Analysis of VH->bb hadronic process

Paul Wegmann

Supervisors: Dr. Adinda de Wit
Hessam Kaveh

Deutsches Elektronen-Synchrotron Hamburg

Abstract

This study analyses the "Higgs Strahlung" process in which a Standard Modell Higgs Boson is produced in association with a vector boson. In the examined decay channel, the Higgs Boson is required to decay into a pair of b quarks whereas the vector boson is required to decay hadronically. This analysis aims to make an estimate about the measurable significance when conducting an experiment *LHC* for which the Asimov Significance (AMS) is used. To approach this task, first the reconstruction of the bosons, using Monte Carlo Data, has been analyzed. Furthermore a High Level Trigger and further selection have been identified which can be used in the experiment. Last, several improvements on the event selection using classifiers, such as Boosted Decision Trees (and neural networks), have been applied. The final AMS value calculated accounts to 1.06. This analysis can be used as a basis for further studies dealing with a similar decay channel.

This documentation is work in progress and not finished yet.

Contents

1. Introduction

The aim of this study is to make a statement about the analyzebility of the production of a Higgs Boson in association with a vector boson. Hereby the Higgs is required to decay into a pair of b quarks and the W/Z boson is anticipated to decay hadronically. To do so Monte Carlo data is used since it allows to know all original features of the data. []

1.1. Motivation

The Higgs Particle, discovered in 2012 [], is being produced in numerous Events at the Large Hadron Collider *LHC* at the European Organization for Nuclear Research *CERN*, Geneva. Its discovery is considered a major success for the Standard Model *SM* as well as for the LHC.

So far the Higgs production has mostly been attested in connection to a leptonic signature. This is due to the fact that leptons are more long-lived than quarks and hence can be reconstructed more easily.

In this thesis a Higgs production in association with a Vector Boson is considered, whereas the Vector Boson decays hadronically and the Higgs Boson into a pair of a b and an anti b quark, which is, with a probability of 58% [**?**], the dominant decay mode. This process will be denoted as VH->bb hadronic. Further discussion on the production as well as the decay will be conducted in section 2.1.

Analyzing and comparing all decay modes and their probability with the predictions made by the Standard Model (SM) is an important step to be done to either validating the SM further or finding indications for new physics beyond the standard model.

In High Energy Physics great quantities of Data are produced. It is therefore an important part of the experiments handle and analyze the the data appropriate. In Experiments like CMS at the LHC at CERN, Geneva, multiple techniques are applied

2. Theoretical Framework

2.1. Higgs Boson

The Higgs mechanism was originally introduced by Peter Higgs to explain the reason for having a massive gauge boson for the weak interaction [Hig64]. Its associates mass with excitation of a scalar field, called the Higgs Field, which is spontaneously broken. These excitations relate also to the production of a Higgs Boson, a scalar particle which has been measured by the LHC [Col12] and is considered to be the prove of the Higgs Field. However the theory was extended explaining also the mass of quarks, electrons, muons and taus.

The Higgs Boson is predicted to be produced in many different ways by the Standard Modell. This study only deals with the H being produced in association with an W Boson. This process is known as Higgs Strahlung. First an excited W boson is produced via quark antiquark annihilation which is then radiating on de-excitation a Higgs Boson. Furthermore, the Higgs Boson is required to decay into a pair of b quarks whereas the W Boson decays hadronically. A Feynman Diagram of this process can be seen in **??**. Depending on the charge of the quarks annihilating the same process can also happen with the Z Boson replacing the W Boson, but this process will be omitted in this study.

The relative production rate of H due to Higgs Strahlung accounts to 4.1 % at a collision energy of 13 TeV. Other important productions are the gluon fusions which have the highest fraction of 87.2 %, vector boson fusion with 6.8 % and top fusion with 1.9 %.

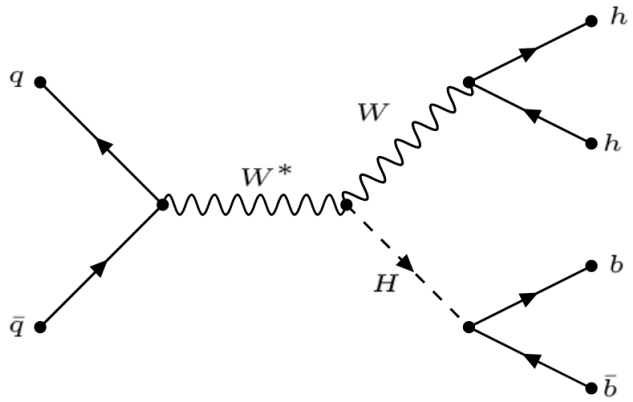


Figure 1: Feynman Diagram of studied H production and decay

The H decays most likely into two b quarks with a probability of 57 % as predicted by the Standard Model. Further decay channels are two W bosons (21 %), two gluons (9 %, two taus (6 %), two charms (3 %), two Z bosons (3 %) as well as other decays being below 1 %.

Independently, multiple theories exist predicting up to eight different Higgs Bosons which will not be discussed in this report.

2.2. CMS Detector

The CMS Detector is one of the four Experiments at the LHC at CERN. Its structure can be seen in figure **??**. The Detector has a cylindrical composition and consists of four layers which are all homocentric. Starting from the hub, the first layer in encountered is a tracking detector, in which charged particle leave a trace which can be reconstructed in 3D. The next mantel is the Electromagnetic Calorimeter, in which photons and electrons/positrons are expected to be stopped completely for measuring their Energy, whereas the deposit is proportional to the measured electrical current. Next up is the Hadronic Calorimeter, which purpose is to measure the Energy of all hadrons. After this layer, all particles except muons and neutrinos, are expected to have been absorbed already. Therefore, the final coat in the Detector is the Muon Chamber in which Muons leave a trace due to their charge. In between the Hadron Calorimeter and the Muon Chamber is a superconducting solenoid which induces a magnetic field of 4 T [Luc19]. In the muon chamber the field is inverted compared to the inner detector system. The magnetic Field helps identifying the particles due to the additional information of the charge - mass ratio.

2.3. Parameters

Due to the high luminosity of the experiments at LHC not all events can be recorded, but must be selected carefully. Therefore, the recording of data has trigger system, which is comprised on hardware triggers and high level triggers *HLTs*, which will be further discussed in section **??**.

The data taken by CMS is processed to extract different parameters such as kinetic arguments which are all described by the CMS Convention [cms19] , in which the z-axis is oriented in the same direction as the beam pipe, the y-axis is pointing towards the sky and the x-axis is facing into the middle of the LHC (see **??**). Coordinates similiar to cylindric coordinates are used, whereas the vector ϕ is used to describe the position in the plane transverse to the beam and θ is the angle to the z-axis.

Instead of θ, it is common to use the pseudorapidity η , which is defined as

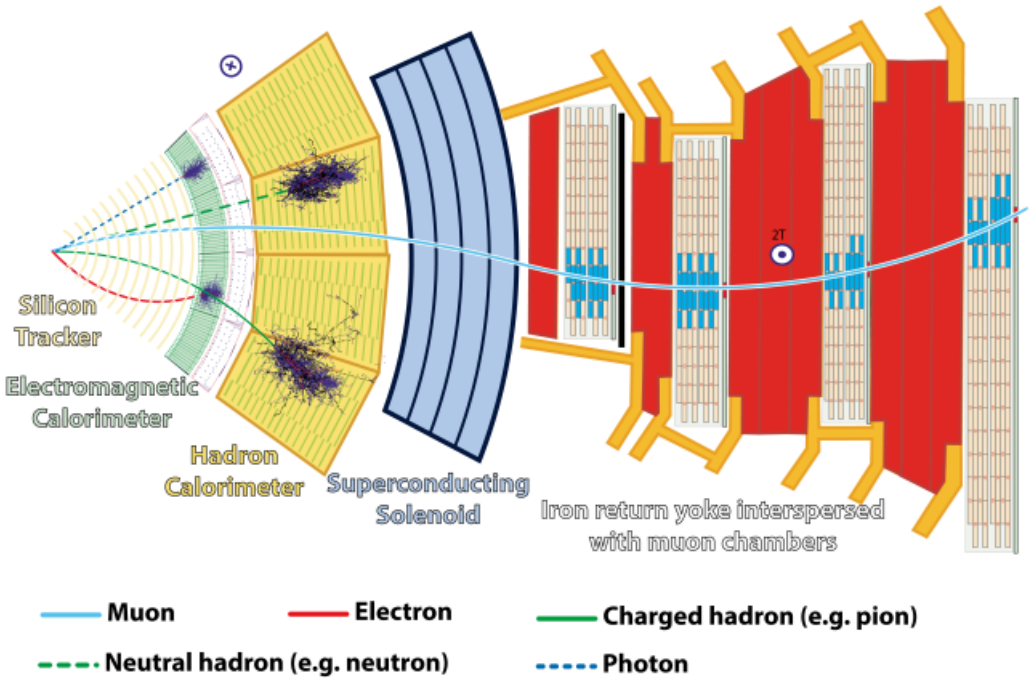


Figure 2: Transverse section of the CMS detector. Different Particles and their corresponding exchange can be seen from [**?**].

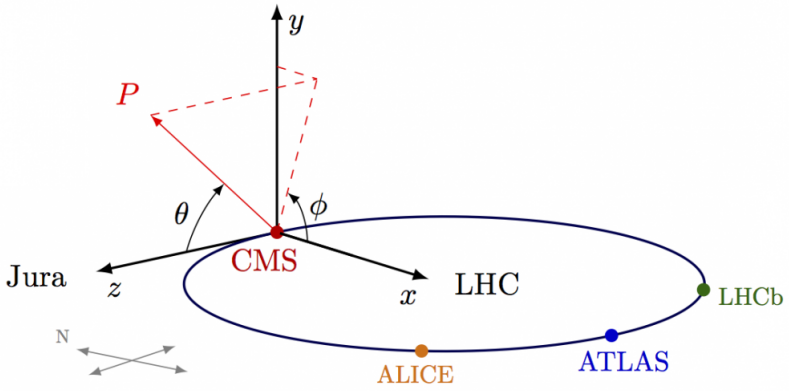


Figure 3: Coordinate system used to describe cms data (from [Luc19])

$$\eta \equiv -\ln\left(\tan\frac{\theta}{2}\right) = \frac{1}{2}\ln\left(\frac{|\vec{p}| + p_L}{|\vec{p}| - p_L}\right) \tag{1}$$

with $\vec{p}$ being the momentum vector and p_L being the momentum in the beam direction of the the particle. For relativistic particles η converges to the rapidity which differences are Lorentz invariant under boosts along the z-axis.

Another important variable is the angular distance

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}, \tag{2}$$

which helps to quantify differences in the direction of two particles.

In the area of the beam pipe no detectors can be placed, thus an important parameter is the transverse momentum p_t displaying the energy exchange as well as being conserved if all particles are taken into account. Due to the very low cross section of the neutrinos the sum of all transverse momenta (denoted as HET) offers an opportunity for indirectly measuring neutrinos.

The invariant mass is a Lorentz invariant scalar and defined as

$$m_0^2 = E^2 - \vec{p}^2. \tag{3}$$

It equals the particles mass in its rest frame within an uncertainty. It also can be derived with p_t, η and ϕ for relativistic particles.

The parameter "BtagDeepB" denotes an output of a neural network trained for identifying jets originating from a b-quark. It has proven itself to be a convincing parameter reaching b-tagging efficencies of up to 70% and will be used in this analysis (for further reading see [Col13]).

2.4. Monte Carlo Data

Monte Carlo *MC* Data plays an important role in state of the art research. In general it can be described as a method which relies on random sampling. Especially in physics MC is useful for analyzing problems with many degrees of freedom, because through event sampling MC is approximating the expected outcome of an experiment through statistical convergence(, by the law of large numbers.)

In high energy physics MC refers to simulated data of the used detector. It is generated by simulating the collision events first (gen-level). The resulting particles and jets are then fed into an artificial detector which produces data as described in section **??** which is then reconstructed as normal data to have comparable result in the end (recon-level). These results are used for analysis to make an estimate on whether or not the data is in accordance with the underlying theory.

Actual detector data from CMS does not consist solely out of one clean interaction point where two protons are scattered but of multiple collisions generating multiple jets. This is because of the high luminosity of the LHC. This effect makes data analysis harder, however it is needed to be able to make significant statistical data analysis. Otherwise the time needed for data recording would prolong dramatically due to the high rarity of the analyzed processes.

Therefore multiple interaction points called vertices emerge in each bunch crossing and jets and particles are matched to reconstructed interaction points []. The one with the highest squared sum transverse momentum of all physics-object is called primary interaction vertex and is considered to be the most interesting vertex. This is because it has the highest measurable

energy deposit and it can be reconstructed best due to the high transverse component. ?and therefore the highest cross-section?

The events from the analyzed MC data already only consists of such primary interaction vertices. That is why for the analysis only jets from one interaction vertex are considered,however there are still background processes in the selected data because of wrong assignation of jets as well as background processes taking place in the same collision which is known as pileup and is also considered in the artificial data.

Finally, in the used data sample the gen-level and the recon-level of each event are given, making it possible to create a connection between the original particle responsible for the signature in the data, this will be further discussed in section 3.2.

The used Monte Carlo data is considering two signal processes and one production of background data. The processes of interest is the production of a Higgs Boson by Higgs production as discussed in section 2.1, hereby only W+/W- and no Z processes are considered. For background data is caused by Quantum Chromo Dynamics (QCD) scattering.

In total the original data includes signal data including 1 000 000 different events, 500 000 for each W boson, and 12 000 000 events for QCD background. It has to be considered that this is the number of events generated and does not correspond to the number of events happening at CMS. Hence, the Events have to be weighted to have the expected numbers of events happening. Since the data is already split into processes and energy levels each bunch of data has to be reweighted to a given value. The weights were calculated with the following formula:

$$w = \frac{L \cdot \sigma}{N_{entries}} \tag{4}$$

Here, σ corresponds to the cross section in the studied energy range, $N_{entries}$ is the number of events concluded and L is to the integrated luminosity, which is equal to $58\,830\,\mathrm{fb}^{-1}$ corresponding to the integrated luminosity of CMS in 2018.

Incorporating weights results in 70 000 signal events and in $2.15 \cdot 10^{11}$ background events.

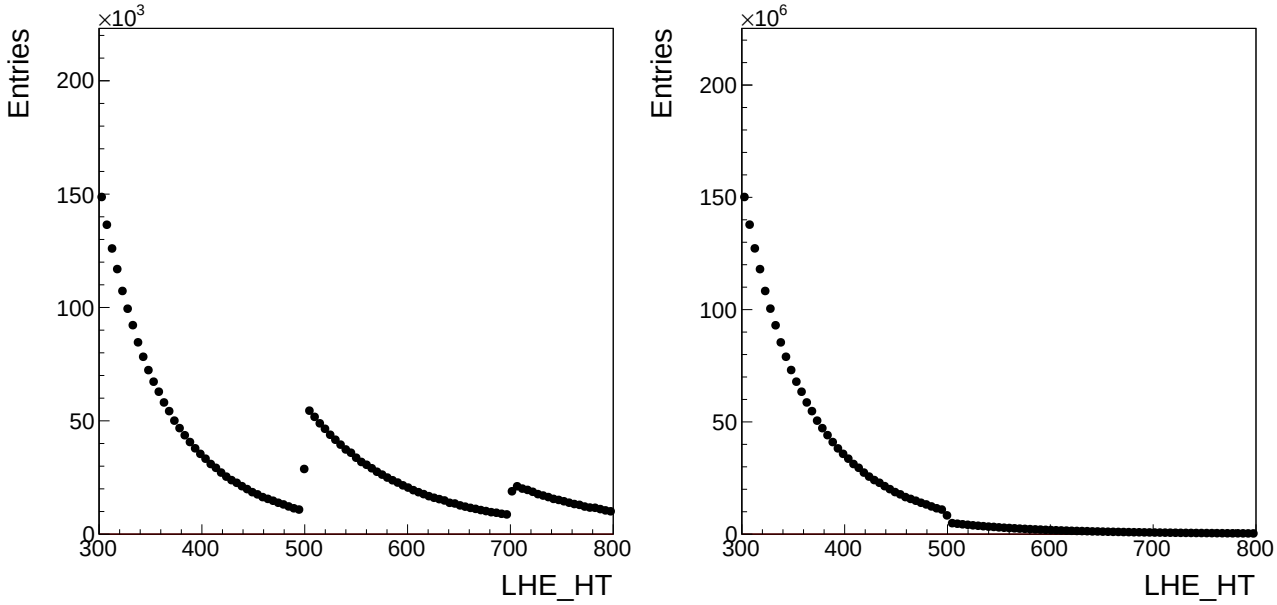
The importance of this procedure is illustrated in figure **??**. The parameter `LHE_HT` parameterizes the squared sum of the transverse energy of all jets and particles detected. Nevertheless there is discrepancy regarding the shape at 500 GeV whose origin is not understood but will be neglected since it is a small deviation which is expected to have not a big impact on the final results. In the following section only the weighted events will be discussed.

2.5. AMS

In this analysis the Asimov Significance (AMS) value is used to make an estimate about separation of data as well as broaching a value for an estimate for the significance of a measurable deviation in background(for a discovery).

In particle physics one wants to quantify the significance of a new discovery. This is usually done by making a statement about the probability of the outcome of an experiment. Normally a number of Events n is measured from which the p-value which assumes a no-signal (null) hypothesis, is calculated. It is equal to the consequential probability of measuring n or more (less) Events. In figure **??** an example for calculating the double tailed p-value for a measurement x_{obs} of a Gauss distribution $P(x, \mu)$ can be seen, with μ being the expectation value. The Z-value is given by $|x_{obs} - \mu| = z \cdot \sigma$, whereas σ conforms the standard deviation of the Gaussian distribution. The corresponding formula for $x_{obs} > \mu$ is:

$$p = \int_{x_{obs}}^{\infty} P(x, \mu) dx + \int_{-\infty}^{\mu - x_{obs}} P(x, \mu) \tag{5}$$



(a) Transverse energy before reweighting (b) Transverse energy after reweighting

Figure 4: Proportion of jets

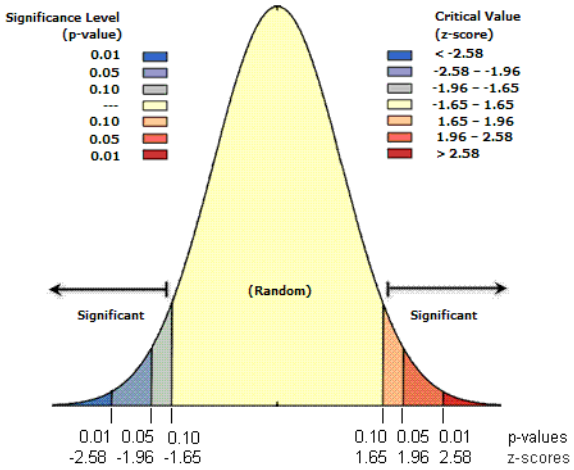


Figure 5: double tailed p-value and corresponding Z-value for a Gaussian distribution

Expecting a Gauss distribution of events measured it is meaningful to consider the double tail event p-value instead of the single tailed.

If there is an expected value for signal events the median Z value can be calculated to considering a null hypothesis.

The AMS value can be deduced from the likelihood function by assuming a Poisson counting experiment. As discussed in [Cow12] the likelihood with known expected background events b corresponds to a Poisson distribution

$$L(s) = \frac{(s+b)^2}{n!} \cdot e^{-(s+b)}. \tag{6}$$

In general the Likelihood can be extended with nuisance parameters θ including further assumptions. The calculated likelihood ratio is defined as

$$\lambda(s) = \frac{L(s, \hat{\hat{\theta}}(s)}{L(\hat{s}, \hat{\theta})}. \tag{7}$$

Hereby, the single-hat variables refer to maximum likelihood estimators whereas the double-hat implies that the likelihood maximizes under the specific value s. As derived in [] the likelihood ratio can be used for a test statistic q_0 given by:

$$q_0 = \begin{cases} -2 \cdot \ln\lambda(0) & \text{if } \hat{s} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The following formula is used in this study as a test statistic and is based on the previous formulas. It can be derived by assuming a large number of events [Cow12]:

$$Z = \sqrt{2\left(n \cdot \ln\frac{n}{b} + b - n\right)} \tag{9}$$

If $s << b$ the test statistic can be approximated by

$$Z = \frac{s}{\sqrt{b}}(1 + \mathcal{O}(s/b)) \tag{10}$$

which is equal to the common formula $s/\sqrt{b}$ used for discoveries. However, as it can be seen in figure **??**, the used formula results in a better estimate for the median Z-value even for small backgrounds. The dots in the plot are produced from Monte Carlo simulations. As discussed in [Cow12] the odd structure is caused by the discrete nature of the data.

In this study no errors are taken into account and the expected background value is put on the level of the value of background extracted from the Monte Carlo value.

The discussed test statistic offers a good possibility for evaluating the separation of signal and background if the background is known. However it is an approximation for large data samples, so one needs to be careful for small samples.

2.6. Classifiers

Classifiers deal with assigning classes by using input variables called features. In this study machine learning algorithms are applied to approach this task. Machine learning implicates that the classifier practices on training data to generalize so that it can predict the label with the given features as good as possible. To rate the efficiency, the algorithm is tested on independent data. This done to check for overtraining, which is a phenomenon occurring when the classifier memorizes events rather than generalizing.

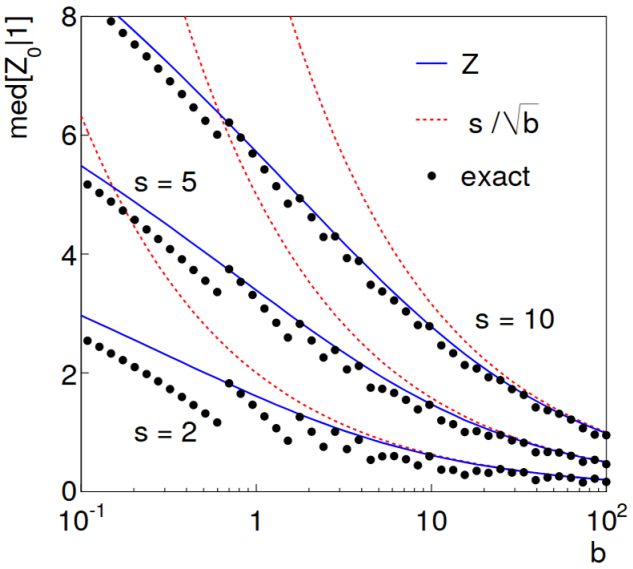


Figure 6: Comparison of used Z formula versus $s/\sqrt{b}$ of estimating the expected Z-median, from [Cow12]

Here,the machine learning algorithms are used for binary classification (only two possible labels), since it is used for estimating if an event can be assigned Higgs Strahlung as discussed in section 2.1. For improving the selection while training a loss function is used. It quantifies the error of the decision tree and is minimized via back propagation. There are many different loss function which all lead to different results during training. However, normally it is a matter of trial and error to figure out a suitable loss function. This study only considers supervised learning, therefore the labels of the training and test data sets are known.

2.6.1. Boosted Decision Tree

For the implementation of a Boosted Decision Tree (BDT) the ROOT library TMVA [HST+10] is used. The foundation of this machine learning algorithm is a decision tree made up of nodes which separates the events according to their features. Each node only applies one cut, therefore leading to a binary separation at each node. Multiple nodes are applied consecutively in order to check various features. The maximum of successive nodes, also known as the depth of the tree, can be set with the option `MaxDepth` in TMVA. An example can be seen in figure **??**. In total three features are checked to make a prediction, namely the sex, the age and the number of family member aboard. The predicted label of the tree is not always right, in this example the accuracy, which is the fraction of rightly predicted events over all events, corresponds to $0.73 \cdot 0.36 + 0.83 \cdot 0.61 + 0.94 \cdot 0.02 + 0.89 \cdot 0.02 = 0.806$.

To improve the selection further multiple trees are trained with boosted data and their output is combined to conclude the final prediction. When boosting, the weight of misclassified events is increased to raise their importance when training the next decision tree, making them more likely to be classified rightly. This leads to an ensemble of trees, known as forest, that can make more precise predictions. The number of total trees within the forest can be set with the option `NTrees`. High numbers of trees are likely to result in overfitting. This can be avoided by setting a minimum percentage of events ending up in final leafs by using the option`MinNodeSize`. The bigger the value the less a BDT tends to overfit, since it has to generalize more. Another option is `BoostType` which defines the formula used for reweighting

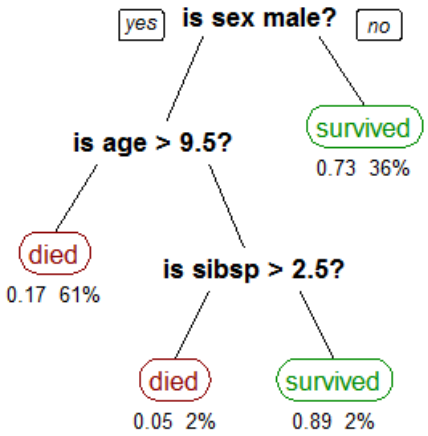


Figure 7: Decision tree for survivors of the titanic (from [Wik19]), "sibsp" is the number of family members aboard. The right numbers by the final nodes show the number of affected cases and the number on the left corresponds to the actual probability of survival.

Table 1: Table of options varied for BDT

option	description	predefined
`NTrees`	Number of trees trained and used for making predictions.	- •

the events as well as `SeparationType` defining the criterion used for the split value in each node. TMVA also offers the option `VarTransform` standing for variable transformation. This transformation is applied on all of the data before training and testing. There are five options predefined: `N, G, D, P, U`. They stand for normalization, so all features used have the same variable range, Gaussian Transformation which will be further discussed, Decorrelation, PCA and Uniform, respectively. Not all possible feature of TMVA have been tested in this study, due to limited time. A summary of all used parameters can be seen in table **??**.

To estimate the goodness of the final decision tree multiple metrics are conducted. What should be tested is the ability to transfer from training data to unseen test data. So, to make a meaningful statement about the quality of the BDT this analysis is only calculated on test data since the BDT is biased by the data used for training.
First the binned AMS value is calculated on the output of the BDT. It is used as an quantity about the the separation of signal and background. To keep track of overtraining the normalized congruent area of the output of test and train data is calculated for signal and background each. The two values, both having a maximum value of 1, are added up to form the final metric. Therefore, the maximum value possible is 2 and the minimum 0. Models with values close to 2 are considered to be not overtrained. Another metric used is the receiver operating characteristic curve (ROC-Curve). For a fixed threshold value c x- and y-coordinates are calculated. All values higher than the threshold value are classified as signal and the rest is assorted to be background. Thereby, the signal efficiency and background rejection can be derived which are defined as the fraction of rightfully classified signal/background and all signal/background events. So, when the signal efficiency drops the background rejection rises. Perfect separation of signal and background results in 1 for both metrics. To plot the ROC-

Curve the threshold gets varied over the whole output range. The integral of the curve is calculated and used as a metric. The closer the integral is to one the better the classifier.

3. Work

3.1. Selection of Data

In an modern analysis in high energy physics plenty of selections have to be done to reduce the amount of data as much as possible. In this case the initial selection, refereed to as preselection, done in this study will be discussed first. The goal is to achieve a similar basis for the data for further analysis. This selection can already filter out many background events having an non-suiting signature for the studied final state, as well as signal events which are not detected properly.

Several eligibility criteria can easily be deducted from the studied process. First four jets in total are required to agree with the expected final state of four hadronic particles. At least two of the resulting jets are b-jets which are identified via the `btagDeepB` value [Col13]. This value.... Another requirement is that the `Jet_eta` value needs to be smaller than 2.5 This is due to the architecture of the detector which only covers the tracking of charged particles in a range of $|\eta| \leq 2.5$. Although the calorimeters cover a wider range, the tracking information is crucial for the `btagDeepB` determination, thus these jets have to be neglected.

Next, events with highly energetic leptons, in this case electrons and muons, are neglected as they are not expected in the process and are therefore an indication of dominant background. Therefore an identification of electrons and muons has to be done. The data already concludes variables dealing with lepton identification using for example multivariate analysis (mva) as well as cut based quantities. For the electron recognition a mva Classifier is used with a relatively high cut requiring a response of 90 %. The corresponding variable `Electron_mvaFall17V2Iso_WP90` contains whether or not the respective particle passed the threshold of the classifier. For the muon identification `Muon_tightId` is used which is a cut based ranking being, as emphasized by the name, a strict criteria. Finally, `recon_Hmass` needs to be in between 90 GeV and 155 GeV, this will be discussed further in section 3.2.

In total each jet considered needs to have a transverse momentum above a threshold of 20 GeV, since high transverse jets can be reconstructed more accurate and it is favorable to work with jets having a sharp signal.

In summary the preselection conducted is:

- At least four Jets in total
- Two jets with |`Jet_eta`|>2.5 and `Jet_pt`>20 GeV and `btagDeepB`>0.4981
- None electron with `Electron_pt`>20 and classified as `Electron_mvaFall17V2Iso_WP90`
- None muon with `Muon_pt`>20 and classified as `Muon_tightId`
- 90 GeV<`recon_Hmass`<155 GeV

13 000 signal events are selected as well as $3.2 \cdot 10^9$ QCD background events, corresponding to around 20 % of signal and 1.5 % of background being picked.

3.2. H and W Reconstruction

This section deals with the criteria used to select the b-jets from the Higgs boson (denoted as H-jets) and the hadronic jets from the W boson (denoted as W-jets) to make an estimate about

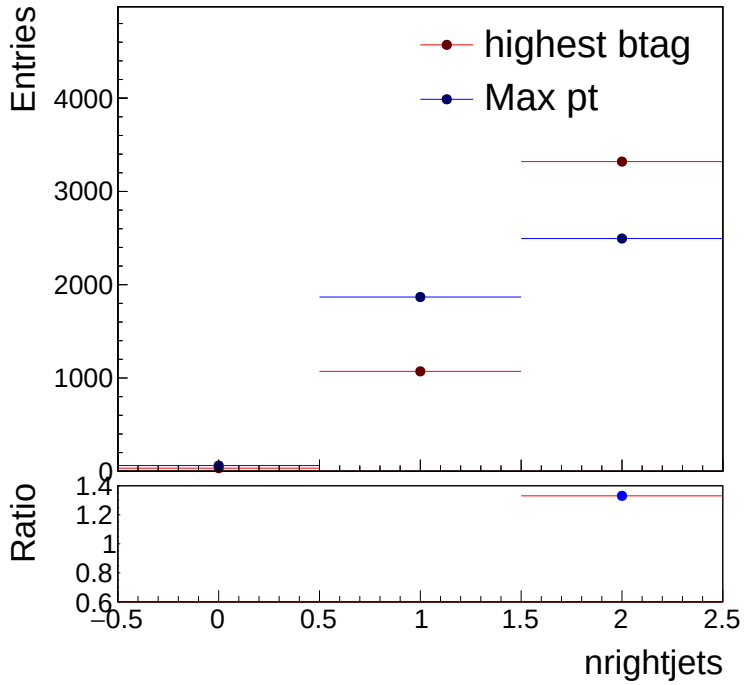


Figure 8: Comparison of two selection criteria for reconstructing the Higgs particle

the reliability of the extracted values.

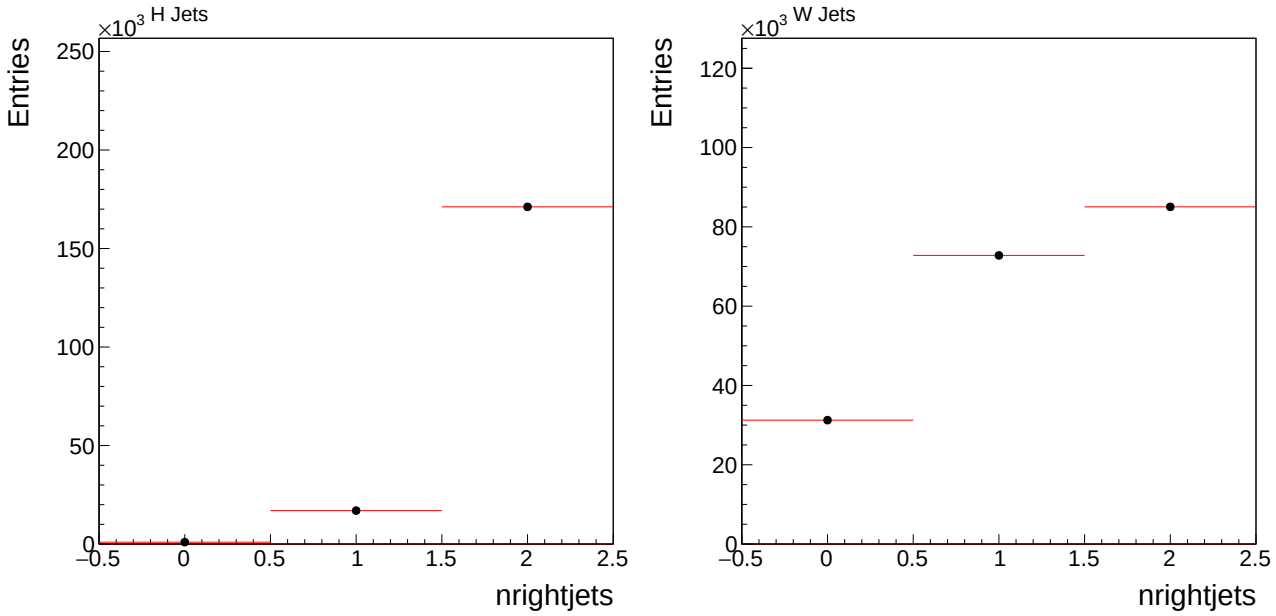
First a connection between GenJets and ReconJets needs to be established to be able to make a meaningful statement about whether the selected jets where identified rightly.

To determine the GenParticles originating from the Higgs and W boson (out of the gen-level data) the information "genPartIdxMother" was used, which stores the kind of the mother particle. Next is to link this particle with a detected jet which is done via the feature "genJetIdx", which already matches the recon-Jets with gen-Jets. However, not always the described linking is accurate therefore a radial distance lower than 0.3 between the recon-jet and the associated gen-jet is required.

The analysis for the selection of the H- and the W-jets was done only on data containing the VH->bb hadronic process. Multiple single selection criteria were tested as well as combinations. The most convenient criteria is picking jets with the highest p_t since the Higgs production is expected to occupy a lot of energy as well as picking the jets with the highest `bTagDeepB` value. In figure **??** the results of the two different selection criteria are compared. Certainly in this figure only signal events with multiple possible selections after the selection discussed in section 3.1 are considered which corresponds to approximately 1400 Events of 13000 selected Events (≈11 %).

The used selection is picking the highest b-tagged value and for the H-jets and the two jets which combined invariant mass is the closest to 80 GeV for the W-jets. The proportion of events with zero, one or two jets assigned properly can be seen for the Higgs and the W boson in figure 9a and 9b respectively. The proposed selection is rather efficient for the H-jets regarding the fact that there are rarely b-jets in the W decay. However, the criteria used for the W-jets performance poorly.

Last, the the signal and background distribution can be seen in figure 10a. The AMS value is calculated bin-wise, meaning that for each bin the AMS is calculated respectively resulting



(a) Rightly assigned H-jets

(b) Rightly assigne W-jets

Figure 9: Proportion of jets

in the given final value which is the square root of the squared sum of all values. However the displayed plots show normalized distributions but the AMS value is calculated using weighted events. As expected the resulting AMS value is very low as the ratio of signal to background events is very low. Nevertheless distinct differences can be seen in the curve shape of both distribution. The Higgs mass can be counted back to the peak of the signal curve which lies in between 115 GeV and 120 GeV. Surprisingly the maximum of the distribution is not around the given Higgs mass of 125 GeV but is slightly shifted to lower energies. This can be ascribed to missed jets in detection, wrong reconstructions of jets as well as wrong assignation of jets done in the here prescribed analysis. The derived AMS values will be used as reference values for further analysis.

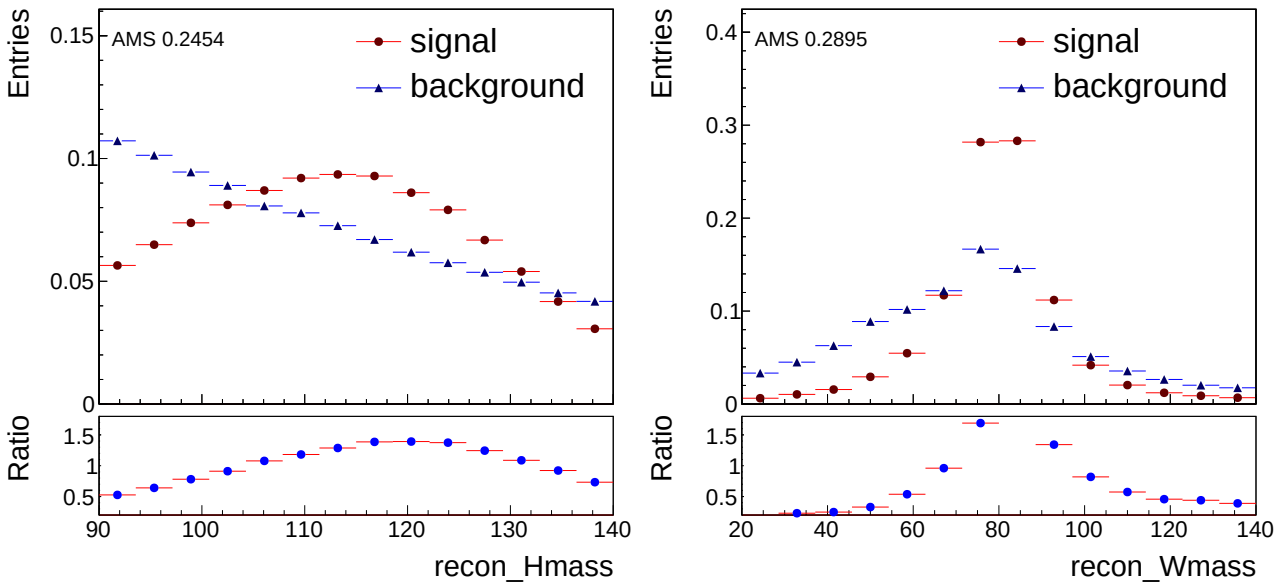
In figure 10b the reconstructed W mass can be seen. A very pronounced peak around 80 GeV can be seen. Considering the selection procedure, however, this is to be expected. For that reason the reconstructed H mass will be used to make an estimate about the significance. Comparing the AMS values, however, it is higher for the distribution of the W mass although the shapes look more congruent.

Additionally, the choice of binning has a slight influence on the outcome of the significance, so the binning of the reconstructed Higgs mass for calculating the AMS value has to be consistent throughout the analysis. Here 14 bins with evenly distances between 90 GeV and 140 GeV has been picked.

Looking at the ratios shows a stronger divergence for low and high values. Additionally there is a deformity of background in lower energies, probably because of...

3.3. High Level Trigger

Picking a High Level Trigger (HLT) is a crucial step for conducting experiments at LHC. This is due to high luminosity and an event rate of 1 GHz [Xab19] which would otherwise lead to vast amounts of data, impossible to process. The High Level Trigger is the second and final level of the trigger system at CMS. The first level is hardware based as it needs to be as fast as possible, whereas the second trigger has more time to consider saving events due to the reduced



(a) Reconstructed Higgs mass

(b) Reconstructed W mass

Figure 10: Normed distribution of signal and background and the corresponding AMS value

Table 2: Fractions of the weighted signal and background reduction. The selected triggers for further analysis are highlighted.

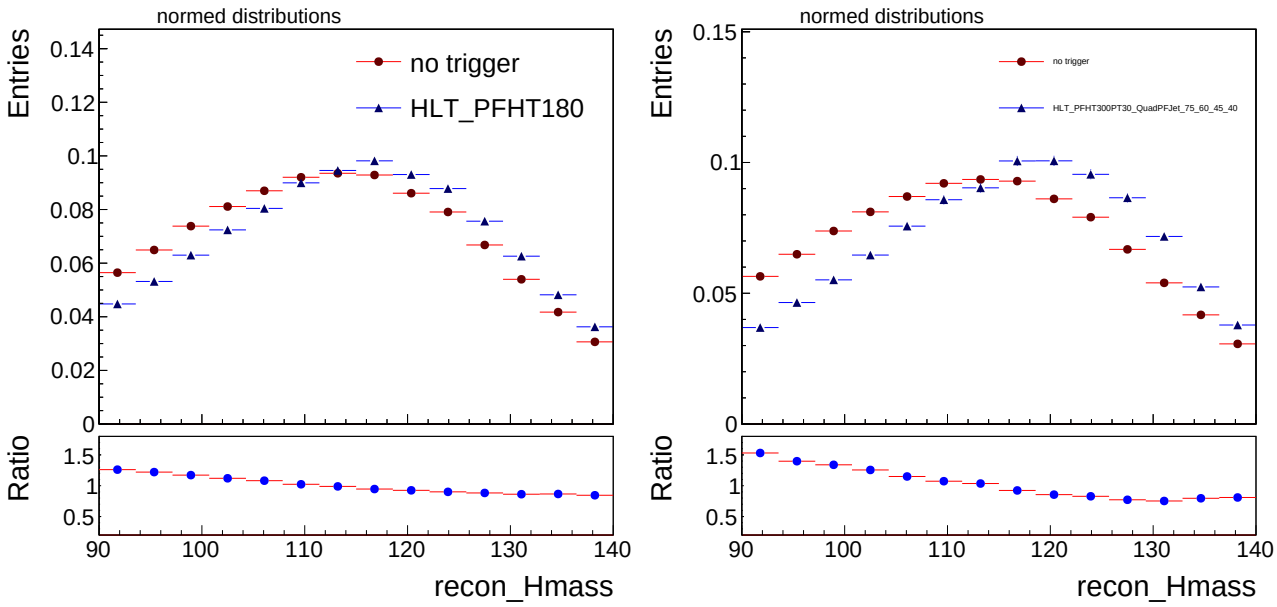
Trigger	Signal	Background
HLT_AK8PFJet80	0.66	0.27
HLT_PFJet80	0.54	0.17
HLT_PFHT300PT30_QuadPFJet_75_60_45_40	0.15	0.01
HLT_DiPFJetAve35_HFJEC	0.69	0.48
HLT_PFHT180	0.60	0.13
HLT_DoublePFJets40_CaloBTagCSV_p33	0.69	0.32
HLT_AK4PFJet80	0.55	0.17
HLT_AK4CaloJet80	0.59	0.19

the data flow and therefore can use more complex algorithms and reconstructions. Nevertheless, the final reconstruction, used in this analysis, still has differences.

Here, multiple triggers and their sensitivity for the MC Data regarding the VH->bb hadronic process were studied. In total eight triggers were selected and compared regarding their sensitivity for signal and background events. The ratios of accepted signal and total signal, as well as for background respectively, can be seen in table **??**.

AK stands for the Anti-kt algorithm described in section **??**

Considering a high reduction in background without loosing to much signal `HLT_PFHT180` offers a good trigger although `HLT_PFHT300PT30_QuadPFJet_75_60_45_40`, which will be refereed to as quad trigger, has the highest relative decrease. Both triggers apply a cut on the HT, the summed transverse energy, furthermore the second HLT requires for Jets with transverse momentum above 75, 60, 45, 40 respectively resulting in a fewer events for signal but also a significant reduction in background. The abbreviation PF stands for particle flow, which is the algorithm used for reconstruction of the jets.



(a) `HLT_PFHT180` (b) `HLT_PFHT300PT30_QuadPFJet_75_60_45_40`

Figure 11: Proportion of jets

Although there is a threshold for the transverse energy, there are still some events which are below that critical value due to the HLT happening online, meaning that it has to make a decision fast on whether or not to save a signal. For that reason the reconstruction procedure for the HLT needs to be much faster than for the offline reconstruction resulting in differences between the HLT cuts and offline cuts applied.

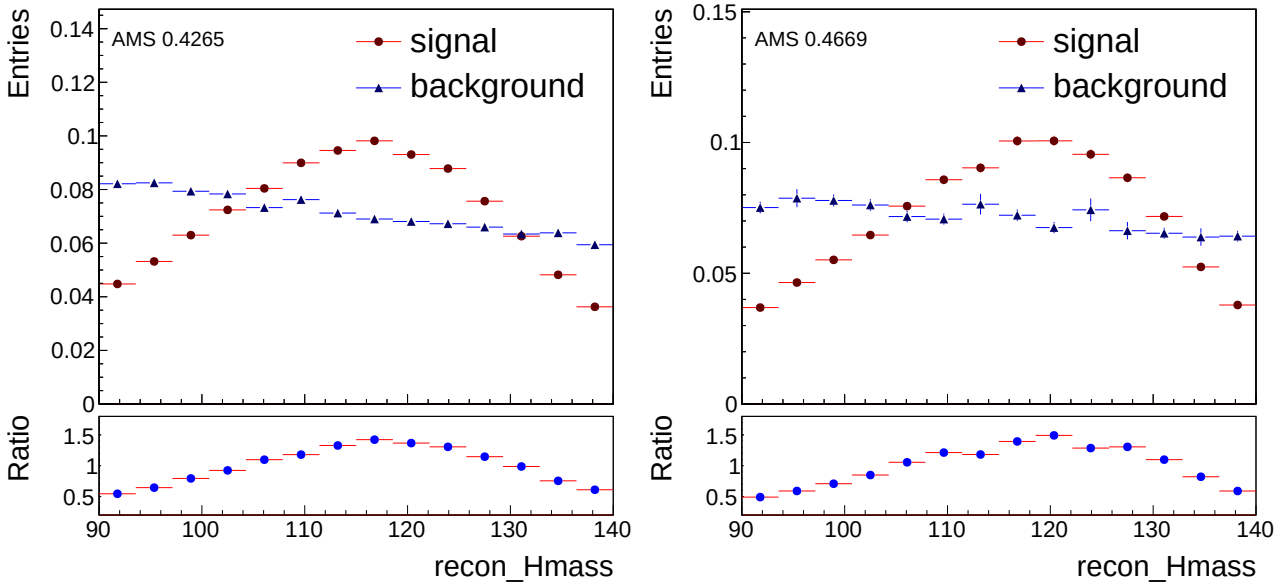
For verifying that the High Level Trigger does not drastically change the distribution of the signal and therefore does nothing unexpected, a shape comparison of the reconstructed Higgs before and after applying triggers can be seen in figure 11a and 11b. A slight tendency towards higher masses can be seen for both triggers, nonetheless it does not have to be considered further.

In figure 12a and 12b the distribution of the background and signal as well as the corresponding AMS value after applying the given triggers can be seen. As described in section 2.5 the AMS value is used for making an estimate about the current significance of the separation applied. Here the shift of the distribution to a bigger H mass can also be seen by comparing both plots. Furthermore the error bars of the background in 12b and fluctuations are more dominant which is caused by the relatively large reduction of background. The AMS values account to 0.427 for the trigger requiring a minimum transverse energy of 180 GeV and 0.467 for requiring four jets above the given threshold. Compared to the AMS value of section 3.1 already an improvement is made, although this value is still to small for making a meaningful statement.

Looking at the errorbars there can be seen fluctuation for the `Quad` trigger. The errors are automatically calculated by root and account to the statistical uncertainty. Since in the data are events with rather big weights (up to 50000), these can have a major impact on error bars as well as fluctuations, when reducing data. In this case the more rigorous trigger almost omits almost all strongly weighted events. The remaining events then cause the observed shape.

The number of events picked and accounts to around 2000 Events for signal and 22 500 000 for background using the `Quad` trigger and to 8000 and 400 000 000 for the `HT180` trigger.

Both triggers have different selected data which needs to be taken into account for further analysis. However it offers a good opportunity for comparing both evaluations.



(a) `HLT_PFHT180` (b) `HLT_PFHT300PT30_QuadPFJet_75_60_45_40`

Figure 12: normed distributions of reconstructed H mass after applying triggers

3.4. Further improvement of event selection

So far the discussion dealt with reconstructing the particles and preparing the data to be able to result in an AMS value. The next sections will deal with improving this value. It starts with discriminating events by applying further cuts as well as analyzing distributions for identifying useful parameters for classification which are then used to train classifiers.

3.4.1. Cuts on data

By cutting off regions with dominant background and vanishing signal the significance can be improved further. This task is approached by examining the distribution of various variables, looking for discriminant regions.

In figure ?? examples for distributions of variables can be seen. Starting with `deltaR_reconHW` which complies to the angular distance between the reconstructed Higgs and W Boson. A peak for background and signal can be seen around an angle of 3. Going to higher angles shows a faster reduction of signal than of background, indicating the jets of W-jet reconstruction are more likely to be boosted, considering the quad trigger, resulting in a favorable region for an upper limit. This is due to the fact that there is not a high cut on the transverse momentum for the reconstruction of the W Boson. Looking at figure 13b, where the corresponding AMS value for a range of upper thresholds is displayed, this is reflected as well. There is a peak around 3.38 improving the original AMS value from 0.467 to 0.514.

Next, the distribution of `deltaR_W` which is the angular distance between the two jets used for the W reconstruction, shows a clear maximum at 0.5 for signal whereas the background is relatively constant decreasing for towards high angles. Another peak of the signal can be seen around 5.5 Due to the high overlap of data for small angles the best AMS value is calculated tight cut.

The last example is `Jet_pt` dealing with the transverse momentum of jets. Figure 13e is showing the jets with larger transverse momentum assigned to originate from the W Boson. The signal distribution is shifted more towards higher transverse momentum which can also be seen in the cut-plot. However there is no improvement when omitting lower energies which

Table 3: AMS values for cuts on various variables

cutted variables	HLT_PFHT300PT30_QuadPFJet_75_60_45_40		HLT_PFHT180	
	AMS	cut at	AMS	cut at
recon_Wmass>	0.439	54.04 GeV	0.482	59.70 GeV
recon_Wmass<	0.459	102.42 GeV	0.504	102.42 GeV
recon_Wpt>	0.432	7.58 GeV	0.508	176.77 GeV
deltaR_reconHW<	0.483	3.38	0.514	3.38
Jet_btagDeepB W-jet, low pt<	0.432	0.89	0.473	0.88
Jet_btagDeepB W-jet, high pt<	0.436	0.81	0.477	0.80

could be ascribed to loosing too much signal.

The distributions of the HT180 trigger will not be discussed here but can be looked up in the appendix A since there are no significant differences in the discussion.

Nonetheless one has to be careful whenever a lot of data is cut off, because the derived formula from section 2.5 used as an estimate does not consider errors since it assumes well known estimate of the background. Therefore, regions with some signal and no background lead to high AMS values. In practice, the background has to be estimated by calibration with the help of further measurements leading to uncertain background estimates. Therefore diminishing background is connected with even bigger errors. This has to be considered for calculating the significance.

In table ?? all discussed variables which improved the AMS from section 3.3 value are summarized and the resulting AMS values. All cuts improving the AMS value are combined together leading to a final value of 0.602 for HT180 and 0.658 for Quad. After applying all cuts together only 1003 signal and 3548292 background events are selected. The final distributions of signal and background of recon_Hmass can be seen in figure ??.

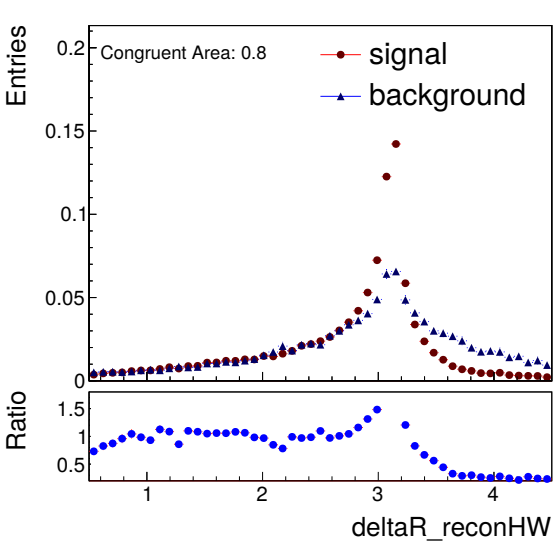
3.4.2. Boosted Decision Tree (BDT)

Here, the procedure for developing a BDT, as discussed in section 2.6.1, using TMVA [HST+10] and the results will be discussed. The binned AMS value is calculated on the output of the BDT distribution to have a numeral parameterizing for the separation of data. It can not be compared with the AMS values discussed before. Additionally to not have a biased estimation the AMS value is not calcualted using all of the data but only on the test data on which the Boosted decision tree is trained on. The split of test and training data is set to have a equal number of tests and training events. However, the weights are not applied when considering this selection, but due to the selection being random there are no preferences in splitting data therefore the effects of picking unweighted data can be neglected.

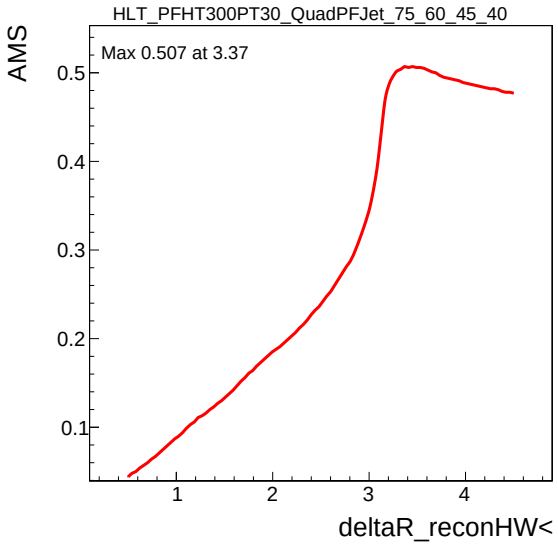
To figure out suiting options for the BDT some parameters are varied individually with the other options set to the default values displayed in table ??. The results can be seen in table ??.

Varying NTrees shows that there is no improvement when exceeding 400 Trees, in addition, test showed that the BDT becomes more and more likely to overtrain. The variable MinNodeSize as well as MaxDepth have a great influence on each other and are also very likely to lead to overtraining if picked too fine.

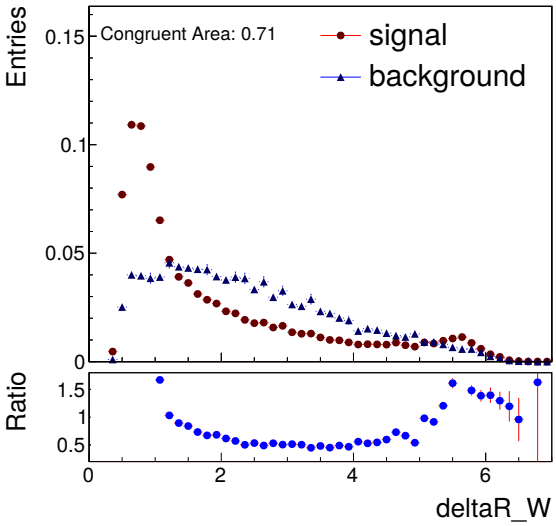
A very interesting parameter to look at is VarTransform which has no impact on the architecture of the BDT but on the data as discussed in section 2.6.1. The best results are achieved using the Gaussian Transformation. Although Multiple combinations of variables transforma-



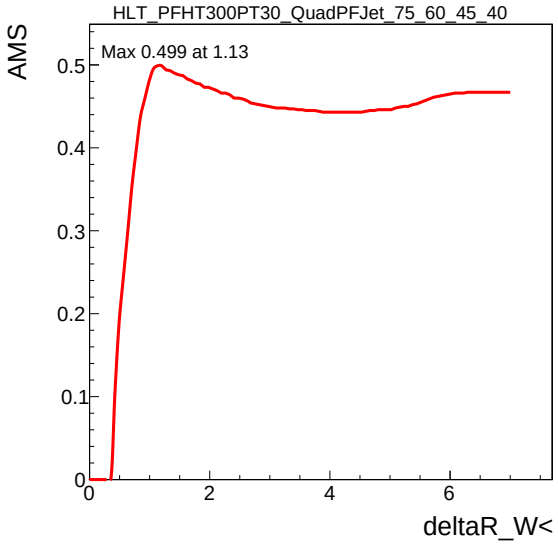
(a) Distribution



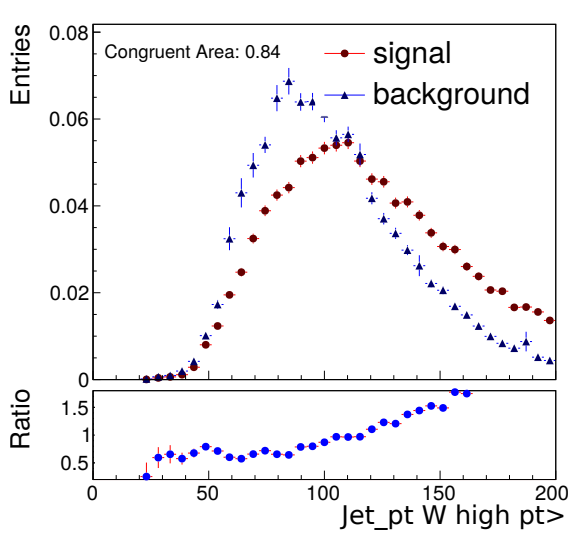
(b) AMS values for upper limit



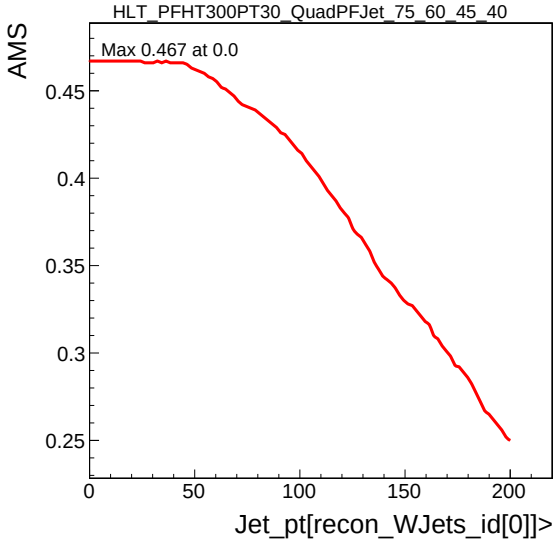
(c) Distribution



(d) AMS value for upper limit

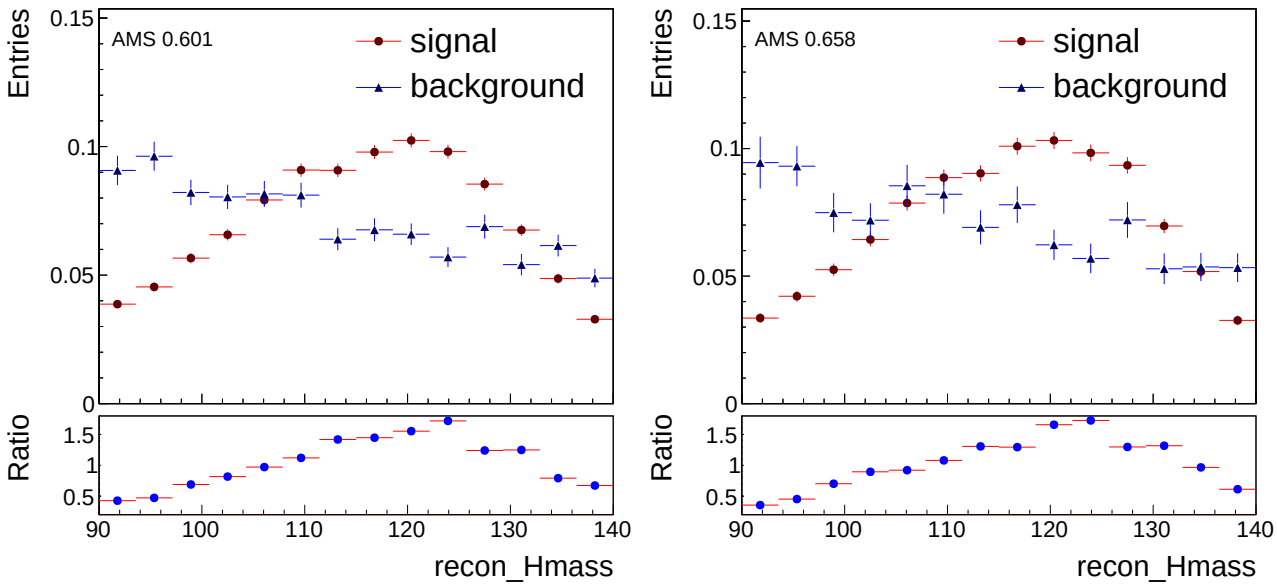


(e) Distribution



(f) AMS value for lower limit

Figure 13: Analysis of cut values for quad jet trigger. Each row deals with one variable. On the left normalized signal and background distributions regarding one variable can be seen. On the right is the AMS value plotted when cutting at the corresponding value on the x-axis.



(a) `HLT_PFHT180` (b) `HLT_PFHT300PT30_QuadPFJet_75_60_45_40`

Figure 14: Normed distributions of reconstructed H mass after applying all cuts

Table 4: Default values of the varied options for the BDT

Option	Default
`NTrees`	500
`MinNodeSize`	5.0 %
`BoostType`	`AdaBoost`
`SeperationType`	`GiniIndex`
`VarTransform`	None

Table 5: Analysis procedure of BDT options for HLT_PFHT300PT30_QuadPFJet_75_60_45_40

options	AMS	MAD	MSE	ROCIntegral
HLT_PFHT300PT30_QuadPFJet_75_60_45_40				
MinNodeSize=1%	1.010	0.111	0.019	0.832
MinNodeSize=2.5%	0.994	0.107	0.018	0.833
MinNodeSize=5%	0.860	0.104	0.017	0.830
MinNodeSize=7.5%	0.832	0.108	0.018	0.825
MinNodeSize=10%	0.803	0.105	0.016	0.823
BoostType=AdaBoost	0.856	0.104	0.017	0.829
BoostType=RealAdaBoost	0.759	0.474	0.226	0.817
BoostType=Bagging	0.316	0.864	0.812	0.737
BoostType=Grad	0.745	0.540	0.375	0.819
SeparationType=CrossEntropy	0.855	0.107	0.018	0.826
SeparationType=GiniIndex	0.811	0.106	0.017	0.829
SeparationType=GiniIndexWithLaplace	0.860	0.106	0.018	0.831
SeparationType=MisClassificationError	0.855	0.102	0.016	0.830
SeparationType=SDivSqrtSplusB	0.518	0.321	0.150	0.752
NTrees=100	0.803	0.205	0.063	0.825
NTrees=250	0.799	0.141	0.030	0.829
NTrees=400	0.833	0.120	0.022	0.827
NTrees=600	0.879	0.098	0.015	0.830
NTrees=1000	0.925	0.076	0.009	0.828
VarTransform=G	0.943	0.108	0.018	0.831
VarTransform=N	0.816	0.109	0.018	0.829
VarTransform=D	0.869	0.102	0.016	0.821
VarTransform=P	0.749	0.102	0.016	0.814
VarTransform=U	0.916	0.104	0.017	0.831

tion can be applied, norovement of the AMS value for using Gaussian Transformation have been achieved. Therefore this option is always set to G in this study.

4. Summary

In this study selection criteria as well as a High Level Trigger have been identified for Higgs production, whereas the Higgs Boson is required to decay into a pair of b-quarks and the Vector Boson is required to decay hadronically. A good method for reconstructing the Higgs Boson has been presented and various selection improvements have been discussed. The final AMS value achieved is equal to 1.06.

Next steps in this analysis address further improvement of the AMS value. For this purpose several options can be discussed. For example the reconstruction of the W Boson can be improved further. Also, different classifiers or alternative architectures for the studied classifiers can be examined. Especially, a neural network offers a great possibility for further refinement. Another interesting opportunity is the consideration of "Fat Jets" which can result in a higher signal strength.

Table 6: Analysis procedure of BDT options for HLT_PFHT180

options	AMS	MAD	MSE	ROCIntegral
HLT_PFHT180				
MinNodeSize=1%	0.978	0.126	0.024	0.839
MinNodeSize=2.5%	0.936	0.125	0.024	0.838
MinNodeSize=5%	0.844	0.117	0.021	0.832
MinNodeSize=7.5%	0.811	0.128	0.025	0.828
MinNodeSize=10%	0.739	0.121	0.023	0.819
BoostType=AdaBoost	0.884	0.130	0.026	0.834
BoostType=RealAdaBoost	0.739	0.475	0.227	0.808
BoostType=Bagging	0.242	0.747	0.652	0.753
BoostType=Grad	0.771	0.512	0.343	0.833
SeparationType=CrossEntropy	0.843	0.125	0.024	0.834
SeparationType=GiniIndex	0.864	0.123	0.023	0.832
SeparationType=GiniIndexWithLaplace	0.909	0.125	0.024	0.832
SeparationType=MisClassificationError	0.866	0.127	0.025	0.834
SeparationType=SDivSqrtSplusB	0.447	0.344	0.163	0.732
NTrees=100	0.752	0.215	0.070	0.820
NTrees=250	0.823	0.165	0.041	0.826
NTrees=400	0.842	0.134	0.027	0.832
NTrees=600	0.905	0.112	0.019	0.832
NTrees=1000	0.917	0.094	0.014	0.832
VarTransform=G	0.959	0.129	0.025	0.839
VarTransform=N	0.848	0.121	0.023	0.831
VarTransform=D	0.790	0.121	0.022	0.831
VarTransform=P	0.772	0.119	0.022	0.826
VarTransform=U	0.910	0.127	0.024	0.838

Appendices

A. Distribution and cuts for `HLT_PFHTand180`

See **??**

References

[cms19] `https://wiki.physik.uzh.ch/cms/latex:tikz`, 2019. [Online; accessed on 21-August-19].

[Col12] The ATLAS Collaboration. Observation of a new particle in the search for the standardmodel higgs boson with the atlas detector at the lhc. *Physics Letters B*, 2012.

[Col13] The CMS Collaboration. Identification of b-quark jets with the cms experiment. *Journal of Instrumentation*, 2013.

[Col16] The CMS Collaboration. The cms trigger system. *Journal of Instrumentation*, 2016.

[Cow12] Glen Cowan. Discovery sensitivity for a counting experiment with background uncertainty. 2012.

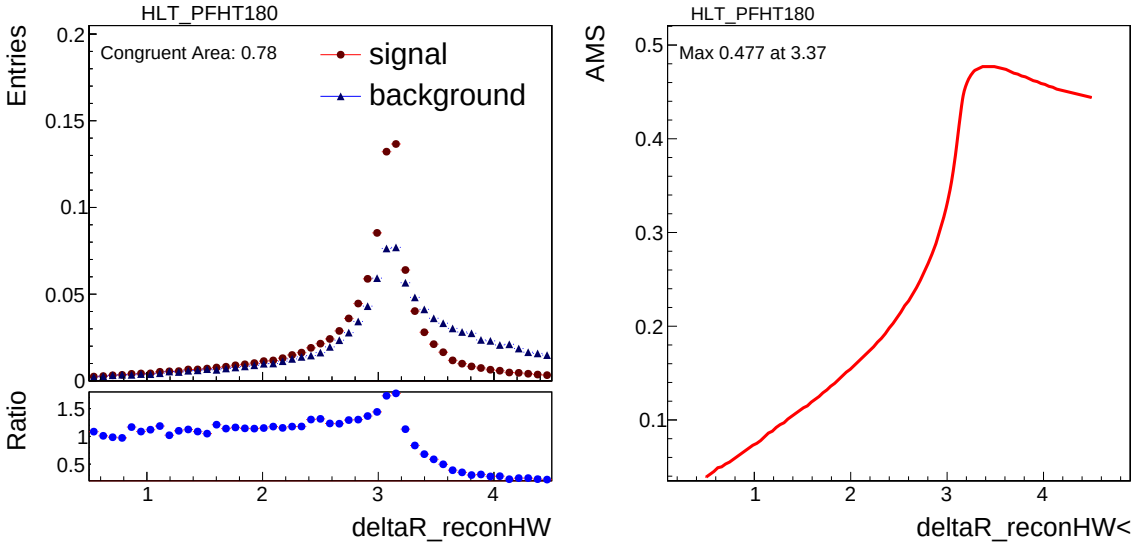
[Hig64] Peter Higgs. Broken symmetries and the masses of the gauge bosons. *Physical Review Letters*, 1964.

[HST+10] A Höcker, Jana Stelzer, Fredrik Tegenfeldt, Helge Voss, K Voss, Asen Christov, Sophie Henrot Versille, M Jachowski, A Krasznahorkay, Y Mahalalel, Xavier Prudent, and P Speckmayer. Tmva - toolkit for multivariate data analysis with root:users guide. 01 2010.

[Luc19] Lucas Taylor. Superconducting magnet. `http://cms.web.cern.ch/news/superconducting-magnet`, 2019. [Online; accessed on 20-August-19].

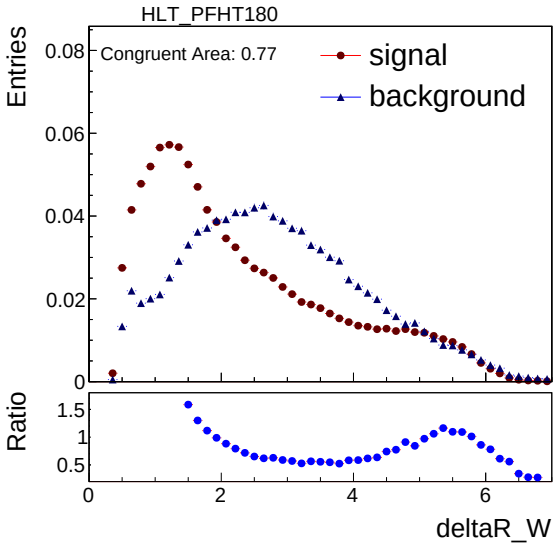
[Wik19] Wikipedia contributors. Decision tree learning — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=912094993`, 2019. [Online; accessed 31-August-2019].

[Xab19] Xabier Cid Vidal, Ramon Cid Manzano. Lhc trigger - taking a closer look at lhc. `https://www.lhc-closer.es/taking_a_closer_look_at_lhc/0.lhc_trigger`, 2019. [Online; accessed 02-September-2019].

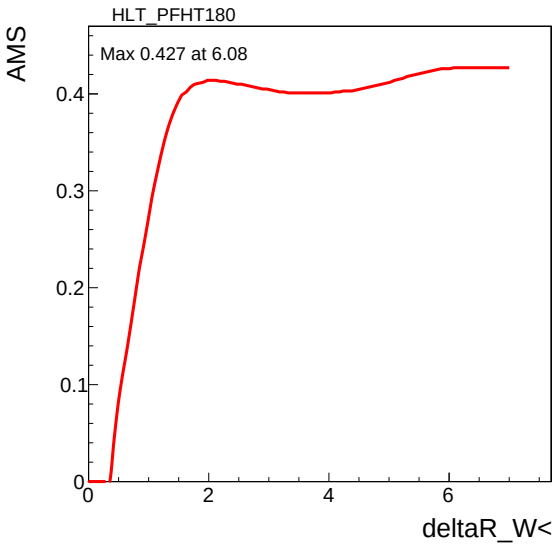


(a) Distribution

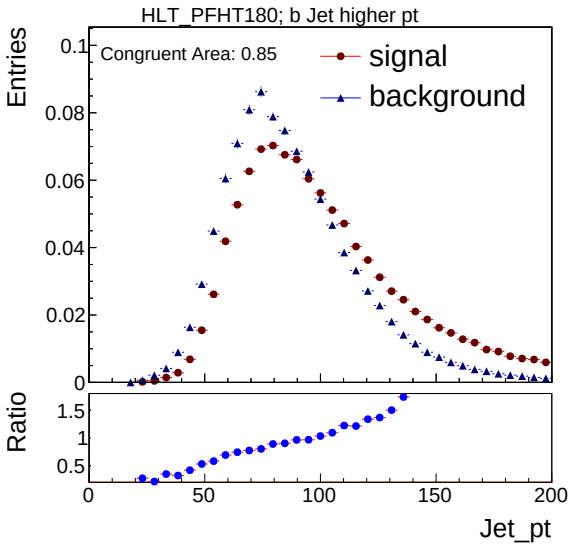
(b) AMS values for upper limit



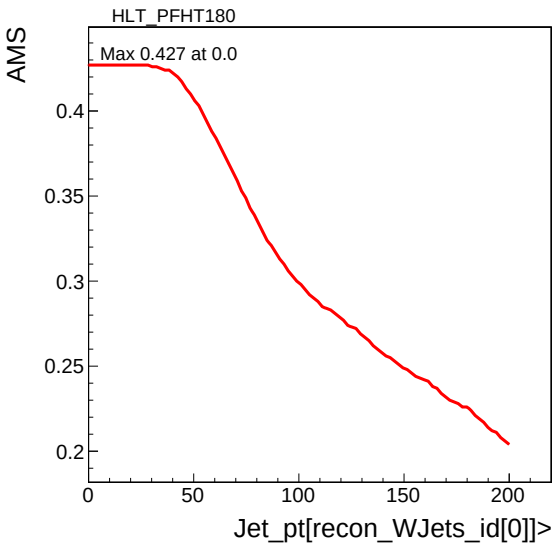
(c) Distribution



(d) AMS value for upper limit



(e) Distribution



(f) AMS value for lower limit

Figure 15: Analysis of cut values for HT180 jet trigger. Each row deals with one variable. On the left normalized signal and background distributions regarding one variable can be seen. On the right is the AMS value plotted when cutting at the corresponding value on the x-axis.