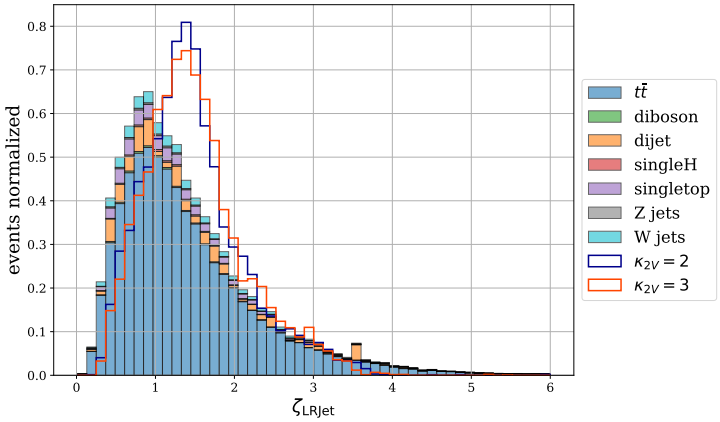


Figure 6.12: Distribution of the ratio of leading and subleading large-radius jet masses, ζ_{LRjet}.

The ratio of the leading and subleading forward jet transverse momenta, ζ_{SRjet}, given by

$$\zeta_{\text{SRjet}} = \frac{p_T^{\text{leading}}}{p_T^{\text{subleading}}} \tag{6.4}$$

is shown in Figure 6.13. This feature tests for symmetry or asymmetry in the transverse momenta of the forward jets. As shown in Figure 6.13, there is no significant difference between the signal and background distributions, suggesting that this feature may not improve model performance. No significant differences are observed between signal and background distributions.

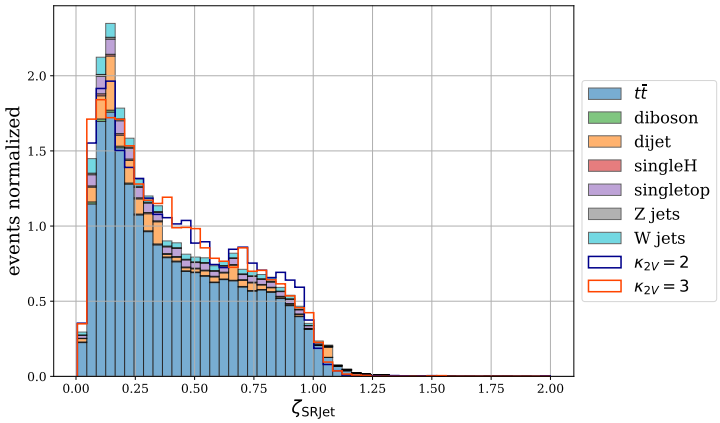


Figure 6.13: Distributions of the ratio of leading and subleading forward jet transverse momenta, ζ_{SRjet}.

Finally, the transverse mass of the leptonically decaying W boson

$$m_T^{W_{\mathrm{lep}}} = \sqrt{(E_T^{\mathrm{lep}} + E_T^{\mathrm{miss}})^2 - (p_x^{\mathrm{lep}} + p_x^{\mathrm{met}})^2 - (p_y^{\mathrm{lep}} + p_y^{\mathrm{met}})^2}, \tag{6.5}$$

as shown in Figure 6.14 is implemented. E_T^{miss} refers to the missing transverse energy attributed to undetected neutrino, and p_x^{met} and p_y^{met} are derived from the relations in Equation 4.6. Since the neutrino's 4-vector cannot be fully reconstructed, the transverse mass $m_T^{W_{\mathrm{lep}}}$ is used. Due to preselection cuts, the leptonically decaying W boson of the signal decay channel is often off-shell. For $t\bar{t}$ decays, the W_{lep} mass is assumed to be $m_W = 80\,\mathrm{GeV}$, while for the signal, it is approximately $m_H - m_{W_{\mathrm{had}}} = 45\,\mathrm{GeV}$. The distributions in Figure 6.14 reflect the expected behaviour, with the signal peaking below 80 GeV and the $t\bar{t}$ distribution peaking at 80 GeV. For the signal, the W boson decays off-shell, causing the transverse mass to peak below 80 GeV, while the $t\bar{t}$ distribution peaks at 80 GeV.

With the baseline features, physically reasonable features, and new composite features now defined, the next step is to train the neural network and boosted decision trees to evaluate whether these theoretical considerations hold in practice. For a better clarity, the feature sets are again summarised in Table 6.1

6.2 Model Evaluation and Optimisation

To assess the performance of the machine learning models, both a neural network (NN) and a boosted decision tree (BDT) are trained and evaluated, with particular attention to avoiding overfitting and ensuring robust generalisation.

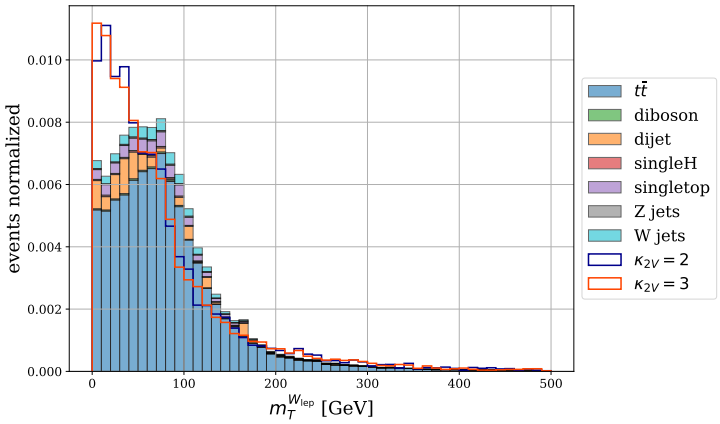


Figure 6.14: Number of normalised events as a function of the transverse mass of the leptonically decaying W boson.

Feature set	**Constituents**
baseline	m, η and p_T of all jets (SR jets, forward jets, LR jets, H_{bb} and W_{had} candidates) and the lepton, E_T^{miss}, $N_{2\text{f}}$, N_{SR}
new features	ΔR_{2f}, $\Delta R_{\text{lep, subleadLR}}$, m_{forward}, ζ_{LRjet}, ζ_{SRjet}, $m_T^{W_{\text{lep}}}$
physically reasonable features	m, η and p_T of the forward jets and LR jets and the lepton, E_T^{miss}, $N_{2\text{f}}$

Table 6.1: Constituents of the different feature sets.

The neural network is discussed first. The learning rate is set to 0.0001, which is lower than the default value of 0.001 for the Adam optimiser. The network architecture consists of 5 hidden layers, each with 100 neurons, following the guideline that the hidden size should be more than twice the number of input features (32 baseline features + 6 new features). The model is trained for 30 epochs, which is deemed sufficient for the given requirements, and a batch size of 500 events per batch is used.
The boosted decision trees are configured with a learning rate of 0.01 and a maximum of 10,000 trees. However, this maximum is rarely reached, as training automatically terminates, if the loss function for the validation dataset does not decrease for five consecutive boosting rounds. The evaluation of both models is conducted exclusively on the test dataset.

6.2.1 Baseline evaluation for both models

The BDT and the NN are evaluated for the training on the baseline feature set.
In Figure 6.15 and in Figure 6.16, the training histories for the BDT and NN are shown. The loss function value for the training and validation sample is plotted as a function of

the number of trees for the BDT and as a function of the epoch for the NN.

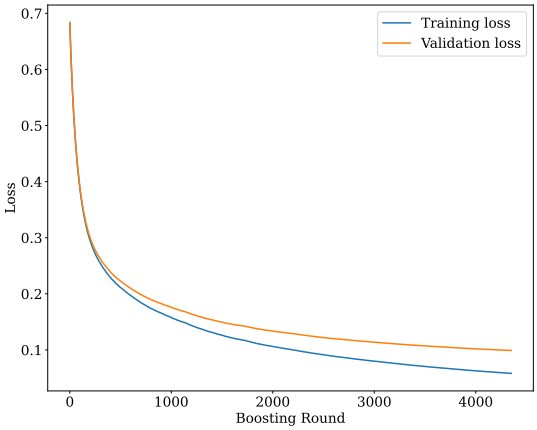


Figure 6.15: BDT loss function over the number of trees for the validation and training datasets.

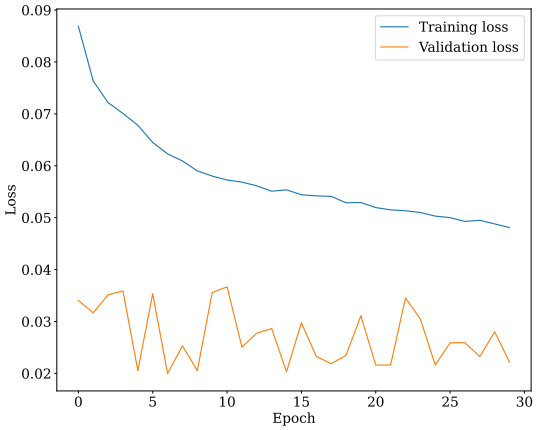


Figure 6.16: NN loss function over the number of epochs for the validation and training datasets.

For the BDT, the training and validation loss decrease smoothly. This can be attributed to the definition of the loss function, which adds a term for each tree. For the NN, the training loss decreases while the validation loss fluctuates and decreases more slowly. Setting the value of the learning rate to even lower values would likely suppress these fluctuations but also degrade performance, as it would take much longer to converge to a (local) minimum of the loss function. This would then need more computation time. The fluctuations can be explained by the use of batches. When the total number of batches is small (i.e. when the batch size is large), the loss function behaves more stable because the gradient is computed from a limited number of aggregated components. However, this also leads to reduced network performance, as the variability between individual samples is less effectively represented in the training process. The training history and the ROC curve comparison for the NN trained on the baseline feature set with a high batch size of 50000 are shown in Figure 6.17 and Figure 6.18 respectively.
The plots in Figure 6.19 and Figure 6.20 display the normalised number of events with ordered outputs of the BDT and NN respectively between [0,1] for each event.
The BDT demonstrates better performance on the baseline features, as evidenced by output distributions. The BDT's output values at 0 and 1 peak much higher, indicating greater confidence in its classifications.
In Figure 6.21, the signal efficiency, also known as the true positive rate (TPR),

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{6.6}$$

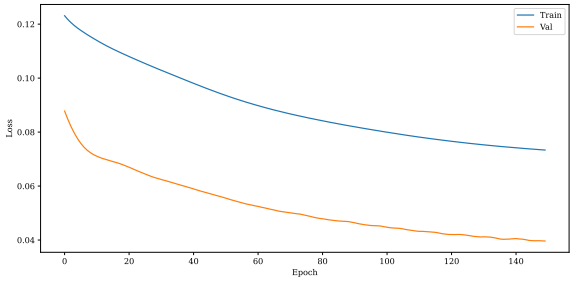


Figure 6.17: Training and validation loss over the number of epochs for a NN trained on the baseline feature set with large batch size ($bs = 50000$).

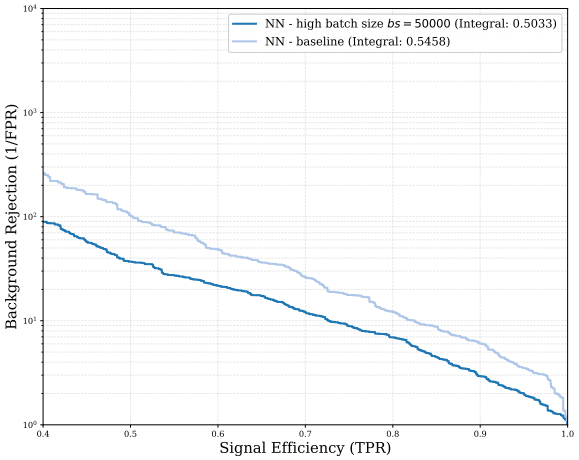


Figure 6.18: Comparison of the background rejection as a function of signal efficiency for the NN trained on the baseline feature set for default batch size $bs = 500$ and $bs = 50000$.

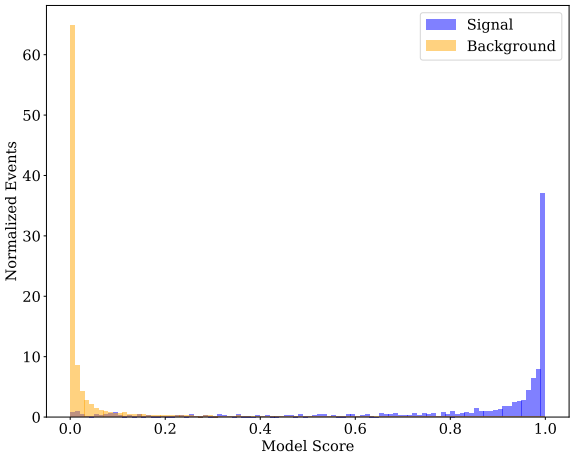


Figure 6.19: Event outputs for the BDT.

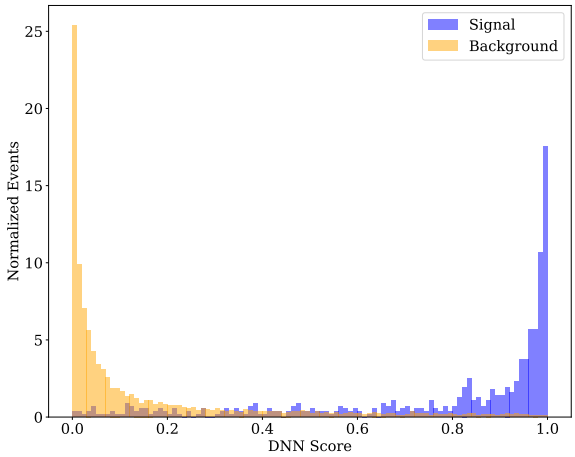


Figure 6.20: Event outputs for the NN.

where TP is the number of true positives (correctly labelled signal events) and FN is the number of false negatives (incorrectly labelled signal events), is plotted against the background rejection, defined as the inverse of the false positive rate (FPR),

$$\frac{1}{\text{FPR}} = \frac{\text{TN} + \text{FP}}{\text{FP}}, \tag{6.7}$$

where FP is the number of false positives (incorrectly labelled background events) and TN is the number of true negatives (correctly labelled background events). A higher curve

indicates better performance for given value of TPR, as the goal is to achieve both high signal efficiency and high background rejection. The given integral values are calculated as the area under curve for 1-FPR over TPR in the interval of [0.4,1].

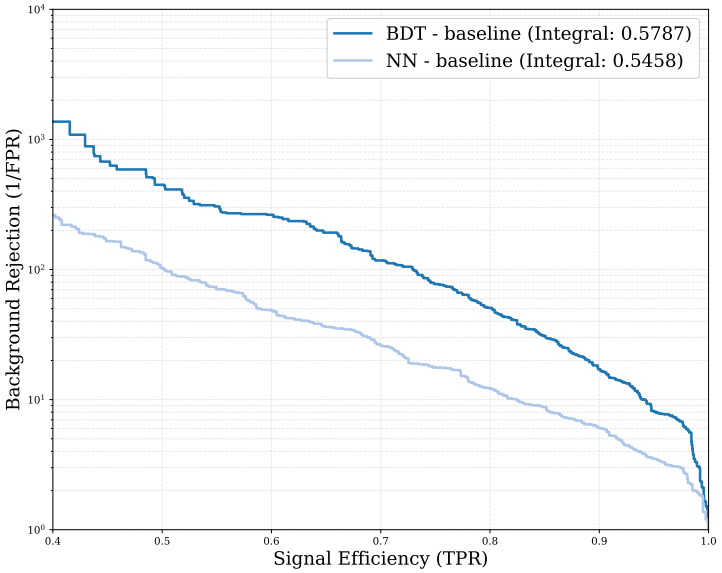


Figure 6.21: Background rejection over signal efficiency for the BDT and NN trained on the baseline features in the interval [0.4, 1]. The integral value represents the area under the curve in this interval for (1-FPR) over TPR.

Overfitting is a common concern for machine learning models. To evaluate this, the Kolmogorov-Smirnov (KS) test is employed, which measures the difference between the distributions of the test and training datasets. Two values are calculated: the statistical value T,

$$T := \sup_{z \in \mathbb{R}} |F_n(z) - G_m(z)|, \tag{6.8}$$

where F_n and G_m are the distributions for the training and test data with n and m samples being the test and training samples respectively. T measures the maximum absolute difference between the two distributions. The p-value gives the probability that the observed difference between the two distributions is random, assuming $F = G$. Ideally, the test and training distributions should be similar, resulting in a small T and a large p-value. A sign of overfitting would be a large value of T and a small p-value, since the training sample would have more distinct outputs. The default threshold of $p = 0.05$ is chosen to limit the false positive rating to a probability of 5%. For the neural network, $T_{\mathrm{NN}} = 0.007$ and $p_{\mathrm{NN}} = 0.709$. For the boosted decision trees, $T_{\mathrm{BDT}} = 0.012$ and $p_{\mathrm{BDT}} = 0.128$. The KS test does not indicate clear signs of overfitting in either case.

Another method to check for overfitting is to examine the training histories of the machine

learning models shown in Figure 6.15 for the BDT and in Figure 6.16 for the NN. For the BDT, there is no indication of overfitting, as both curves decrease throughout the entire training process. Due to the heavy fluctuations, the training history of the NN does not provide a definitive indication of overfitting. But since the KS test yielded small differences in the test and training sample distributions, it is reasonable to assume that the NN is not overfitting.

Now, considering the individual features in more detail: The feature correlation matrices for the given samples are shown in Figure 6.22.

Figure 6.22: Correlation matrix for BDT and NN trained on baseline feature set.

The higher correlation values can be categorised into three groups. The first group includes correlations between the transverse momenta p_T and masses m. For the forward jets and

the small radius jets, these correlations have no obvious physical reason. For the LR jets on the other hand, the correlations can be explained, since the LR jet with higher mass is more likely to be the H_{bb} candidate, which has more likely a higher p_T value. The second group includes correlations of p_T or the mass m of the leading and subleading large-radius (LR) jets with the H_{bb} and W_{had}.The connections between the subleading LR jets and H_{bb}, as well as between the leading LR jets and W_{had} primarily originate from the $t\bar{t}$ decay rather than the signals, where they are not expected to be highly correlated. This type of correlation was anticipated and is the reason for training on only the physically reasonable features. The last group of correlated features includes the pseudorapidities η for the leading and subleading forward jets, as previously discussed in Figure 6.6, or the correlation between the number of small-radius jets N_{SR} and the number of forward jet pairs $N_{2\text{f}}$.

Next to the feature correlations, it is useful to know how important the individual features are for the training of the applied machine learning models. This not only helps to optimise the machine learning models, but also to compare the physical expectations with the actual training. Starting with the feature importances of the BDT: Each importance is calculated via the mean gain of applying the feature in all trees. The most important features are $p_T^{\text{leading, LR}}$, $m_{H_{bb}}$, $m_{W_{\text{had}}}$, and $m_{\text{leading, LR}}$. All these variables share the property that the signal and background distributions are sufficiently different to separate a large number of events via a simple cut (Chapter 8).

The feature importances of the NN, calculated by permuting the features for the trained NN and measuring the resulting performance loss, show a different pattern. The ranking of feature importances is shown in Figure 6.23.

As can be seen in Figure 6.23, the most important features are $p_T^{\text{leading, LR}}$, $p_T^{W_{\text{had}}}$, $m_{H_{bb}}$, and $p_T^{\text{leading, f}}$ (Chapter 8). Surprisingly, $p_T^{W_{\text{had}}}$ features no significant difference in the distributions for signals and backgrounds. Due to the method of calculation, some feature importances have negative values. This is the case, for example, for E_{lep} or $p_T^{H_{bb}}$. Negative importances indicate that interchanging these features with other variables in the input layer actually improves performance. Thus, these features negatively affect the NN. On the other hand, if a feature is highly correlated, changing it could even be beneficial. Although this is a borderline case, this effect could be possible for features like $p_T^{H_{bb}}$, where the distributions appear promising.

6.2.2 Optimisation via Features

To now optimise the performance of the models, the impact of using physically reasonable features and newly implemented features is evaluated. All comparisons are made relative

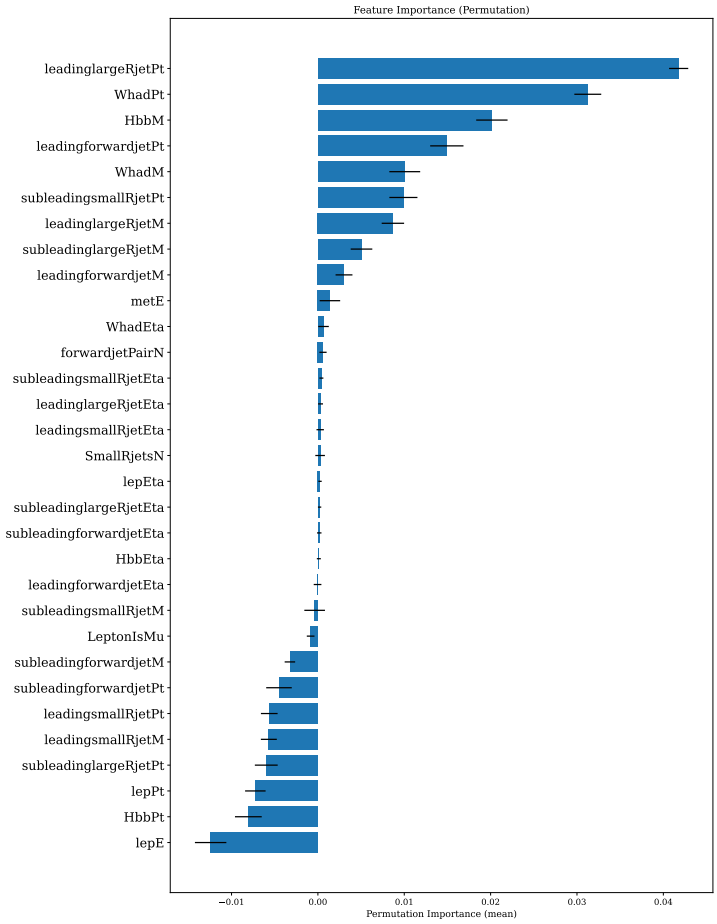


Figure 6.23: Feature importances for the NN trained on baseline features.

to the baseline feature set, which serves as the reference for assessing improvements or degradations in model performance.

The ROC curve for the neural network (NN) trained on the physically reasonable features is compared to the baseline in Figure 6.24. The baseline feature set yields higher performance, indicating that the inclusion of all features is more effective than restricting to physically reasonable ones. A similar trend is observed for the boosted decision tree (BDT), as shown in Figure 6.25, where the background rejection as a function of the signal efficiency (ROC curves) for both feature sets are compared. The newly implemented features are added to the physically reasonable feature set. For the BDT, as shown in Figure 6.24, this addition improves performance, though not sufficiently to match the baseline. For the NN, the ROC curves in Figure 6.25 indicate no significant performance increase with the inclusion of the new features.

The physically reasonable feature set does not enhance the performance of either the NN or the BDT in comparison to the baseline. Consequently, the BDT and NN are

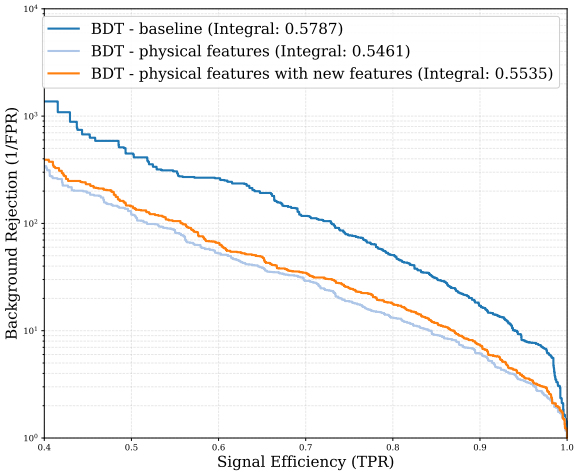


Figure 6.24: Background rejection as a function of the signal efficiency for the BDT trained on the baseline, physically reasonable, and physically reasonable feature sets with new features.

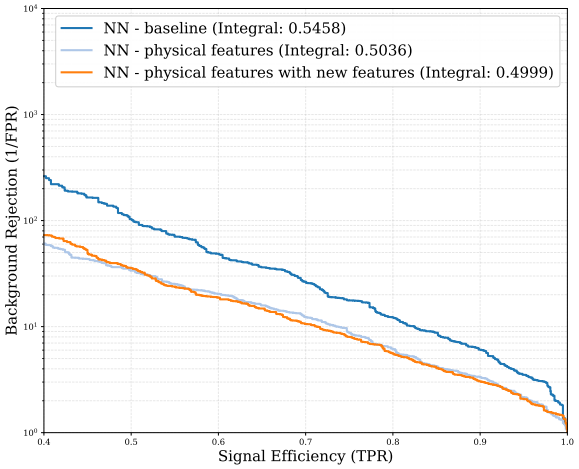


Figure 6.25: Background rejection as a function of the signal efficiency for the NN trained on the baseline, physically reasonable, and physically reasonable feature sets with new features.

evaluated using the baseline feature set with the newly implemented features. As shown in Figure 6.26, the BDT exhibits improved performance with the new features, whereas the NN shows no significant change, as shown in Figure 6.27.

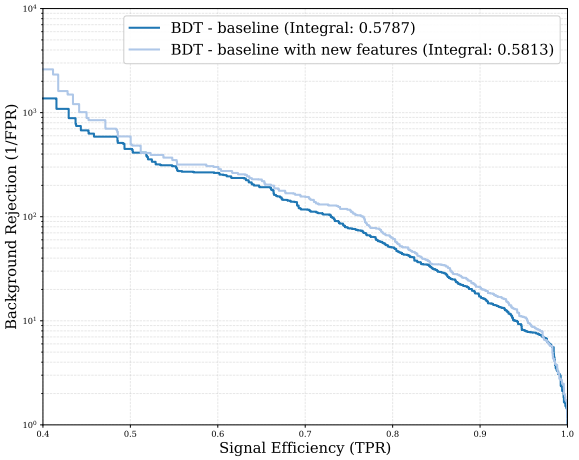


Figure 6.26: Background rejection as a function of the signal efficiency for the BDT trained on the baseline feature set and the baseline feature set with new features.

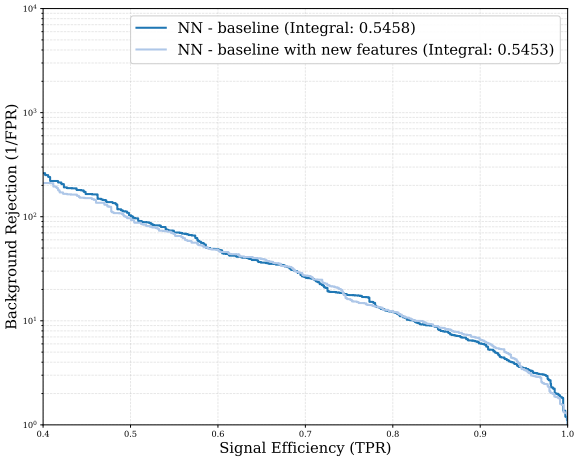


Figure 6.27: Background rejection as a function of the signal efficiency for the NN trained on the baseline feature set and the baseline feature set with new features.

Given that the BDT operates through a series of optimised cuts, it is reasonable to hypothesise that its performance would improve if only high-quality features were used. To test this, the BDT is trained on the top 12 features, as determined by the gain of feature importance from the physically reasonable feature set with new features. This set includes (in order of importance) m_{forward}, $m_{\text{leading, LR}}$, $p_T^{\text{leading, LR}}$, $m_{\text{sublead, LR}}$, $\delta_{\text{lep, }\mu}$, $p_T^{\text{sublead, LR}}$, $p_T^{\text{leading, f}}$, ζ_{LRjet}, $m_T^{W_{\text{lep}}}$, p_T^{lep}, $\Delta R_{\text{lep, subleadLR}}$, and ΔR_{2f} (in order of importance). Notably, five of the six new features (excluding ζ_{SRjet}) are among the top 12, with m_{forward} being the most important. However, the ROC curves in Figure 6.28 reveal that the BDT trained on these top 12 features performs worse than when trained on the physically reasonable features with the new features. This suggests that a larger feature set generally leads to better results for the used samples.

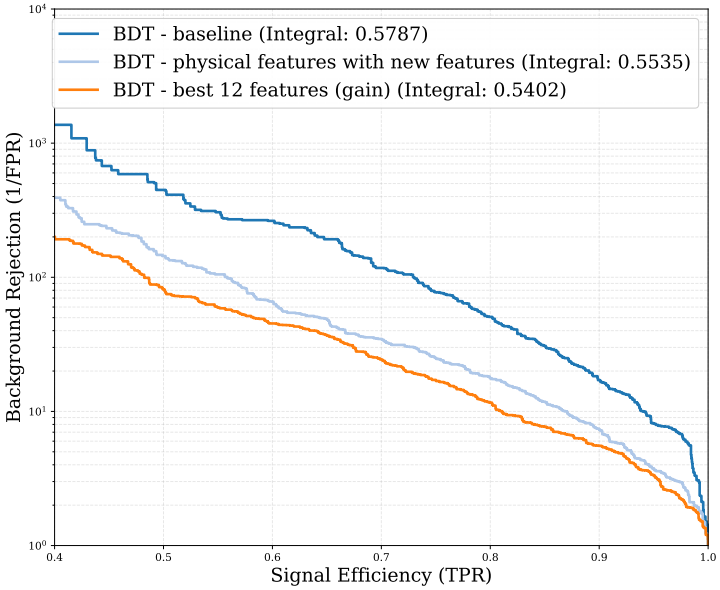


Figure 6.28: Background rejection as a function of the signal efficiency for the BDT trained on the best 12 features, the baseline feature set, and the physically reasonable feature set with new features.

6.2.3 Hyperparameter Optimisation

As stated before, the right choice of hyperparameters is essential to optimise a machine learning model. Since the NN performance was in general lower than the performance of the BDT, only the BDT hyperparameter optimisation is conducted. For that, the impact of varying the learning rate for the BDT was examined. The maximum number of trees was not adjusted, as it was never reached in previous training sessions. The Background rejection as a function of the signal efficiency for the BDT trained on the baseline features

with learning rates of 0.01, 0.05, and 0.001 is shown in Figure 6.29. There is no significant performance difference between learning rates of 0.01 and 0.05, but performance degrades at 0.001. This degradation occurs because the training process is prematurely stopped after 10,000 trees, before convergence is achieved. Since adjusting the number of trees to a higher value would lead to more computation time, a learning rate of 0.01 is optimal.

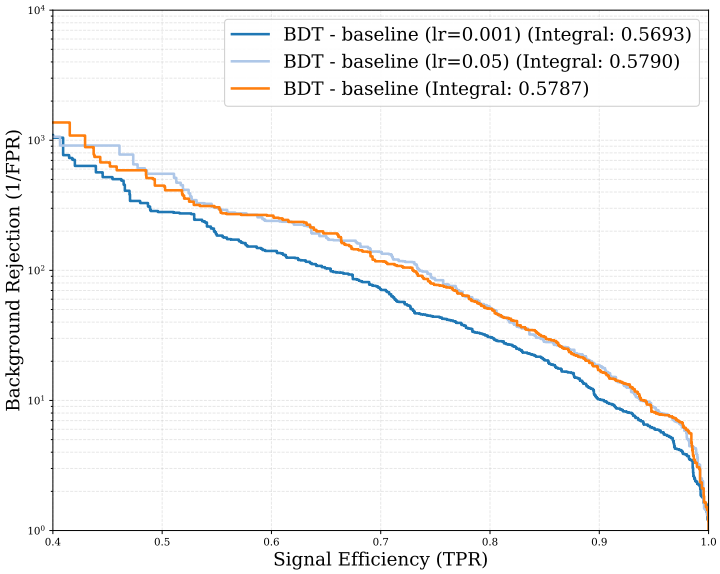


Figure 6.29: Background rejection as a function of the signal efficiency for the BDT trained on the baseline feature set with different learning rates.

6.3 Results

Having optimised the machine learning models through systematic feature selection and hyperparameter tuning, the best-performing configuration is identified as the Boosted Decision Tree (BDT) with a learning rate of $lr = 0.01$, trained on the baseline features augmented with the newly implemented variables. For clarity, this is shown in Table 6.2.

machine learning model	BDT
features to train on	baseline features + new features
learning rate	0.01
maximal number of trees	10000

Table 6.2: Optimised model parameters

To further validate the robustness of this BDT model, its performance is evaluated on extended test datasets. First, signal events with $\kappa_{2V} = 0$, $\kappa_{2V} = 0.5$ and $\kappa_{2V} = 1.5$ are

incorporated into the test dataset, which previously included only $t\bar{t}$ events and signal data with $\kappa_{2V} = 2, 3$. This leads to 31351 signal events in total. As shown in Figure 6.30, the BDT's performance slightly improves with the inclusion of these additional signal events. This adaptability is particularly advantageous for applications of ATLAS measurements, as it demonstrates the model's ability to generalise across different Beyond Standard Model (BSM) predictions without degradation in performance.

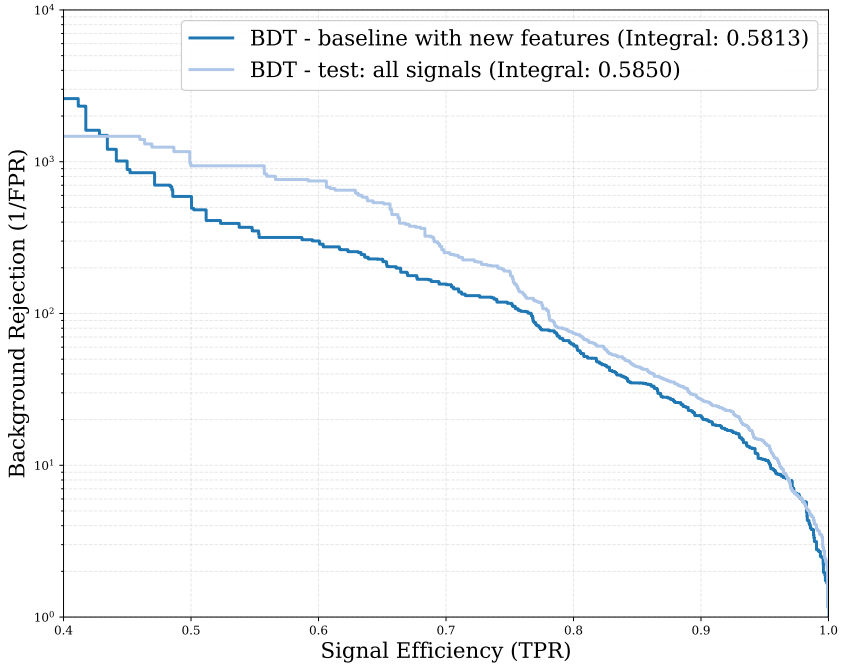


Figure 6.30: ROC curves of the BDT, trained on the baseline feature set with new variables, evaluated on the original test dataset and an extended test dataset including signal events with $\kappa_{2V} = 0, 0.5, 1.5$. The model's performance slightly improves with the inclusion of additional signal events.

Next, the BDT model is tested on a dataset that includes additional background channels: single Higgs boson decay, single top quark decay, Z boson + jets, W boson + jets, and diboson decay. As shown in Figure 6.31, the BDT model's performance decreases when evaluated on all these background samples.

The means of the outputs from the individual background are shown in Table 6.3. The background class is assigned to the BDT output value of 0. The $t\bar{t}$ events have the lowest mean value, whereas all other background processes have higher values, with diboson decay being the worst predicted background process.

While the degradation of the BDT's performance for testing on all background samples can be partially mitigated by retraining the model on the expanded dataset, including both the new background and signal events, as depicted in Figure 6.32, this approach

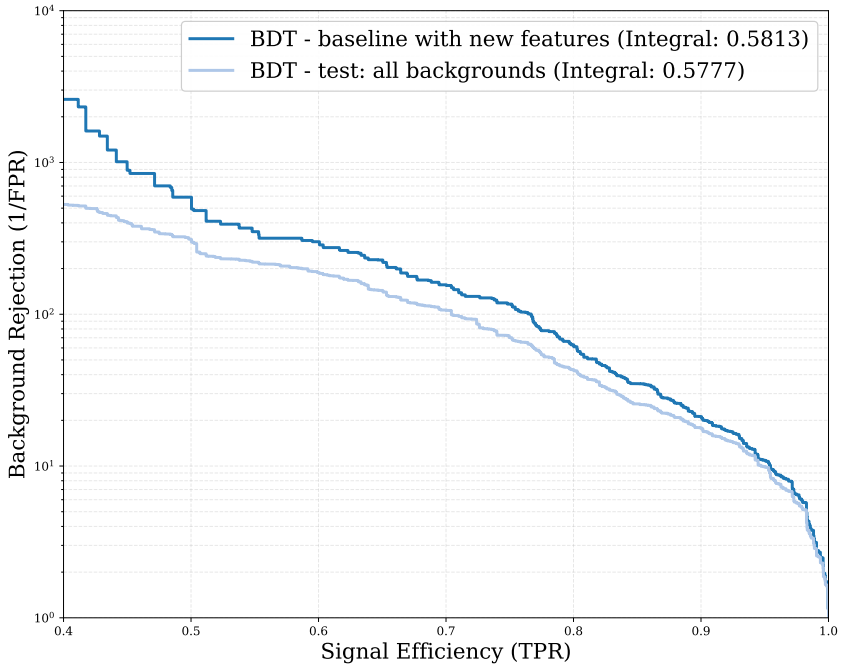


Figure 6.31: Background rejection as a function of signal efficiency for the BDT, trained on the baseline feature set with new features, evaluated on a test dataset that includes all additional background channels.

Background	**Events**	**Weighted Mean**
diboson	68266	0.080853
single H	69437	0.066829
Z + jets	74310	0.065782
single top	7513	0.063940
W + jets	143692	0.057973
dijet	1166	0.045940
$t\bar{t}$	95570	0.030956

Table 6.3: Mean BDT scores for all background processes.

incurs a significant computational cost, as the number of training events increases by a factor of 3.

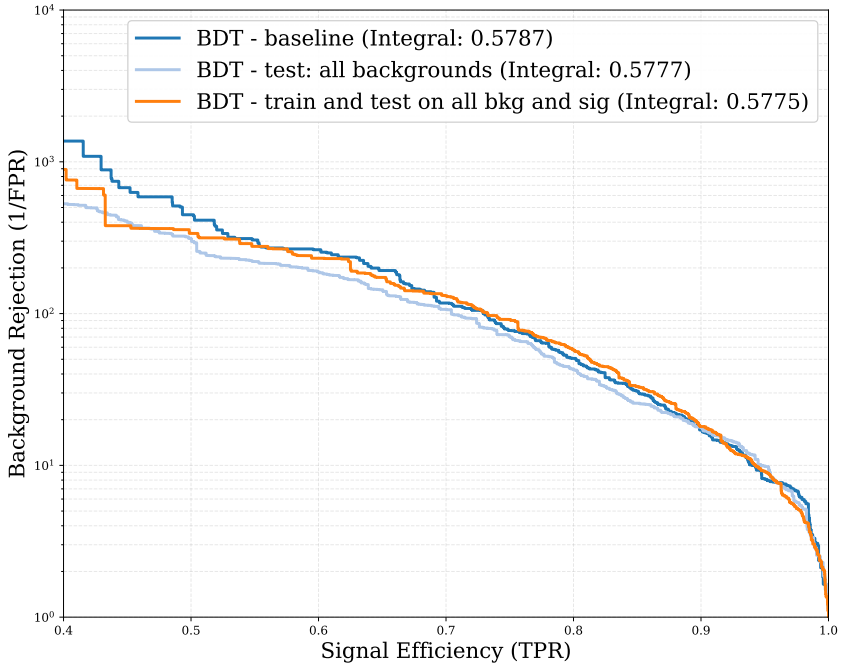


Figure 6.32: Background rejection as a function of signal efficiency for the BDT trained and tested on the baseline feature set with new features, including all available signal and background data.

7 Conclusion

In this thesis, the VBF production of Higgs boson pairs decaying via the $HH \to b\bar{b}WW^*$ channel, with one lepton in the final state, was investigated.

For that, two machine learning models were proposed to separate signal from background events for the investigated decay channel. The two machine learning models, namely a feed forward neural network and boosted decision trees, were both trained on Monte-Carlo simulated samples with $t\bar{t}$ decays serving as background events and signal events using BSM coupling modifier $\kappa_{2V} = 2, 3$. The simulated dataset includes a set of features and both the NN and the BDT were optimised using different subsets and extensions of this feature set.

For that, six new features were implemented, both to link different objects and to boost the quality of the feature set. These new features are $\Delta R_{qq'}$, $\Delta R_{\text{lep, subleadLR}}$, $m_{q\bar{q}}$, R_{LRjet}, R_{SRjet} and $m_T^{W_{\text{lep}}}$. All but the R_{SRjet} were important features to increase the performance, especially for the BDT.

The performance of the NN did not significantly increase when adding the new features and even dropped when the feature variables, that were not part of the physically reasonable set, were taken away.

The BDT generally performed better, when more features were used. Overall, the BDT had a better performance than the NN for every tested feature combination, and the best model included in this research was the BDT trained on the baseline feature set and all newly implemented features. Therefore, this model was tested on all available background and signal data from the Monte Carlo simulation. In case of adding additional background sources in the test data set (single Higgs boson decay, single top quark decay, Z boson + jets, W boson + jets, and diboson decay), the performance slightly decreased again, whereas for added signal samples with $\kappa_{2V} = 0\,, \kappa_{2V} = 0.5$ and $\kappa_{2V} = 1.5$ to the test dataset, the performance even improved, likely because of more signal statistics to train on in total numbers.

During the research process, the BDT had several other advantages. It was easier to handle, the results were more understandable, and the computation time was much lower than for the NN. However, the low performance of the NN can be explained by the low

signal statistics.
It then may be helpful to give the background samples more weight, in order to ensure, that the background samples different from $t\bar{t}$ are also well separated from the signal events. If computational resources permit, it may be advantageous to include additional background samples in the training process, such as diboson decay samples or even generate more statistics for the signal and background processes to train the machine learning models on. Especially the NN is highly dependent on the amount of data which is provided to it, and therefore might increase its performance with more statistics to match the BDT's performance.
The reliable separation of signal from background events is crucial for the precise determination of the coupling constants within the SM, thereby contributing to a deeper understanding of the underlying theory and enabling increasingly precise tests of the theoretical predictions with experimental observations. The precise identification and separation of signal and background processes in this channel explicitly are essential for the measurement of Higgs boson self-coupling modifiers, which directly probe the Higgs potential and therefore constitute a critical test of the SM.
Machine-learning-based approaches to signal discrimination are becoming increasingly important, as they generally outperform traditional cut-based methods. Both machine learning models that were investigated in this thesis performed well, but the BDT is recommended for further research and also for the later applications on real measurements.

Bibliography

[1] ATLAS Collaboration. Observation of a new particle in the search for the standard model higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1–29, 2012.

[2] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B*, 716:30–61, 2012.

[3] P. W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13:508–509, 1964.

[4] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, 1964.

[5] ATLAS and CMS Collaborations. Combined measurement of the higgs boson mass in pp collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS experiments. *Phys. Rev. Lett.*, 114:191803, 2015.

[6] Universität Zürich, Physik-Institut. Feynman diagrams in LaTeX. `https://wiki.physik.uzh.ch/cms/latex:feynman`, 2024. Accessed: 2026-02-16.

[7] S. L. Glashow. Partial-symmetries of weak interactions. *Nucl. Phys.*, 22(4):579, 1961.

[8] A. Salam. Weak and electromagnetic interactions. *Conf. Proc.*, C680519:367, 1968.

[9] S. Weinberg. A model of leptons. *Phys. Rev. Lett.*, 19:1264–1266, 1967.

[10] M. E. Peskin and D. V. Schroeder. *An Introduction to Quantum Field Theory*. Westview Press, Boulder, CO, USA, 1995.

[11] P. Langacker. Introduction to the standard model and electroweak physics. In T. Han, editor, *TASI 2008: The Dawn of the LHC Era*, pages 3–48, Singapore, 2009. World Scientific. arXiv:0901.0241 [hep-ph].

[12] K. Abeling. *Search for resonant Higgs boson pair production in the $b\bar{b}WW^*$ decay channel in the boosted 1-lepton final state using the full Run 2 ATLAS dataset*. PhD thesis, Georg-August-Universität Göttingen. II.Physik-UniGö-Diss-2022/01.

[13] ATLAS Collaboration. Measurement of the w-boson mass in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 78(2):110, 2018.

[14] P. A. Zyla and others (Particle Data Group). Review of particle physics. *Prog. Theor. Exp. Phys.*, 2024(083C01), 2024.

[15] CERN Yellow Report 4 Working Group. Standard model higgs boson branching ratios and total decay widths. Yellow report, CERN, 2016.

[16] LHC Higgs Cross Section Working Group. Handbook of LHC higgs cross sections: 4. deciphering the nature of the higgs sector. *CERN Yellow Report*, CERN-2016-002-X, 2017. Comprehensive SM cross sections (ggF, VBF, VH, ttH) including uncertainties.

[17] M. B. Green et al. *Superstring Theory: 25th Anniversary Edition*. Cambridge Monographs on Mathematical Physics. Cambridge University Press, 2012.

[18] J. L. Feng. Dark matter candidates from particle physics and methods of detection. *Annu. Rev. Astron. Astrophys.*, 48:495, 2010.

[19] LHC Higgs Cross Section Working Group. Handbook of LHC higgs cross sections: 3. higgs properties and 4. deciphering the nature of the higgs sector. 2013–2016. CERN Yellow Reports.

[20] F. Bishara, R. Contino, J. Rojo, et al. Higgs pair production in vector-boson fusion at the LHC and beyond. *Eur. Phys. J. C*, 77:481, 2017.

[21] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2 edition, 2009.

[22] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. Chapter 6: Feedforward Neural Networks.

[23] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2015.

[24] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[25] TIBCO Software Inc. What is a neural network? `https://www.spotfire.com/glossary/what-is-a-neural-network`, 2024. Accessed: 2026-02-16.

[26] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

[27] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[28] L. R. Evans and P. Bryant. LHC machine. *JINST*, 3, 2008.

[29] ATLAS Collaboration. The ATLAS experiment at the CERN lhc. *JINST*, 3:S08003, 2008.

[30] CMS Collaboration. The CMS experiment at the CERN LHC. *JINST*, 3:S08004, 2008.

[31] ALICE Collaboration. The ALICE experiment at the CERN LHC. *JINST*, 3:S08002, 2008.

[32] LHCb Collaboration. The LHCb detector at the LHC. *JINST*, 3:S08005, 2008.

[33] ATLAS Collaboration. Performance of the ATLAS trigger system in 2015. *Eur. Phys. J. C*, 77(5):317, 2017.

[34] M. Cacciari, G. P. Salam, and G. Soyez. The anti-k_t jet clustering algorithm. *J. High Energy Phys.*, 2008(04):063, 2008.

[35] ATLAS Collaboration. Measurement of the top quark mass with the ATLAS detector using $t\bar{t}$ events with a high transverse momentum top quark. *Phys. Lett. B*, 867:139608, 2025. ATLAS Collaboration most precise single-channel top quark mass measurement.

[36] ATLAS Collaboration. Search for new phenomena with the ATLAS detector at the LHC. *arXiv preprint*, 2024.

[37] ATLAS Collaboration. Jet measurements and acceptance in pp collisions with the ATLAS detector. *Eur. Phys. J. C*, 75:17, 2015.

8 Appendix 1 - features

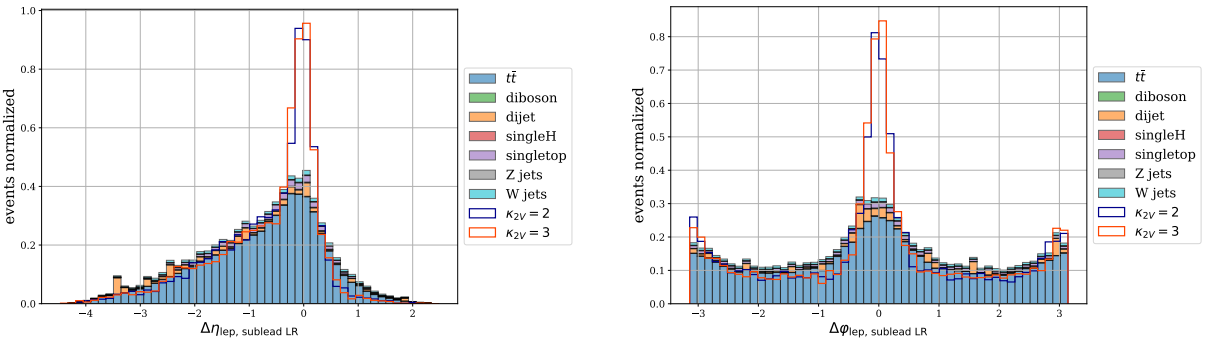


Figure 8.1: Number of normalised events as a function of $\Delta\eta$ and $\Delta\varphi$ between the lepton and the subleading LR jet

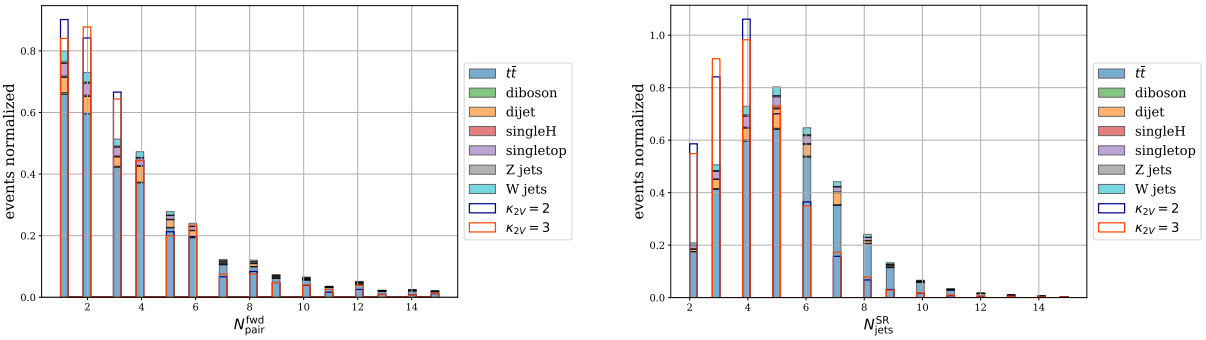


Figure 8.2: Number of normalised events as a function of the number of forward jet pairs (left) and the number of small radius jets (right)

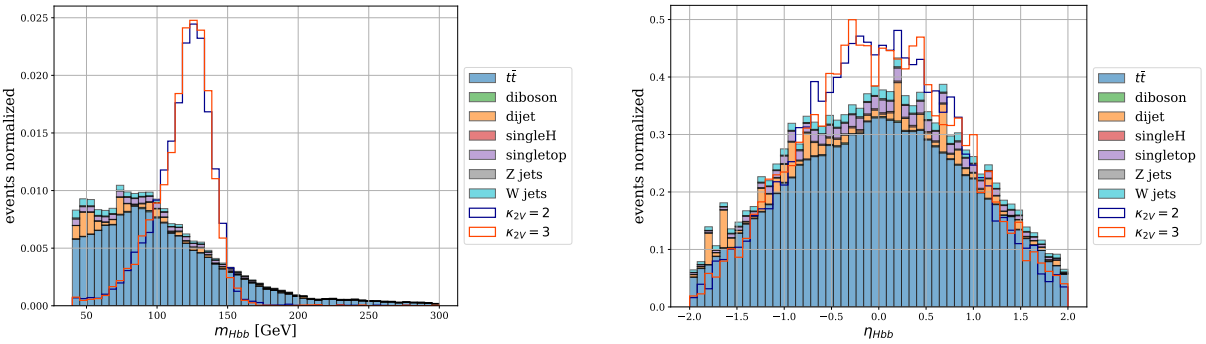


Figure 8.3: Number of normalised events as a function of the mass and η of the H_{bb} candidate

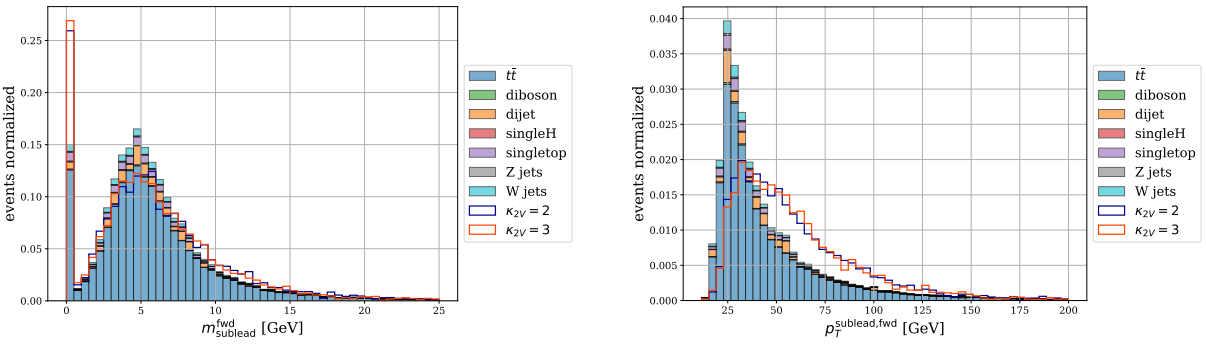


Figure 8.4: Number of normalised events as a function of the mass and transverse momentum of the subleading forward jet

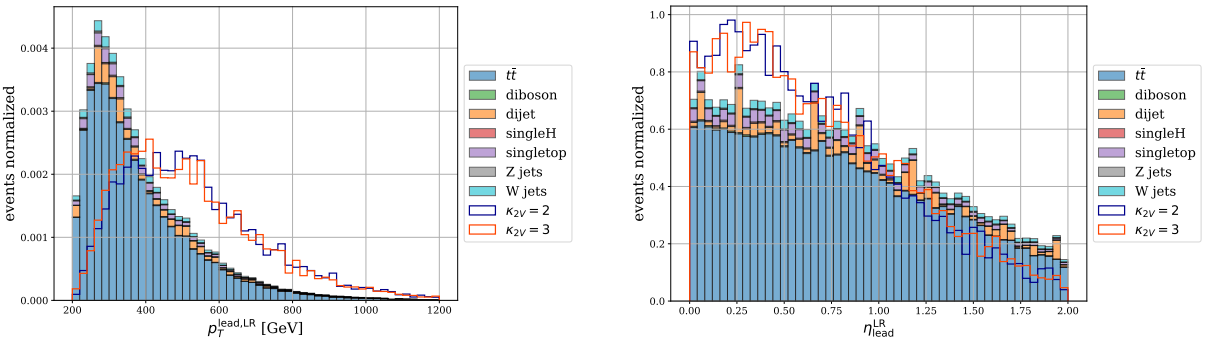


Figure 8.5: Number of normalised events as a function of the transverse momentum and η of the leading LR jet

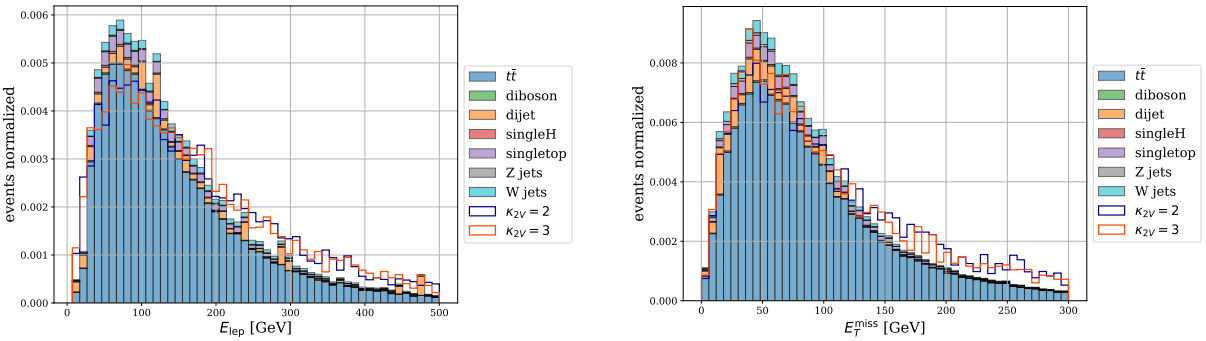


Figure 8.6: Number of normalised events as a function of the energy of the lepton (left) and the missing transverse energy (right) corresponding to the neutrino

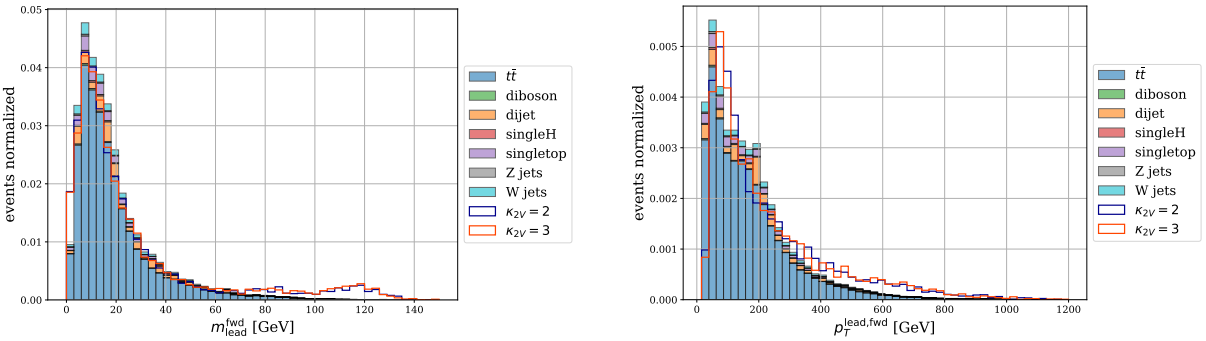


Figure 8.7: Number of normalised events as a function of the mass and transverse momentum of the leading forward jet

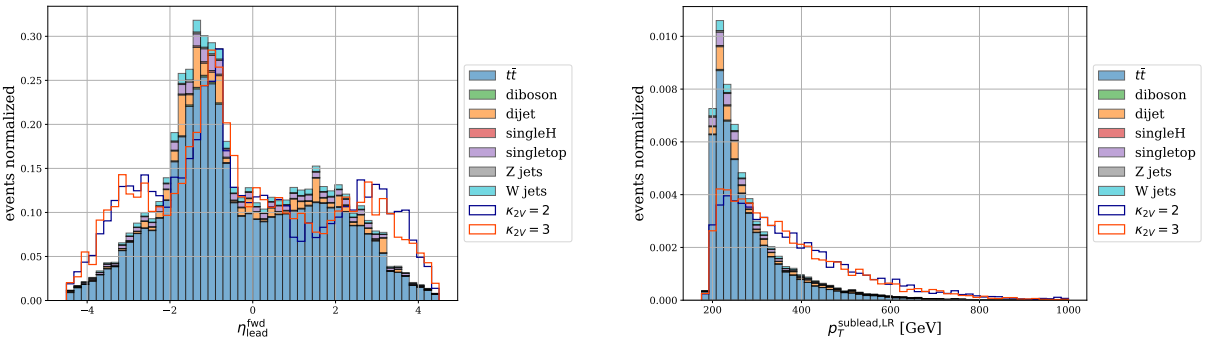


Figure 8.8: Number of normalised events as a function of η of the leading forward jet (left) and the transverse momentum of the subleading LR jet (right)

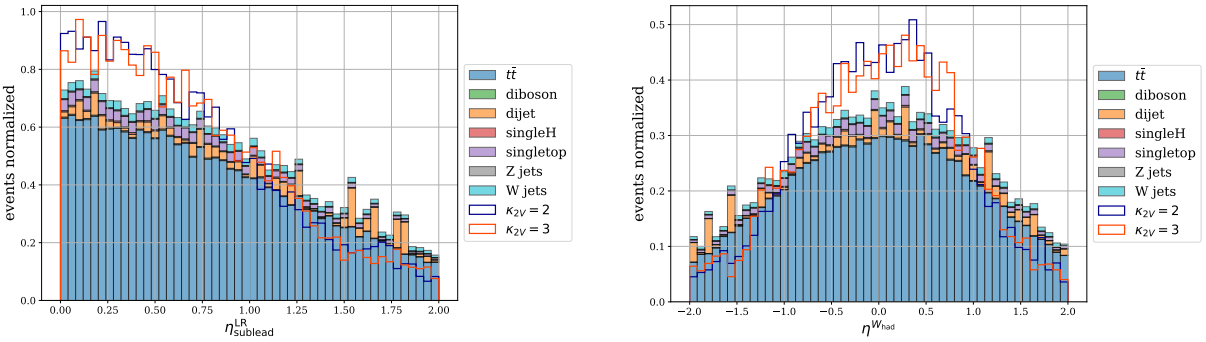


Figure 8.9: Number of normalised events as a function of η of the subleading LR jet (left) and the W_{had} candidate (right)

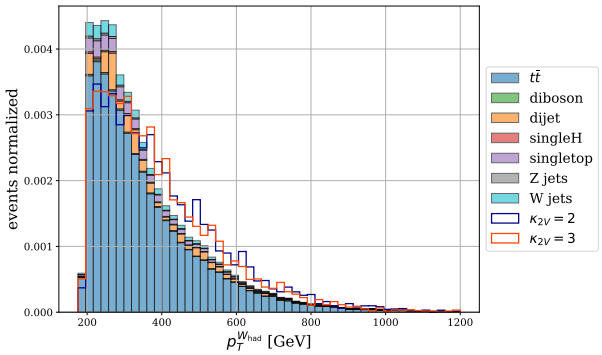


Figure 8.10: Number of normalised events as a function of the transverse momentum of the W_{had} candidate

9 Appendix 2 - additional figures and plots

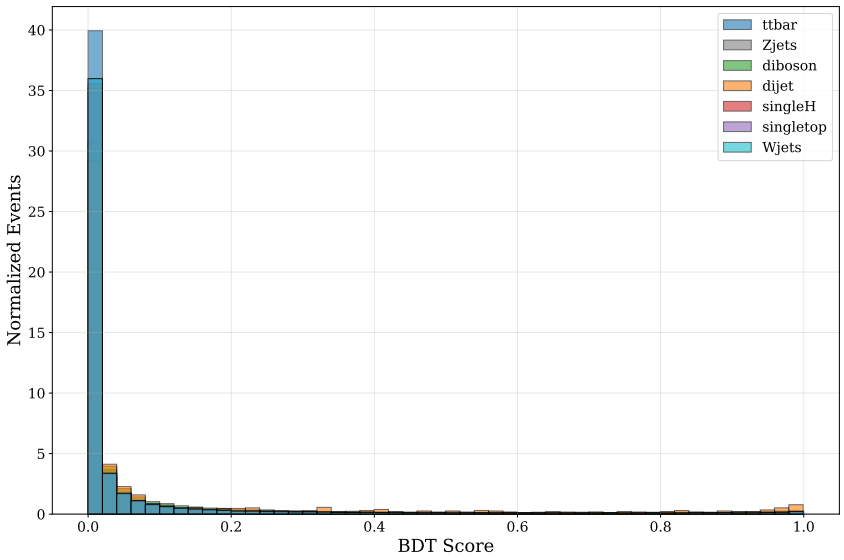


Figure 9.1: Score distributions for different backgrounds from the BDT trained on the baseline feature set with new implemented features

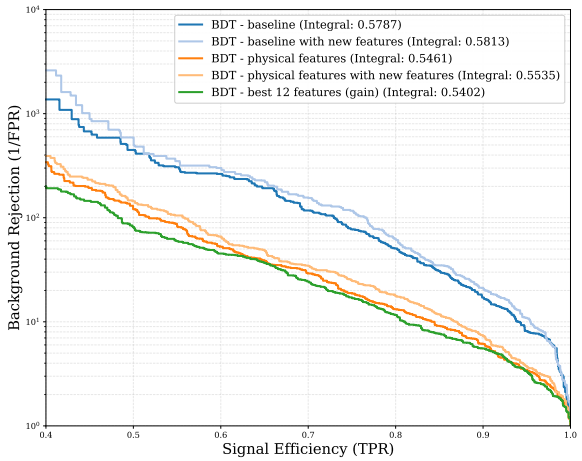


Figure 9.2: Comparison of the background rejection as a function of signal efficiency for the BDT.

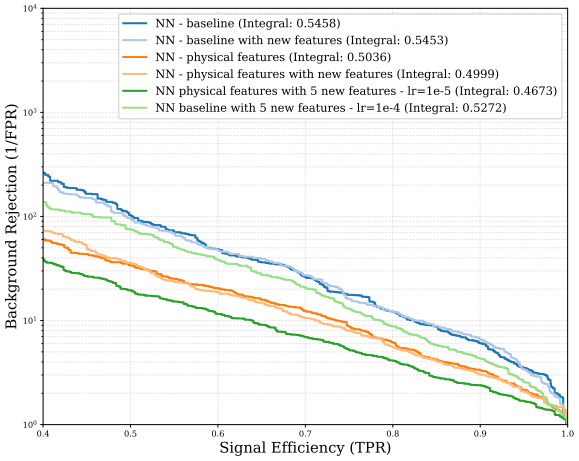


Figure 9.3: Comparison of the background rejection as a function of signal efficiency for the NN, partially without ζ_{SRjet}.

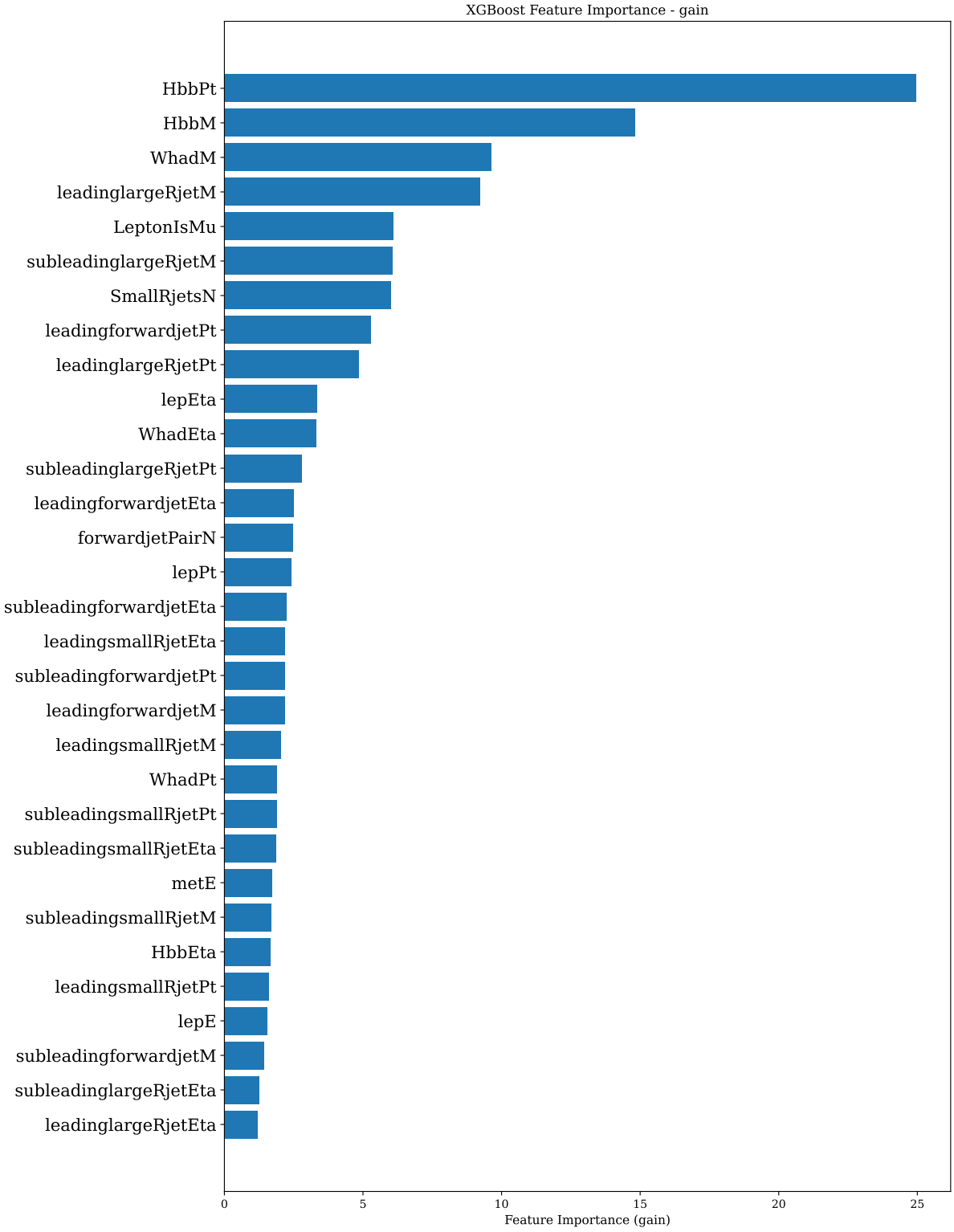


Figure 9.4: Feature importances calculated via gain for BDT trained on baseline feature set.

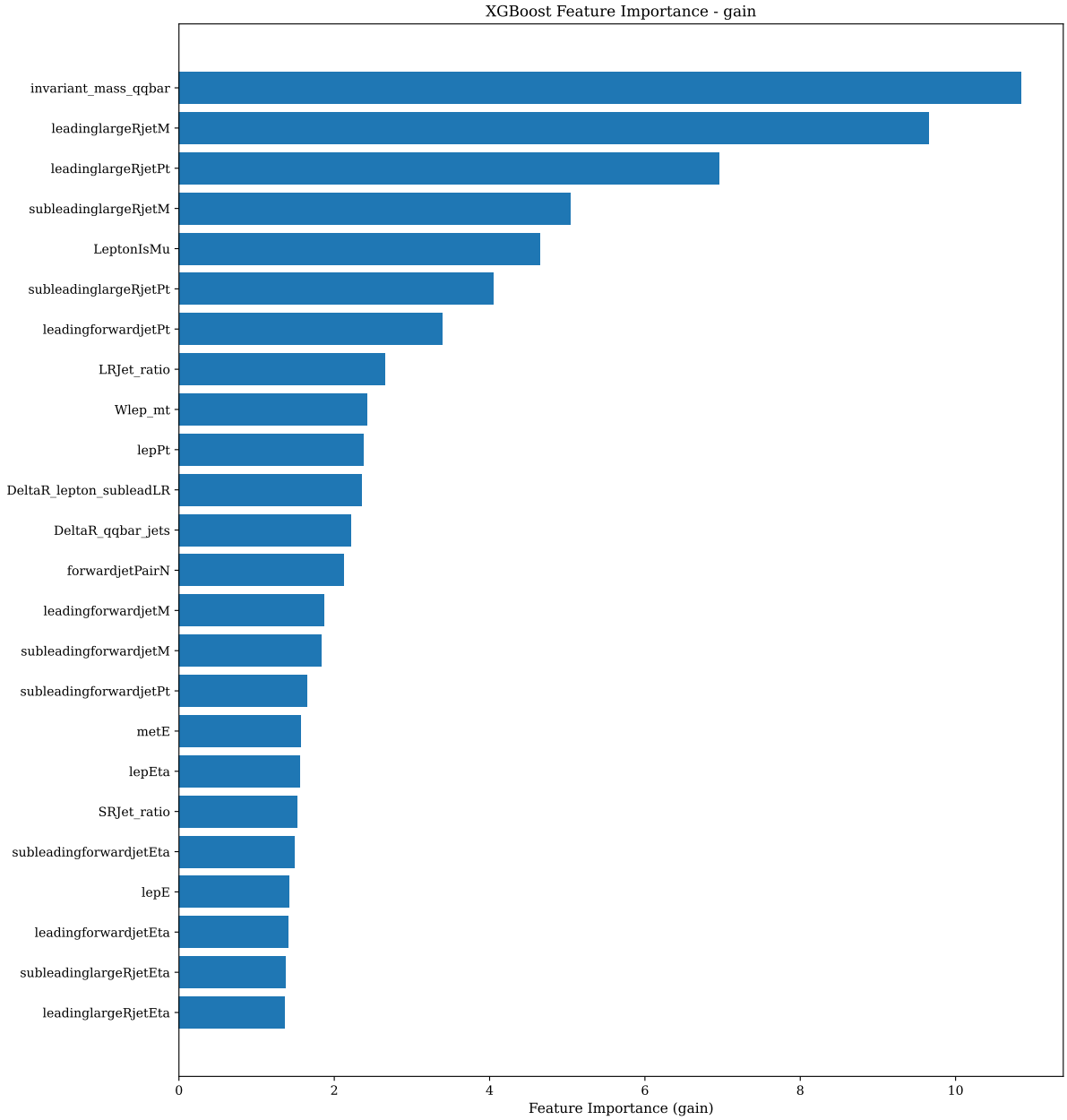


Figure 9.5: Features importances for the BDT trained on physically reasonable features with new features.

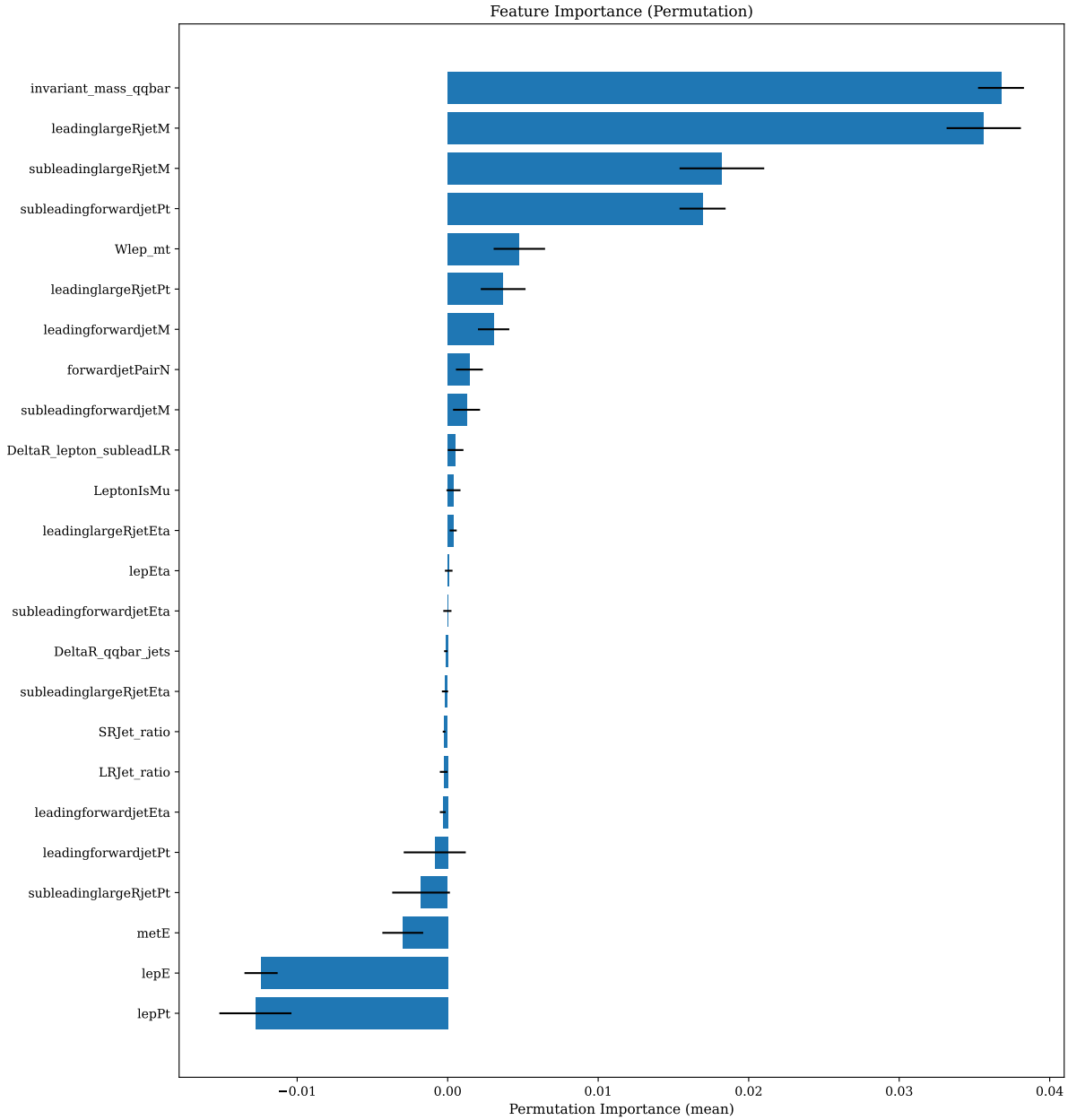


Figure 9.6: Features importances for the NN trained on physically reasonable features with new features.

Danksagung

Ich möchte mich herzlich bedanken! In den letzten Wochen habe ich so viel lernen dürfen. Inhaltlich natürlich, dafür bin ich dem 2. Physikalischen Institut dankbar. Aber auch strukturell, wie der naturwissenschaftliche Betrieb funktioniert, wie man eigenständig forscht (oder es zumindest versucht). Letzteres wurde möglich durch die Offenheit und Geduld während der sehr strukturierten Betreuung durch Prof. Stan Lai, dem ich dafür ganz herzlich danken möchte. Dieser Dank geht natürlich in gleicher Weise (wenn nicht sogar ein bisschen mehr) an Kira Abeling, die mir mit ebenso viel Geduld, Zeit und Freundlichkeit begegnet ist. Ich habe sie immer als extrem fleissig und hilfsbereit erlebt, was mir die Arbeit sehr erleichtert hat. Auch an die Münchner Arbeitsgruppe rund um Celine Strauch und Valerio D'Amico ein grosser Dank, dafür, dass sie so viel Arbeit für mich investiert haben.
Und als Letztes möchte ich mich auch sehr bei meiner Familie bedanken, die mich bei meiner nerdigen Leidenschaft so sehr unterstützt.

Erklärung nach §13(9) der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen:
Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe.
Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, im Rahmen einer nichtbestandenen Prüfung an dieser oder einer anderen Hochschule eingereicht wurde.
Erklärung zur Nutzung von ChatGPT und vergleichbaren Werkzeugen im Rahmen von Prüfungen: In der hier vorliegenden Arbeit habe ich ChatGPT oder eine andere KI wie folgt genutzt:
1. zur Optimierung oder Umstrukturierung von Software-Quelltexten

2. zum Korrekturlesen oder Optimieren

Ich versichere, alle Nutzungen vollständig angegeben zu haben.

Göttingen, den 4. Juni 2026

(Friedrich Konrad Hoppe)