Improving Higgs boson categorisation with Boosted Decision Trees

Petković, Andro

University of Split
Faculty of Science

Improving Higgs boson categorisation with Boosted Decision Trees

Master thesis

Andro Petković

Split, September 2021.

I would like to thank my supervisor Toni Šćulac for guidance in every step of this project. I would also like to thank Matteo Bonanomi for help with the procedure of extracting signal strength modifier.

Temeljna dokumentacijska kartica

Sveučilište u Splitu
Prirodoslovno – matematički fakultet
Odjel za fiziku
Ruđera Boškovića 33, 21000 Split, Hrvatska

Diplomski rad

Unaprjeđenje kategorizacije Higgsovog bozona koristeći ubrzana stabla odluke

Andro Petković

Sveučilišni diplomski studij Fizika, smjer Astrofizika i fizika elemntarnih čestica

Sažetak:

Higgsov bozon otrkrile su CMS i ATLAS kolaboracije 2012. godine na CERN-u. Od tada su napravljena razna mjerenja njegovih svojstava i sva su se pokazala konzistentnima s predviđanjima Standardnog Modela unutar intervala pogreške. Međutim, potrebno je napraviti preciznija mjerenja kako bi se odredilo je li čestica uistinu Higgsov bozon iz Standardnog Modela. Svako odstupanje od očekivanih vrijednosti otvorilo bi prozor u novu fiziku. Higgsov bozon može se proizvesti na više načina u velikom hadronskom sudaraču. U ovom diplomskom radu predstavljene su dvije kategorizacije produkcijskih mehanizama Higgsovog bozona. Prva je već postojeća kategorizacija koja je otprije implementirana u CMS analizi. Druga kategorizacija, bazirana na algoritmu strojnog učenja - ubrzanom stablu odluke, razvijena je u sklopu ovog diplomskog rada. Poboljšanje u klasifikaciji je ostvareno te je ubrzano stablo odluke povećalo čistoću kategorija u rasponu 6-18%, ovisno o promatranom produkcijskom mehanizmu, u odnosu na postojeću kategorizaciju. Koristeći kombinaciju klasifikacije bazirane na ubrzanom stablu odluke i otprije implementirane klasifikacije, određena je očekivana vrijednost modifikatora jačine signala te pripadajuća standardna devijacija na Asimovom skupu podataka.

Ključne riječi: Higgsov bozon, ubrzano stablo odluke, modifikator jačine signala, kategorizacija

Rad sadrži: 33 stranice, 26 slika, 7 tablica, 20 literaturnih navoda. Izvornik je na engleskom jeziku.

Mentor: doc. dr. sc. Toni Šćulac

Ocjenjivači: doc. dr. sc. Toni Šćulac,
doc. dr. sc. Marko Kovač,
doc. dr. sc. Petar Stipanović

Rad prihvaćen: 16. rujna 2021.

Rad je pohranjen u Knjižnici Prirodoslovno – matematičkog fakulteta, Sveučilišta u Splitu.

Basic documentation card

University of Split
Faculty of Science
Department of Physics
Ruđera Boškovića 33, 21000 Split, Croatia

Master thesis

Improving Higgs boson categorisation with Boosted Decision Trees

Andro Petković

University graduate study programme Physics, orientation Astrophysics and Elementary Particle Physics

Abstract:

The Higgs boson was discovered in 2012 by CMS and ATLAS Collaborations at the Large Hadron Collider in CERN. Since then, properties of Higgs boson are being measured and within the uncertainity intervals, the latest analysis results are in agreement with predictions of the Standard model (SM). However, more precise measurements have to be performed to establish whether this particle has all the properties of the SM Higgs. In this thesis, two categorizations of Higgs boson production mechanisms are presented. The first categorization is already implemented in CMS analysis. The second categorization, based on a machine learning algorithm - boosted decision tree (BDT), is developed as a part of this project. The improvement in the categorization efficiency has been achieved and the BDT categorization outperformed the existing classification by 6-18%, depending on a specific production mechanism. Using a combination of boosted decision tree and existing classification, signal strength fit was performed on Asimov dataset. Expected value of the signal strength modifier has been determined as well as the corresponding standard deviation.

Keywords: Higgs boson, boosted decision tree, signal strength modifier, categorization

Thesis consists of: 33 pages, 26 figures, 7 tables, 20 references. Original language: English.

Supervisor: Assist. Prof. Dr. Toni Šćulac

Reviewers: Assist. Prof. Dr. Toni Šćulac,
Assist. Prof. Dr. Marko Kovač,
Assist. Prof. Dr. Petar Stipanović

Thesis accepted: September 30th, 2021

Thesis is deposited in the library of the Faculty of Science, University of Split.

Contents

1 Introduction

The Higgs boson was discovered in 2012 by CMS and ATLAS Collaborations at the Large Hadron Collider by observing an excess of events at a mass of approximately 125 GeV with a statistical significance of five standard deviations above the background expectations [1, 2]. Within the uncertainity intervals, all the latest analysis results are in agreement with Standard model (SM). However, more precise measurements of Higgs boson properties have to be performed to establish whether this particle has all the properties of the SM Higgs. Any deviation in the measurement would imply new physics beyond Standard Model.

In this thesis, the analysis is constrained to the so-called "golden channel" ($H \rightarrow ZZ^* \rightarrow 4l$), in which the Higgs boson decays into 2 Z bosons, which subsequently decay into 4 leptons (4 electrons or 4 muons or 2 electrons and 2 muons). Higgs boson production mechanisms with the largest cross section, i.e., gluon fusion, vector boson fusion, W and Z associated production and production in association with top quarks are studied. At the CMS, Higgs boson cannot be directly observed. Instead, measurements of variables related to objects involved in its production and decay can be performed. Using those variables, the goal is to deduce by which production mechanism Higgs boson has been produced. Two approaches to categorization are being considered: cut-based, which is already implemented in the CMS analysis, and the boosted decision tree categorization, which is developed in this project. Both categorizations are then used in the procedure of signal strength fitting, where the signal strength modifier is defined as the ratio of the signal cross section to it's Standard Model theoretical value [3]. An improved classification may lead to smaller uncertainity intervals in the expected values of signal strength, which can result in more precise measurements with real data.

This thesis is organised as follows. Brief introduction to the particle physics is given in Section 2. Additionally, the main production mechanisms and decay modes of the Hisggs boson are discussed. In Section 3, characteristics of the Compact Muon Solenoid are described. In Section 4, a machine learning algorithm of boosted decision tree is explained as well as the procedure for obtaining the values of signal strength modifiers. In Section 5, the efficiencies of cut-based and BDT categorization are reported. Furthermore, the values of the expected signal strength and corresponding uncertainty intervals are given for both classifications.

2 The Higgs boson

2.1 Particle world

The world appears to be built from only a few different particles. Atoms consist of positively charged nuclei and negatively charged orbiting electrons, which are bound to the atom by an electromagnetic force. In nuclei, protons and neutrons are bound by a strong nuclear force. All the previously mentioned particles also interact via weak nuclear force which is responsible for β decay. As a product of β decay, a nearly massless particle, a neutrino, is produced. Almost all problems in physics, except the ones on larger scales, can be explained by electrons, protons, neutrons and neutrinos interacting by electromagnetic, weak and strong nuclear force. However, at higher energy scales, a richer structure is observed. Protons and neutrons are not fundamental, but consist of up and down quarks. Quarks are the only elementary particles that experience a strong nuclear force, which is governed by the laws of Quantum Chromodynamics (QCD). In QCD, the strong interaction is realized by the exchange of gluons, carriers of the strong nuclear force[4]. Up and down quarks have electric charge $\frac{2}{3}e$ and $-\frac{1}{3}e$. Quarks are never observed in isolation. Instead, they are confined to bound states called hadrons[4]. Quarks carry the colour charge, the QCD equivalent of the electric charge, while the hadrons are colourless. When an object containing colour charge fragments, to obey confinement, a narrow cone of hadrons is created in the process of hadronization[4]. These cones are called jets and are a crucial part of the analysis in this thesis. Just as the quarks interact via the exchange of gluons, every particle with electric charge interacts by exchanging photons. In particle physics, these interactions are usually depicted using Feynamnn diagrams, e.g., the one shown in figure 1:

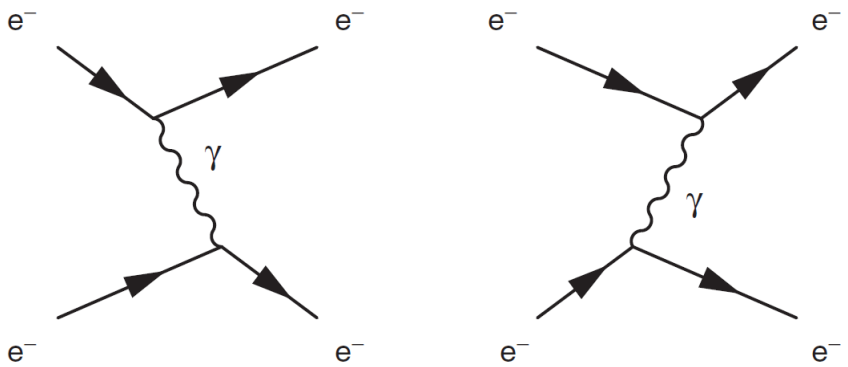

Figure 1: *2 time orderings for electron scattering via exchange of photon. Taken from [4].*

In Feynmann diagrams, time runs from left to right. In the left diagram of the figure 1, upper electron emmits the photon which is absorbed by the lower electron. In the right diagram of figure 1, lower electron emmits a photon which is absorbed by the upper electron. In both cases, the electromagnetic force is realized by the photon carrying a momentum.

Following the electromagnetic and strong nuclear force, the weak nuclear interaction is mediated by the exchange of W^+, W^- bosons via charged-current and Z bosons via neutral current. Force-carrying particles are called gauge bosons and share the same spin-1 [4]. Hypothesized graviton, a carrier of a gravitational force, should have spin-2. Gravity, which currently cannot be explained by the principles of quantum field theory, has a neglibible impact at distance scales in particle physics and is therefore not being considered. The relative strengths of the 4 forces in the range of 10^{-15} m are reported in table 1:

Fundamental forces	
Force	Strength
Strong	1
Electromagnetic	10^{-3}
Weak	10^{-8}
Gravitational	10^{-37}

Table 1: *The relative strengths of the 4 forces in the range of* 10^{-15} *m, based on [4].*

It is observed that the electron neutrino, up quark, down quark and electron are not the only elementary particles. Instead, they constitute the first generation. The second generation consists of a muon neutrino, strange quark, charm quark and muon while the third generation consists of tau neutrino, top quark, bottom quark and tau lepton [4]. The only difference between these generations is in the increasing mass of elementary particles. The final element of particle physics is the Higgs boson, which was discovered by the ATLAS and CMS experiments at the Large Hadron Collider (LHC) in 2012. It provides the mechanism by which all other particles acquire mass and is the only spin-0 particle [4]. All known elementary particles with their mass, charge and spin are shown in figure 2:

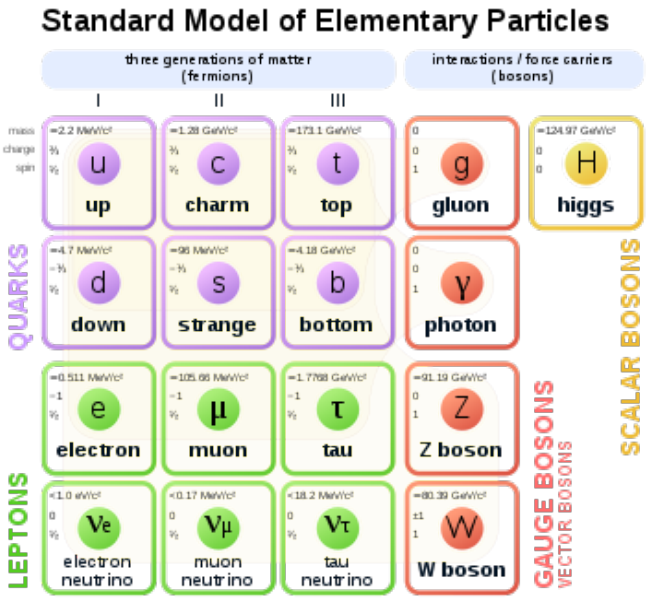

Figure 2: *Elementary particles of the Standard Model. Taken from [5].*

2.2 Lagrangian of the Standard model

Elementary particle physics is described by the Quantum Field Theory of the Standard model. In the Quantum Field Theory (QFT), particles are represented as excitations of quantum fields and their dynamics is determined by the corresponding Lagrangian [4]. In the Standard model, the local gauge symmetry of the gauge group $SU(3)_C \times SU(2)_L \times U(1)_Y$ is imposed on the Lagrangian where C, L and Y stand for color, left-handed and hypercharge respectively [6]. Here, the $SU(n)$ group represents the Lie group (informally, a group of symmetries where symmetries are continous) of $n \times n$ unitary matrices with determinant 1. The recipe for introducing an interaction in QFT is replacing the partial derivative ∂_μ in the free Lagrangian:

$$\mathscr{L}_{free} = i\bar{\psi}\gamma^\mu\partial_\mu\psi \tag{2.1}$$

with the corresponding covariant derivative D_μ.
The electroweak interaction is introduced by requiring $SU(2)_L \times U(1)_Y$ symmetry. Since the weak-charged current couples to left-handed (LH) chiral particle states and right-handed (RH) chiral antiparticle states , RH particle and LH antiparticle states are placed into isospin singlets ψ_R, while LH particle and RH antiparticle states are placed into isospin doublets ψ_L [4].
In the following expressions, a summation over repeating letters is assumed.
Weak isospin singlets transform as [6]:

$$\psi'_R = e^{-i\beta(x)\frac{Y}{2}}\psi_R, \tag{2.2}$$

while the doublets transform as [6]:

$$\psi'_L = e^{-i\alpha_i(x)\frac{\sigma_i}{2} - i\beta(x)\frac{Y}{2}}\psi_L, \tag{2.3}$$

where σ_i, $\alpha_i(x)$ and $\beta(x)$ are Pauli matrices (generators of $SU(2)$ group), parameters of $SU(2)_L$ and $U(1)_Y$ group, respectively. To achieve the desired symmetry, the new gauge fields B_μ and W_μ are introduced with the coupling constants g and g_w. Now, the covariant derivatives can be defined as [6]:

$$D^L_\mu = \partial_\mu + ig_w\frac{\sigma_i}{2}W^i_\mu + ig\frac{Y}{2}B_\mu, \tag{2.4}$$

$$D^R_\mu = \partial_\mu + ig\frac{Y}{2}B_\mu, \tag{2.5}$$

Gauge fields transform as [6]:

$$B_\mu \to B_\mu + \frac{1}{g}\partial_\mu\beta(x), \tag{2.6}$$

$$W^i_\mu \to W^i_\mu + \alpha^j(x)\epsilon^{ijk}W^k_\mu + \frac{1}{g_w}\partial_\mu\alpha^i(x), \tag{2.7}$$

and the electroweak Lagrangian can be written as:

$$\mathscr{L}_{EW} = i\bar{\psi_L}\gamma^\mu D^L_\mu \psi_L + i\bar{\psi_R}\gamma^\mu D^R_\mu \psi_R - \frac{1}{4}W^{\mu\nu}_i W^i_{\mu\nu} - \frac{1}{4}B^{\mu\nu}_i B^i_{\mu\nu}, \tag{2.8}$$

with $W^{\mu\nu}_i W^i_{\mu\nu}$ and $B^{\mu\nu}_i B^i_{\mu\nu}$ corresponding to the trilinear and quadrilinear interactions among gauge bosons and the summation over all doublets and singlets implied.
Quantum Chromodynamics is obtained by requiring Lagrangian invariance under $SU(3)_C$ group, which means that the quark fields transform as [6]:

$$q' = e^{-i\alpha_i(x)\frac{\lambda_i}{2}}q, \tag{2.9}$$

where $\alpha_i(x)$ and λ_i correspond to parameters of $SU(3)_C$ group and Gell-Mann matrices (generators of $SU(3)$ group). Following the same procedure as for the electroweak sector, new gauge fields G_μ, associated with gluons, are introduced and the QCD covariant derivative is defined as [6]:

$$d_\mu = \partial_\mu + ig_s\frac{\lambda_i}{2}G^i_\mu, \tag{2.10}$$

with strong coupling constant g_S and $SU(3)$ structure constants f^{abc}.
Gauge fields transform as [6]:

$$G^i_\mu \to G^i_\mu + \alpha^j(x)f^{ijk}G^k_\mu + \frac{1}{g_S}\partial_\mu\alpha^i(x), \tag{2.11}$$

and the QCD Lagrangian is given by:

$$\mathscr{L}_{QCD} = i\bar{q}\gamma^\mu d_\mu q - \frac{1}{4}G^{\mu\nu}_i G^i_{\mu\nu}, \tag{2.12}$$

with the last term describing the trilinear and quadrilinear interactions between gluon fields and the summation over the quark fields assumed.
This theory, however, requires massless bosons and fermions to preserve local gauge invariance. Solution was proposed in three independent papers from Englert and Brout [7], Higgs [8], and Guralnik, Hagen and Kibble [9], by introducing the mechanism of spontaneous symmetry breaking. A new field, which is symmetric under the gauge transformations of the Standard model gauge group, which acquires a nonzero expectation value in the vacuum state and breaks the electroweak symmetry, is introduced as $SU(2)_L$ doublet of complex scalar fields [6]:

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix}\phi^1 + i\phi^2 \\ \phi^3 + i\phi^4\end{pmatrix} \tag{2.13}$$

The Higgs Lagrangian is given by [4]:

$$\mathscr{L}_{Higgs} = (D_\mu\phi)^\dagger(D_\mu\phi) - V(\phi) \tag{2.14}$$

where D_μ is electroweak covariant derivative and $V(\phi)$ is Higgs potential:

$$V(\phi) = \mu^2\phi^\dagger\phi + \lambda^2\phi^\dagger\phi \tag{2.15}$$

with constant terms λ and μ^2.
It can be shown that, in terms of Weinberg angle θ_w, photon field A_μ, Z boson field Z_μ and $W^\pm$ boson field $W^\pm_\mu$, Higgs Lagrangian can be written as [6]:

$$\begin{aligned}\mathscr{L}_{Higgs} =& \frac{1}{2}\partial_\mu h\partial^\mu h + \mu^2 h^2 + \frac{g_w^2 v}{4}W_\mu^- W^{+\mu} + \frac{g_w^2 v}{8\cos^2\theta_w}Z_\mu Z^\mu \\ &+ \frac{g_w^2 v}{2}hW_\mu^- W^{+\mu} + \frac{g_w^2}{4}h^2 W_\mu^- W^{+\mu} + \frac{g_w^2 v}{4\cos^2\theta_w}hZ_\mu Z^\mu \\ &+ \frac{g_w^2 v}{8\cos^2\theta_w}h^2 Z_\mu Z^\mu + \frac{\mu^2}{v}h^3 + \frac{\mu^2}{4v^2}h^4\end{aligned} \tag{2.16}$$

where v and h represent the terms corresponding to the vacuum state and the expansion about the vacuum state.
Masses of Z and $W^\pm$ bosons are extracted from (2.16):

$$\begin{aligned} m_Z &= \frac{g_w v}{2\cos\theta_w} \\ m_W &= \frac{1}{2}g_w v\end{aligned} \tag{2.17}$$

and their ratio $\frac{m_W}{m_Z} = \cos\theta_w$ is experimentally verified [4].
The Higgs mechanism also generates fermion masses which are incorporated in Yukawa extension [6]:

$$\mathscr{L}_{Yukawa} = \sum_f -m_f\psi\bar{\psi}(1-\frac{h}{v}) + \sum_{f'} -m_{f'}\psi'\bar{\psi}'(1-\frac{h}{v}) \tag{2.18}$$

where the first sum runs over up-type fermions and the second sum over down-type fermions.
Full Standard model Lagrangian is then given by summing the individual contributions:

$$\mathscr{L}_{SM} = \mathscr{L}_{QCD} + \mathscr{L}_{EW} + \mathscr{L}_{Higgs} + \mathscr{L}_{Yukawa} \tag{2.19}$$

With this Lagrangian, Standard model has passed many experimental tests. However, it still does not explain the baryon asymmetry, neutrino oscillations and neutrino nonzero masses. It also does not incorporate the theory of gravity consistent with general relativity, dark matter and dark energy.

2.3 Higgs boson production and decay modes

Higgs boson can be produced at the Large Hadron Collider (LHC) via many different processes. Feynmann diagrams of the most dominant production mechanisms are shown in figure 3:

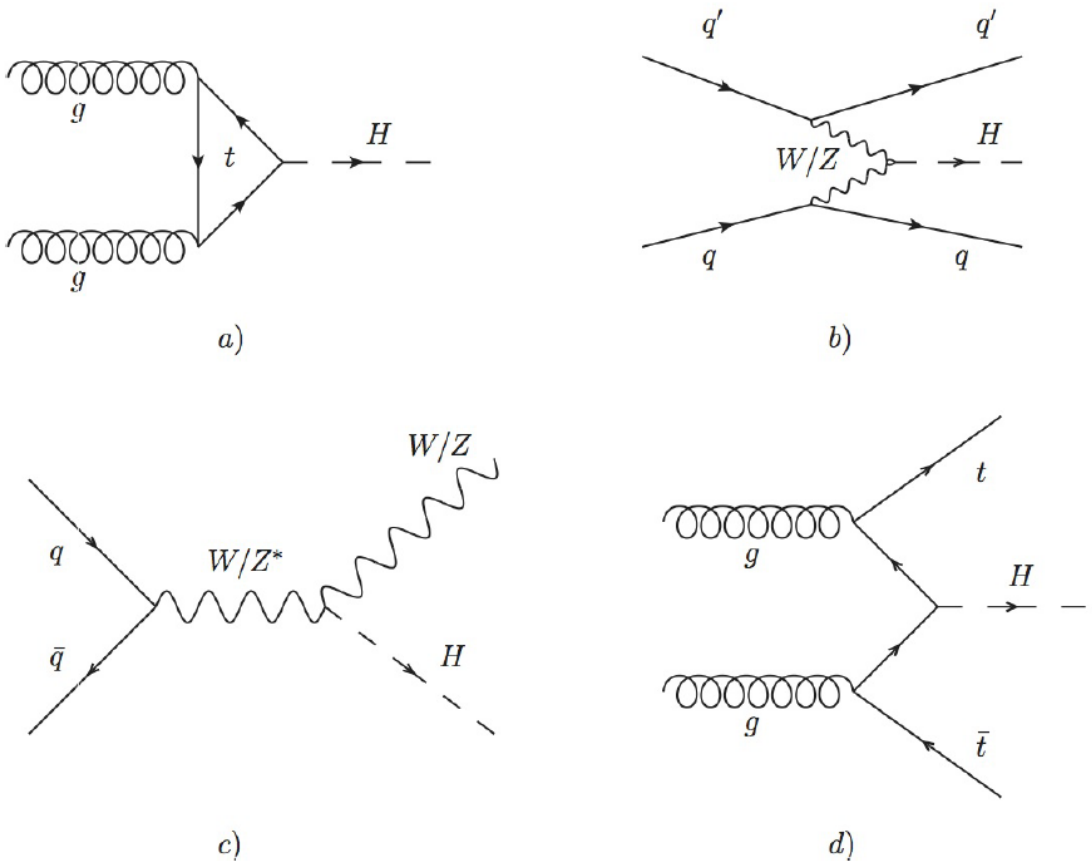

Figure 3: *Most dominant Higgs boson production mechanisms: (a) gluon fusion (ggH), (b) vector boson fusion (VBF), (c) W and Z associated production or Higgsstrahlung (VH) and (d) ttH associated production. Taken from [6].*

The gluon fusion (ggH) via loop of virtual quarks has the largest cross section of all production mechanisms since the protons are being collided at the LHC. Vector boson fusion (VBF), in which two fermions exchange virtual Z or W bosons, which then fuse into the Higgs boson, has the second largest cross section. In the third most dominant process, associated production with a vector boson or Higgsstrahlung (VH), the fermion and anti-fermion collide and produce a W or Z boson which later on radiates a Higgs boson. Of all the production mechanisms depicted in figure 3, associated production with a top quark pair (ttH), in which the quark and antiquark from the gluon decay fuse into the Higgs boson, has the smallest cross section.
Other production mechanisms include a bottom quark pair (bbH) and single top production (tqH). However, these are not included in this thesis due to their small cross sections. Contributions of different processes to the production of Higgs boson, as a function of the total

energy in the centre of mass frame ($\sqrt{s}$), are summarized in figure 4 (where qqH corresponds to vector boson fusion). Since 2015 LHC operates at $\sqrt{s} = 13$ TeV.

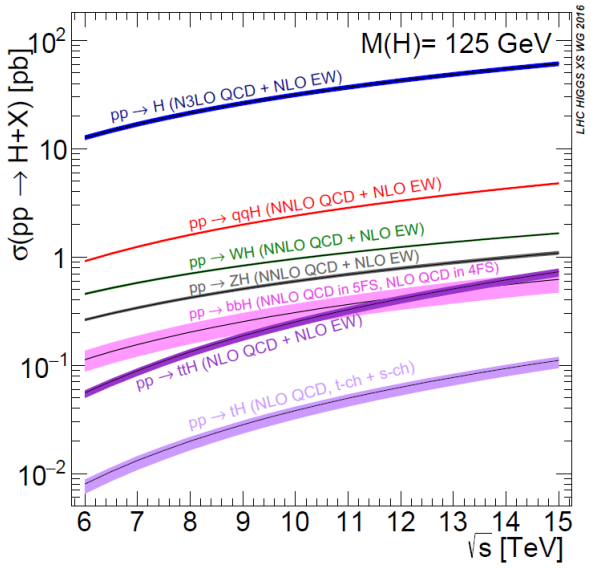

Figure 4: *Cross section for 125 GeV Higgs boson at LHC as a function of $\sqrt{s}$. Taken from [6].*

At the LHC, processes that mimic Higgs production are also observed. They are called background and many efforts are put into discriminating them against the genuine Higgs bosons. The most prominent background processes (and the ones that are considered in signal strength fits section) are shown in figure 5. There are also rare elctroweak backgrounds, but they are not considered in this thesis.

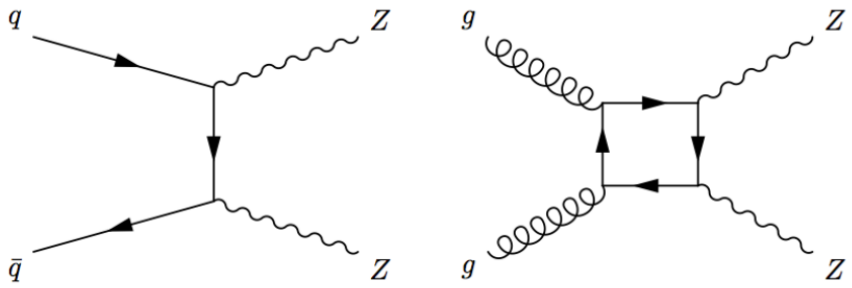

Figure 5: *qqZZ (left) and ggZZ (right) background processes. Taken from [10]*

According to the SM, Higgs boson can decay into all particles. However, due to the proportionality of the coupling to a particle mass, Higgs boson is most likely to decay into massive particles [4]. Corresponding branching ratios are shown in table 2.

Decay modes that involve quarks and gluons, despite having larger branching ratios, are challenging to study due to the high QCD background. In the second most probable process, $H \rightarrow WW^*$, Higgs boson invariant mass cannot be reconstructed due to the presence of neutrinos. In this thesis, the channel $H \rightarrow ZZ^* \rightarrow 4l$ is being studied. Since the CMS detector

Decay channel	Branching ratio
$H \to \gamma\gamma$	2.28×10^{-3}
$H \to ZZ$	2.64×10^{-2}
$H \to W^+W^-$	2.15×10^{-1}
$H \to \tau^+\tau^-$	6.32×10^{-2}
$H \to b\bar{b}$	5.77×10^{-1}
$H \to Z\gamma$	1.54×10^{-3}
$H \to \mu^+\mu^-$	2.19×10^{-4}

Table 2: *The predicted branching ratios of the Higgs boson for m_H = 125 GeV. Taken from [11].*

can reconstruct the final states of the four leptons involved in this process, it leaves a clear experimental signature at LHC, and is therefore considered as a "golden channel".

3 Compact muon solenoid

The LHC is the world's largest and most powerful particle accelerator. It accelerates protons and collides them at four locations around it's ring [12]. The Compact Muon solenoid sits at one of the collision points and is designed to observe new physics that LHC may reveal [12]. CMS has several layers of concentric components and consists of Solenoid, Tracker, Electromagnetic Calorimeter (ECAL) , Hadronic Calorimeter (HCAL) and muon detectors [12].

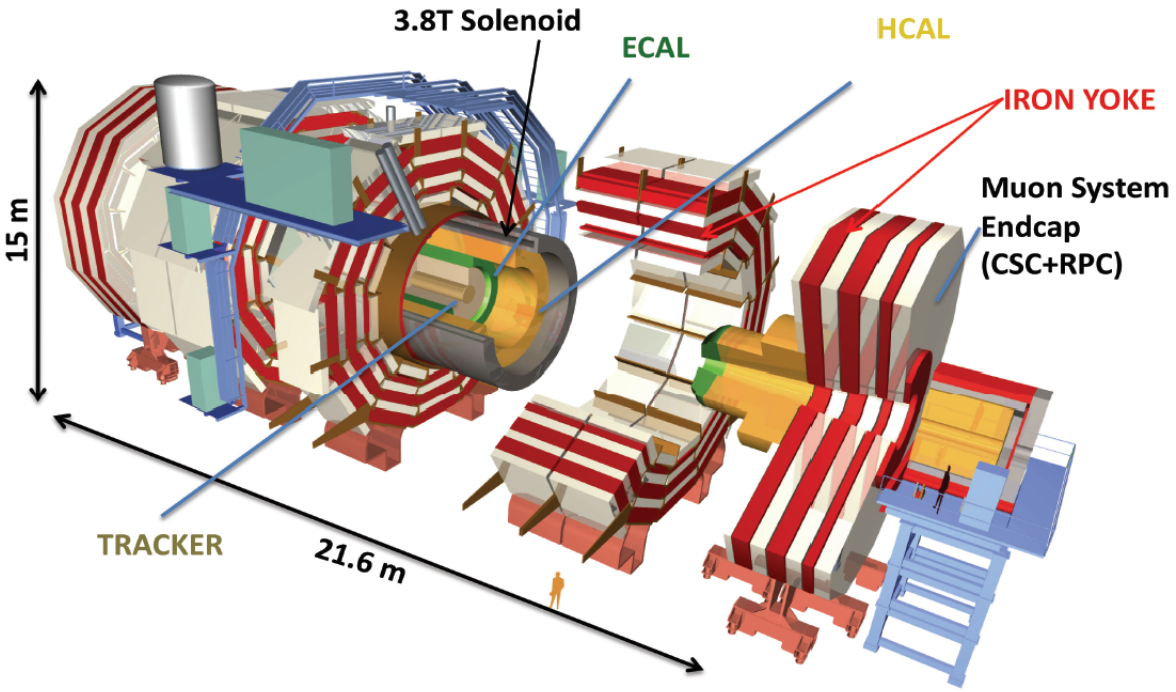

Figure 6: *Schematic view of the CMS detector. Taken from [13].*

3.1 Solenoid

CMS magnet is a solenoid that creates a magnetic field up to 3.8 Tesla [12]. To create such a strong magnetic field, the solenoid is cooled to -268.5 °C in order to achieve superconductivity [12]. Particles emerging from collisions are bent by this uniform magnetic field and the curvature of their path gives a measure of their momentum [12]. Calorimers and the Tracker are located inside the magnet while the muon detectors are placed outside. The magnet weights around 12 000 tonnes and provides most of the experiment's structural support [12].

3.2 Tracker

The CMS Tracker is the innermost part of the detector. It measures the momentum of charged particles by finding their positions at a number of key points [12]. It is made of silicon pixels and silicon microstrip detectors [12]. As the particles travel through the tracker, the pixels and microstrips produce tiny electric signals that are amplified and detected [12]. The tracker also houses sensors with 75 million separate electronic read-out channels [12].

3.3 Electromagnetic Calorimeter

The task of the Electromagnetic Calorimeter is to measure the energy of electrons and photons. It is made of a barrel section and two endcaps [12]. The cylindrical barrel consists of 61 200 lead tungstate crystals formed into 36 supermodules, and is sealed of with the endcaps made up of almost 15,000 further crystals [12]. Lead tungstate is a high density and transparent material [12]. It interacts with photons and electrons, producing light in fast, short, well-defined photon bursts that allow for a precise and fast measurement [12]. ECAL also contains Preshower detectors that are located in front of the endcaps [12]. These allow CMS to distinguish between single high-energy photons and close pairs of low-energy photons [12].

3.4 Hadronic Calorimeter

The Hadronic Calorimeter measures the energy of hadrons and provides an indirect measurement of the presence of neutrinos [12]. It consits of barrel, endcap and forward sections [12]. HCAL finds a particle's position, energy and arrival time using alternating layers of absorber and fluorescent scintillator materials that produce a rapid light pulse when the particle passes through [12]. This light is then readout and summed over many layers of material providing a measure of particle's energy [12].

3.5 Muon detectors

Muons can penetrate several layers of material without interacting and are not stopped by CMS calorimeters [12]. Muon chambers are therefore placed at the outer part of the experiment. The momentum of a particle is measured by fitting a curve to hits among the four muon stations, which sit outside the magnet coil and are interleaved with iron return yoke plates [12].

4 Methods

4.1 Multivariate analysis

A key task in this thesis is to discriminate different Higgs boson production mechanisms exploiting a number of discriminating variables provided by the CMS detector. This problem in statistics is also known as hypothesis testing. Given some data, 4 hypotheses are opposed based on the 4 production mechanisms studied in this thesis (ggH, VBFH, VH and ttH). Multivariate analysis deals with the construction of a test statistic using multidimensional ntuples of data: $x = (x_1, x_2, ..x_N)$. Neymann-Pearson lemma [14] defines the test statistic $t(x)$ which is optimal for the separation of hypotheses A and B:

$$t(x) = \frac{\mathscr{L}(x|A)}{\mathscr{L}(x|B)} \tag{4.1}$$

where $\mathscr{L}(x|A)$ and $\mathscr{L}(x|B)$ denote likelihood functions of hypothesis A and B given some data x. However, since the individual probability density functions of variables $x_1, x_2, ..x_N$ and their correlations are usually not known in practical problems, other expressions for the test statistic have to be used. One of them (which is used in this thesis) is the boosted decision tree (BDT) score.

4.1.1 Boosted decision trees

Decision tree is a machine learning algorithm that classifies data using a series of binary classifiers organized in a tree structure that operate until the stop criterion is reached.
Starting from the root node, BDT first finds the variable that splits the data in the most optimal way. One of the most common separation criteria used is a Gini index, defined as:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{4.2}$$

where n is the number of events and p_i is the probability of an event being classified to a particular class. In the process of adaptive boosting (AdaBoost), the misclassified events are

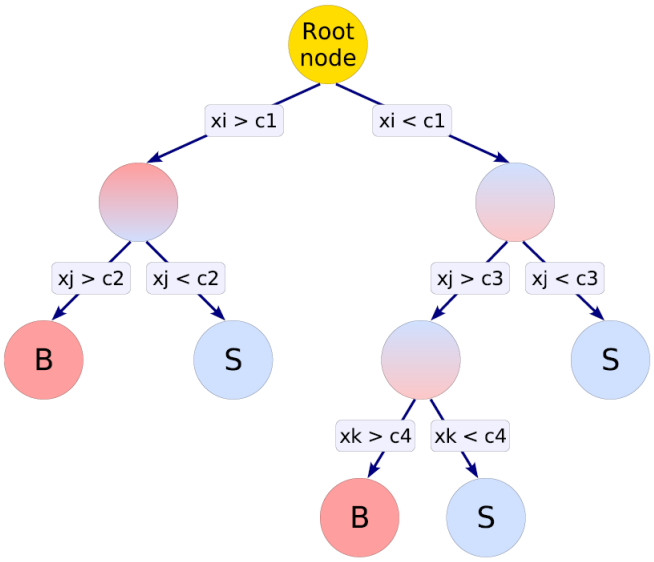

Figure 7: *Schematic view of a decision tree. Taken from [15].*

then given a higher weight in the next iteration of training and the boost weight α is calculated as [15]:

$$\alpha = \frac{1 - err}{err} \tag{4.3}$$

where err represents the missclassification rate. Learning rate of the AdaBoost algorithm can be further configured by introducing AdaBoostBeta parameter β and making substitution $\alpha \to \alpha^{\beta}$ [15]. The boosted event classification $y(x)$ is then given by [15]:

$$y(x) = \frac{1}{N} \sum_{i=1}^{N} \ln(\alpha_i) h_i(x) \tag{4.4}$$

where $h_i(x)$ is a result of an individual classifier that takes the values 1 and -1 for signal and background events, respectively. AdaBoost performs best on weak classifiers which have small maximum depth and are thus much less prone to overtraining [15].

One of the shortcomings of decision trees is their instability with respect to statistical fluctuations in the training sample from which the tree structure is derived [15]. This issue is adressed by constructing a forest of decision trees and classifying an event by a majority vote of the classifications done by each tree in the forest [15].

4.2 Signal strength modifier

Signal strength modifier μ is defined as the ratio of a signal cross section to it's Standard Model theoretical value [3]:

$$\sigma = \mu \sigma_{SM} \tag{4.5}$$

Since any deviation from SM prediction can be a sign of a new physics, the signal strength modifier is a parameter of interest in CMS analysis. The latest measurements of four-lepton mass (m_{4l}) and the observed signal strength modifier published by CMS Collaboration are shown in figures 8 and 9:

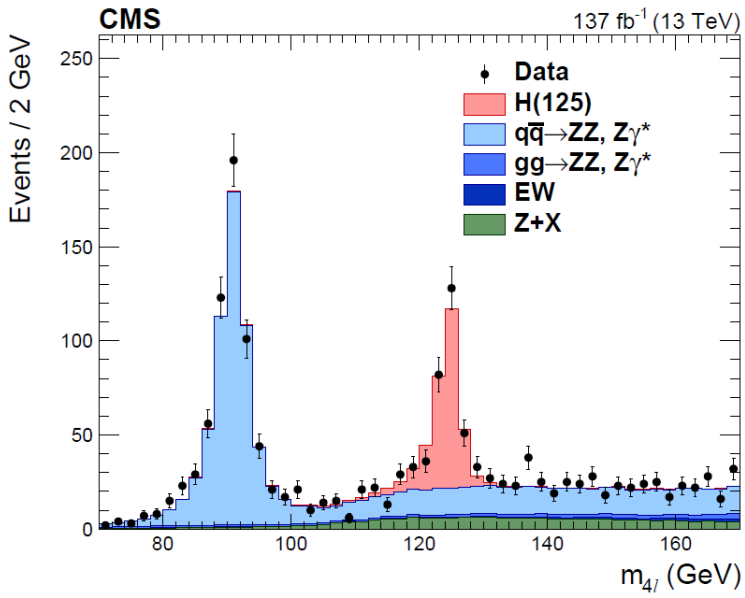

Figure 8: *Points with error bars represent the data and stacked histograms represent the expected distributions for the signal and background processes. The SM Higgs boson signal with $m_H = 125$ GeV, denoted as H(125), the ZZ and rare electroweak backgrounds are normalized to the SM expectation, the Z+X background to the estimation from data. Taken from [16].*

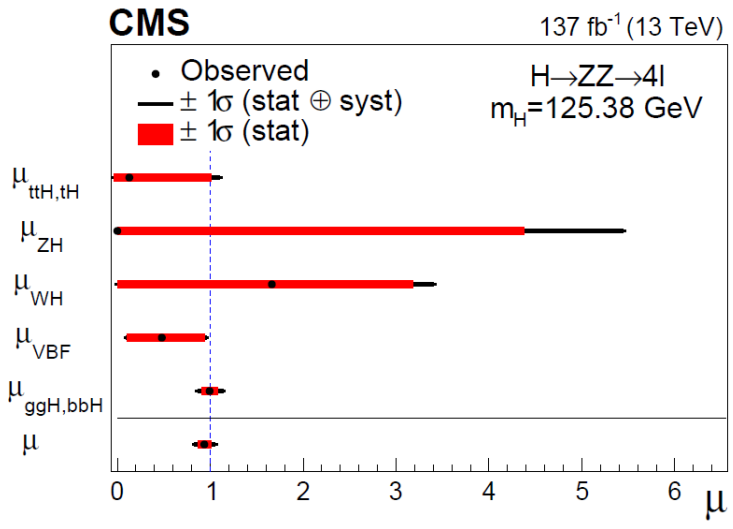

Figure 9: *Observed signal strength modifiers for 5 SM Higgs production mechanisms.The thick black lines indicate the one standard deviation confidence intervals including both statistical and systematic sources. The thick red lines indicate the statistical uncertainties corresponding to the one standard deviation confidence intervals. Taken from [16].*

Signal strength modifier is determined using the maximum likelihood method. Each independent source of systematic uncertainty is assigned a nuisance parameter θ_i, the full set of which is denoted as θ [6]. Systematic uncertainties usually reflect the possible deviations of a quantity from the input value $\tilde{\theta}$ provided by a separate measurement [6]. The likelihood function is now defined as [6]:

$$\mathscr{L}(data, \tilde{\theta}|\mu, \theta) = \prod_c \mathscr{L}_c(data|\mu s(\theta) + b(\theta)) \prod_i p(\tilde{\theta}_i|\theta_i) \tag{4.6}$$

where $s(\theta)$ and $b(\theta)$ stand for the expected SM signal and background yields with the first product running over all channels c and the second over all nuisance parameters i. In 4.6, likelihood function $\mathscr{L}_c$ denotes the product over N events of the probability density functions of observable $\mathscr{O}$ for signal $f_s(\mathscr{O}|\theta)$ and background $f_b(\mathscr{O}|\theta)$, weighted by the total expected signal and background rates $S(\theta)$ and $B(\theta)$ [6]:

$$\mathscr{L}_c(data|\mu s(\theta) + b(\theta)) = \frac{1}{N} \prod_{e=1}^{N} (\mu S(\theta) f_s(\mathscr{O}_e|\theta) + B(\theta) f_b(\mathscr{O}_e|\theta)) e^{-\mu S(\theta) + B(\theta)} \tag{4.7}$$

To obtain the uncertainty intervals, a negative log-likelihood function is defined as [6]:

$$-2\Delta ln\mathscr{L} = -2ln \frac{\mathscr{L}(data, \tilde{\theta}|\mu, \hat{\theta}_\mu)}{\mathscr{L}(data, \tilde{\theta}|\hat{\mu}, \hat{\theta})} \tag{4.8}$$

where $\hat{\mu}$ and $\hat{\theta}$ denote best fit values for signal strength modifier and nuisance parameters while $\hat{\theta}_\mu$ maximizes the numerator for a fixed set of values μ. According to Wilk's theorem, the function defined in 4.8, with n parameters of interest, approaches a chi- square (χ^2) distribution in the limit of a large data sample [17]. For example, when measuring one parameter μ, a 68% confidence interval can be deduced from $-2\Delta ln\mathscr{L} < 1$ condition [6]. Expected results can also be provided for some nominal values of the parameters [6]. This would require to generate a large number of pseudoexperiments and determine their median outcome, but a very good approximation is provided by the Asimov data set, i.e., one single representative data set in which the observed rates and distributions coincide with predictions under the nominal set of nuisance parameters [6]. Probability density functions (PDF) in 4.7 are 2-dimensional and rely on four-lepton mass m_{4l} and background kinematic discriminant D_{bkg}^{kin} [6]:

$$f_{2D}^{\mu}(m_{4l}, D_{bkg}^{kin}) = \mathscr{P}(m_{4l}) \times \mathscr{P}(D_{bkg}^{kin}|m_{4l}) \tag{4.9}$$

Mass PDF used for the signal is Double-sided Crystal Ball function while Bernstein polynomial of order 2 is used for the background [6]. The conditional term is based on 2D histogram templates to which a smoothing procedure is applied [6].

5 Results

Since Higgs boson production mechanisms cannot be directly identified, the objects involed in the $H \to ZZ^* \to 4l$ process are used to categorize events in several exclusive categories. Two types of classifications are presented here: cut-based categorization and a BDT categorization which was developed as a part of this project. Cut-based categorization has already been adopted in the analysis and CMS Collaboration has published the latest measurements of cross sections [16]. The goal is to compare the efficiencies between the two classifications and determine how they reflect the outcome of the measurement of the signal strength modifier. In the categorization sections (5.1, 5.2.1 and 5.2.2) only signal processes are classified, i.e., classification of background events is not performed and the analysis is done in the $118 < m_{4l} < 130$ GeV window. Signal strength fits section extends four-lepton mass range to $105 < m_{4l} < 140$ GeV and treats ggZZ and qqZZ background together with the signal events.

5.1 Cut-based categorization

In the cut-based categorization, the following variables are used for the classification of Higgs boson production mechanisms:

- the number of selected jets
- the number of selected b-tagged jets
- the number of additional leptons
- kinematic discriminators
- number of additional Z bosons

In the CMS analysis, kinematic discriminators are used to distinguish different processes. For example, when considering some arbitrary mechanisms A and B, the kinematic discriminator is defined as:

$$D_{AB} = \left[1 + \frac{P_A}{P_B}\right]^{-1} \tag{5.1}$$

where P_A and P_B stand for probabilities for processes A and B, which are calculated from the SM Lagrangian. Defined this way, higher values of D_{AB} indicate that the event originates from process A, while the lower values indicate that the event originates from process B. In the cut-based categorization, the kinematic discriminators are constructed to separate VBF (with 1 or 2 selected jets), ZH and WH production mechanisms from ggH. They are defined as [6]:

$$D^{ME}_{VBF-1j} = \left[1 + \frac{P_{H+J}(\Omega^{H+J}|m_{4l})}{\int d\eta_j P_{VBF}(\Omega^{H+JJ})|m_{4l})}\right]^{-1} \tag{5.2}$$

$$D^{ME}_{VBF-2j} = \left[1 + \frac{P_{H+JJ}(\Omega^{H+JJ}|m_{4l})}{P_{VBF}(\Omega^{H+JJ})|m_{4l})}\right]^{-1} \tag{5.3}$$

$$D^{ME}_{ZH-hadr.} = \left[1 + \frac{P_{H+JJ}(\Omega^{H+JJ}|m_{4l})}{P_{ZH-hadr.}(\Omega^{H+JJ})|m_{4l})}\right]^{-1} \tag{5.4}$$

$$D^{ME}_{WH-hadr.} = \left[1 + \frac{P_{H+JJ}(\Omega^{H+JJ}|m_{4l})}{P_{WH-hadr.}(\Omega^{H+JJ}|m_{4l})}\right]^{-1} \tag{5.5}$$

where denominators in expressions 5.3, 5.4, 5.5 represent probabilities for VBF, ZH, and WH production, while $P_{H+J}(\Omega^{H+J}|m_{4l})$ and $P_{H+JJ}(\Omega^{H+JJ}|m_{4l})$ represent probabilities for ggH+1 jet and ggH+2 jets producion. Studies using simulation have shown that 40% of the VBF events have less than two selected jets in the $H \to 4l$ analysis, resulting in a VBF jet that is out of the detector acceptance, or that it is not reconstructed, or fails the selection requirements [6]. That is why in expression 5.2, P_{VBF} is simply integrated over the pseudorapidity of the unobserved jet while constraining the transverse momentum of the $4l$ + 2 jets system to be zero [6].
Seperating power of kinematic discriminators and the rest of the variables used in cut-based analysis are shown in figures 10 :

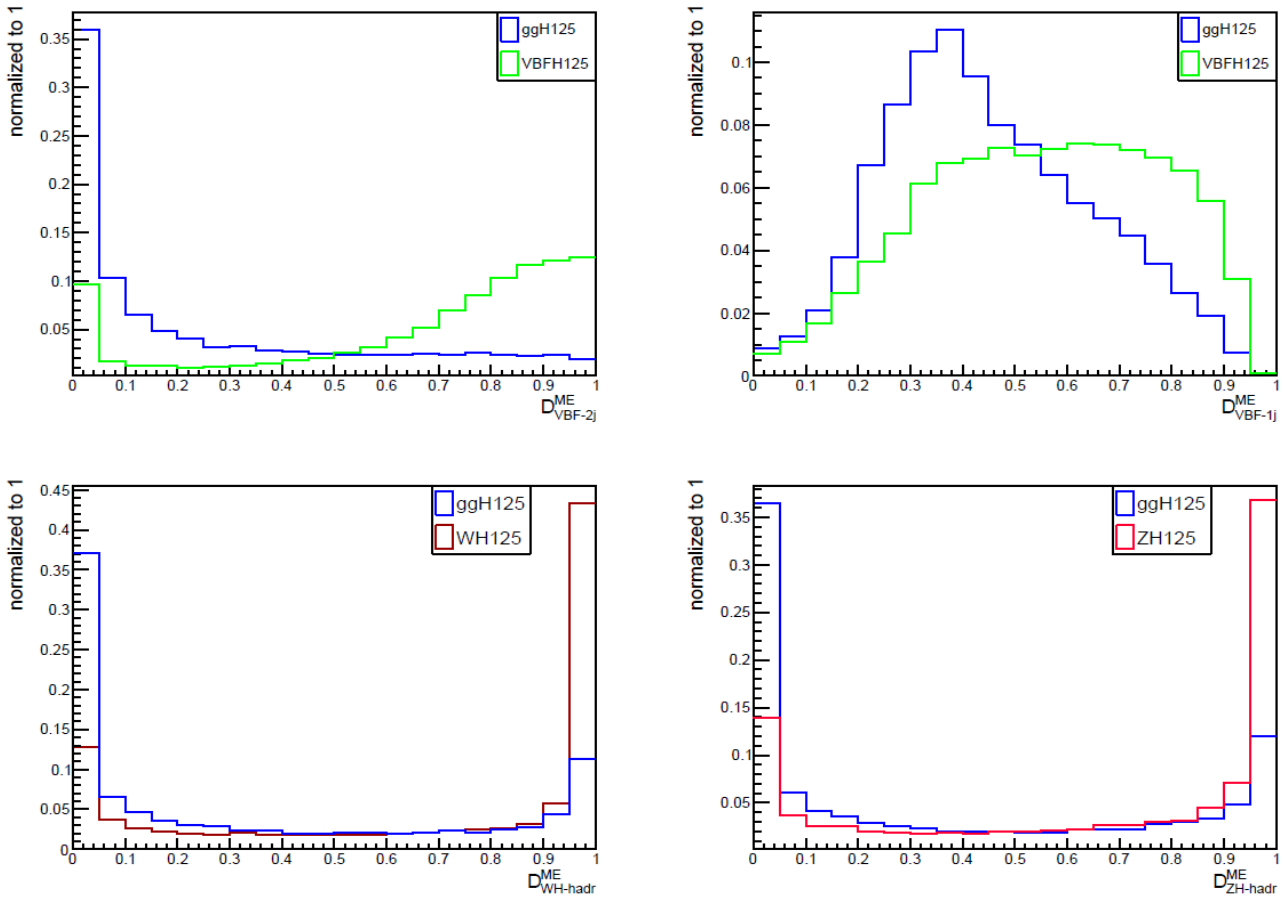

Figure 10: *Distributions of kinematic discriminators in* $118 < m_{4l} < 130$ *GeV window. Distribution of* D^{ME}_{VBF-2j} *with at least 2 selected jets (top left), distribution of* D^{ME}_{VBF-1j} *with exactly 1 selected jet (top right), distribution of* $D^{ME}_{WH-hadr}$ *with at least 2 selected jets (bottom left), distribution of* $D^{ME}_{ZH-hadr}$ *with at least 2 selected jets (bottom right).*

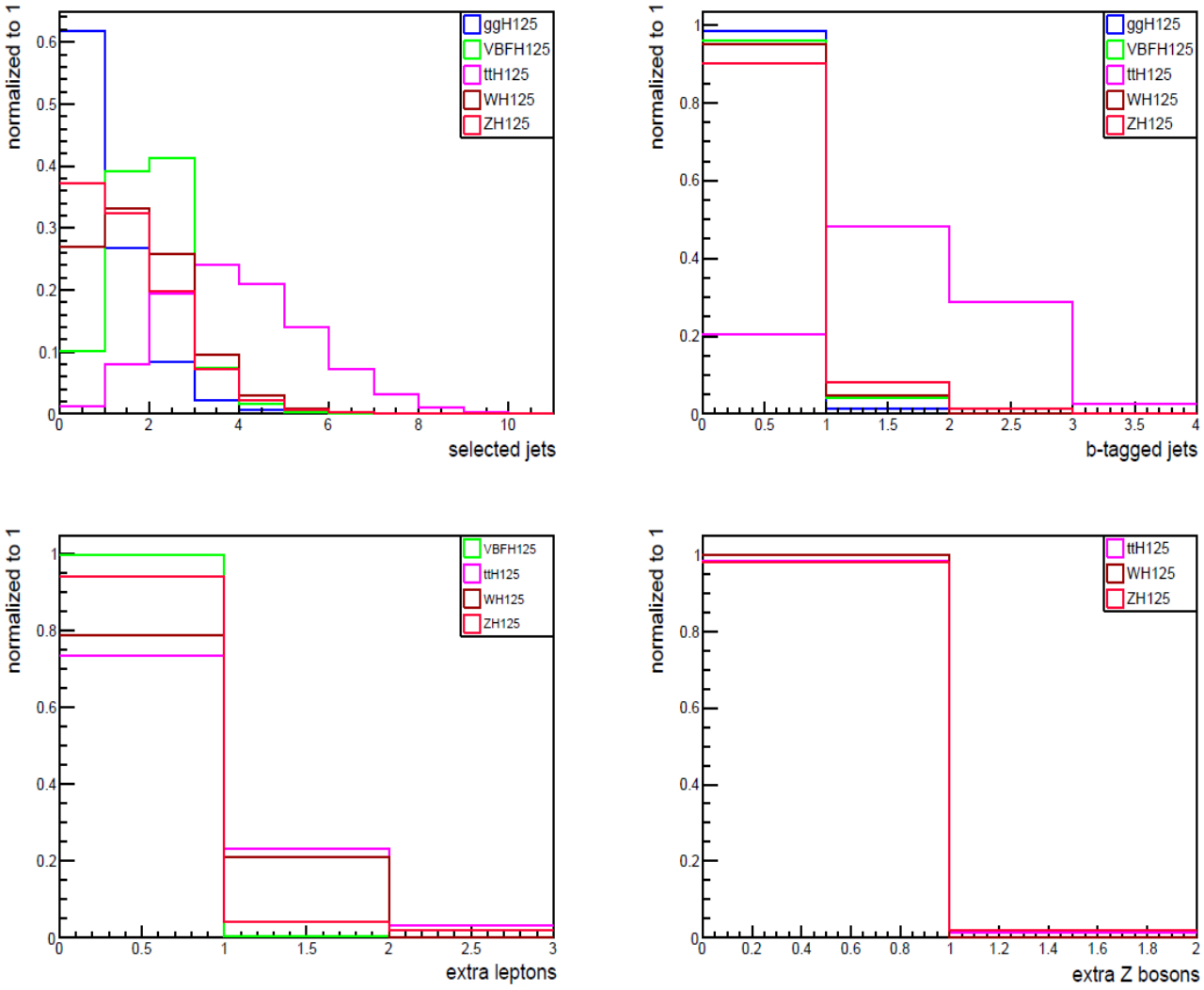

Figure 11: *Distributions of discrete variables in* $118 < m_{4l} < 130$ *GeV window. Distribution of number of selected jets (top left), distribution of number of selected b-tagged jets (top right), distribution of number of extra leptons (bottom left), distribution of number of extra Z bosons (bottom right)*

Seven exclusive categories are now defined as follows [18]:

- The **VBF-2jet-tagged category** requires exactly four leptons. In addition, there must be either two or three jets of which at most one is b-tagged, or four or more jets, none of which are b-tagged. Finally, $D^{ME}_{VBF-2j} > 0.5$ is required.
- The **VH-hadronic-tagged category** requires exactly four leptons. In addition, there must be two or three jets, or four or more jets, none of which are b-tagged. $D^{ME}_{VH-hadr} = max(D^{ME}_{ZH-hadr}, D^{ME}_{WH-hadr}) > 0.5$ is required.
- The **VH-leptonic-tagged category** requires no more than three jets and no b-tagged jets in the event, and exactly one additional lepton or one additional Z boson. This category also includes events with no jets and at least one additional lepton.

- The **ttH-hadronic-tagged category** requires at least four jets of which at least one is b-tagged and there are no additional leptons in the event.
- The **ttH-leptonic-tagged category** requires at least one additional lepton in the event.
- The **VBF-1jet-tagged category** requires exactly four leptons, exactly one jet and $D^{ME}_{VBF-1j} > 0.5$.
- The **Untagged category** consists of the remaining selected events.

Using this categorization, events from the 2018 simulation of 125 GeV Higgs boson production mechanisms ggH, VBF, ttH and VH are now classified and the results are summarized in figure 12:

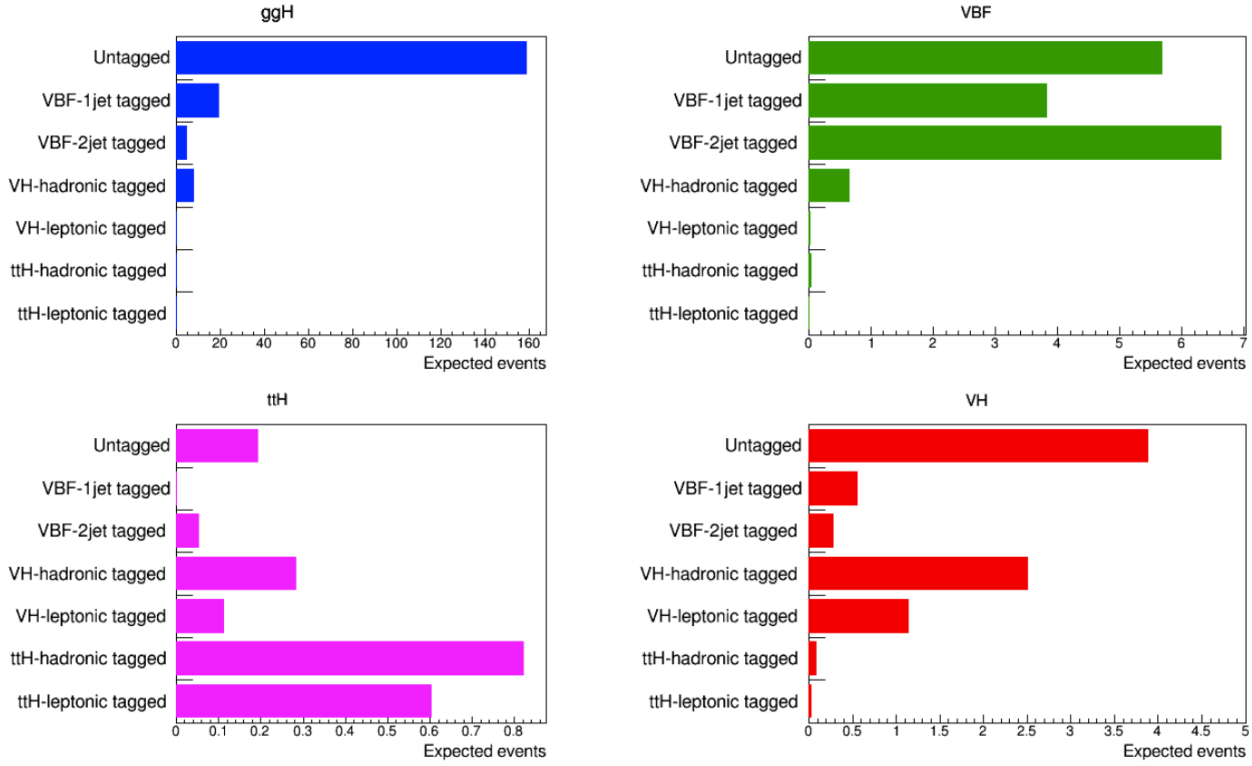

Figure 12: *Distributions of ggH, VBF, ttH and VH signal processes in the seven categories.*

To get a sense how well cut-based classification performs, a confusion matrix is created, where VBF-1jet and VBF-2jet, VH-hadronic and VH-leptonic, ttH-hadronic and ttH-leptonic categories are merged into VBF, VH and ttH respectively.
From 3, it can be seen that substantial ggH contamination is present in VBF and VH categories due to the high relative abundance of Higgs bosons created by gluon fusion and low separation power of the variables used in classification. On the other hand, the ttH production mechanism leaves a cleaner signature at LHC and is thus easier to identify, resulting in 78.2 % purity of the ttH category.

simulation / category	ggH	VBF	ttH	VH
Untagged	94.2%	3.4%	0.1%	2.3%
VBF	68.2%	29.2%	0.2%	2.4%
ttH	11.9%	2.9%	78.2%	7.0%
VH	63.8%	5.3%	3.0%	27.9%

Table 3: *Cut based confusion matrix with contribution percentages of most dominant production mechanims in each category*

Finally, the individual contributions of different Higgs boson production mechanisms in all seven categories and the expected yields (number of events in each category) are shown in figure 13:

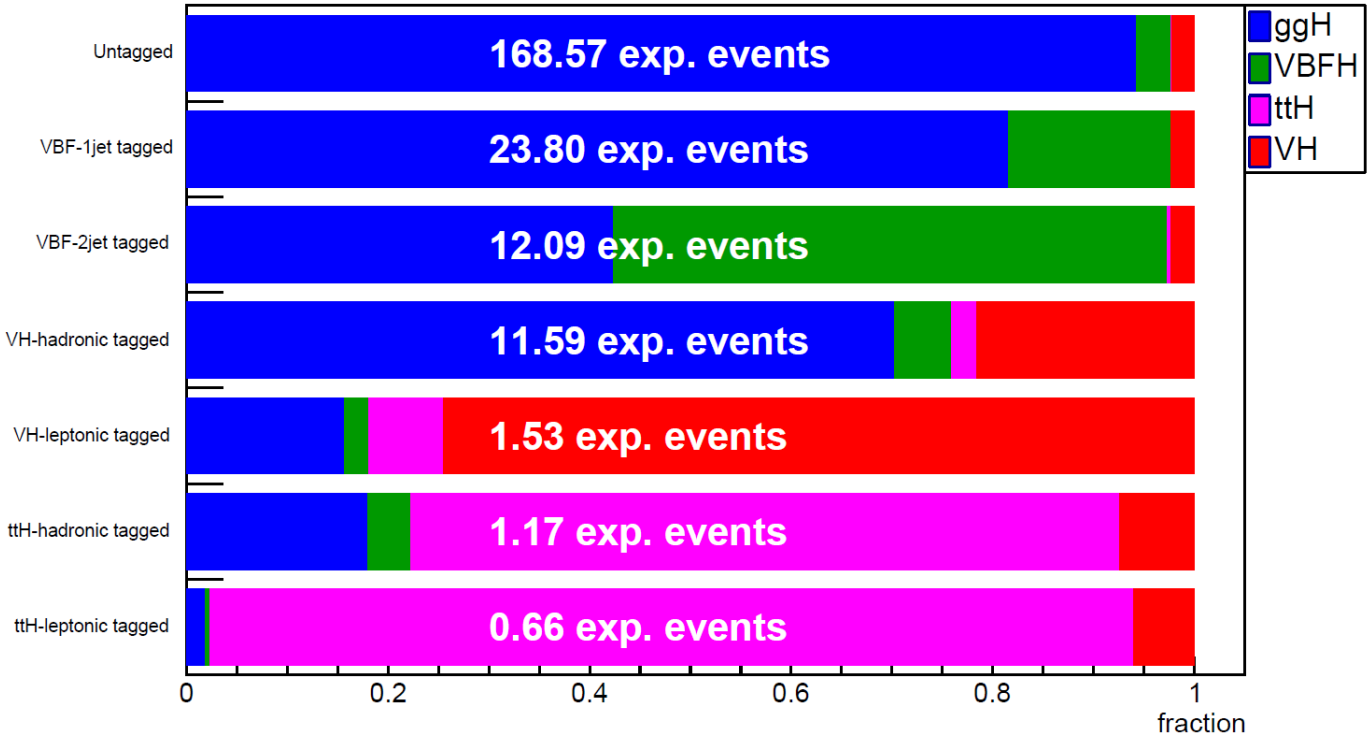

Figure 13: *Contribution of each production process for every category.*

5.2 Boosted decision tree categorization

Boosted decision tree categorization is performed using ROOT's Toolkit for Multivariate Data Analysis (TMVA), which provides a machine learning environment for the processing and evaluation of multivariate classification and regression techniques in high energy physics [15]. Data is categorised using 3 successive binary classifications following the logic described in Section 4.1.1, each containing 1000 trees with maximum depth equal to 3 with AdaBoost

parameter being 0.5. The number of grid points in variable range used in finding the optimal cut in node splitting is set to 20. These parameters are tuned to yield the best possible efficiency while constraining the training time to a reasonable value, i.e., if the change in the parameter results in neglibly better efficiency, while vastly increasing the computational time, it is discarded. Simulated files contain a large number of events to improve the statistical significance of Higgs measurements. To scale them to the number of observed events, they are given individual weights w_{event}. At the beginning of a training, initial event-by-event weights are set according to 5.6:

$$w = \frac{L \cdot \sigma \cdot w_{event}}{\sum w_{event}} \tag{5.6}$$

where $L = 137$ fb^{-1} and σ stand for luminosity and cross section for the selected process while $\sum w_{event}$ stands for the corresponding normalization. Luminosity gives a measure of how many collisions are happening in a particle accelerator, and a number of interactions N is calculated as the product of the luminosity integrated over time and the cross section:

$$N = \sigma \int L(t)dt \tag{5.7}$$

First boosted decision tree is trained on the simulated ROOT files of 4 leading Higgs boson production mechanisms (ggH, VBFH, VH, ttH). Signal class is defined as the one containing the ttH events while the rest of the events are classified as background. The resulting test statistic $BDT^{ttH}_{ggH+VBFH+VH}$ is used to discriminate the ttH production mechanism from the rest of the events. Having filtered out the fusion of top quarks, the second boosted decision tree is trained on ggH, VBFH and VH events, separating VH from VBFH and ggH, resulting in test statistic $BDT^{VH}_{ggH+VBFH}$. Finally, in the last decision tree, vector boson fusion and gluon fusion are being discriminated using the test statistic BDT^{VBFH}_{ggH}. Training outputs of the 3 boosted decision trees are then stored in datasets w1, w2 and w3. Together with cuts on the corresponding BDT scores, they are used to perform BDT categorization as shown in figure 14.

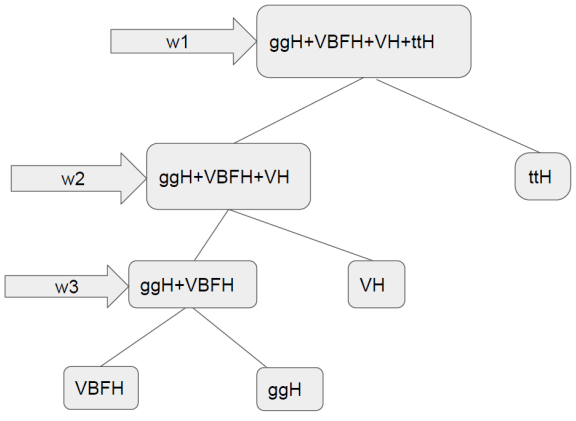

Figure 14: *Schematic view of successive BDT applications.*

5.2.1 Comparison with cut-based categorization

To gain an insight on how much efficiency is gained purely from using BDT, in opposition to cut-based approach, the same variables as in the cut-based analysis are used as the input of BDT. Distributions of boosted decision tree scores are shown in figures 15, 16 and 17. Cuts are taken in the way that both BDT and cut-based categorization yield approximately equal number of signal events in the ggH, VBFH, VH and ttH categories. Since the number of expected events takes discrete values, the choice is to make the yields in BDT slightly larger while keeping the difference as small as possible.

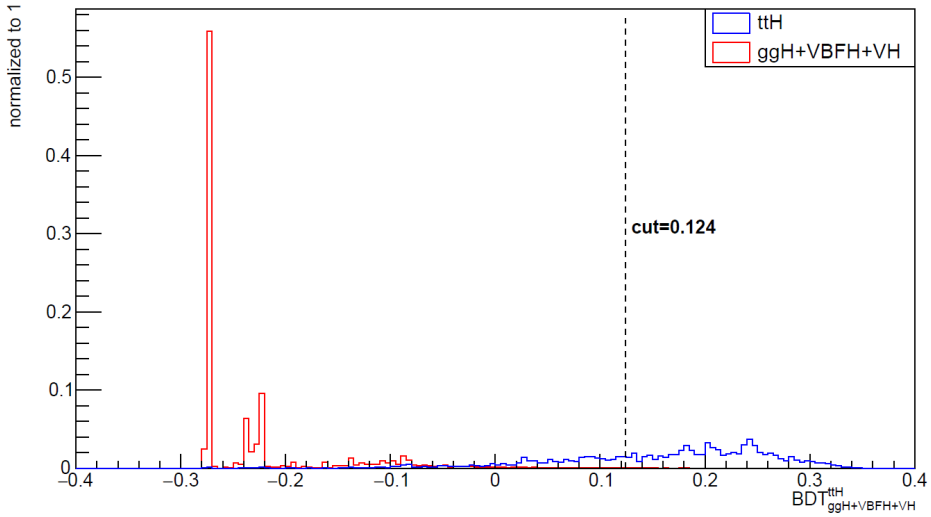

Figure 15: *Distribution of a test statistic used to discriminate ttH and ggH+VBFH+VH production mechanims.*

Events with $BDT^{ttH}_{ggH+VBFH+VH} > 0.124$ are classified into the ttH category, while the remaining events are further analysed.

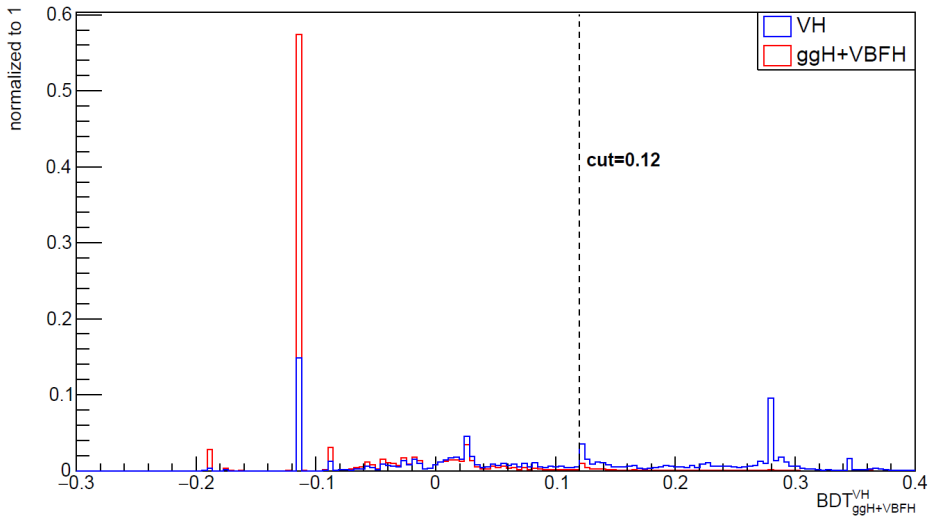

Figure 16: *Distribution of a test statistic used to discriminate VH and ggH+VBFH production mechanims.*

Events with $BDT^{VH}_{ggH+VBFH} > 0.12$ are classified into the VH category. Rest of the events proceed into the final step of the analysis.

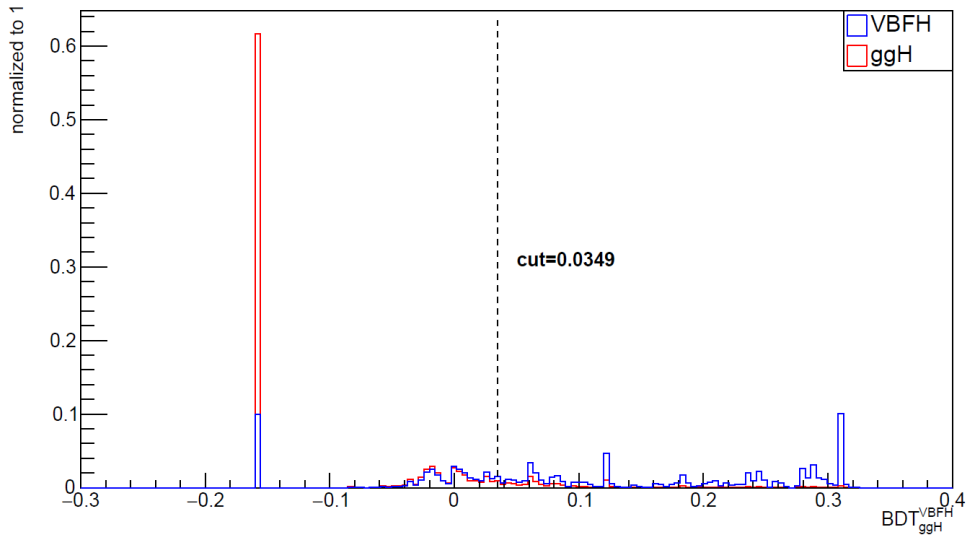

Figure 17: *Distribution of a test statistic used to discriminate VBFH and ggH production mechanims.*

Finally, events with $BDT^{VBFH}_{ggH} > 0.0349$ are classified into VBFH category. Rest of them are placed into Untagged category.

BDT categorization is now applied to Monte Carlo (MC) data and the performance is reported in BDT confusion matrix 4:

simulation / category	ggH	VBF	ttH	VH
Untagged	94.5%	3.3%	0.0%	2.2%
VBF	66.0%	31.1%	0.3%	2.6%
ttH	9.0%	1.6%	82.1%	7.3%
VH	60.4%	6.3%	3.8%	29.5%

Table 4: *BDT confusion matrix with contribution percentages of most dominant production mechanims in each category*

As expected, Untagged category mostly consists of ggH events. In comparison with the cut-based confusion matrix 3, an improvement in identification efficiency is achieved, and the purity increase (calculated by subtracting the diagonal elements of the BDT confusion matrix with diagonal elements of cut-based confusion matrix) in the categories VBF, ttH and VH is 1.9%, 3.9% and 1.6 %, respectively. It is not surprising that the BDT outperforms cut-based classification. To ensure compatibility with the current CMS analysis, which requires 7 categories, VBF, VH and ttH classes are further split into VBF-1jet, VBF-2jet, VH-hadronic, VH-leptonic, ttH-hadronic and ttH-leptonic categories based on the number of additional leptons and the number of jets. Seven exclusive BDT categories are now defined as follows:

- The **ttH-leptonic-tagged category** requires $BDT^{ttH}_{ggH+VBFH+VH} > 0.124$, at least one additional lepton and no more than 3 jets.
- The **ttH-hadronic-tagged category** requires $BDT^{ttH}_{ggH+VBFH+VH} > 0.124$, more than one additional lepton or at least 3 jets.
- The **VH-leptonic-tagged category** requires $BDT^{VH}_{ggH+VBFH} > 0.12$ and at least one additional lepton.
- The **VH-hadronic-tagged category** requires $BDT^{VH}_{ggH+VBFH} > 0.12$ and no more than 1 additional lepton.
- The **VBF-2jet-tagged category** requires $BDT^{VBFH}_{ggH} > 0.0349$ and more than 1 jet.
- The **VBF-1jet-tagged category** requires $BDT^{VBFH}_{ggH} > 0.0349$ and no more than 1 jet.
- The **Untagged category** consists of the remaining selected events.

Results of BDT classification are summarized in figure 18:

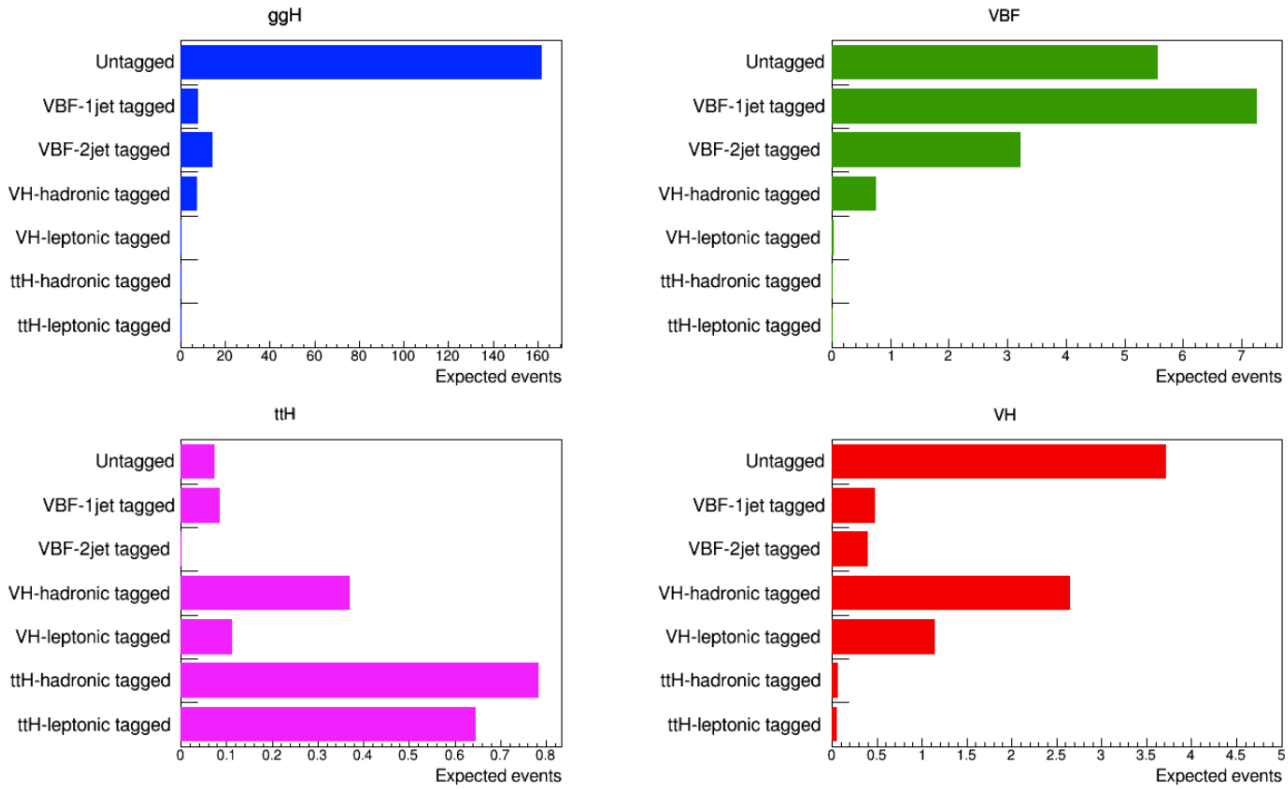

Figure 18: *Distributions of ggH, VBF, ttH and VH signal processes in the seven categories.*

Finally, the individual contributions of different Higgs boson production mechanisms in all seven categories and the expected number of events in each category are shown in figure 19:

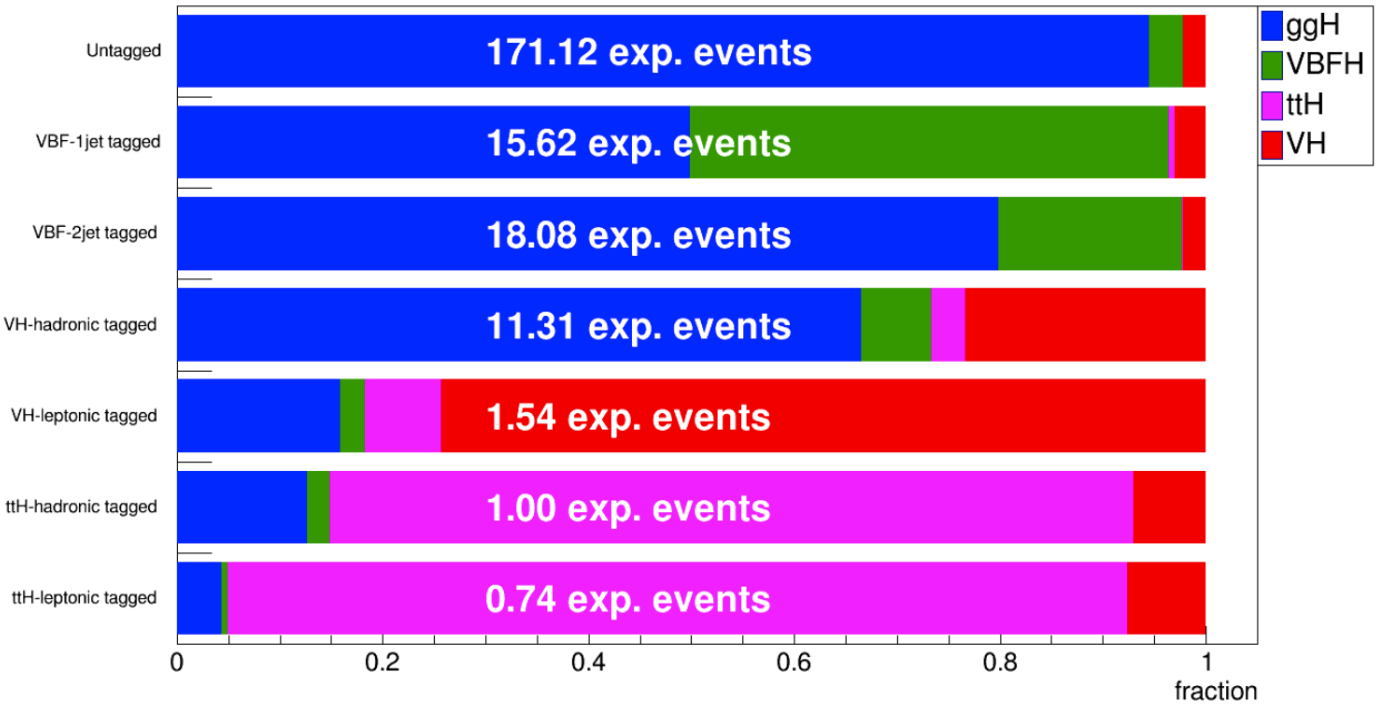

Figure 19: *Contribution of each production process for every category..*

By comparing figures 14 and 19, simillar behaviour is observed. For example, purity is highest in the tth-hadronic and tth-leptonic categories, while the VBF categories show significant contamination with gluon fusion for both types of classification. Additionally, BDT categorization exhibits a slightly better separation power between different classes.

5.2.2 Boosted decision tree with additional variables

The main strength of the boosted decision tree is that it can gain significant performance increase by combining many weak classifiers. This is exploited in this section and a modified BDT with many weak learners is constructed. To perform the categorization of Higgs boson production mechanisms, the following variables are used:

- Variables used in a cut-based analysis
- Individual probabilities from which the kinematic discriminators 5.2, 5.3, 5.4, 5.5 are derived
- Missing transverse momentum (missing momentum in a plane perpendicular to the beam axis)
- Transverse momentum of ZZ pair (combined transverse momentum of 2 Z bosons in a plane perpendicular to the beam axis)
- Dijet mass (mass of 2 jets)

- ZZjj transverse momentum (combined transverse momentum of 2 jets and ZZ pair)
- Dijet pseudorapidity (difference between pseudorapidity of 2 jets, where pseudorapidity is defined as $\eta = -ln\left[\tan\sqrt{\frac{\theta}{2}}\right]$, with θ being angle relative to the beam axis)
- ZZ pair pseudorapidity (difference between pseudorapidity of 2 Z bosons)
- ZZ pair azimuthal angle (difference between azimuthal angle of 2 Z bosons, where azimuthal angle is an angle in spherical coordinate system located in a plane perpendicular to the beam axis)
- Dijet Fisher constant (a constant calculated as a function of various jet variables)

Separating power of additional variables is depicted in figures 20 and 21:

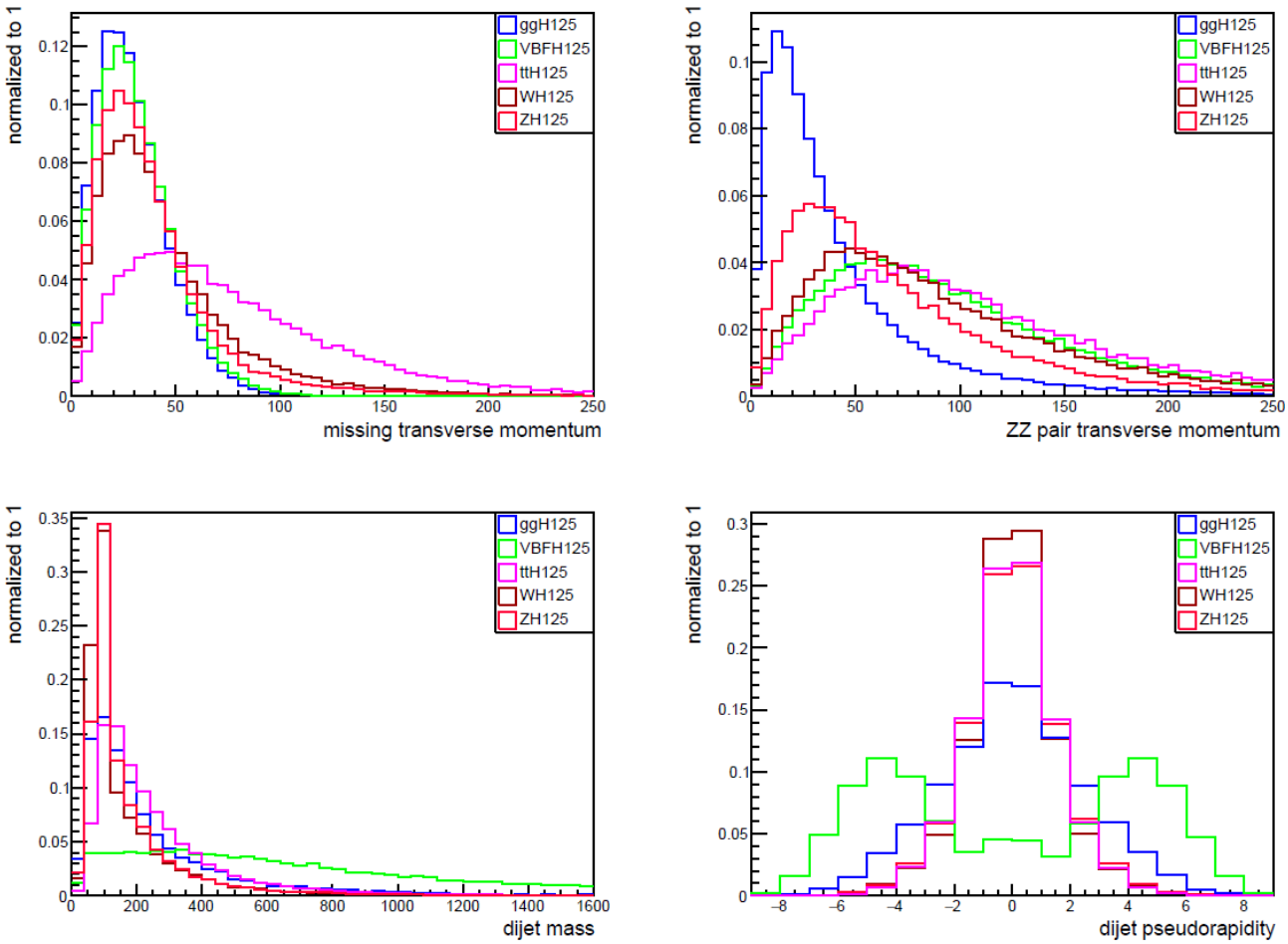

Figure 20: *Distribution of additional variables in* $118 < m_{4l} < 130$ *window. Distribution of missing transverse momentum (top left), distribution of ZZ pair transverse momentum (top right), distribution of dijet mass (bottom left), distribution of dijet pseudorapidity (bottom right).*

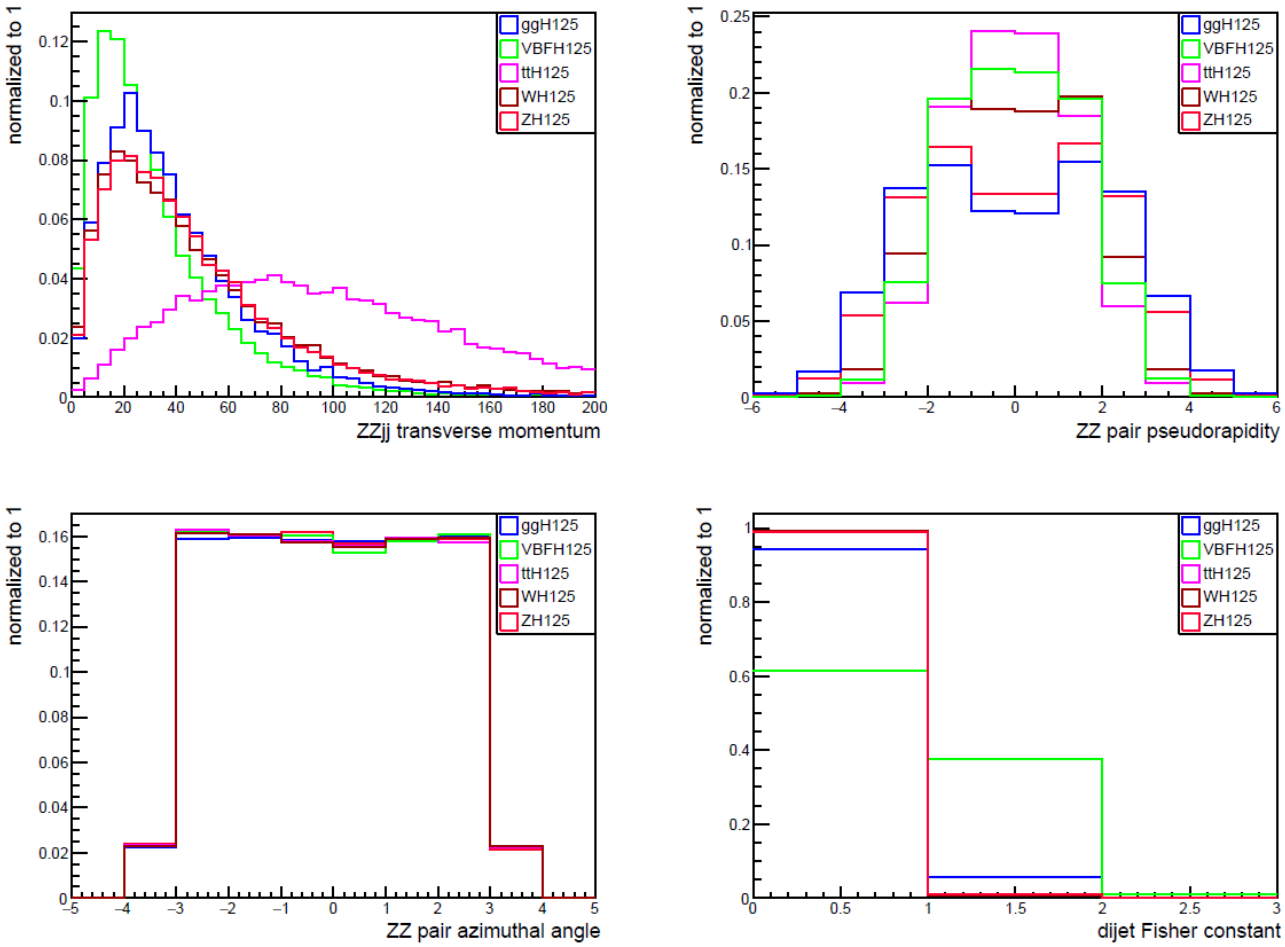

Figure 21: *Distribution of additional variables in* $118 < m_{4l} < 130$ *window. Distribution of ZZjj transverse momentum (top left), distribution of ZZ pair pseudorapidity (top right), distribution of ZZ pair azimuthal angle (bottom left), distribution of dijet Fisher constant (bottom right).*

Analysis is performed in the same way as in the case of the basic BDT. Events are filtered using successive layers of binary classifications, resulting in test statistics $bdt_{ttH}^{ggH+VBFH+VH}$, $bdt_{VH}^{ggH+VBFH}$ and bdt_{VBFH}^{ggH} as shown in figures 22, 23 and 24:

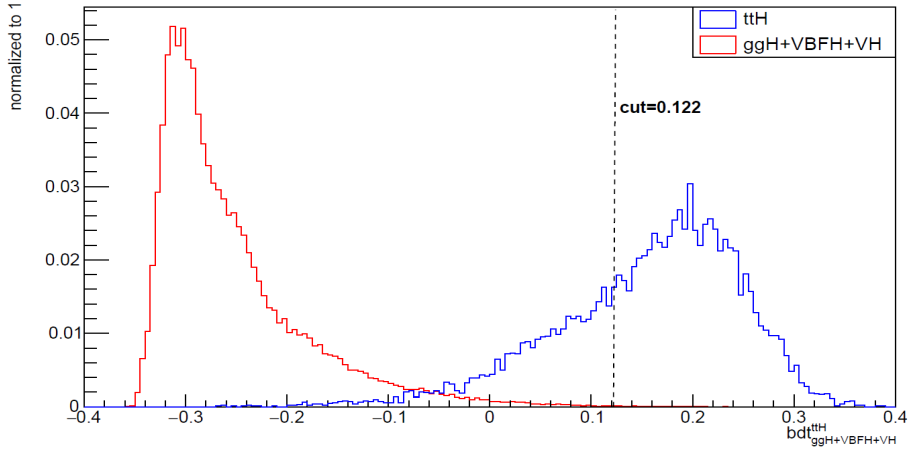

Figure 22: *Distribution of a test statistic used to discriminate ttH and ggH+VBFH+VH production mechanims.*

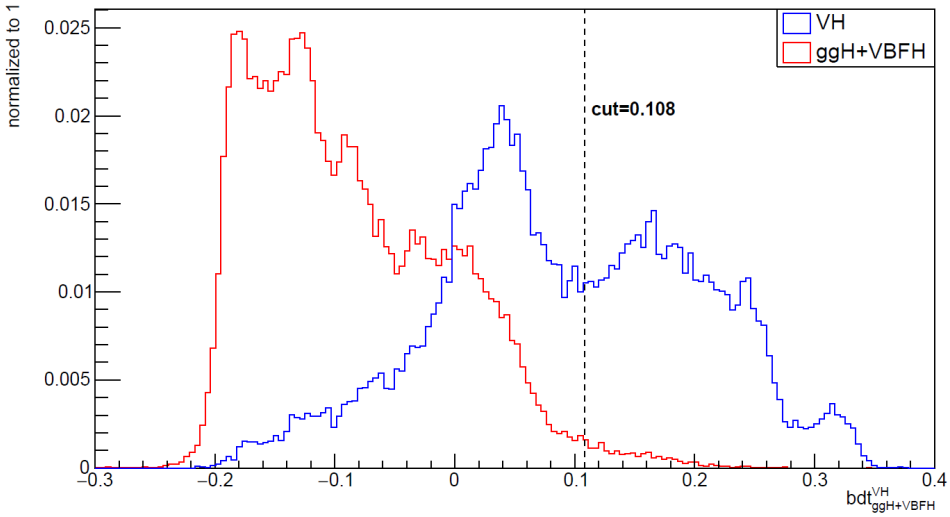

Figure 23: *Distribution of a test statistic used to discriminate VH and ggH+VBFH production mechanims.*

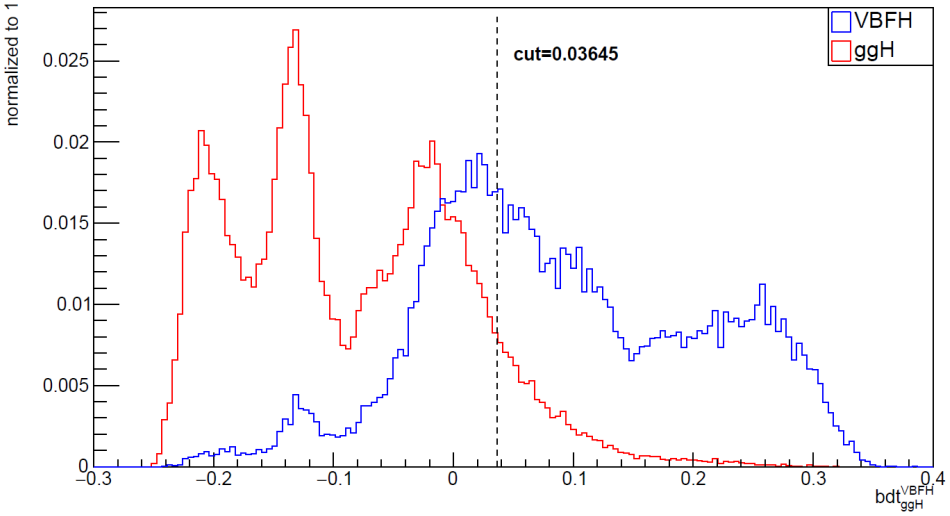

Figure 24: *Distribution of a test statistic used to discriminate VBFH and ggH production mechanims.*

Having computed the distributions of BDT scores, the cuts are tuned in a way that equalizes the signal yields in each category (ggH, VBFH, VH and ttH) to the yields in the cut-based analysis. With this choice, a comparison of the 2 methods can be made by constructing the modified BDT confusion matrix and comparing it to the one derived from the cut-based categorization.

simulation / category	ggH	VBF	ttH	VH
Untagged	94.4%	3.4%	0.0%	2.2%
VBF	61.6%	35.1%	0.3%	3.0%
ttH	4.9%	1.1%	88.7%	5.3%
VH	43.4%	4.7%	5.7%	46.3%

Table 5: *Modified BDT confusion matrix with contribution percentages of most dominant production mechanims in each category*

As expected, with the use of a modified boosted decision tree, the signal efficiency is increased, in opposition to both the basic BDT and cut-based approach. By observing 3 and 5, a modified BDT signal purity gain with respect to the cut-based purity is 5.9%, 10.5% and 18.4% in VBFH, ttH and VH categories. With the increase in dimensionality, a modified BDT also outperforms the basic BDT by 4.0%, 6.7%, 16.8% in VBFH, ttH and VH classes. The increase in performance can be seen by subtracting the diagonal elements of confusion matrices 4 and 5 or by comparing BDT distributions of the basic BDT score (figures 17, 15, 16) with the distributions of the modified BDT score (figures 24, 22, 23). Distributions of the modified BDT scores clearly exhibit better separation between classes, with the largest difference in the test statistic used to distinguish VH and ggH+VBFH production mechanisms. Purity gains from BDT (with and without additional variables) with respect to cut-based categorization are summarized in table 6:

category	basic BDT	modified BDT
VBFH	1.9%	5.9%
VH	1.6%	18.4%
ttH	3.9%	10.5%

Table 6: *Purity gains in VBFH, VH and ttH classes with repect to cut-based categorization.*

VBF, VH and ttH classes are further split into VBF-1jet, VBF-2jet, VH-hadronic, VH-leptonic, ttH-hadronic and ttH-leptonic categories based on the number of additional leptons and the number of jets. Seven exclusive modified BDT categories are now defined as follows:

- The **ttH-leptonic-tagged category** requires $bdt^{ttH}_{ggH+VBFH+VH} > 0.122$, at least one additional lepton and no more than 3 jets.
- The **ttH-hadronic-tagged category** requires $bdt^{ttH}_{ggH+VBFH+VH} > 0.122$, more than one additional lepton or at least 3 jets.
- The **VH-leptonic-tagged category** requires $bdt^{VH}_{ggH+VBFH} > 0.108$ and at least one additional lepton.
- The **VH-hadronic-tagged category** requires $bdt^{VH}_{ggH+VBFH} > 0.108$ and no more than 1 additional lepton.
- The **VBF-2jet-tagged category** requires $bdt^{VBFH}_{ggH} > 0.03645$ and more than 1 jet.
- The **VBF-1jet-tagged category** requires $bdt^{VBFH}_{ggH} > 0.03645$ and no more than 1 jet.
- The **Untagged category** consists of the remaining selected events.

Results of the modified BDT classification are summarized in figure 25:

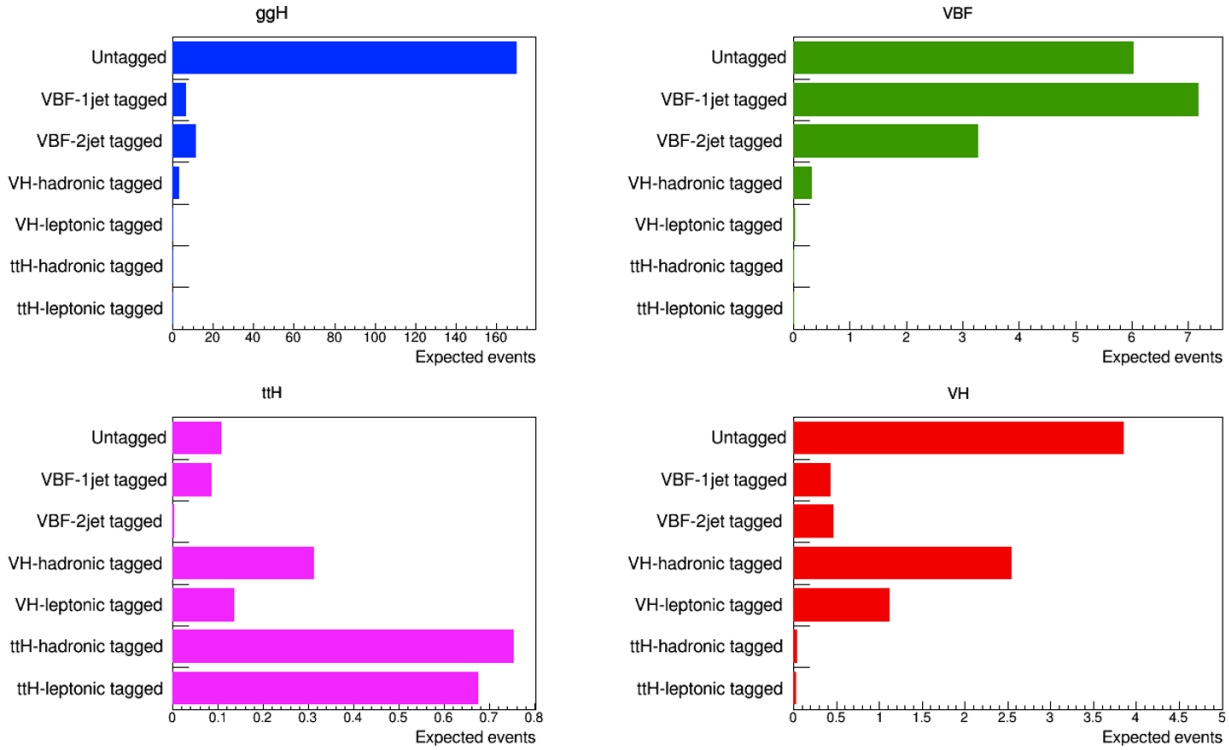

Figure 25: *Distributions of ggH, VBF, ttH and VH signal processes in the seven categories.*

Finally, the individual contributions of different Higgs boson production mechanisms in all seven categories and the expected number of events in each category are shown in figure 26:

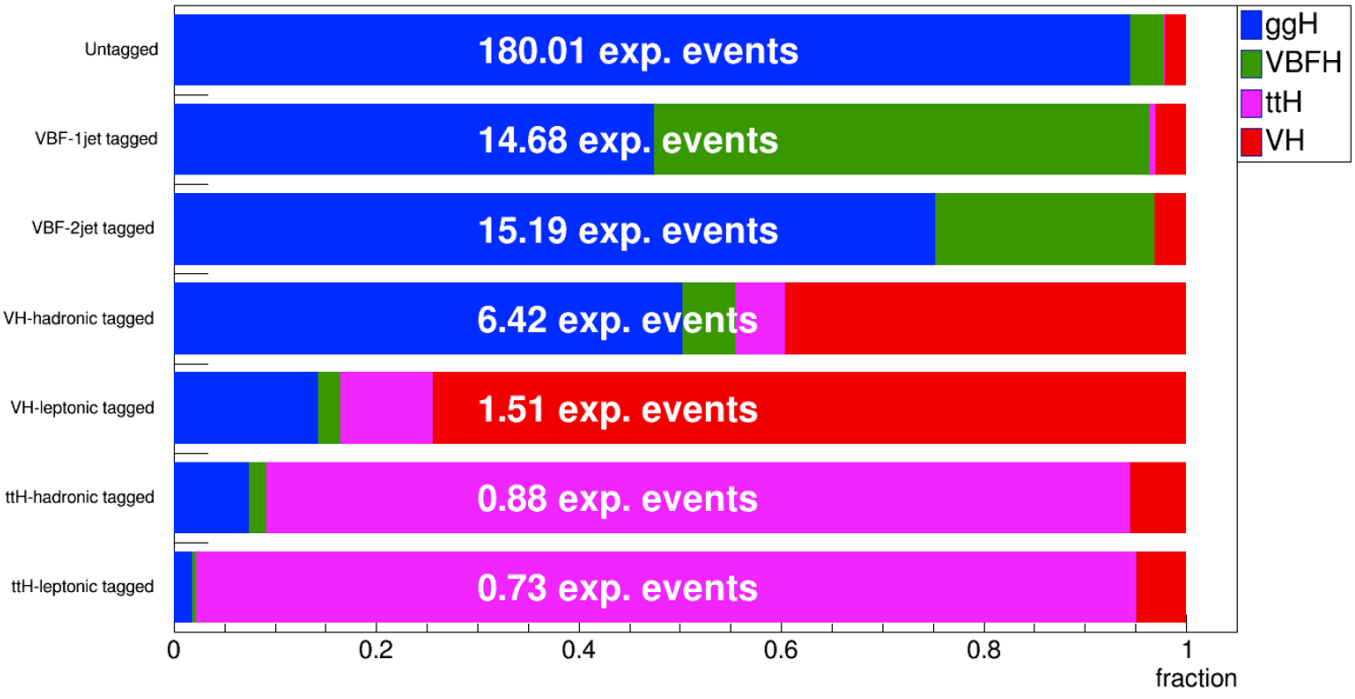

Figure 26: *Contribution of each production process for every category..*

By comparing figures 13, 19 and 26, simillar behaviour is observed for cut-based, basic BDT and modified BDT classification. Categories with the highest purity are ttH-leptonic and ttH-hadronic, while the significant ggH contamination is present in VBF categories. With the addition of new variables, the efficiency of BDT categorization is increased, with the biggest performance leap in VH categories.

5.3 Signal strength fits

Based on a modified BDT and cut-based categorization, the expected signal strength fits (without systematics) are performed on Asimov datasets. The training of a modified BDT is extended from 118-130 GeV m_{4l} to 105-140 GeV to ensure compatibility with current CMS analysis. The resulting test statistic is named $\overline{BDT}$. Furthermore, since the current CMS analysis uses even finer categories, called Simplified template cross sections (STXS), which are created by splitting of the classes defined in 5.1 based on jet kinematics, the full BDT categorization is not used. Instead, the hybrid classification (where ttH and VH are filtered out using BDT score, while VBF is seperated based on cuts) is deployed. The hybrid categories are now defined as follows:

- The **ttH-leptonic-tagged category** requires $\overline{BDT}^{ttH}_{ggH+VBFH+VH} > 0.122$, at least one additional lepton and no more than 3 jets.
- The **ttH-hadronic-tagged category** requires $\overline{BDT}^{ttH}_{ggH+VBFH+VH} > 0.122$, more than one additional lepton or at least 3 jets.
- The **VH-leptonic-tagged category** requires $\overline{BDT}^{VH}_{ggH+VBFH} > 0.08$ and at least one additional lepton.
- The **VH-hadronic-tagged category** requires $\overline{BDT}^{VH}_{ggH+VBFH} > 0.08$ and no more than 1 additional lepton.
- The **VBF-2jet-tagged category** requires exactly four leptons. In addition, there must be either two or three jets of which at most one is b-tagged, or four or more jets, none of which are b-tagged. Finally, $D^{ME}_{VBF-2j} > 0.5$ is required.
- The **VBF-1jet-tagged category** requires exactly four leptons, exactly one jet and $D^{ME}_{VBF-1j} > 0.5$.
- The **Untagged category** consists of the remaining selected events.

The procedure for extracting the values of signal strength modifiers is already developed within CMS framework and uses cut-based categorization. The corresponding code is available at [18, 19, 20]. In this thesis, it is used for signal strength fits by replacing the cut-based with the

hybrid categorization. In this procedure, the hybrid categories are split into STXS, which are then merged to form ggH, VBFH, WH, ZH and ttH categories. As the BDT was trained on simulation data from 2018, the analysis is performed for that year. Systematics are not included since the uncertainties corresponding to the scale factors, which would correct the differences between collision data and simulation, are currently not implemented on additional variables used by BDT. Expected signal strength fits with 68% confidence intervals for every process, as well as inclusive fits, are reported in table 7 for both the hybrid and cut-based categorization.

exp. value	cut-based	hybrid
Inclusive	$1.000^{+0.118}_{-0.115}$	$1.000^{+0.118}_{-0.114}$
μ_{ggH}	$1.000^{+0.150}_{-0.146}$	$1.000^{+0.147}_{-0.140}$
μ_{VBFH}	$1.000^{+0.818}_{-0.624}$	$1.000^{+0.815}_{-0.621}$
μ_{WH}	$1.000^{+2.675}_{-1.000}$	$1.000^{+2.670}_{-1.000}$
μ_{ZH}	$1.000^{+7.441}_{-1.000}$	$1.000^{+6.160}_{-1.000}$
μ_{ttH}	$1.000^{+2.012}_{-0.949}$	$1.000^{+2.092}_{-1.000}$

Table 7: *Expected values for signal strength modifier.*

From table 7, it can be seen that cut-based and hybrid categorization values for the signal strength modifier are very close with the deviation of inclusive fit slightly smaller in hybrid categorization.

6 Conclusion

In this thesis, 2 methods for categorization of Higgs boson production mechanisms were compared: cut-based, which is already a part of CMS analysis, and categorization using the boosted decision tree, which was developed as a part of this project. As expected, BDT outperformed cut-based classification, with the difference being larger when BDT was supplemented with additional variables. Using hybrid categorization, where BDT and cut-based analysis were combined, the signal strength fit to Asimov set was performed. The expected value determined using hybrid categorization yielded a slightly smaller uncertainty interval then cut-based. The result may be further improved by deploying the full BDT classification, which was not used since the current CMS analysis splits categories into STXS subcategories, which are better optimized for cut-based classification. With the introduction of BDT subcategories, which would replace STXS, finer measurements with higher precision may be performed. Furthermore, to include systematics in the modified BDT categorization, work should be done to provide uncertainties on scale factors for additional variables since they are currently not included in the simulation data.

7 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, 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, 2012

[3] Luca Lista, Statistical Methods for Data Analysis in Particle Physics, Springer, 2015

[4] Mark Thomson, Modern Particle Physics, Cambridge University Press, 2013

[5] Wikipedia, Standard Model, URL: https://en.wikipedia.org/wiki/Standard_Model

[6] T. Šculac, "Measurements of Higgs boson properties in the four-lepton channel in pp collisions at centre-of-mass energy of 13 TeV with the CMS detector", Palaiseau, France, 2018.

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

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

[9] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. Global conservation laws and massless particles. Phys. Rev. Lett., 13:585, 1964.

[10] Buschmann, Malte & Goncalves, Dorival & Kuttimalai, Silvan & Schönherr, Marek & Krauss, Frank &Plehn, Tilman, Mass Effects in the Higgs-Gluon Coupling: Boosted vs Off-Shell Production, Journal of High Energy Physics, 2014

[11] The Higgs boson, URL: http://opendata.atlas.cern/books/current/get-started/_book/the-higgs-boson.html

[12] CMS Experiment, URL: https://cms.cern/detector

[13] Ettore Focardi, Status of the CMS Detector, Physics Procedia, Volume 37, 2012.

[14] J. Neyman, E. Pearson, On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lond. Ser. A 231, 289–337 (1933)

[15] Höcker, A & Stelzer, Jana & Tegenfeldt, Fredrik & Voss, Helge & Voss, K & Christov, Asen & Henrot Versille, Sophie & Jachowski, M & Krasznahorkay, A & Mahalalel, Y & Prudent, Xavier & Speckmayer, P. TMVA User’s Guide (2010).

[16] CMS Collaboration, Measurements of production cross sections of the Higgs boson in the four-lepton final state in proton-proton collisions at $\sqrt{s} = 13$ TeV, 2021

[17] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statist., 9(1):60–62, 03 1938.

[18] GitHub, ZZAnalysis, URL: https://github.com/CJLST/ZZAnalysis

[19] GitLab, Datacards 13 TeV, URL: https://gitlab.cern.ch/HZZ4l/Datacards13TeV/-/tree/HIG-19-JES

[20] GitHub, Higgs Analysis-Combined Limit, URL: https://github.com/meng-xiao/HiggsAnalysis-CombinedLimit

A Links to the code and training datasets

The code for training of the boosted decision tree, BDT categorization and various plots is available in the master branch of my Master thesis repository. Training datasets for basic BDT, modified BDT and BDT on extended $105 < m_{4l} < 140$ range are also located in that repository and are available at basic BDT, modified BDT and extended BDT respectively. The base code for the procedure of extracting the values of the signal strength modifier is available at [18, 19, 20].