VILNIAUS UNIVERSITETAS
FIZIKOS FAKULTETAS
TEORINĖS FIZIKOS IR ASTRONOMIJOS INSTITUTAS

Monika Venčkauskaitė

HIGSO IR W BOZONŲ VIENALAIKIO SUSIDARYMO ĮVYKIŲ IŠSKYRIMAS CMS DETEKTORIUMI ESANT 13 TEV PROTONŲ SUSIDŪRIMO ENERGIJAI

Magistrantūros studijų baigiamasis darbas

Teorinės fizikos ir astrofizikos studijų programa

Studentas	Monika Venčkauskaitė
Darbo vadovas	dr. Andrius Juodagalvis
Konsultantas	dr. Adrian Perieanu (Hamburgo Univ.)
Recenzentas	dr. Thomas Gajdosik
Katedros vedėjas	prof. habil. dr. Leonas Valkūnas

Vilnius 2016

VILNIUS UNIVERSITY
FACULTY OF PHYSICS
INSTITUTE OF THEORETICAL PHYSICS AND ASTRONOMY

Monika Venčkauskaitė

IDENTIFICATION OF THE W-BOSON ASSOCIATED HIGGS BOSON PRODUCTION EVENTS WITH THE CMS DETECTOR AT 13 TEV PROTON COLLISION ENERGY

Master thesis

Theoretical Physics and Astrophysics studies

Student	Monika Venčkauskaitė
Supervisor	dr. Andrius Juodagalvis
Advisor	dr. Adrian Perieanu (Hamburg Univ.)
Reviewer	dr. Thomas Gajdosik
Head of department	prof. habil. dr. Leonas Valkūnas

Vilnius 2016

Table of Contents

Introduction

The Standard Model of particle physics was developed through 1960s and 1970s [1, 2]. It is a well tested theory that is able to explain a lot of phenomena in the particle colliders. The recent discovery of the Higgs boson-like particle in 2012 [3, 4] at the Large Hadron Collider [5], has given further credibility to this theory. Following this discovery, the Nobel Prize in Physics in 2013 was awarded to François Englert and Peter Higgs, who laid the theoretical foundation for the understanding the origin of the mass of elementary particles with the mechanism of electroweak symmetry breaking [6, 7].

Since the discovery of the Higgs boson it was shown that this particle interacts and decays in many ways predicted by the Standard Model. However, more studies are required to confirm that the discovered particle has the properties matching to those predicted by the Standard Model [8].

The study of the fermionic Higgs boson decays provides information about the Higgs boson coupling to the fermions. The Higgs boson decay channels $h \to \tau^+\tau^-$ and $h \to \mu^+\mu^-$ are a part of the wide Higgs boson research. In this analysis we look at the part of the $h \to \tau^+\tau^-$ channel where τ leptons decay to muons and neutrinos and the $h \to \mu^+\mu^-$ channel. This analysis focuses on W-associated Higgs boson production mechanism, which means that the Higgs boson is produced together with the W boson. The selected decay channel for the W boson is also into a muon. Therefore, in the final state there are three muons in both cases. The Wh decay channel to three muons in the final state seems to be unreported in public literature.

The purpose of this analysis was to find the discriminating variables between the signal process Wh and the background process WZ which has a similar final state particle signature and use them to train a multivariate analysis (MVA) method. The trained MVA method will be used to perform the Higgs search in the $h \to \tau^+\tau^-$ and $h \to \mu^+\mu^-$ channel data in the Large Hadron Collider Run II at 13 TeV.

The goal was achieved by the following **tasks**:

1. Using simulated signal and background events to find the discriminating variables for Wh and WZ processes.

2. Training a multivariate analysis method boosted decision trees (BDT) for separating Wh and WZ processes.

3. Testing trained BDT method with Monte Carlo data.

This thesis supersedes the preliminary study reported in [9]. It has a following structure: in the first section, the overview of the Large Hadron Collider and the Compact Muon Solenoid experiments is given. Then, in the second section the basic structure of the Standard Model and the Higgs boson particle are introduced. Third section encompass the explanation of methods that were used for this analysis. In the fourth section the training of the MVA and its results for both signal cases are presented. Last section concludes the results of this analysis.

1 Large Hadron Collider and CMS experiment

1.1 Large Hadron Collider

The Large Hadron Collider (LHC) [5] is the largest and most powerful particle accelerator in the world located at the particle laboratory operated by European Organization for Nuclear Research (derived from the french name *Conseil Européen pour la Recherche Nucléaire*[1], CERN). The LHC is a 27-kilometer ring of superconducting magnets with a number of accelerating structures. Protons or heavy ions are pre-accelerated in a series of linear accelerators and synchrotrons before being filled into the main LHC ring as shown in Fig. 1.

The protons in the LHC are obtained by ionizing hydrogen gas in an electric field. They are first accelerated in the Linac 2 accelerator, then injected into the Proton Synchrotron Booster (PSB), which accelerates them further. Then, at the Proton Synchrotron (PS), these protons are arranged into bunches. At the Super Proton Synchrotron (SPS) these bunches are accelerated to 450 GeV and injected into the two LHC beam pipes. In the LHC, the beams acquire their maximum energy.

Inside the accelerator, two high-energy particle beams travel close to the speed of light before they are brought into a collision. The beams travel in opposite directions in separate beam pipes. The bunch crossing time is 25 ns and there are around 1 billion proton-proton collisions per second (assuming 25 collisions per bunch crossing). Bunches of protons are made to collide at four locations around the accelerator ring, corresponding to the positions of main four particle detectors – ATLAS, CMS, ALICE and LHCb. Two of them, the ATLAS experiment [10] and the Compact Muon Solenoid (CMS) [11], are large, general purpose particle detectors. One of their purposes is the study of the Higgs boson. The analysis of this thesis is aimed at the CMS experiment.

1.2 Compact Muon Solenoid Experiment

Compact Muon Solenoid (CMS) contains subsystems which are designed to measure the energy and momentum of photons, electrons, muons, and other particles produced in the collisions. The innermost layer is a silicon-based tracker, surrounded by a scintillating crystal electromagnetic calorimeter and a sampling calorimeter for hadrons. The tracker and the calorimetry fits inside the Solenoid which generates a magnetic field of 3.8 T. Outside the magnet are the muon detectors positioned as shown in Fig. 2. The CMS experiment is described in detail in [11].

1.2.1 The interaction point and inner tracking system

The interaction point is the point where the proton-proton collisions occur between the two counter-rotating beams of the LHC. At each end of the detector, magnets focus the beams into the

[1] European Council for Nuclear Research

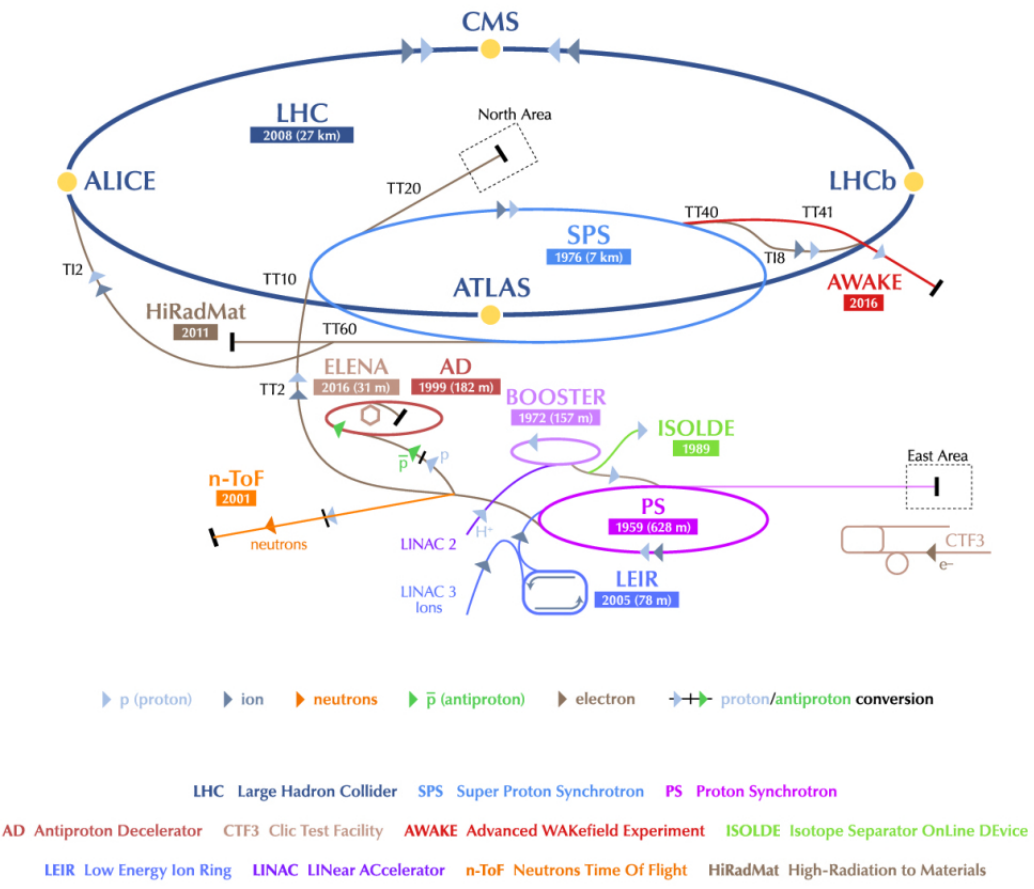

Fig. 1: The CERN accelerator complex with the main accelerator ring LHC and the pre-accelerators: Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS). At the SPS the W and Z bosons were discovered in 1983 [12].

interaction point. The collisions occur at a centre of mass energy of 13 TeV since May 2015. This corresponds to 6.5 TeV energy per beam.

The inner tracking system is in the innermost layer of the detector. It is made entirely out of silicon. The inner part of the tracking system is made out of silicon pixel modules and the outer part is made out of silicon microstrips. The CMS silicon tracker consists of 13 layers in the central region and 14 layers in the endcaps.

The particles traveling through the inner tracking system produce electric signals that are amplified and detected. Because the tracker is made out of multiple layers the tracks of charged particles can be reconstructed as illustrated in Fig. 3. The tracker can reconstruct the paths of high-energy muons, electrons and hadrons coming from the main interaction point or reconstruct the tracks coming from the secondary vertices of the decaying particles.

The particle momentum can be calculated and the charge of particle determined from the radius of it's track and knowing the value of the magnetic field.

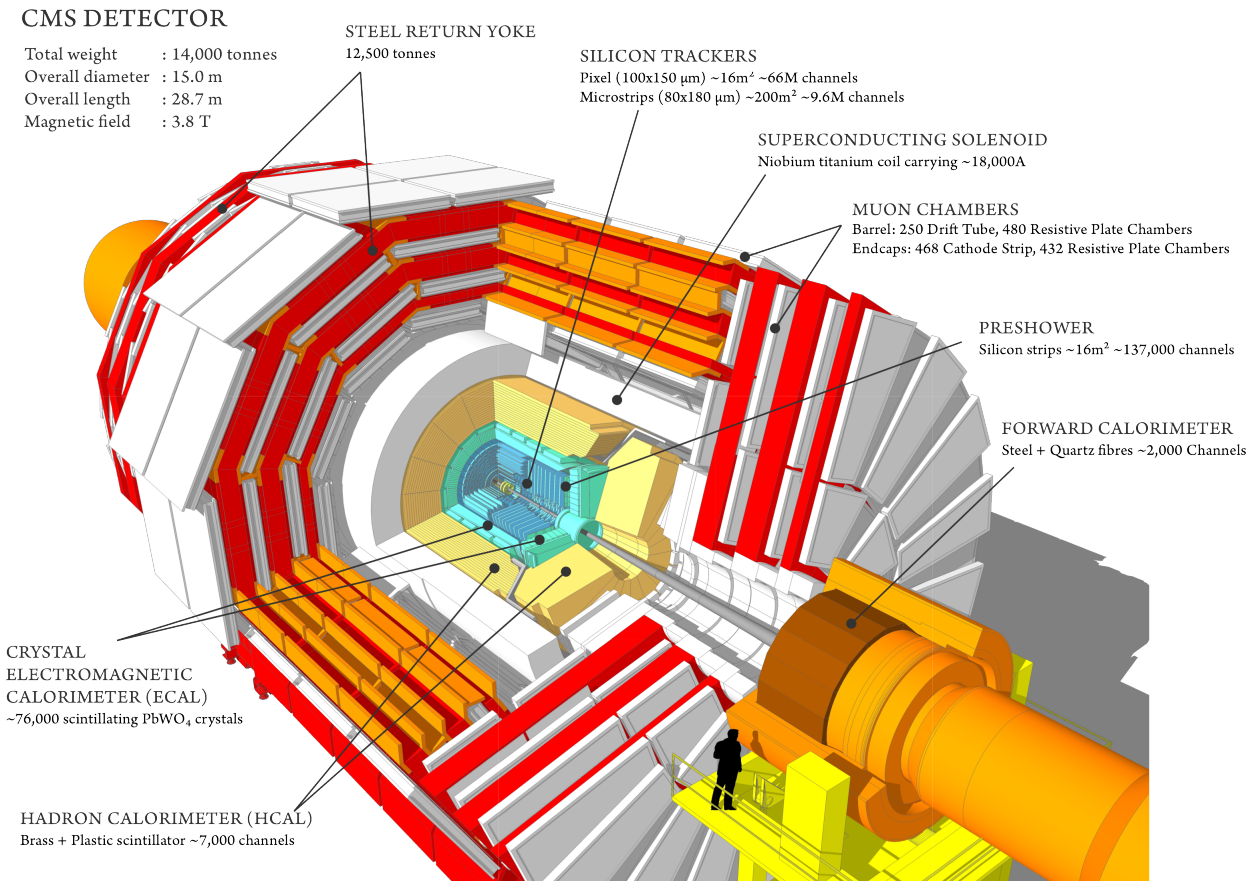

Fig. 2: The Compact Muon Solenoid experiment [13].

1.2.2 The Electromagnetic calorimeter

The purpose of the Electromagnetic Calorimeter (ECAL) is to measure the energies of electrons and photons. It is made out of lead tungstate ($PbWO_4$) crystals which are highly transparent. They scintillate when electrons and photons pass trough them. This results in an electromagnetic shower, caused by bremsstrahlung and pair production processes.

The bremsstrahlung process is a process where electromagnetic radiation is produced by the deceleration of a charged particle when deflected by another charged particle. In the ECAL it is an electron deflected by an atomic nucleus. The moving electron loses kinetic energy. The energy is conserved by converting this energy loss into a photon and emitting it.

Pair production in the ECAL takes place if the photon is near an atomic nucleus and has sufficient energy. The energy of a photon can be converted into an electron-positron pair.

The energy of the particle is deposited in the calorimeter material via Compton scattering and the photo-electric effect. The energy deposit is proportional to the energy of the original particle.

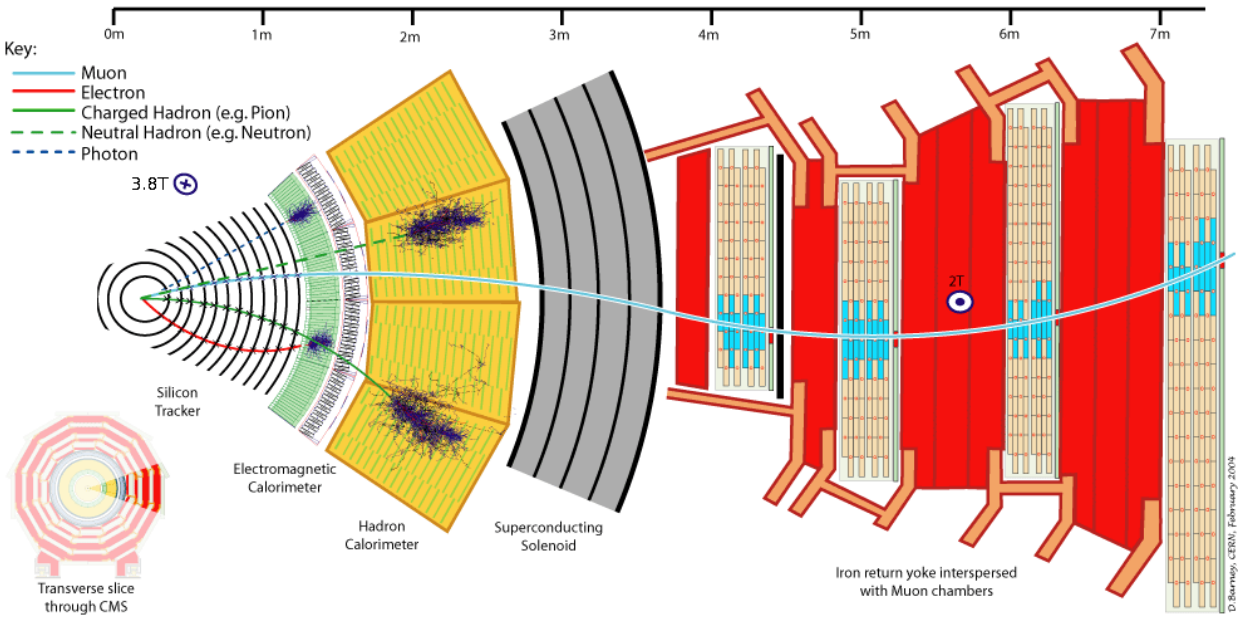

Fig. 3: The Compact Muon Solenoid view transverse to the beam [14].

1.2.3 The Hadronic calorimeter

The Hadronic Calorimeter (HCAL) measures the energy of the hadrons (which are particles made of quarks and gluons: protons, neutrons, pions, kaons).

The HCAL is made out of layers of steel and brass (which act as absorbers) and tiles of plastic scintillators in between them.

Hadronic particles interact with the absorber producing secondary particles. These particles form hadronic showers when they pass through the absorber layers. Showers cause the active scintillator layers to emit light. The wavelength of this light is shifted to the longer length by the wavelength shifting fibres. The light signals are detected by hybrid photodiodes and digitized.

The HCAL can measure neutral and charged hadrons and also the missing transverse energy which comes from undetectable particles (neutrinos or other possible exotic particles).

1.2.4 The Muon system

Detecting muons is one of the Compact Muon Solenoid's (CMS) most important tasks. Muons usually are not stopped by calorimeters. The muon chambers are placed at the outermost part of the experiment (outside the solenoid) where muons are the only particles likely to induce a signal.

To identify muons and measure their momenta, CMS uses three types of detectors: drift tubes (DT), cathode strip chambers (CSC) and resistive plate chambers (RPC) as one can see in Fig. 4.

The DTs are used for precise trajectory measurements in the central barrel region, while the CSCs are used in the endcaps. The RPCs provide a fast signal when a muon passes through the muon detector, and are installed in both the barrel and the end caps.

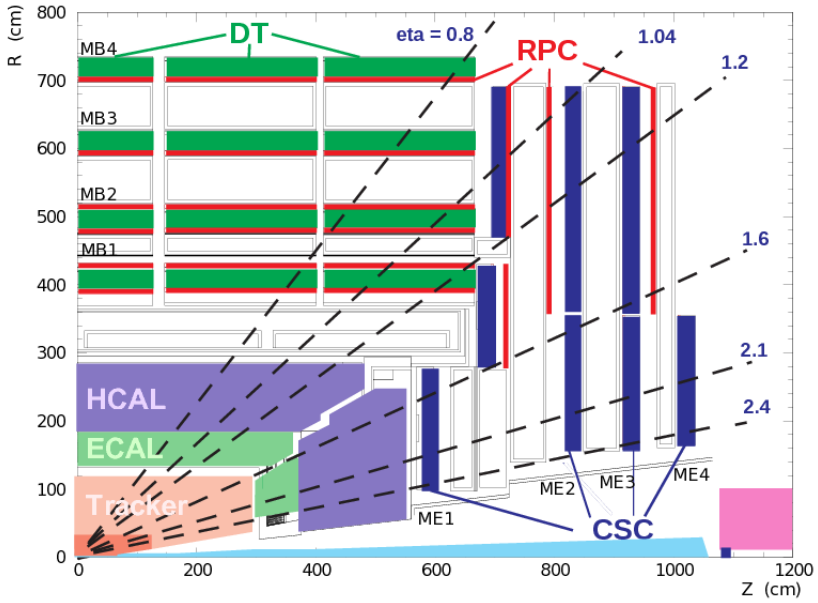

Fig. 4: The Muon System inside the Compact Muon Solenoid [15].

The drift tube (DT) system measures muon positions in the barrel part of the detector. Each tube contains a stretched wire within a gas volume. When a muon or any charged particle passes through the volume it ionizes the gas. The electrons from the ionized gas make an electric current which flows to the positively charged wire. By registering where electrons hit the wire and by calculating the muon's original distance away from the wire DTs give two coordinates for the muon's position: the coordinate along the beam direction and the perpendicular coordinate.

Cathode strip chambers (CSC) are used in the endcap disks where the magnetic field is uneven and particle rates are high. CSCs consist of arrays of positively charged anode wires perpendicular to negatively charged copper cathode strips within a gas volume. When muons pass through, they ionize the gas atoms and the electrons flow to the anode wires. Positive ions form a current from the wire and towards the copper cathode. This induces a flow to the strips, at certain angles to the wire direction. This way, the two position coordinates for each passing particle can be measured. This kind of measuring has a fast enough response for triggering. There are six layers in each module so muons can be identified accurately and their tracks can be matched to those in the tracker.

Resistive plate chambers (RPC) are fast gaseous detectors that provide a muon trigger system simultaneously with those of the DTs and CSCs. They allow for a quick measurement of the muon momentum. RPCs consist of two parallel plates, a positively charged anode and a negatively charged cathode, both made of a very high resistivity plastic material and separated by a gas volume. When a muon passes through the chamber, electrons are knocked out of gas atoms. These electrons in turn hit other atoms causing an avalanche of electrons. The electrodes are transparent to the signal (the electrons), which are instead picked up by external metallic strips after a small but precise time delay. The pattern of hit strips gives a quick measure of the muon momentum, which is then used by the trigger to make immediate decisions about whether the data are worth keeping.

2 Theory

2.1 Standard Model of Particle Physics

Throughout the history of physics scientists tried to determine the fundamental constituents of matter and to explain the forces that govern our universe. There are four conventionally accepted fundamental forces of nature: gravitational, electromagnetic, strong and weak. The Standard Model formulated in the 1970's [1,2] is able to explain the building blocks of matter and three of the fundamental interactions. While it is clear that the Standard Model does not explain all the interactions in the universe and does not answer all the physics questions, it is a fundamental and well-tested physics theory. It has successfully explained almost all experimental results and precisely predicted a wide variety of phenomena. All the particles can be divided into two groups according to their spin: bosons and fermions. Bosons have an integer spin, fermions have half-integer spin.

2.1.1 Matter particles

In the Standard Model there are two types of elementary fermions: quarks and leptons. Each group consists of six particles, which can be divided further into three generations.

As it can be seen in Table 1, the lightest and most stable particles – the up quark and the down quark form the first generation. In the second generation there are the less stable charm quark and strange quark. Finally, in the third generation there are top quark and bottom quark. Quarks also have one of the three possible colour charge values.

The six leptons also form three generations. The electron and electron neutrino belongs to the first generation, the muon and the muon neutrino form the second generation and the tau lepton and the tau neutrino make the third generation . The electron, the muon and the tau all have an electric charge. The neutrinos are electrically neutral and compared to other leptons have very small masses.

	Leptons		Quarks	
Generation	Flavour	Charge	Flavour	Charge
1st	e	-1	d	$-1/3$
	ν_e	0	u	$2/3$
2nd	μ	-1	s	$-1/3$
	ν_μ	0	c	$2/3$
3rd	τ	-1	b	$-1/3$
	ν_τ	0	t	$2/3$

Table 1. Fermions in the Standard Model.

Boson	Force
γ	Electromagnetic
g	Strong
Z	Weak
$W^{\pm}$	Weak

Table 2. The force-carrier particles in the Standard Model.

2.1.2 Force carriers

Gravity is the weakest force of all the fundamental forces and has infinite range. The other three of them: the electromagnetic, the weak and the strong forces are described in the Standard Model as resulting from the exchange of the force-carrier particles between matter particles. These force-carriers are bosonic and each particle corresponds to a different force.

The electromagnetic force has infinite range and is much stronger than gravity. This force is carried by photon.

The weak and strong forces dominate only at the level of elementary particles. The strong force is the strongest of all the fundamental interactions and is carried by gluon. The weak force is stronger than gravity, but weaker than electromagnetic force. It's carried by the Z and W bosons.

The Standard Model at the scale of individual particles still works well because the effect of the gravity at these scales is very weak. The effect of gravity for collider physics experiments is negligible. The force-carrier particles are summarized in the Table 2.

2.2 Higgs Boson

The Higgs particle is a massive scalar elementary particle theorized by Robert Brout, François Englert [6], Peter Higgs [7], Gerald Guralnik, Carl Richard Hagen and Tom Kibble [16] in 1964 and is a key building block in the Standard Model [1]. It has no intrinsic spin (spin value is zero). It doesn't have electric charge nor colour charge.

The Higgs field explains why all the elementary particles except photon and gluon are massive. The question of the Higgs field's existence has been the last unverified part of the Standard Model of particle physics.

On 4 July 2012, the discovery of the Higgs boson at the LHC with a mass of 125 GeV/c^2 was announced [3, 4]. Since then, the particle has been shown to behave, interact, and decay in many of the ways predicted by the Standard Model. In this thesis we examine one of the Higgs boson production channels which doesn't have the 5σ statistical significance yet.

2.2.1 Production of the Higgs at LHC

There are four main processes for the Standard Model Higgs boson production at the Large Hadron Collider: the gluon fusion, the vector boson fusion, the vector boson associated production

and the top pair associated production [17]. The Feynman diagrams of these processes are depicted in Fig. 5.

In the case of the LHC the Higgs production from the gluon fusion is the most probable process (it has the highest cross-section of all the Higgs production processes [18]). Cross section in particle physics is the probability that a given atomic nucleus or subatomic particle will exhibit a specific reaction in relation to a particular species of particle. Cross section is expressed in terms of area. In particle physics the used units are barns, $1\mathrm{b} = 10^{-28}\mathrm{m}^2$ and is approximately the cross-sectional area of a uranium nucleus. If the bombarding particle hits a circular area of this size perpendicular to its path and centered at the target nucleus or particle, the given reaction occurs and, if it misses the area, the reaction does not occur. When the hadrons (protons in the case of the LHC) collide it is actually their constituents that interact. When the two gluons that are binding the hadrons together collide they can form a loop of virtual quarks. The coupling of particles to the Higgs boson is proportional to their mass, so the most probable fermions to form this loop are either the virtual top quark or virtual bottom quark. From this loop the Higgs boson is produced. The gluon fusion process is the dominant process of Higgs boson production. However, since only the Higgs boson is produced in this channel it does not have any other distinct signature exept the decay products of the Higgs itself.

The next most probable Higgs production process (which has the second largest production cross section) after the gluon fusion is the vector boson fusion (VBF) process. In this process two quarks from the colliding protons radiate a massive vector boson each. The radiated two bosons fuse to produce the Higgs. The two outgoing quarks form hadrons and create two jets. These two jets, called tagging jets, are a distinguishing feature of VBF events.

Higgs bosons can be radiated from a W or Z boson. After that, there are either Higgs and Z bosons or Higgs and W bosons. Both of them decay and their decay products are detected. Even if the cross sections for the associative Higgs production are low compared to the total Higgs cross section, the associative Higgs production is important to test, because it has a clear signature in the final state. In this thesis we examine this production process.

The Higgs boson can also be produced from a top quark pair. The production cross section of the $t\bar{t}$ associated Higgs production is very small therefore the Higgs searches in this channel are really challenging.

2.2.2 Higgs boson decay modes

The Higgs boson interacts with all the massive elementary particles of the Standard Model, therefore there are many different processes through which it can decay [17]. Each of these possible processes has it's own probability, expressed as the branching ratio. The branching ratio of certain decay is the fraction of that decay in relation to the total number of decays. The Standard Model predicts these branching ratios as a function of the Higgs mass.

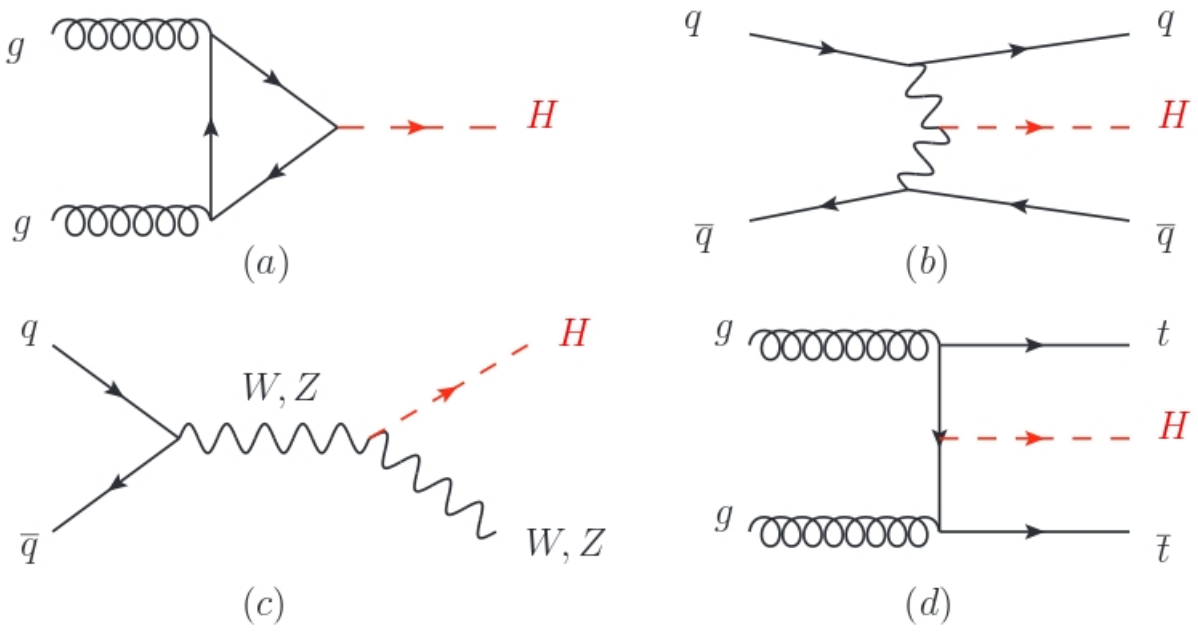

Fig. 5: Generic Feynman diagrams contributing to the Higgs production at the LHC: (a) gluon fusion, (b) vector boson fusion, (c) Higgs-strahlung (or associated production with a gauge boson) and (d) associated production with top quarks. In this analysis the Higgs-Strahlung process is examined. These processes are described in a greater detail in [17].

Higgs couplings to gauge bosons and fermions depend on their mass. Therefore, the branching ratio of the Higgs boson strongly depends on it's own mass relative to the masses of the decay products as it is listed in [19]. The branching ratios for Higgs can be calculated for different Higgs boson masses as shown in Fig. 6 for the main decay channels at the Large Hadron Collider.

For the Higgs mass of 125 GeV/c^2 the Standard Model predicts that the most common decay is into a bottom–antibottom quark pair as seen in Table 3, which happens in 57.7% of all the decays: $h \to b\bar{b}$.

Another high possibility is for the Higgs to split into a pair of massive gauge bosons. The biggest possibility among them is for the Higgs to decay into a pair of W bosons, which happens about 23.1% of all the decays. However, the decays in this channel are hard to distinguish due to the W boson ability to decay into quarks (which are hard to distinguish from the background) or undetectable neutrino presence in the W boson decays to leptons.

The next most common (second most common fermionic decay) at that mass is a tau–antitau pair, which happens in about 6% of all the decays: $h \to \tau^+\tau^-$. Other decays for this value of the Higgs mass are much less likely. $h \to \tau^+\tau^-$ and $h \to \mu^+\mu^-$ are the selected decay modes for the current analysis.

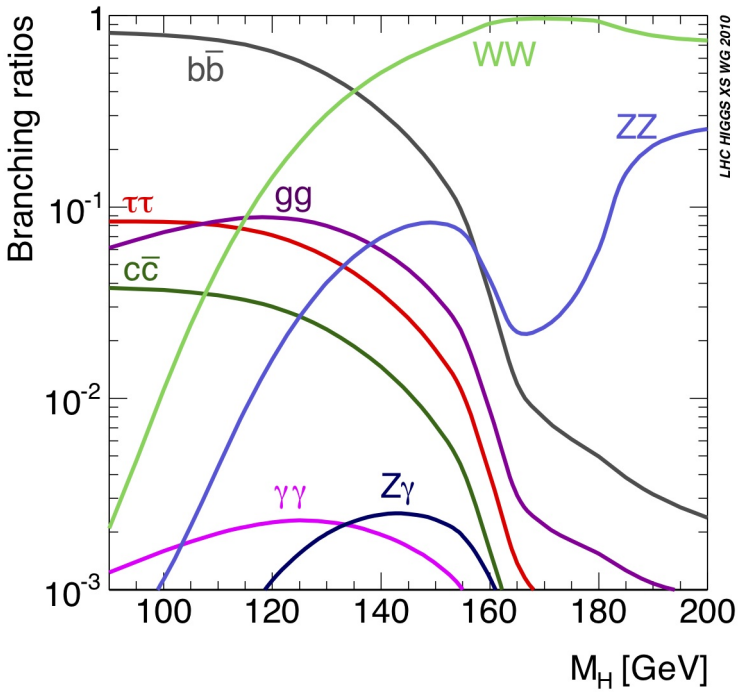

Fig. 6: The branching fraction dependence from mass for the Higgs boson as shown in [19].

Decay mode	Branching ratio
$b\bar{b}$	57.7%
W^+W^-	23.1%
gg	8.6%
$\tau^+\tau^-$	6.3%
$c\bar{c}$	2.7%
ZZ^*	2.6%
$\gamma\gamma$	$2.3\% \times 10^{-1}$
$\mu^+\mu^-$	$2.2\% \times 10^{-2}$
γZ	$1.5\% \times 10^{-1}$

Table 3. Branching ratios for Higgs boson decays as listed in [17].

3 Methods

3.1 Analysis Workflow

The analysis is carried in three main steps, following the usual CMS event simulation and reconstruction sequence [20]. Firstly, the events for different processes as the signal and the most important backgrounds are generated using Monte Carlo simulation software PYTHIA 8.2 [21,22]. The τ lepton decays are simulated by the package TAUOLA [23] interfaced to PYTHIA. Then the generated events undergo the digitization process with the software package Geant4 [24] and detector reconstruction with CMSSW software [20]. Finally, the analysis is performed with a private code. The selected muons in the analysis are called *tight* muons. That means that the requirements applied for the selected muons are [25]:

1. The candidate is reconstructed as a global-muon. That means that the muon has a track both in a tracker (tracker track) and in a muon system (standalone-muon track). For each standalone-muon track, a matching tracker track is found by comparing parameters of the two tracks propagated onto a common surface. A global-muon track is fitted combining hits from the tracker track and standalone-muon track.

2. $\chi^2/\mathrm{N_{d.o.f.}} < 10$ where $\mathrm{N_{d.o.f.}}$ is the number of degrees of freedom, χ is for the global-muon track fit.

3. At least one muon chamber hit included in the global-muon track fit.

4. Muon segments (i.e., a short track stub made of DT or CSC hits) in at least two muon stations.

5. Its tracker track has transverse impact parameter $|d_{xy}| < 2$ mm with respect to the primary vertex.

We also required the muon to be isolated so that it would not come from a jet.

3.1.1 Monte Carlo event generation with Pythia

The PYTHIA program described in [21,22] is a standard tool for the generation of high-energy collisions. It is able to comprise a set of physics models for the evolution from a few-body hard process to a complex multihadronic final state. It contains a library of hard processes (hard scattering processes at the LHC are those involving a momentum transfer that is large compared to the proton mass) and models for initial- and final-state parton showers, multiple parton-parton interactions, beam remnants, string fragmentation and particle decays. The PYTHIA 8.2 is a C++ software.

PYTHIA is widely used software for LHC physics studies. Due to the complicated nature of the collisions and decay processes, the analytic approach is not suitable for this analysis. For the generation of proton collisions it's important to specify which Parton Distribution Function

(PDF) [1,26], is being used. The PDF is the momentum distribution function of the partons within the proton. For this analysis we chose the commonly used CTEQ 5L [27] function.

The high-energy collisions between elementary particles normally give rise to complex final states, with large multiplicities of hadrons, leptons, photons and neutrinos. The relation between these final states and the underlying physics description is also complex. Therefore, complete events are generated with Monte Carlo methods. The complexity is mastered by a subdivision of the full problem into a set of simpler separate tasks. All main aspects of the events are simulated, such as hard-process selection, initial- and final-state radiation, beam remnants, fragmentation, decays, and so on, the details are explained in [22]. Therefore events generated by PYTHIA should be directly comparable with experimentally observable ones. The programs can be used to extract physics from comparisons with existing data or to study physics at future experiments.

3.1.2 Matching between the generated and reconstructed event levels

The analysis was performed having two levels of events. First level was the generated level. Events at this level were obtained using MC generator (PYTHIA 8.2). The second level in the analysis was the reconstructed level. The generated events were reconstructed with CMS software (CMSSW). The same software is used to reconstruct the data detected with the CMS detector.

Then each muon from the event at the generated was matched to the his counterpart at the reconstructed level for each event. In order to do this matching, the angular separation ΔR of the muon counterparts was calculated:

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

where $\eta = -\ln(\tan\theta/2)$, θ is a particle's polar angle with respect to the beam, and ϕ is it's azimuthal angle.

A generated-reconstructed muon pair is considered to be matched if $\Delta R < 0.3$.

The pair with the minimum value of the angular separation was identified as the matching pair. In the Fig. 7 the angular separation distribution for the generated and reconstructed muon pairs is plotted. The angular separation in most cases is less than 0.3 for the muons from either W, Higgs or Z boson.

In the figures 8, 9 and 10 there are the muon momentum distribution at both levels (generated and reconstructed) for each boson (Higgs, W and Z). We can see that generated muons are reconstructed with correct momentum. Moreover, it is evident that muons from Higgs and Z bosons tend to have lower momentum than muons from W boson. Furthermore, transverse momentum of the muon from W boson decay is affected by the final state radiation more than muons from tau lepton decays (final state particles loose energy radiating photons when they are decelerated) so its reconstructed transverse momentum is more likely to be smaller than generated transverse momentum.

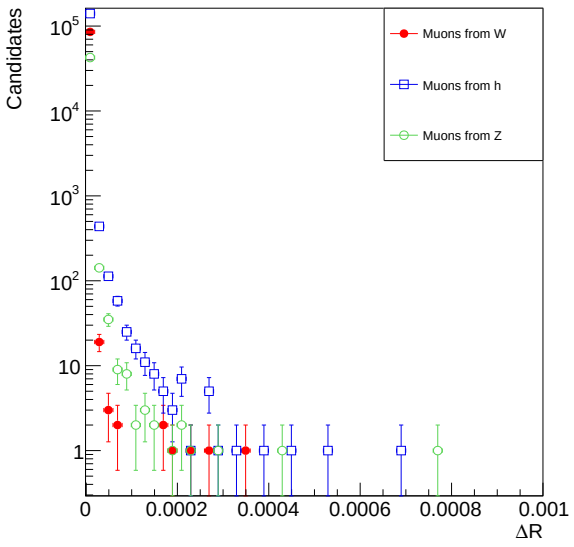

Fig. 7: Angular separation distribution for generated and reconstructed muon pairs.

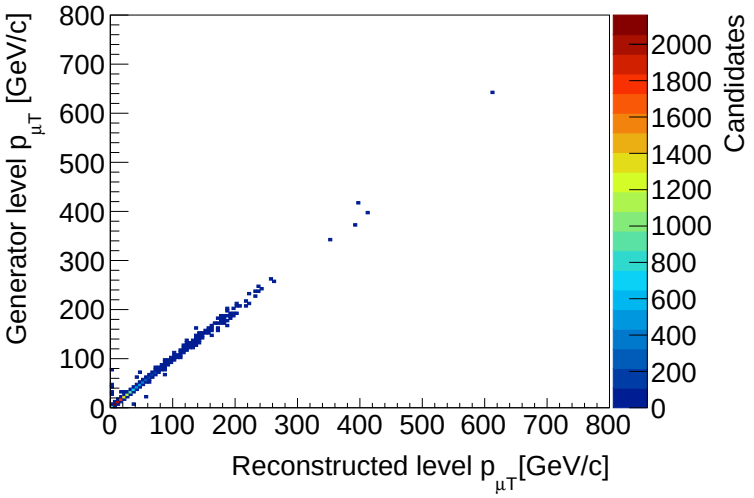

Fig. 8: The reconstruction of muon p_T from $h \to \tau\tau$ with $\tau \to \nu\mu\nu$ decay.

All the three muons from the analyzed signal process of $W \to \mu\nu$ and $h \to \tau\tau$ with $\tau \to \nu\mu\nu$ will have different kinetic energies corresponding to their origin particle. Therefore, we can expect the muons from W boson more likely to have higher energy than muons from h or Z bosons. This can be used to identify muons at the reconstructed level. This energy distribution will be investigated in the following paragraph.

3.1.3 Identifying muon origin at reconstructed level

One of the important quantities in particle physics is transverse momentum. The transverse momentum is the momentum of a particle in perpendicular direction to the beam or: $p_T = \sqrt{p_x^2 + p_y^2}$ where z axis stands for the direction of the beam, x and y - directions perpendicular to the beam.

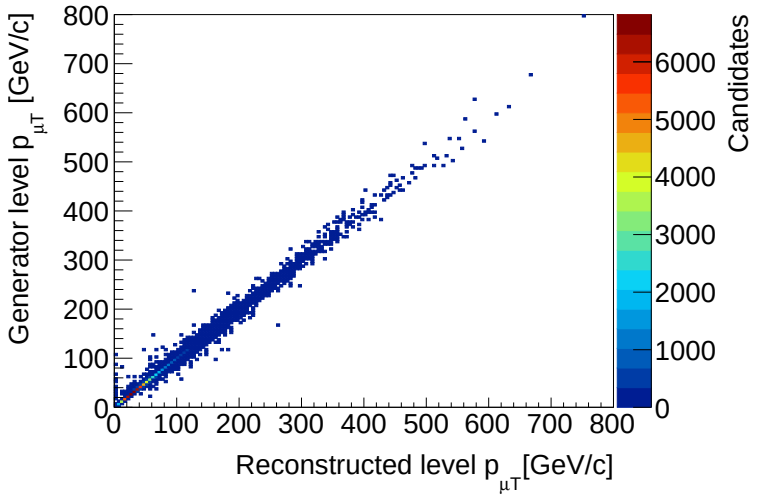

Fig. 9: The reconstruction of muon p_T from $W \to \mu\nu$ decay.

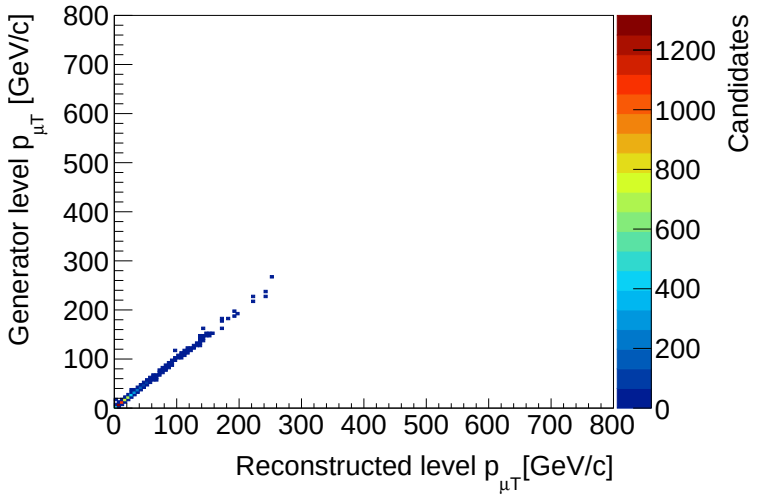

Fig. 10: The reconstruction of muon p_T from $Z \to \tau\tau$ with $\tau \to \nu\mu\nu$ decay.

From the W boson decay we have only one muon, whereas from the Higgs decay we have two. The highest transverse momentum p_T muon in this process has the highest probability to be coming from W boson and lowest p_T muon most often comes from Higgs boson. This is illustrated in Fig. 11 muon p_T distribution by muon origin (*left*) and relative p_T spread of muons by their origin (*right*) .

At the reconstructed level we identify the lowest p_T muon and next to lowest p_T muon with the opposite charge to be from Higgs boson (or Z boson) and the last muon to be from W boson. This classification is true for:

- 80% of cases for the Wh, h $\to \tau\tau$;
- 80% of cases for the WZ, Z $\to \tau\tau$;
- 66% of cases for the Wh, h $\to \mu\mu$;

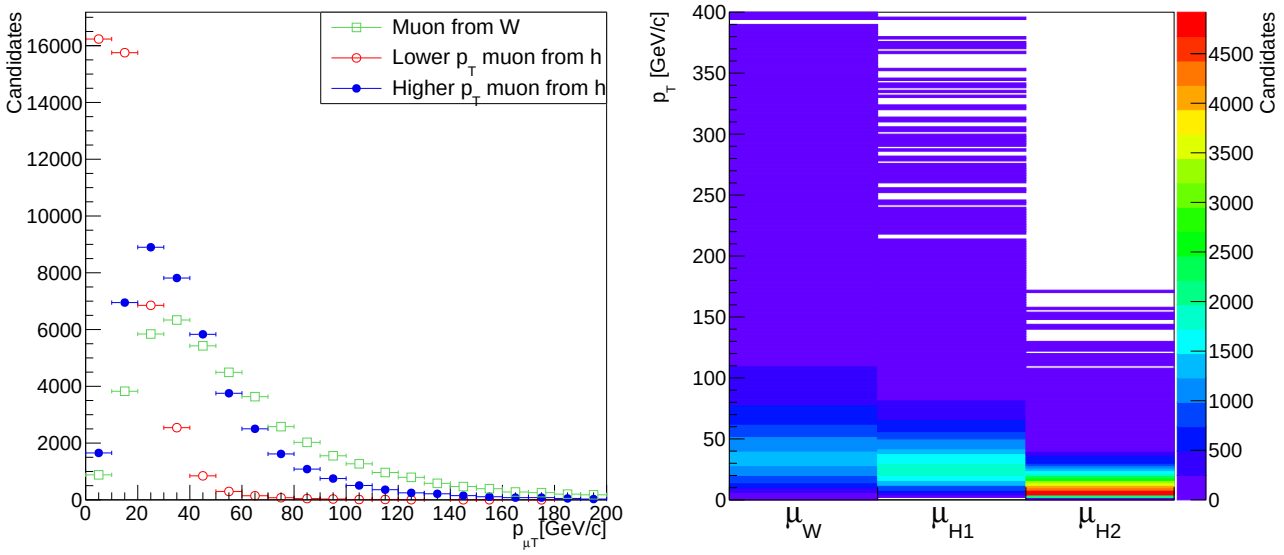

Fig. 11: Muon p_T distribution in a Wh process for muons originating from the W boson and the Higgs boson (*left*) and a relative p_T spread visualisation (*right*).

- 61% of cases for the WZ, $\mathrm{Z} \to \mu\mu$.

3.2 Signal and Background Processes

The selected decay channels for this analysis are:

1. $h \to \tau\tau, \tau \to \mu\bar{\nu}_\mu\nu_\tau, W \to \mu\bar{\nu}_\mu$;
2. $h \to \mu\mu, W \to \mu\bar{\nu}_\mu$.

They both have three muons in the final state. So any process, that is possible in the proton collisions and also has three muons in the final state is a background process. The main background processes for the selected channel are the following:

1. WZ production;
2. ZZ production;
3. WW production;
4. Drell-Yan bottom quark pair $\mathrm{b\bar{b}}$ production;
5. Top quark pair $t\bar{t}$ production;

Process	σ	Estimated N_{events}, $\mathcal{L} = 100$ fb^{-1}
Wh with W$\rightarrow \mu\bar{\nu}_\mu$ and h$\rightarrow \tau\tau$, $\tau \rightarrow \mu\bar{\nu}_\mu\nu_\tau$	0.2796 fb	28
WZ with W$\rightarrow \mu\bar{\nu}_\mu$ and Z$\rightarrow \tau\tau$,$\tau \rightarrow \mu\bar{\nu}_\mu\nu_\tau$	3.9958 fb	400
Wh with W$\rightarrow \mu\bar{\nu}_\mu$ and h$\rightarrow \mu\mu$	0.0320 fb	3
WZ with W$\rightarrow \mu\bar{\nu}_\mu$ and Z$\rightarrow \mu\mu$	131.6725 fb	13167

Table 4. Calculated number of events for signal processes and irreducible background processes.

The number of signal events can be estimated from the production cross section σ and the decay branching fractions ([17], [18], [28], [29]) for the selected channel by using the following equation:

$$N_{\text{signal}} = \mathcal{L}_{\text{int}}\sigma_{Wh}\Gamma_{h\rightarrow\tau^+\tau^-}\Gamma^2_{\tau\rightarrow\mu\bar{\nu}_\mu\nu_\tau}\Gamma_{W\rightarrow\mu\bar{\nu}_\mu},$$

where $\mathcal{L}_{\text{int}}$ is the integrated luminosity. In scattering theory and accelerator physics, luminosity $\mathcal{L}$ is the ratio of the number of events detected N in a certain time t to the interaction cross-section σ:

$$L = \frac{1}{\sigma}\frac{dN}{dt}$$

and integrated luminosity L_{int} is the integral of the luminosity with respect to time. $\sigma_{Wh} = 1.373$pb is the production cross section of Wh process at 13 TeV, $\Gamma_{h\rightarrow\tau^+\tau^-} = 6.32 \times 10^{-2}$ is the branching fraction for the $h \rightarrow \tau^+\tau^-$ decay, $\Gamma_{\tau\rightarrow\mu\bar{\nu}_\mu\nu_\tau} = 1.741 \times 10^{-1}$ for the $\tau \rightarrow \mu\bar{\nu}_\mu\nu_\tau$ process, and $\Gamma_{W\rightarrow\mu\bar{\nu}_\mu} = 1.063 \times 10^{-1}$ for the $W \rightarrow \mu\bar{\nu}_\mu$ process accordingly. By multiplying the production cross-sections with the corresponding signal branching fractions, we get the signal cross-section $\sigma_{Wh\rightarrow\mu\tau\tau\rightarrow 3\mu} = 0.0002796$ pb. In the same way, $\sigma_{WZ} = 36.8$ pb is the production cross section of WZ process at 13 TeV, $\Gamma_{Z\rightarrow\tau^+\tau^-} = 3.37 \times 10^{-2}$ and the irreducible background cross-section $\sigma_{WZ\rightarrow\mu\tau\tau\rightarrow 3\mu} = 0.0039958$ pb.

The numbers of the signal and background process events for luminosity of $\mathcal{L} = 100$ fb^{-1} are shown in the Table 4. The branching fractions $\Gamma_{h\rightarrow\mu^+\mu^-} = 2.2 \times 10^{-4}$ and $\Gamma_{Z\rightarrow\mu^+\mu^-} = 3.366 \times 10^{-2}$ are used attaining the numbers in the table.The number of signal events relatively to the background is quite small. However, most of the backgrounds have a signature that is different from the signal process so they are reducible. The most difficult to separate from the signal is the part of WZ background in which the Z boson decays to tau-antitau pair which decays further to muons. The signature of this background process is almost identical to the signal process. However, the mass of the Higgs boson (125 GeV/c^2) is significantly higher than the mass of the Z boson (91 GeV/c^2) therefore there are differences in the dynamics of particles produced in these processes.

Since this background of $W \rightarrow \mu\bar{\nu}_\mu$ and $Z \rightarrow \tau\tau$ with $\tau \rightarrow \mu\bar{\nu}_\mu\nu_\tau$ is the most difficult to identify from the signal process $W \rightarrow \mu\bar{\nu}_\mu$ and $h \rightarrow \tau\tau$ with $\tau \rightarrow \mu\bar{\nu}_\mu\nu_\tau$, it is important to find the possible differences between them that would allow us to separate these processes as it was done in [9] and the list of these differences was expanded in this thesis.

3.3 Multivariate Data Analysis

3.3.1 ROOT system

ROOT system is a data analysis framework [30] dedicated for large scale data analysis. It is written in C++ and contains an efficient hierarchical Object Oriented database, a C++ interpreter and advanced statistical analysis tools.

The user can interact with ROOT in three ways: using graphical interface, command line or with a batch scripts. In this analysis the batch script was written, compiled and dynamically linked in ROOT.

The structure of ROOT is a layered class hierarchy. Most of the classes inherit from a common base class TObject. The data is defined as a set of objects called Trees and the specialized storage methods are used to get direct access to certain attributes of the selected objects without the need to read the entire data. The Tree architecture extends the concept of the Ntuple (tabular where each event consists of a fixed length row of data) to all complex objects or data structures found in Raw Data or Event Summary Data.

In ROOT there are histograming methods in an arbitrary number of dimensions, curve fitting, function evaluation, minimization, graphics and visualization classes that are convenient for the setup of data analysis system. Due to the built-in C++ interpreter cling, the command, the scripting and the programming language are all C++. ROOT is an open system that can be dynamically extended by linking external libraries.

3.3.2 ROOT TMVA

In many high-energy physics searches with small signals in large data sets multivariate analysis became an important tool. Multivariate classification methods are based on machine learning techniques. TMVA (Multivariate analysis toolkit) [31] is integrated in to the analysis framework ROOT and it has a large variety of multivariate classification algorithms. All of these algorithms belong to the group of *supervised learning* algorithms. They use the training events for which the output is known and determine the mapping function that describes a decision boundary (for classification of events). TMVA enables the training of the classification algorithm, testing, performance evaluation and application procedures.

The use of TMVA is shown in a diagram in Fig. 12. The training and testing are performed using the ROOT trees or text files provided by the user. These input files are produced with the user training script and contain various discriminating variables for signal and background. The user selects which variables to use for the MVA training and adds them to the TMVA factory. Each event in the data can have an individual weight. After that, the selected MVA methods are booked, trained, tested on independent sample and evaluated. The testing and training samples are both obtained from the input data by splitting all the events into the training and test ROOT trees so the evaluation of the MVA algorithms is statistically independent.

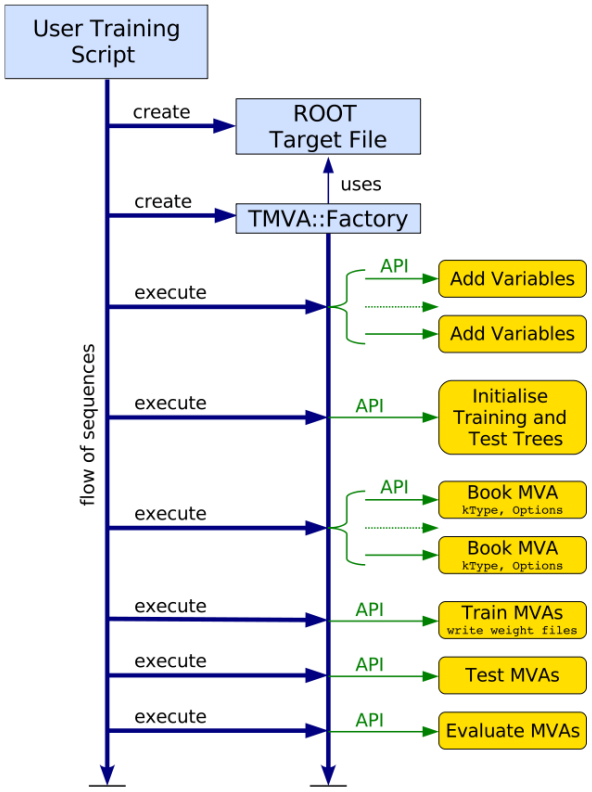

Fig. 12: TMVA flow diagram [31].

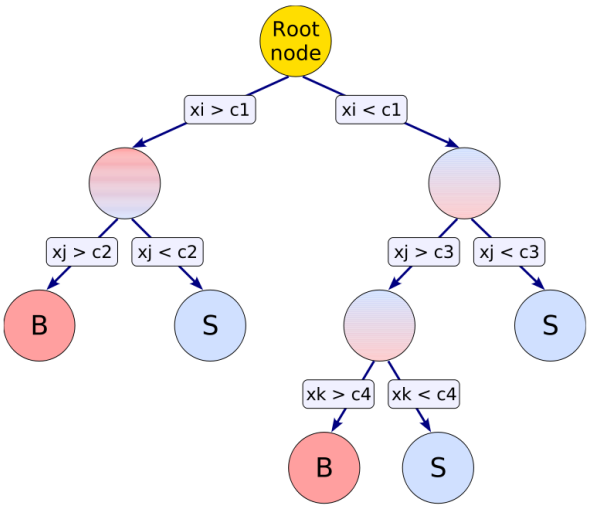

Fig. 13: Schematic view of a decision tree [31].

For this analysis the Boosted Decision Trees method [32], [33] was chosen. It is widely used in the CMS analyses. Moreover, the BDT method performed the best out of all TMVA methods at the signal Wh and background WZ event classification.

3.3.3 Boosted Decision Trees

Boosted decision trees (BDT) is one of the most common methods used for the data analysis in CMS. The classification of events is made by making decisions in multiple tree nodes. BDT algorithm enables to use multiple decision trees (so called decision tree forest). Different decision trees are trained using the original training data set with re-weighted events. Adaptive boosting method is used (AdaBoost, [34] more methods are also avalaible), which gives misclassified events bigger weight in the training of the next tree.

Each tree in this forest is derived like in Fig. 13 from the same training sample provided by the user. At each node, starting from the root node, the data is split into two branches according to the properties of the events that make them more *signal-like* or *background-like*. Splitting is made by using discriminating variables *xi* and separating events that have their *xi* value smaller or larger than a certain value *c1*. Each split uses the variable that at this node gives the best separation between signal and background when being cut on. User can determine the maximum depth of the tree. Too large depths can lead to overtraining. The leaf nodes at the bottom are classified in two categories:

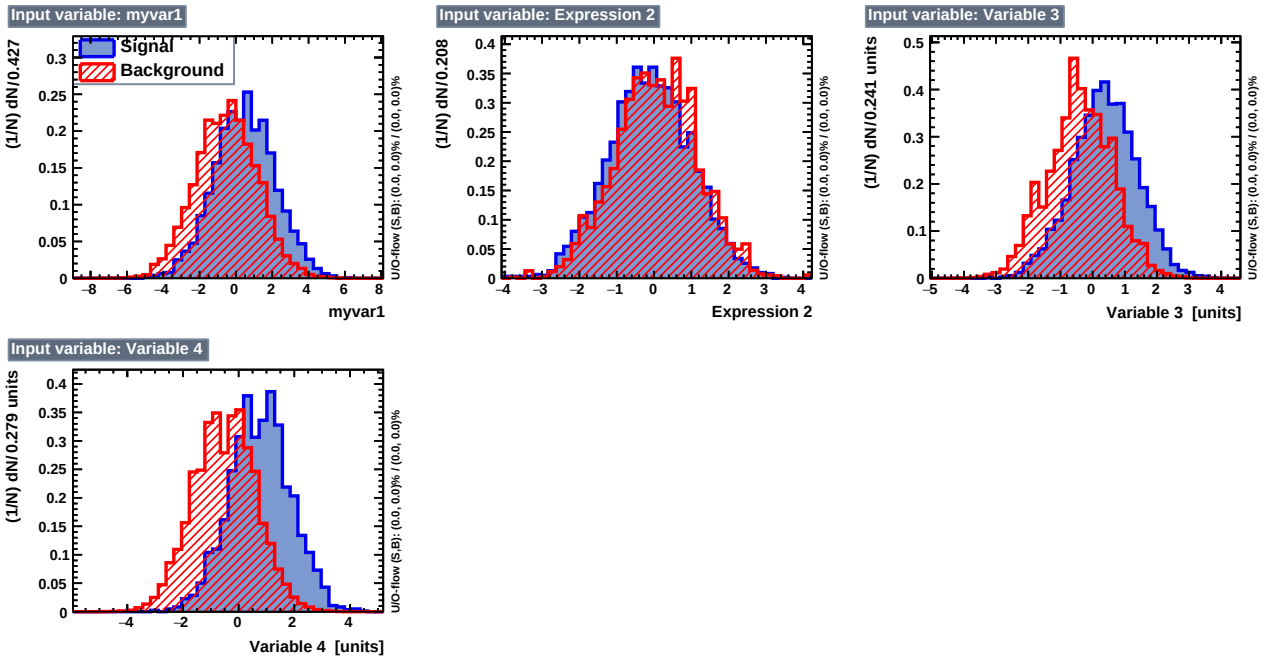

Fig. 14: Input variables used for training from TMVA sample files.

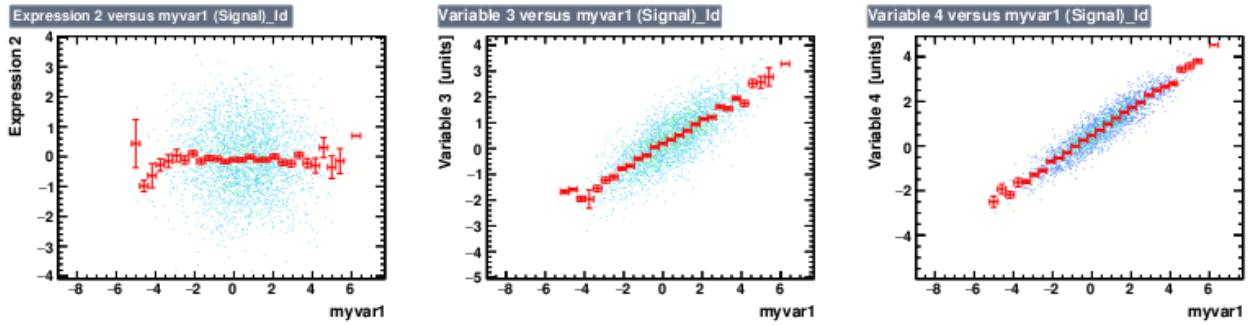

Fig. 15: Correlations for the signal between the input variable number 1 and other variables: expression number 2 and input variables number 3 and 4. TMVA sample files are used.

S is for signal and *B* is for background depending on the majority of events that are there.

BDT algorithm comes with TMVA package and sample files that can be used for training as an example. There are four training variables for background and signal in the sample as shown in a Fig. 14.

TMVA creates *weight* files containing the method-specific training results. It also produces various performance and control plots. It could be useful to calculate correlations between variables. TMVA can produce linear correlation scatter plots for each variable pair as in Fig. 15 and calculate linear correlation coefficients as in Fig. 16.

TMVA can also perform various transformations on input variables in order to get better separation for signal and background. It allows to perform a linear decorrelation transformation of the input variables prior to the MVA training. It could be useful when the selected MVA method doesn't use correlations as it is for the BDT. The result of such decorrelation is shown in Fig. 17.

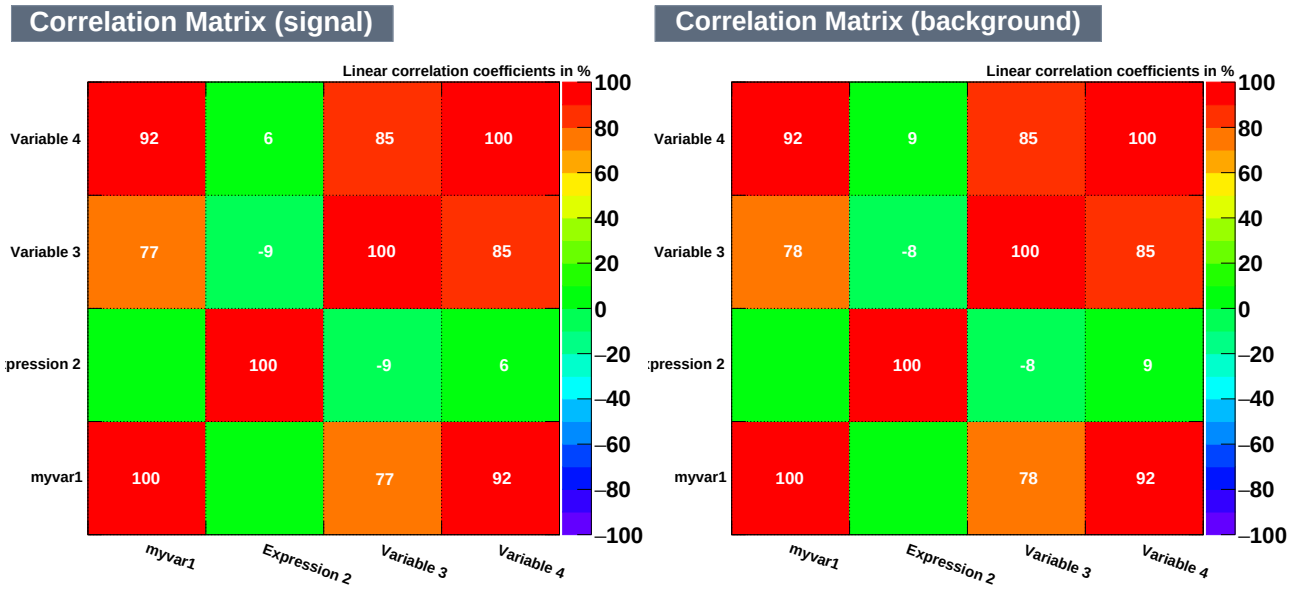

Fig. 16: Linear correlation coefficients for input variables from TMVA sample. *Left* plot shows the coefficient matrix for the signal and the *right* plot shows for the background.

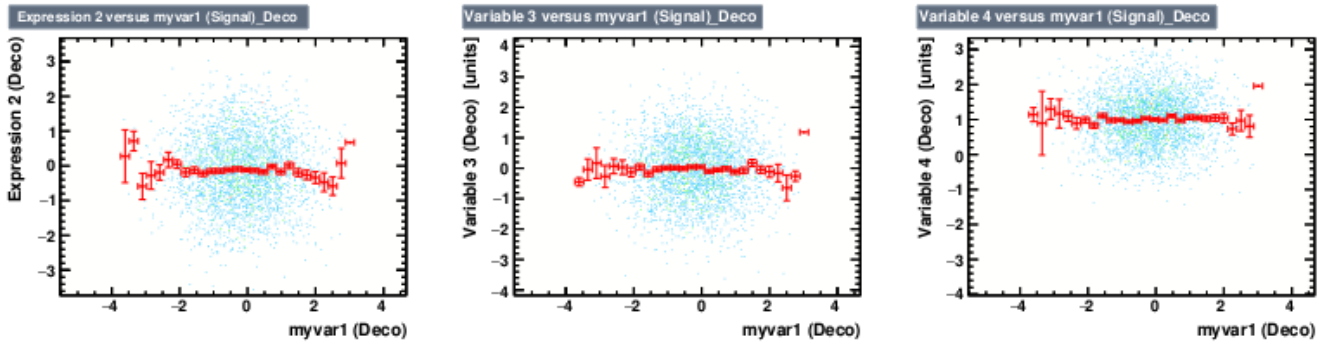

Fig. 17: Correlations for the signal between the decorrelated input variable number 1 and other decorrelated variables: expression number 2 and input variables number 3 and 4.

Decorrelation of input variables can improve the classification of some MVA algorithms.

The example of one of the trees in BDT method for the test sample can be seen in Fig. 18. The classification is made according to variable *var4* value and the sum of *var1* and *var2* values. Signal significance is also calculated in each node.

TMVA draws classifier output distributions for signal and background events for each selected method. This distribution for BDT method can be seen in Fig. 19. By TMVA convention, signal events accumulate at large response values, and background events at small values. Each BDT response value corresponds to the certain signal and background efficiencies. According to these values cuts on the output can be made. Increasing cut value reduces signal efficiency but increases it's purity. This relation can be seen in Fig. 20, which is a Receiver Operating Characteristic diagram. It shows the signal efficiency versus background rejection for the different possible cut points. Signal is described by following variables:

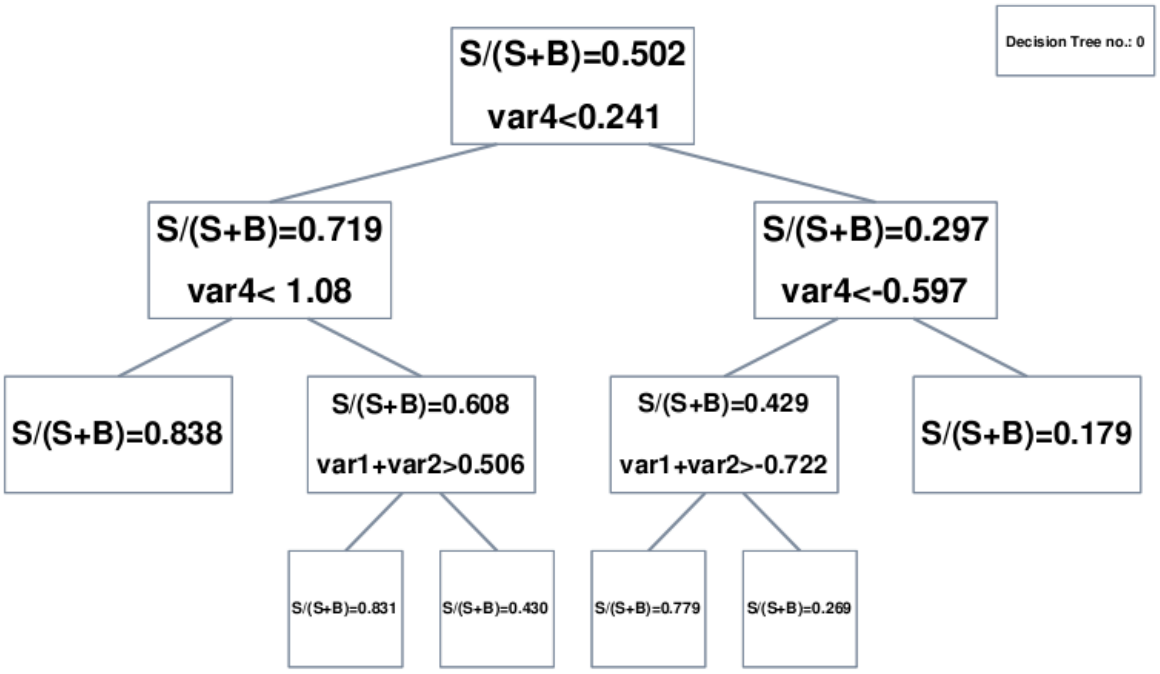

Fig. 18: One of the trees from BDT forest created from example sample files.

1. Signal efficiency is a number of weighted signal events that passed the cuts relative to the total number of weighted signal events.

2. Signal purity is a number of weighted signal events that passed the cuts relative to the all weighted number of events that passed the cuts (which is the sum of the numbers of weighted signal and background events).

Similarly, background is described by:

1. Background efficiency is a number of weighted background events that passed the cuts relative to the total number of weighted background events.

2. Background rejection is the inverse of background efficiency.

3.4 Possible Discriminating Variables

This section explains the search for the discriminating variables as in [9]. The both signal cases $h \to \tau\tau$ and $h \to \mu\mu$ are studied.

3.4.1 Invariant mass

In particle physics, the invariant mass m_0 is equal to the mass in the rest frame of the particle. It can be calculated by the particle's energy E and it's momentum p measured in any frame, by the

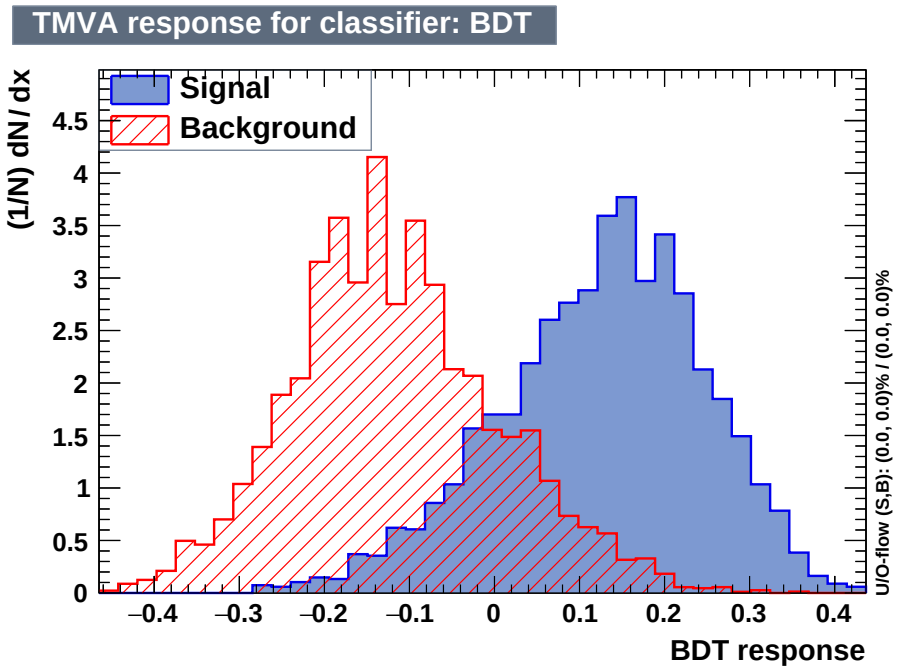

Fig. 19: Classifier output distribution for signal and background events for BDT method from training with example test sample.

energy–momentum relation:

$$m_0^2 c^2 = \left(\frac{E}{c}\right)^2 - \|\mathbf{p}\|^2$$

The invariant mass remains the same in every frame. The invariant mass can be reconstructed from the decay products of the particle.

For example, the invariant mass for the Wh production process with $W \to \mu\nu$ and $h \to \mu\mu$ and for the WZ production process with $W \to \mu\nu$ and $Z \to \mu\mu$ at the reconstructed level can be calculated as it was done in my previous work [9] from the muon pair coming from either Higgs or Z boson as it can be seen in Fig. 21. From the graph we can see, that the invariant mass from $h \to \mu\mu$ peak corresponds to the value around 125 GeV/c^2 which is the mass of the Higgs boson and the invariant mass from $Z \to \mu\mu$ peak corresponds to the value of Z boson mass accordingly (91 GeV/c^2).

In the signal process Wh, with $W \to \mu\nu$ and $h \to \tau\tau$ with $\tau \to \nu\mu\nu$, and the WZ background process, with $W \to \mu\nu$ and $Z \to \tau\tau$ with $\tau \to \nu\mu\nu$, the invariant mass of the Higgs and Z bosons can be calculated from the τ leptons as shown in Fig. 22.

Due to the way τ lepton decays with spin correlations are implemented in PYTHIA 8.1, there are two levels of generated τ leptons in our analysis. It's the *first* $\tau\tau$, which are the ones that appear directly from Higgs or Z boson decay and *last* $\tau\tau$, which are the last in the decay chain before decaying into muons.

However, τ leptons cannot be fully reconstructed in the detector. We can only detect their visible

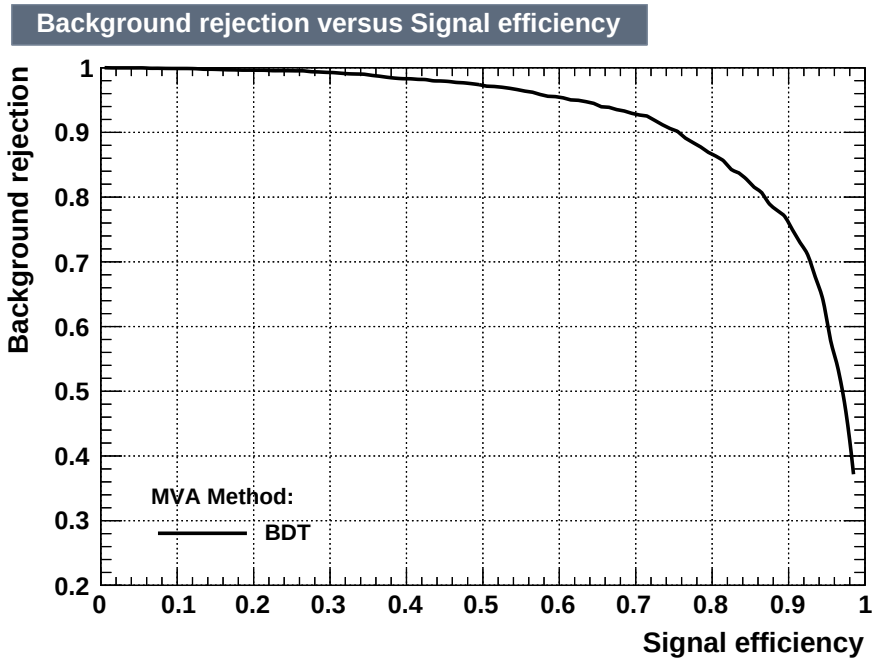

Fig. 20: Background rejection versus signal efficiency on the classifier outputs for the example test sample.

decay products, which are muons in our signal process case. The neutrinos from the $\tau^- \rightarrow \nu_\tau \mu^- \bar{\nu}_\mu$ and $\tau^+ \rightarrow \bar{\nu_\tau} \mu^+ \nu_\mu$ decays are not measured. The invariant mass for muons can be calculated at both generated and reconstructed levels. The comparison between these levels is presented in Fig. 23.

It is clear that due to the neutrinos being present in the decay we can not reconstruct the exact invariant mass of the Higgs or Z bosons.

In the Fig. 24 we can see that there is a slight difference between reconstructed invariant mass from muon pair for Wh and WZ processes. Moreover, comparing invariant mass distributions in Fig. 21 and Fig. 24 we can see that the reconstructed invariant mass is a good discriminant between $h \rightarrow \tau\tau$ with $\tau \rightarrow \nu\mu\nu$, $h \rightarrow \mu\mu$ and $Z \rightarrow \mu\mu$ processes.

3.4.2 Missing transverse energy

The initial momentum of particles traveling transverse to the beam axis is zero, so any net momentum of the sum of all particles produced in our event in the transverse direction indicates that there are some energy that is not detected. This energy is called the missing transverse energy (MET).

Due to the presence of undetected particles, e. g., neutrinos, in the studied decays there is missing energy in the reconstructed event. Using the neutrinos from the generated level this energy can be estimated as shown in Fig. 25.

However, the missing energy in the event is not only due to the neutrinos. The difference between generated and reconstructed level MET can be traced back to the level of understanding of

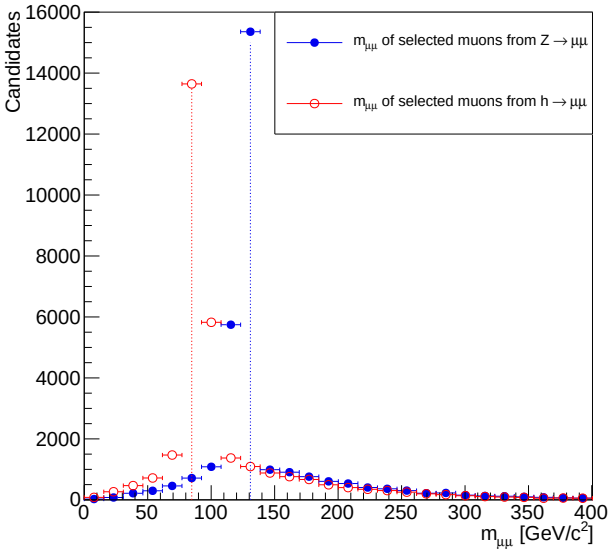

Fig. 21: Invariant mass of the dimuon pair at the reconstructed level for the Z and the Higgs bosons for $h/Z \rightarrow \mu\mu$ prcess.

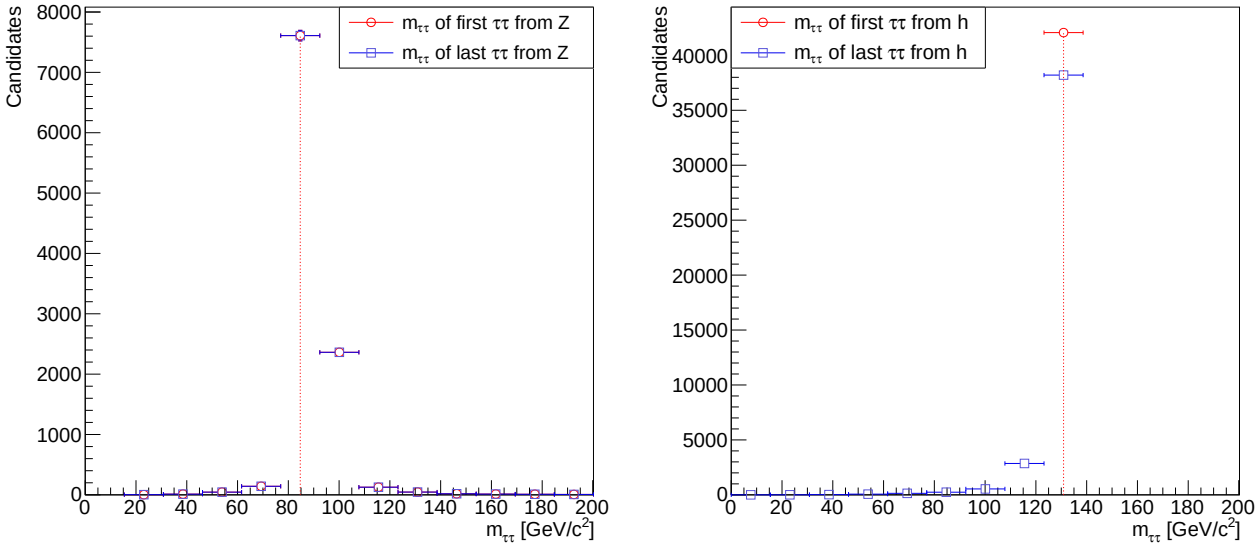

Fig. 22: Invariant mass of the ditau pair at the generator level for the Z boson (left) and the Higgs boson (right) for $h/Z \rightarrow \tau\tau$ process.

the energy scale of the jets originating from quarks and gluons.

Including the detector effects, a small difference between MET values for the signal Wh and main background process WZ which can be observed in Fig. 26 could be used in addition to other discriminants when determining the irreducible background.

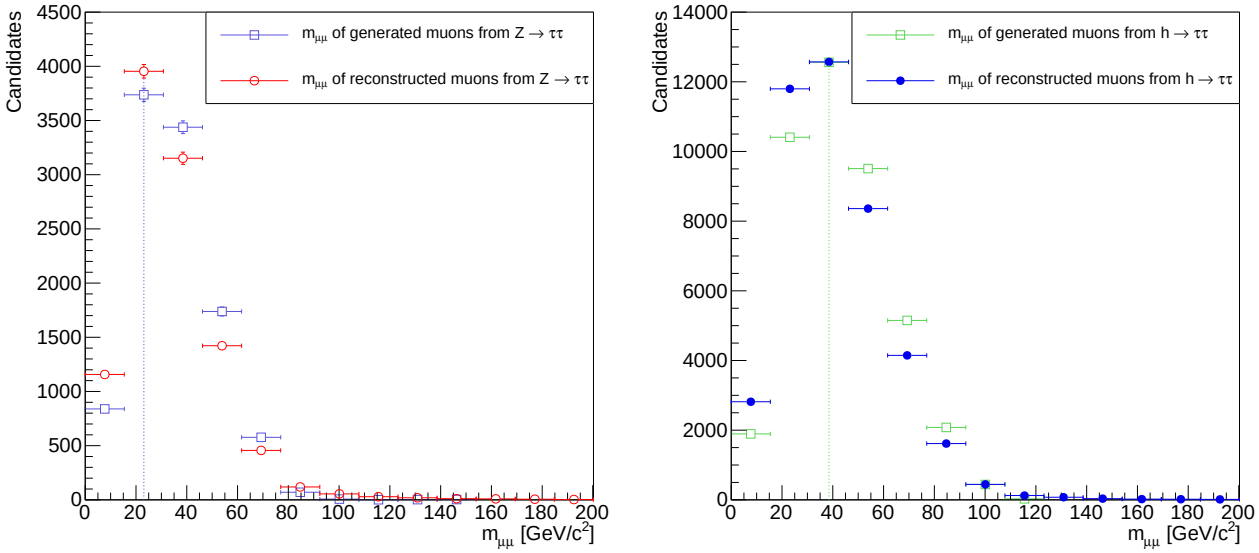

Fig. 23: Invariant mass of the dimuon pair. The invariant mass is reconstructed from the muons from Z boson on the *left*, and from the Higgs boson on the *right*. Reconstruction is done for the $h/Z \to \tau\tau$ process.

3.4.3 Angles between muons

Another possible discriminant for the signal and irreducible background process WZ is angles between produced muons.

We can determine the angle between outgoing muons from each process. Since we expect the highest transverse momentum muon to be coming from the W boson, we look at the angles between the highest transverse momentum muon and the plane of lower transverse momentum muons (which should be from either Higgs or Z bosons). Due to the different Higgs and Z boson masses there we could expect a certain discrimination power. As plotted in Fig. 27 there exists some difference between these processes.

Similarly, we can see a slight difference in Fig. 28 between the processes Wh and WZ when examining the angle between the muon from the W boson and the muon with higher transverse momentum from the Higgs or Z boson decays.

Finally, if we look at the angle between two muons from either Z or Higgs bosons we can see in Fig. 29 slight differences between these processes too. Despite that these differences can not be used on their own to determine which process it is: Wh or WZ, they would be useful when providing the discriminating information to the multivariate analysis training. These and additional variables including also the correlations between them will be used for the MVA training described in the *Multivariate analysis results* section.

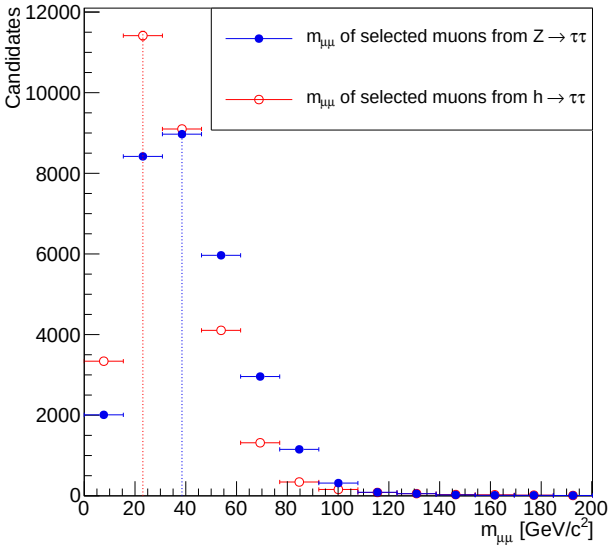

Fig. 24: Invariant mass of the dimuon pair at the reconstructed level for the Z boson (green) and the Higgs boson (red) for $h/Z \to \tau\tau$ process.

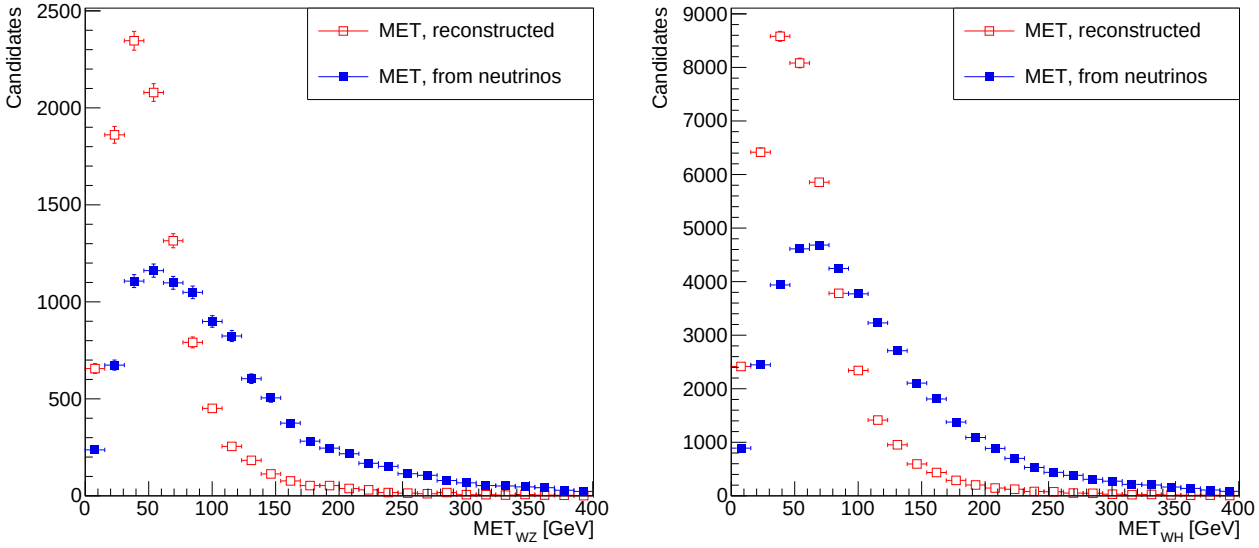

Fig. 25: Missing energy distributions for the signal WZ (left) and background process Wh (right) for $h/Z \to \tau\tau$ process.

3.5 MVA Overtraining and Kolmogorov-Smirnov Test

Kolmogorov-Smirnov test [35] is used to evaluate the overtraining of the selected MVA method. When too many model parameters of an algorithm are adjusted to too few data points in machine learning overtraining can occur. TMVA can calculate Kolmogorov-Smirnov test probability value for the signal and background processes. Too small value of the probability indicates that the method is overtrained.

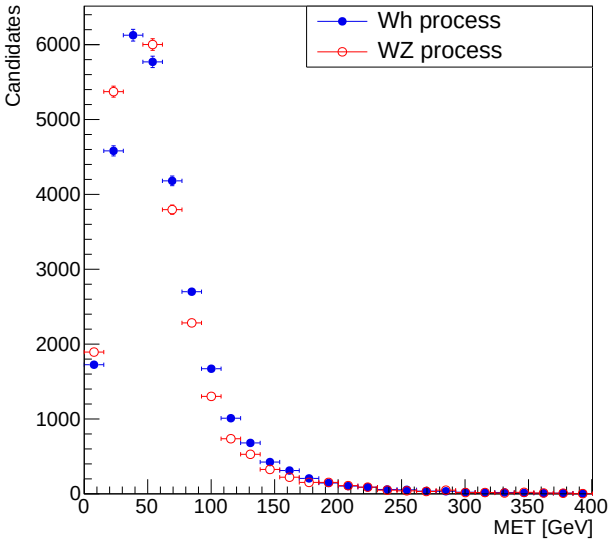

Fig. 26: Missing energy distributions for the signal Wh and background process WZ at reconstructed level, normalized to 1.

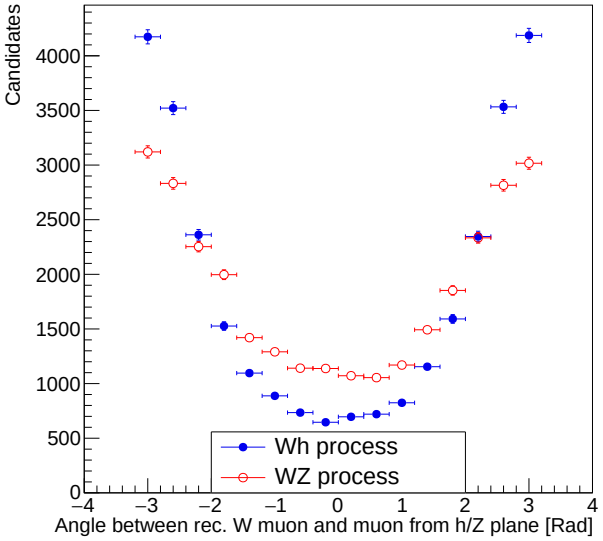

Fig. 27: Angle between the muon from the W boson decay and the plane of muons from the Higgs or Z boson decays for $h/Z \to \tau\tau$ process.

When Kolmogorov-Smirnov probability value is smaller than 0.01 the user has to take action to reduce the overtraining of the trees, otherwise the method would pick up on statistical fluctuations of the training sample and will have seemingly better performance on the training events than the independent (test or data) sample. Overtraining for BDT can be reduced by a few methods: pruning (which is automatic for Adaptive Boost type boosted decision trees), reducing the number of the trees in a forest, increasing minimal number of events requested by a leaf node or changing pruning

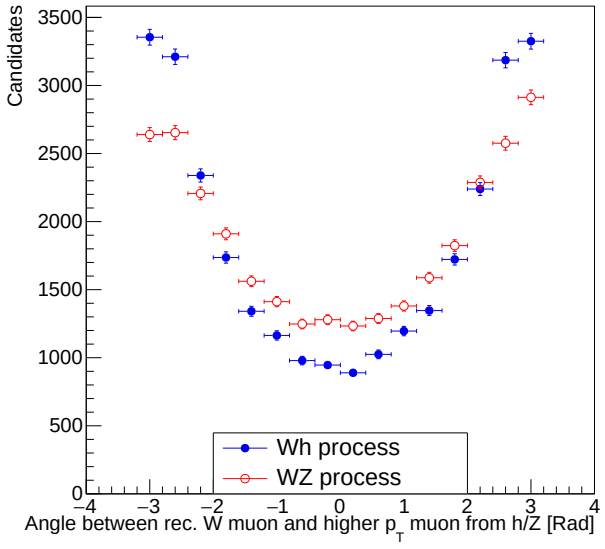

Fig. 28: Angle between the muon from the W boson decay and the muon with higher transverse momentum from the Higgs or Z boson for $h/Z \rightarrow \tau\tau$ process.

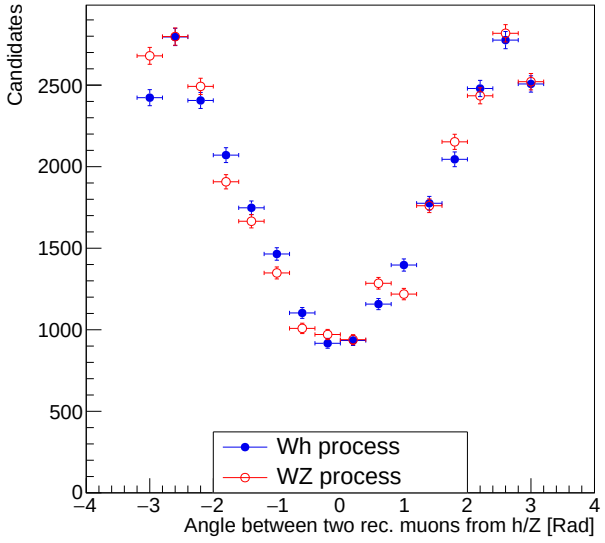

Fig. 29: Angle between two muons from the Higgs or Z boson decays for $h/Z \rightarrow \tau\tau$ process.

strength. Default parameter values can be seen in Fig. 30 and can be adjusted to avoid overtraining.

The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (ECDF). Given N ordered data points $Y_1, Y_2, ..., Y_N$, the ECDF is defined as:

$$E_N = \frac{n(i)}{N},$$

where $n(i)$ is the number of points less than Y_i, Y_i are ordered from smallest to largest value. This

Option	Array	Default	Predefined Values	Description
NTrees	–	800	–	Number of trees in the forest
MaxDepth	–	3	–	Max depth of the decision tree allowed
MinNodeSize	–	5%	–	Minimum percentage of training events required in a leaf node (default: Classification: 5%, Regression: 0.2%)
nCuts	–	20	–	Number of grid points in variable range used in finding optimal cut in node splitting
BoostType	–	AdaBoost	AdaBoost, RealAdaBoost, Bagging, AdaBoostR2, Grad	Boosting type for the trees in the forest (note: AdaCost is still experimental)
AdaBoostR2Loss	–	Quadratic	Linear, Quadratic, Exponential	Type of Loss function in AdaBoostR2
UseBaggedBoost	–	False	–	Use only a random (bagged) subsample of all events for growing the trees in each iteration.
Shrinkage	–	1	–	Learning rate for GradBoost algorithm
AdaBoostBeta	–	0.5	–	Learning rate for AdaBoost algorithm
UseRandomisedTrees	–	False	–	Determine at each node splitting the cut variable only as the best out of a random subset of variables (like in RandomForests)
UseNvars	–	2	–	Size of the subset of variables used with RandomisedTree option
UsePoissonNvars	–	True	–	Interpret UseNvars not as fixed number but as mean of a Possion distribution in each split with RandomisedTree option
BaggedSampleFraction	–	0.6	–	Relative size of bagged event sample to original size of the data sample (used whenever bagging is used (i.e. UseBaggedBoost, Bagging,)
UseYesNoLeaf	–	True	–	Use Sig or Bkg categories, or the purity=S/(S+B) as classification of the leaf node -> Real-AdaBoost

Fig. 30: BDT parameters that can be adjusted by the user to manage the behavior and performance of trees [31].

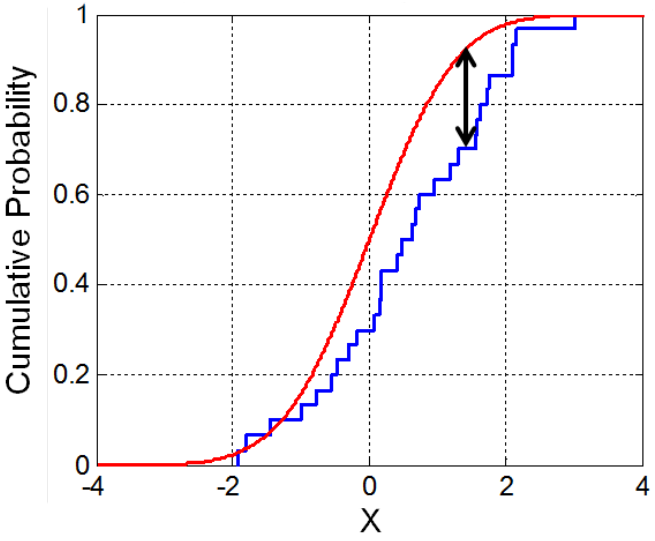

Fig. 31: Illustration of the Kolmogorov–Smirnov statistic. Red line is Cumulative Distribution Function $F(Y_i)$, blue line is an ECDF, and the black arrow is the K–S statistic $D_{\max}$.

Signal efficiency:	Test sample	Error	Training sample
Bkg. eff. 0.01	0.34	$\pm$ 0.08	0.39
Bkg. eff. 0.10	0.76	$\pm$ 0.07	0.78
Bkg. eff. 0.30	0.93	$\pm$ 0.04	0.93

Table 5. Signal efficiency values for test and training samples at different background efficiency values.

is a step function that increases by $\frac{1}{N}$ at the value of each ordered data point as seen Fig. 31.

The Kolmogorov-Smirnov statistic is calculated in the following way:

$$D_{\max} = \max\left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)\right), 1 \leq i \leq N,$$

where $F(Y_i)$ is Cumulative Distribution Function (CDF).

Null hypothesis H_0 is that the data follow a specified distribution. The H_0 regarding the distributional form is rejected if the test statistic D is greater than the critical value.

Kolmogorov-Smirnov signal and background probabilities as seen in Fig. 32 are the probabilities that Kolmogorov's test statistic will exceed the value of z assuming the H_0, where:

$$z = D_{\max}\sqrt{N_{\text{events}}}.$$

Kolmogorov–Smirnov probability should not be smaller than 0.01. However, for overtraining check it is also important that training and independent test sample would have efficiencies that are equal within statistical errors, in order to minimize statistical fluctuations in different samples.

From the Table 5 it is evident, that efficiency values for test and independent training samples in this example are within the statistical errors so the BDT are not significantly overtrained.

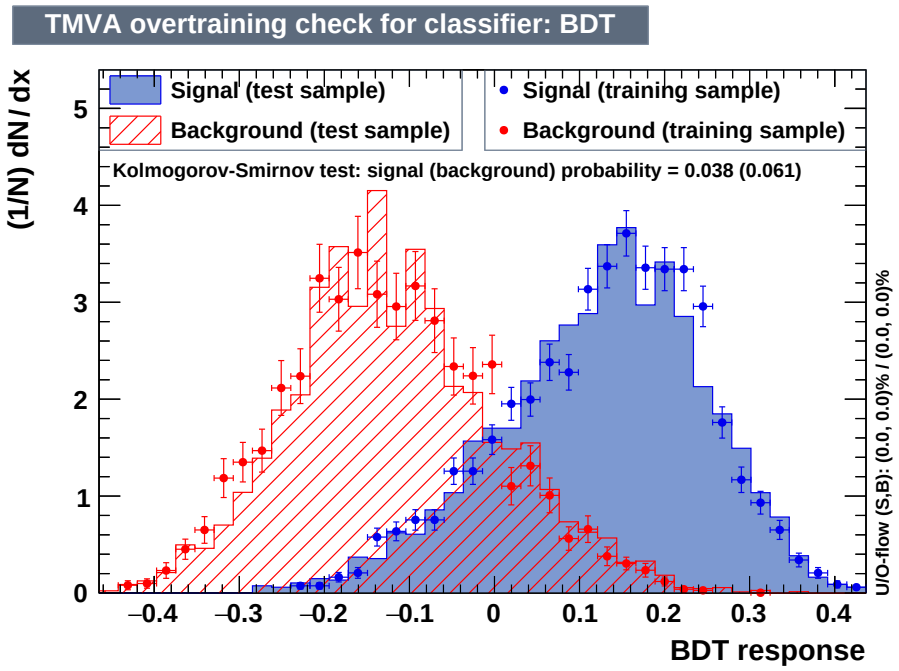

Fig. 32: Overtraining check and classifier output distribution for the signal and background events. The BDT method is applied to the test and training samples.

4 Multivariate Analysis Results

This section describes the results obtained analysing Wh and WZ data samples. Section 4.1 discusses the direct decay of the Higgs boson to two muons. Section 4.2 presents the results for Higgs boson decays to two tau leptons each decaying into a muon.

4.1 MVA for Wh and WZ with $h/Z \to \mu\mu$

4.1.1 Setting up the MVA

Firstly, the sample files for the Wh signal (99800 events) and WZ background (99692 events) are produced, where $W \to \mu^+\nu$, $Z \to \mu^+\mu^-$ and $h \to \mu^+\mu^-$. The *tight* muons (definition explained in chapter 3.1 *Analysis Workflow*) are selected and matched to the generated level muons as listed in Table 6. There are noticeably less muons in the acceptance region (the region where the particles can be detected due to the geometry of the detector) for WZ process (74.54%) than for the Wh process (79.8%). For this analysis the reconstructed *tight* muons that have a matching muon at the generator level are used. Selecting *tight* muons reduces the backgrounds arising from jets. However, only 39.86% of generated muons for the WZ process and 42.92% for the Wh process pass these requirements. Most of them (99% that passed the *tight* requirement) have a match among generated muons. Similarly, if we investigate the number of events that passed the requirements, there are only 25.11% Wh events and 25.67% WZ events that were reconstructed with three *tight* muons in the final state. Furthermore, since we use the matched muons for the analysis, there are only 28.72% of all the generated Wh events and 22.94% of WZ events with three matched muons in the final state as listed in Table 7. There are more events with matched three muons in the final state than reconstructed with three muons in the final state for Wh process. This difference arises

Muons	Wh, N_{muons}	Wh, %	WZ, N_{muons}	WZ, %
Generated	308157	100	307834	100
In acceptance region	246127	79.8	229780	74.64
Reconstructed tight	132256	42.92	122689	39.86
Matched	131159	42.56	121557	39.49

Table 6. Number of various types of muons in the generated Wh and WZ samples, where $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$.

Events	Wh, N_{events}	Wh, %	WZ, N_{events}	WZ, %
Generated	99800	100	99692	100
Reconstructed with at least 1 muon	98994	99.30	98925	99.35
Matched with at least 1 muon	56161	56.34	55164	55.40
Reconstructed with 3 muons	25063	25.11	25594	25.67
Matched with 3 muons	28862	28.72	22843	22.94

Table 7. Number of various types of events in the generated Wh and WZ samples, where $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$.

Rank	Variable	Description	Importance
1	tInvMass3	Invariant mass of three muons.	1.922×10^{-1}
2	tV3mt	m_t of the vector sum of three muons.	8.389×10^{-2}
3	tHmt	m_t of the highest p_t muon.	7.641×10^{-2}
4	tV2mt	m_t of the vector sum of two h/Z muons.	7.608×10^{-2}
5	tLpt	p_t of the lowest p_t muon.	7.285×10^{-2}
6	tAngleWH12	Angle between h/Z muons.	6.133×10^{-2}
7	tAngleWH	Angle between W muon and h/Z muon plane.	5.698×10^{-2}
8	tMETW	Angle between W muon and missing E_t.	4.662×10^{-2}
9	tDRH12	Angular separation between h/Z muons.	4.586×10^{-2}
10	tHeta	Angle between W muon and beam direction.	4.426×10^{-2}
11	tSmt	Scalar sum of all the three muon m_t.	4.344×10^{-2}
12	tInvMassMET	Mass of dimuon h/Z vector and missing E_t vector sum.	4.213×10^{-2}
13	tWH12eta	Angle between H muon plane and beam direction.	4.152×10^{-2}
14	tAngleWH1	Angle between W muon and higher p_t h/Z muon.	4.062×10^{-2}
15	tMET	Missing E_t.	3.882×10^{-2}
16	tMETH12	Angle between h/Z muon plane and missing E_t.	3.701×10^{-2}

Table 8. Discriminating variables for Wh and WZ processes, where $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$.

from the fact that in some events there are more than three reconstructed muons and three of these muons have their matching muons at the generated level. For this analysis the events that have three matched muons in the final state after reconstruction (*Matched with 3 muons* in Table 7) are used to calculate various discriminating variables. More of the Wh events than the WZ events pass this selection. Discriminating variables are calculated from them and used for Boosted Decision Trees (BDT) training.

For the multivariate analysis for separating Wh and WZ processes sixteen discriminating variables were calculated. The names and descriptions of variables can be seen in Table 8. All the variables are are listed by rank (according to their importance). Invariant mass of all the three muons has the highest importance of 1.922×10^{-1} (total importance for all the variables is equal to one). Next most important variables for separating signal and background are the m_t (transverse mass) of the vector sum of energy-momentum 4-vectors of all the three muons in the final state, m_t of the highest p_T (transverse momentum) muon and the m_t of the sum of two 4-vectors of muons that are expected to be from either Higgs or Z boson. Each variable is plotted in Fig. 33.

Linear correlation coefficients for all the input variables for signal and background can be seen in the correlation matrix in Fig. 34. The most important discriminating variables happened to be noticeably correlated: *tInvMass3* is linearly correlated to the *tV2mt* and *tV3mt* with correlation coefficients of 78 and 64 for the signal Wh process and 83 and 65 for the background WZ process accordingly. Since BDT can use decorrelated variables the initial correlations between variables does not necessarily significantly reduce their importance for the signal and background separation.

BDT trees were trained with following parameters: 150 trees in a forest (more trees enable

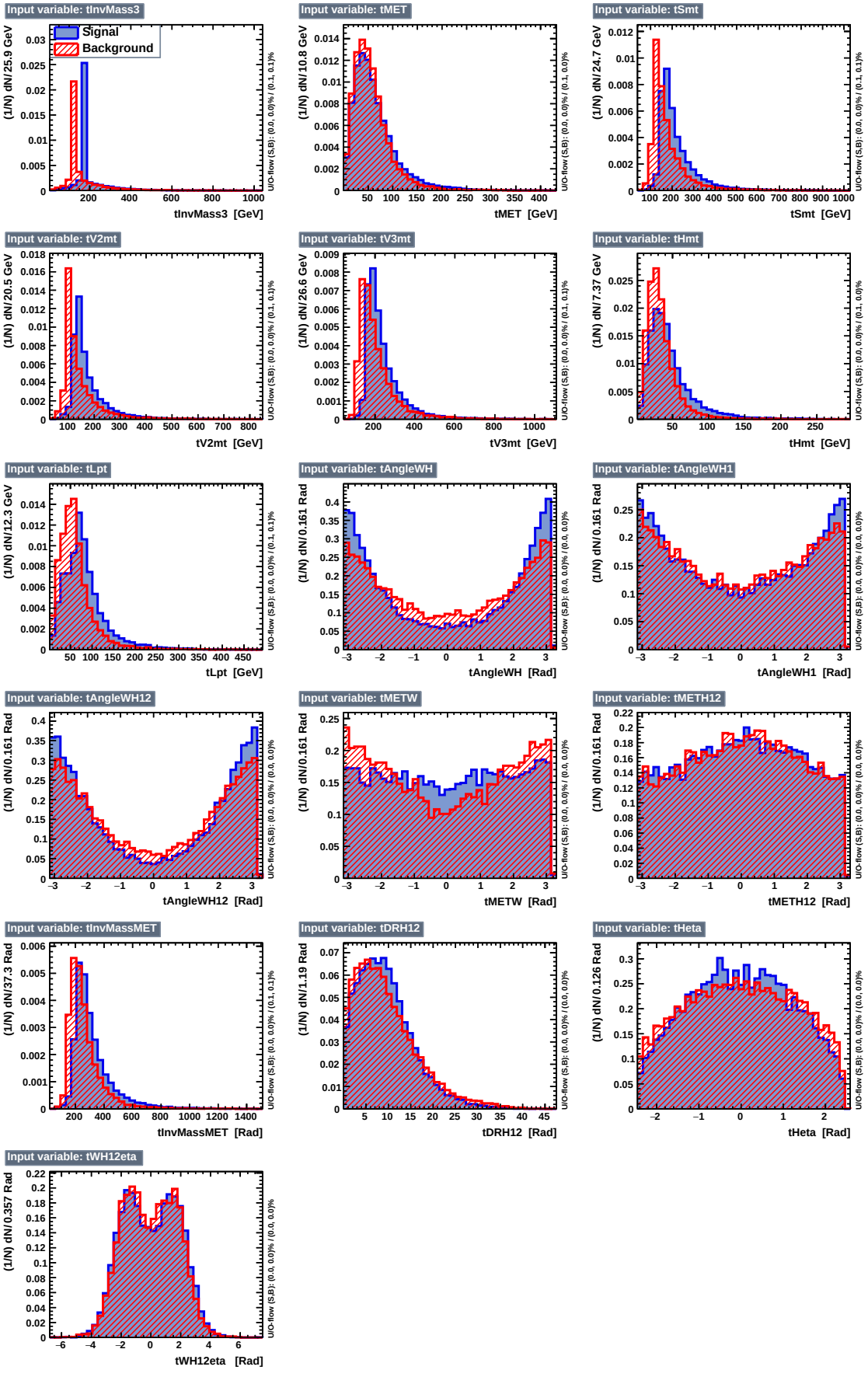

Fig. 33: Discriminating variables for Wh and WZ processes, where $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$.

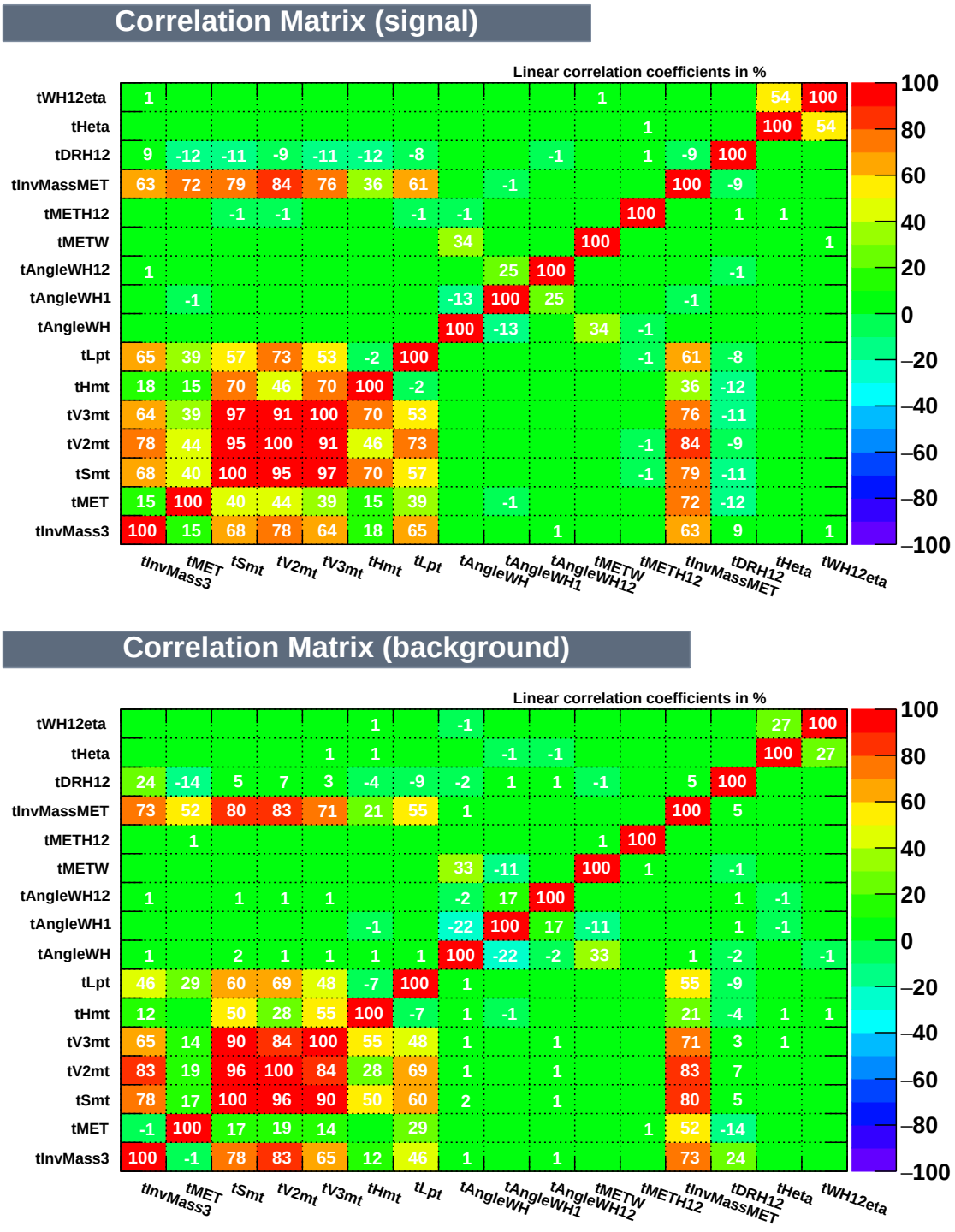

Fig. 34: Correlation matrices for the signal Wh (*top*) and the background WZ (*bottom*) processes, where $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$.

Signal efficiency:	Test sample	Error	Training sample
Bkg. eff. 0.01	0.45	± 0.04	0.45
Bkg. eff. 0.10	0.83	± 0.03	0.83
Bkg. eff. 0.30	0.95	± 0.01	0.95

Table 9. Signal Wh efficiency values for test and training samples at different background WZ efficiency values, where $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$.

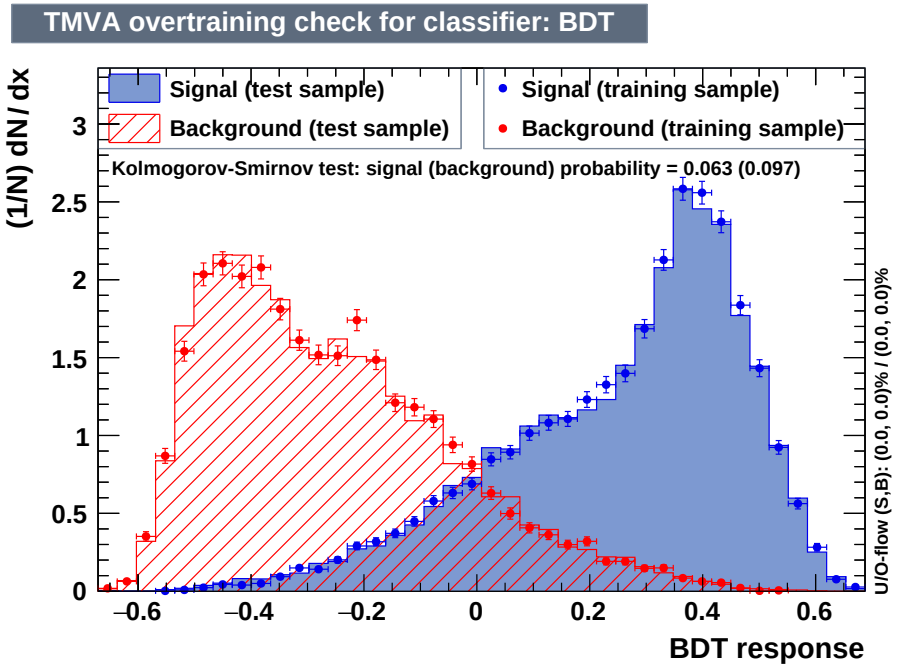

Fig. 35: Overtraining check for BDT for the signal Wh (*top*) and the background WZ (*bottom*) processes, where $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$.

to classify events with higher accuracy, however larger forest increases the risk of overtraining), nodesize of 5%, which means that at least 5% of all the events should be in each tree node after event classification. If there are less than 5% of all the events the node is considered insignificant and is removed by the tree pruning. The depth of a tree parameter is set to 3, so there is a maximum number of 3 levels in each tree. The check for overtraining was done according to Kolmogorov-Smirnov probability seen in Fig. 35 and making sure that signal and background efficiencies are equal within the error limits for training and independent test samples as seen in Table 9. The comparison of the BDT classification for training and test samples also can be seen in Fig. 35. It is important, that BDT response of the test sample would closely resemble the training sample response which indicates correctly performed training. Moreover, BDT response for signal Wh mostly accumulates at the negative BDT parameter values and BDT response for background WZ events accumulates around positive values. This already indicates quite good signal and background separation which can be quantitatively evaluated from the receiver operating characteristic (ROC) diagram.

The resulting ROC curve for two methods: *Cuts* (performing optimal cuts on input variables

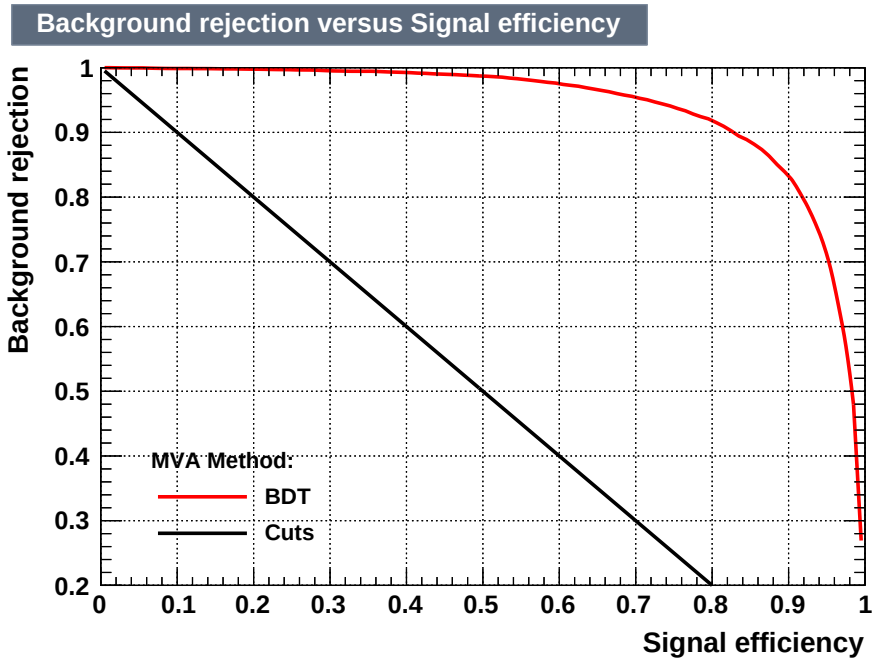

Fig. 36: Receiver operating characteristic diagram for BDT trained for Wh and WZ processes with $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$.

for signal and background processes in order to maximize the signal efficiency at given background efficiency) and BDT can be seen in Fig. 36. The BDT method demonstrates good background rejection (as explained in chapter 3.3.3), for example, for the 50% of signal efficiency background rejection is 98% (meaning that loosing half of the signal events we have only 2% of the background events) while for the signal efficiency of 80% it is 92%. However, the expected number of signal events is much smaller than the number of background events as it was calculated in Table 5. The MVA results from the test sample adjusted by the luminosity, production cross-sections and branching ratios are investigated in the following chapter.

4.1.2 MVA results for Wh and WZ with $h/Z \to \mu\mu$

The simulated signal and background samples have to be reweighed to match the expected number of events for a given value of the integrated luminosity. The background suppression is evaluated by the expected significance: $S/\sqrt{S+B}$, where S is the number of the expected signal events and B is the number of the background events in the signal region.

The resulting signal efficiency, background efficiency and significance dependance on the BDT output cut value can be seen in Fig. 37. From this efficiency dependence on cut value the optimal cut value can be selected. The chosen cut value is $BDT > 0.35$. At this value the signal efficiency of 40% can be expected and background efficiency less than 5%. The reduction of background events for the test sample of 100 fb^{-1} integrated luminosity after making the cut on the BDT output is represented in the Fig. 38. The number of signal events after the cut is the same as before making

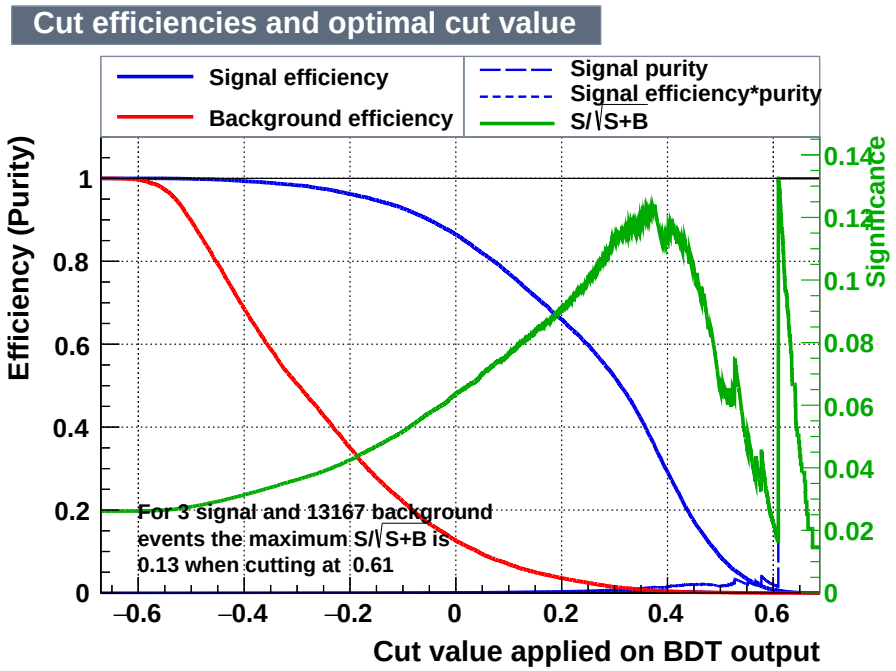

Fig. 37: Signal efficiency dependence on BDT output cut for Wh and WZ processes with $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$, for adjusted by the expected number of events test sample.

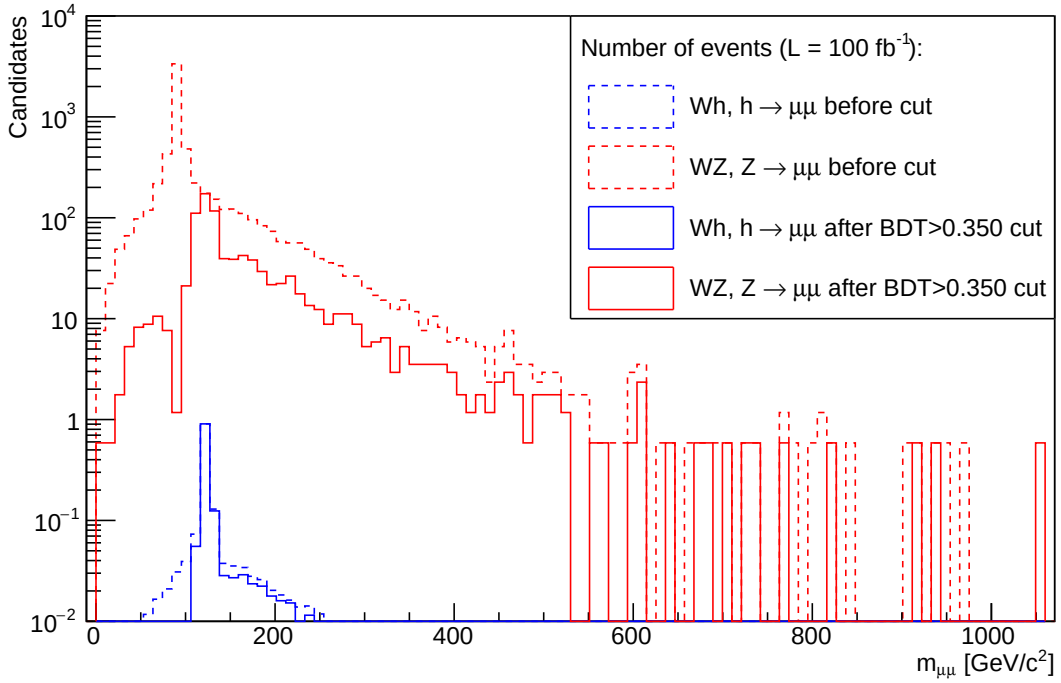

Fig. 38: Number of signal and background events before and after making the cut on the BDT output value of 0.35 for Wh and WZ processes with $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$.

Muons	Wh, N_{muons}	Wh, %	WZ, N_{muons}	WZ, %
Generated	582006	100	618566	100
In acceptance region	463264	79.60	618566	74.68
Reconstructed tight	225519	38.75	229519	37.11
Matched	223327	38.37	227155	36.72

Table 10. Number of various types of muons in the generated Wh and WZ samples, where $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$.

Events	Wh, N_{events}	Wh, %	WZ, N_{events}	WZ, %
Generated	188108	100	200000	100
Reconstructed with at least 1 muon	186276	99.14	198162	99.20
Matched with at least 1 muon	100459	53.47	102457	51.29
Reconstructed with 3 muons	47446	25.47	51432	25.75
Matched with 3 muons	42859	22.81	33680	16.86

Table 11. Number of various types of events in the generated Wh and WZ samples,where $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$.

the cut. Despite low background efficiency observing the signal is still complicated due to the low signal event number relative to the high background.

4.2 MVA for Wh and WZ with $h/Z \to \tau\tau \to \mu\mu + neutrinos$

4.2.1 Setting up the MVA

In the same way as for the Wh and WZ with $W \to \mu\nu$, $Z \to \mu\mu$ and $h \to \mu\mu$ case, the sample files for the Wh signal (188108 events) and WZ background (200000 events) were produced, where $h \to \tau^+\tau^-, \tau \to \mu\bar{\nu}_\mu\nu_\tau, W \to \mu\bar{\nu}_\mu$. The muon selection was the same, also listed in Table 10. In this case the muons from Wh process were also more likely to be in the detector acceptance region: 79.60% of all the generated events while for WZ the probability is 74.68%. Reconstruction of muons is slightly worse than in $h/Z \to \mu\mu$ case: 38.75% and 37.11% accordingly. Event layout is in Table 11. The fraction of reconstructed events with 3 muons is about the same for this case: 25.47% for Wh and 25.75% for WZ. However, there are noticeably less events with 3 matched muons: only 22.81% for the Wh and 16.86% for the WZ. This means, that for the $h/Z \to \mu\mu$ processes we reconstruct the events with three muons in the final state better than for the $h/Z \to \tau\tau$ process. Since the latter processes are harder to distinguish, more signal and background events were used in the analysis for $h/Z \to \tau\tau$ case resulting in 42859 signal events and 33680 background events that passed the selection (while in $h/Z \to \mu\mu$ case 28862 signal and 22843 background events were used).

Input variables used in MVA are listed in Table 12. Their ranking is different than in training for $h \to \mu\mu$ and $Z \to \mu\mu$ processes. The most important discriminating variable is the *tSmt*, a scalar sum of all the three muon m_t with 1.710×10^{-1} importance. Next most important variable is the *tV3mt*, the m_t of the vector sum of energy-momentum 4-vectors of all the three muons in the

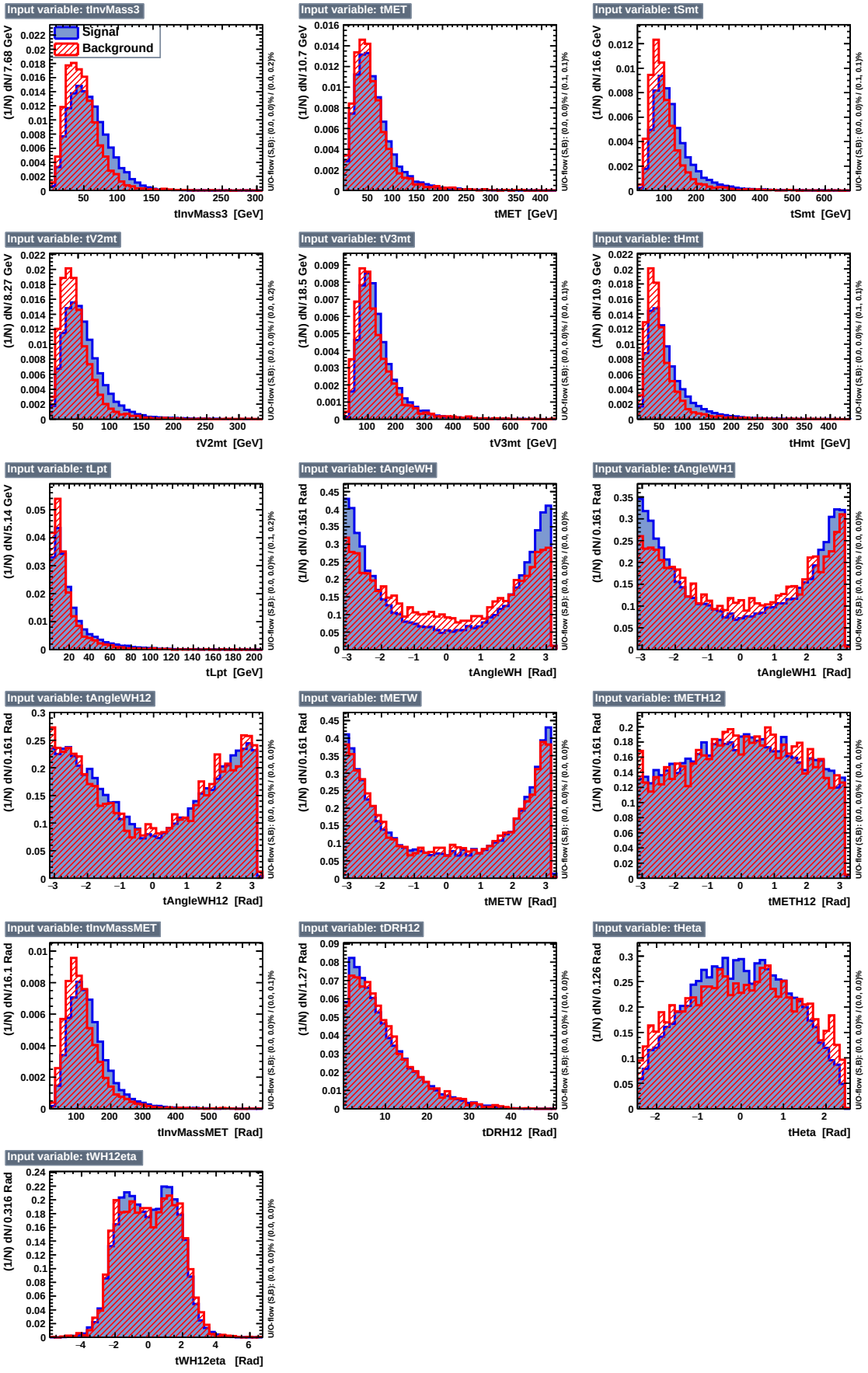

Fig. 39: Discriminating variables for Wh and WZ processes, where $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$.

Rank	Variable	Description	Importance
1	tSmt	Scalar sum of all the three muon m_t.	1.710×10^{-1}
2	tV3mt	m_t of the vector sum of three muons.	1.613×10^{-1}
3	tHmt	m_t of the highest p_t muon.	8.307×10^{-2}
4	tAngleWH	Angle between W muon and h/Z muon plane.	7.343×10^{-2}
5	tHeta	Angle between W muon and beam direction.	7.225×10^{-2}
6	tWH12eta	Angle between H muon plane and beam direction.	6.854×10^{-2}
7	tInvMassMET	Mass of dimuon h/Z vector and missing E_t vector sum.	6.184×10^{-2}
8	tV2mt	m_t of the vector sum of two lower p_t muons.	5.500×10^{-2}
9	tInvMass3	Invariant mass of three muons.	5.267×10^{-2}
10	tDRH12	Angular separation between h/Z muons.	4.841×10^{-2}
11	tAngleWH12	Angle between h/Z muons.	4.169×10^{-2}
12	tAngleWH1	Angle between W muon and higher p_t h/Z muon.	3.384×10^{-2}
13	tMETH12	Angle between h/Z muon plane and missing E_t.	2.410×10^{-2}
14	tMETW	Angle between W muon and missing E_t.	1.945×10^{-2}
15	tMET	Missing E_t.	1.817×10^{-2}
16	tLpt	p_t of the lowest p_t muon.	1.520×10^{-2}

Table 12. Discriminating variables for Wh and WZ processes, where $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$.

Signal efficiency:	Test sample	Error	Training sample
Bkg. eff. 0.01	0.06	$\pm$ 0.01	0.06
Bkg. eff. 0.10	0.31	$\pm$ 0.03	0.33
Bkg. eff. 0.30	0.61	$\pm$ 0.03	0.63

Table 13. Signal efficiency values for test and training samples at different background efficiency values for Wh and WZ, where $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$.

final state with 1.613×10^{-1} importance. m_t of the highest p_t muon and angle between W muon and plane of muons, which a recognized as coming from either Higgs or Z boson are also highly ranked. Their distributions can be seen in Fig. 39. The difference between signal and background is not so clear as in $Z \to \mu\mu$ case. We can already see, that due to the presence of four more neutrinos in the final state, the reconstructed invariant mass for signal and background processes differ much less than in $h \to \mu\mu$ and $Z \to \mu\mu$ case. Forest contained 30 trees for the $h/Z \to \tau\tau$ separation BDT training and the node size was 5%. The actual differences between Wh signal and WZ background are much smaller for the $h/Z \to \tau\tau$ than for the $h/Z \to \mu\mu$ processes.

The linear correlation coefficients for input variables of Wh and WZ processes, where $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$ are listed in Fig. 40.

This means that the risk of BDT overtraining due to the fluctuations is much higher so the number of trees for this case was selected much smaller. This allowed to avoid overtraining as seen by Kolmogorov-Smirnov probability in Fig. 41. The BDT response values for signal and background are also accumulating at values closer to zero resulting in much more overlap for signal and background. This indicates that Wh and WZ processes, where $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$ are not classified with high accuracy. Efficiency obtained from training and test samples are also

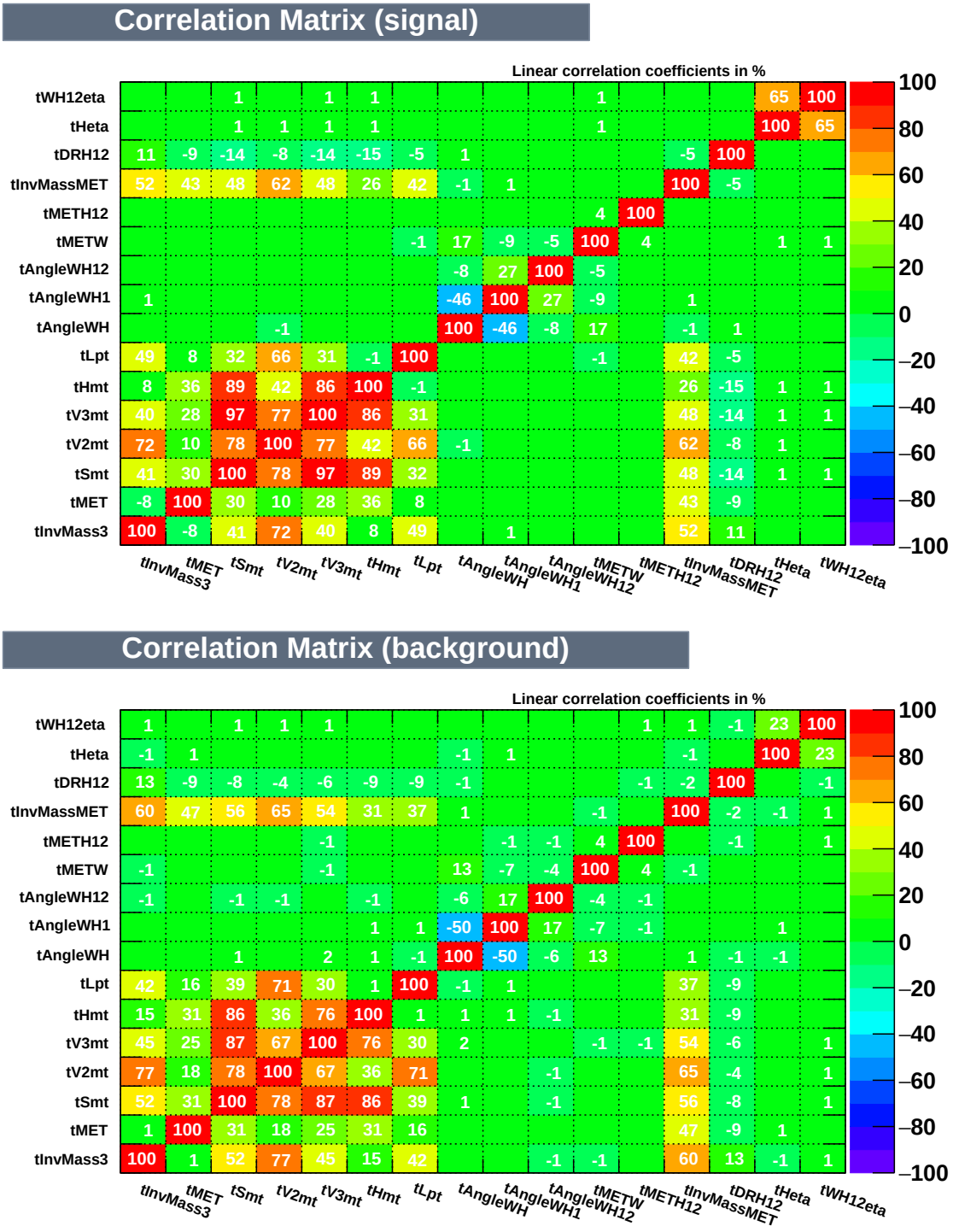

Fig. 40: Correlation matrices for the signal Wh (*top*) and the background WZ (*bottom*) processes, where $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$.

within their error limits so the BDT should not be signifantly overtrained as seen in Table 13.

From BDT response in Fig. 42 we can see as expected, that the signal and background separation for $h \to \tau\tau$ and $Z \to \tau\tau$ case is worse than $h \to \mu\mu$ and $Z \to \mu\mu$ case. By comparing the signal efficiencies we see that in this case we have 78% background rejection for the 50 % signal efficiency (as opposed to the 98% for the $h/Z \to \mu\mu$) and 48% background rejection for the 80% signal efficiency (92% for the $h/Z \to \mu\mu$). The MVA results from the test sample adjusted by the luminosity, production cross-sections and branching ratios are investigated in the following chapter.

4.2.2 MVA results for $h/Z \to \tau\tau \to \mu\mu + neutrinos$

The same as for the $h/Z \to \mu\mu$ case, for the $h/Z \to \tau\tau \to \mu\mu + neutrinos$ processes signal and background events can be adjusted according to their production cross-section, so the expected significance can be calculated.

The resulting signal efficiency, background efficiency and significance dependance on the BDT output cut value can be seen in Fig. 43. The optimal cut value can be selected. Since the signal significance tends to be stable around the optimal cut value, the BDT cut can be chosen at higher signal efficiency. The chosen cut value is $BDT > -0.1$ for 10 fb^{-1} and 100 fb^{-1} luminosities. This corresponds to 75% signal efficiency and the 50% background efficiency. The reduction of background events for the test sample of corresponding luminosities after making the cut on the BDT output is represented in the invariant mass histograms in Fig. 44.

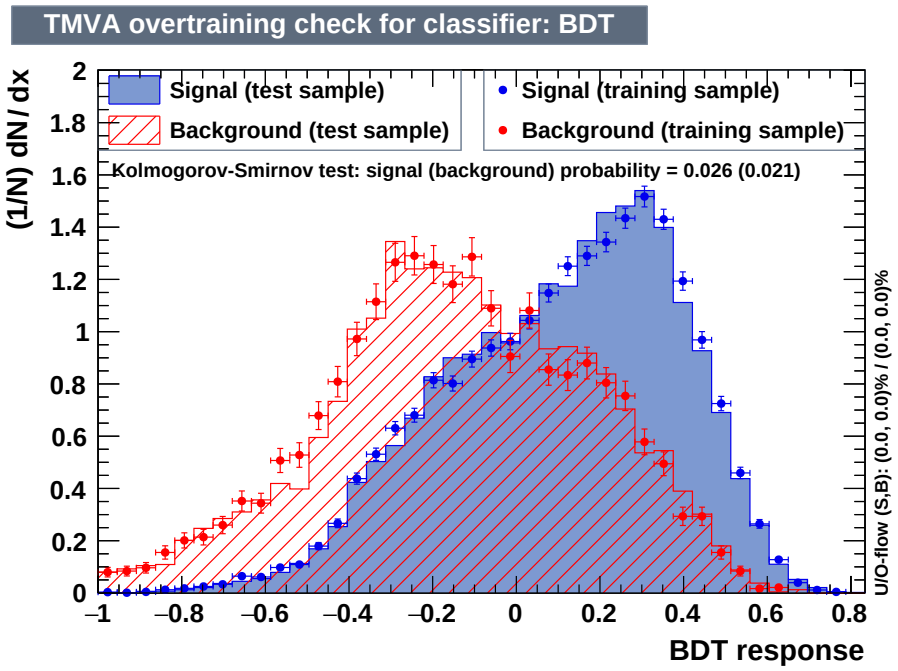

Fig. 41: Overtraining check for BDT for Wh and WZ processes, where $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$.

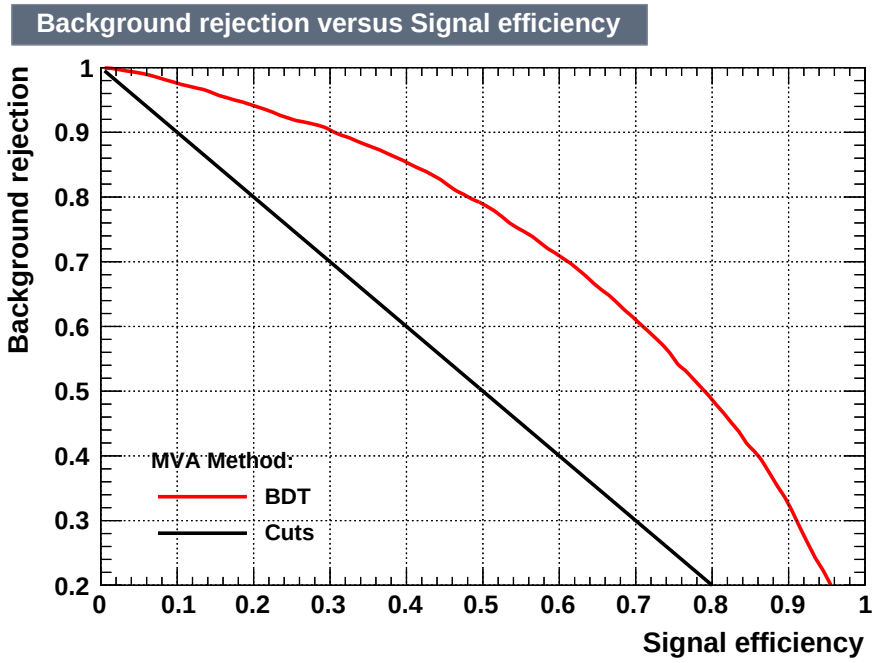

Fig. 42: Receiver operating characteristic diagram for BDT trained for Wh and WZ processes with $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$.

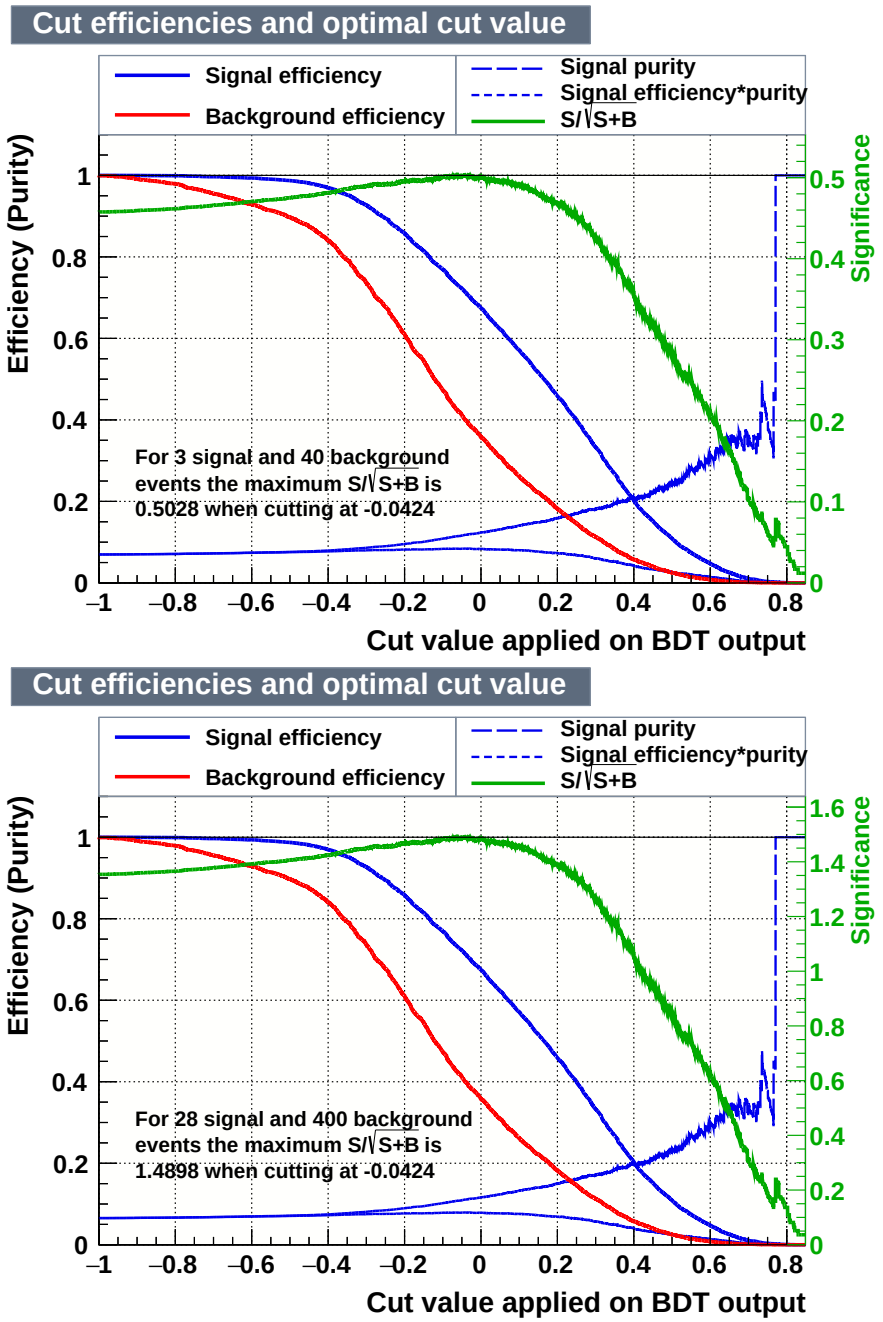

Fig. 43: Signal efficiency dependence on BDT output cut for Wh and WZ processes with $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$, for adjusted by the expected number of events test sample. The luminosities are 10 fb^{-1} (*top*) and 100 fb^{-1} (*bottom*).

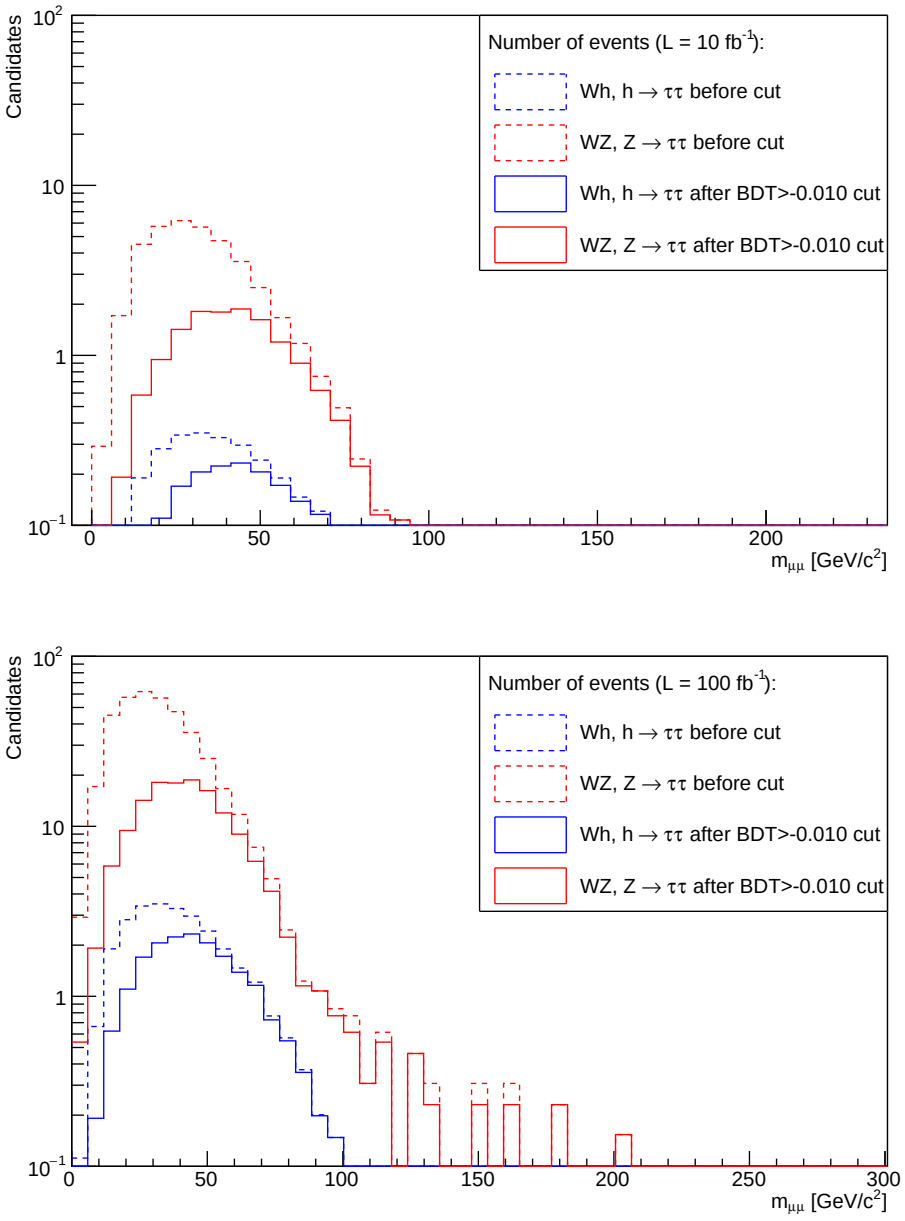

Fig. 44: Number of signal and background events before and after making the cut on the BDT output value. The *top* plot is for the cut value of -0.1 for 10 fb^{-1} luminosity. The *bottom* plot is for the cut value of -0.1 for 100 fb^{-1} luminosity. Both graphs represent Wh and WZ processes with $W \to \mu\nu$, $Z \to \tau\tau$ and $h \to \tau\tau$.

5 Conclusions

Most background processes have a different signature than the signal Wh with $W \to \mu\bar{\nu}_\mu$, $h \to \tau^+\tau^-$ with $\tau \to \mu\bar{\nu}_\mu\nu_\tau$ and signal Wh with $W \to \mu\bar{\nu}_\mu$, $h \to \mu\mu$. So, reducing them in the upcoming data analysis should not be a problem. However, there is a part of the background that is irreducible. These processes are highly similar to the signal process: WZ with $W \to \mu\bar{\nu}_\mu$, $Z \to \tau^+\tau^-$ with $\tau \to \mu\bar{\nu}_\mu\nu_\tau$ and $Z \to \mu^+\mu^-$ respectively.

By analyzing data samples produced with PYTHIA Monte Carlo simulation we found sixteen discriminating variables for signal and background processes. The analyzed discriminant variables provide useful information about the differences between the signal and this irreducible background process. However, each of the discriminants on their own is not powerful enough to separate these two processes with high accuracy. There is a need to combine these discriminants in a multivariate analysis.

We used calculated variables to train multivariate analysis method Boosted Decision Trees and applied them to the independent test samples adjusted by the luminosity, respective cross-sections and branching fractions. Due to the small number of expected signal events the search for signal is complicated. However, trained BDT reduced the WZ with $W \to \mu\bar{\nu}_\mu$, $Z \to \tau^+\tau^-$ and WZ with $W \to \mu\bar{\nu}_\mu$, $Z \to \mu^+\mu^-$ backgrounds.

The **conclusions** from this thesis findings are the following:

1. Identifying the Wh with $W \to \mu\bar{\nu}_\mu$, $h \to \tau^+\tau^-$ with $\tau \to \mu\bar{\nu}_\mu\nu_\tau$ from the background WZ with $W \to \mu\bar{\nu}_\mu$, $Z \to \tau^+\tau^-$ with $\tau \to \mu\bar{\nu}_\mu\nu_\tau$ is more complicated than the Wh with $W \to \mu\bar{\nu}_\mu$, $h \to \mu^+\mu^-$ from WZ with $W \to \mu\bar{\nu}_\mu$, $Z \to \mu^+\mu^-$ due to larger number of neutrinos in the final state (five neutrinos versus one neutrino).

2. Boosted decision trees are the most suitable method for this analysis due to their relatively short training time and good classification performance (they outperformed the Cuts and other classification methods. For the 50% signal efficiency the BDT reached 98% and 78% background rejections for each signal case while the Cuts method reached only 50% background rejection in each case).

3. Reducing the number of the trees in a BDT forest is important if the trees in the forest are being overtrained (for the $h \to \mu^+\mu^-$ signal case 150 trees were used while for $h \to \tau^+\tau^-$ signal case due to smaller differences between the signal and the background the number was reduced to 30 trees to avoid overtraining), however it can result in reduced classification performance.

4. Optimal cut value on the BDT output has to be chosen in order to achieve the maximum possible signal significance, however due to the small number of signal events the selection of BDT output cut value also depends on the expected number of signal events at that cut value

(for $h \to \mu^+\mu^-$ the chosen cut value is $BDT > 0.35$, at this value the signal efficiency is 40% and background efficiency is less than 5%. For $h \to \tau^+\tau^-$ case the cut value $BDT > -0.1$ was chosen which corresponds to 75% signal efficiency and the 50% background efficiency).

Monika Venčkauskaitė

IDENTIFICATION OF THE W-BOSON ASSOCIATED HIGGS BOSON PRODUCTION EVENTS WITH THE CMS DETECTOR AT 13 TEV PROTON COLLISION ENERGY

Summary

Determining the Higgs boson properties is of high importance in particle physics today.

For this analysis we selected the Higgs boson decay channel $h \to \tau^+\tau^-$ with τ leptons decaying each into a muon and neutrinos and $h \to \mu^+\mu^-$ channel. The selected Higgs boson production mechanism is Higgs boson production in association with the W boson. Therefore, the most complicated background process to discriminate is the Z boson production with the W boson where the signature is very similar: $Z \to \tau^+\tau^-$ with τ leptons decaying each into a muon and neutrinos and $Z \to \mu^+\mu^-$. The W boson in all the mentioned processes decays into a muon and a neutrino. The goal of this analysis was to find discriminating variables for these two processes and use them to train the multivariate analysis method. We performed the analysis with the Monte Carlo simulations of the signal and the most important background processes. We found sixteen discriminating variables between the Wh and WZ processes. These discriminants and the correlations between them were used to train multivariate analysis method Boosted Decision Trees. The BDT were tested with independent test samples adjusted by the luminosity, respective cross-sections and branching fractions.

These trained boosted decision trees will be used to analyze the data in proton-proton collisions at 13 TeV detected during the LHC Run II.

Monika Venčkauskaitė

HIGSO IR W BOZONŲ VIENALAIKIO SUSIDARYMO ĮVYKIŲ IŠSKYRIMAS CMS DETEKTORIUMI ESANT 13 TEV PROTONŲ SUSIDŪRIMO ENERGIJAI

Santrauka

Higso bozono savybių nustatymas yra vienas iš svarbiausių šiandieninės dalelių fizikos tikslų, įgyvendinamų CERN laboratorijoje Didžiojo hadronų greitintuvo eksperimentuose.

Šiame darbe analizuojami Higso bozono skilimo kanalai $h \to \tau^+\tau^-$, o kiekvienas τ leptonas skyla į miuoną ir neutrinus, bei $h \to \mu^+\mu^-$. Pasirinktas Higso bozono vienalaikio susidarymo kartu su W bozonu procesas. Sudėtingiausiai atskiriamas foninis procesas yra Z bozono susidarymas kartu su W bozonu. Šio foninio proceso skilimo grandinėlės labai panašios į signalo: $Z \to \tau^+\tau^-$ (čia kiekvienas τ leptonas skyla į miuoną ir neutrinus) bei $Z \to \mu^+\mu^-$. Visais atvejais pasirinktas W bozono skilimas į miuoną ir neutriną. Tiriamo proceso tikimybė yra mažesnė nei foninių procesų.

Šios analizės pagrindinis tikslas: surasti diskriminuojančius kintamuosius šiems dviems procesams ir panaudoti juos mašininio mokymosi daugiadimensinės analizės metodo apmokymui. Analizė atlikta su Monte Carlo simuliacijų duomenimis abiems signalo procesams ir svarbiausiems foniniams procesams. Surasta šešiolika diskriminantinių kintamųjų Wh ir WZ procesams. Pastarieji kintamieji ir jų koreliacijos buvo panaudoti daugiadimensinės analizės metodo - sprendimų medžių apmokymui. Sprendimų medžiai pasirinkti dėl greito jų apmokymo ir geros klasifikacijos. Šis daugiadimensinės analizės metodas yra jautrus fliuktuacijoms, tad svarbu jį suderinti - parinkti medžius ir jų mišką aprašančius parametrus tokius, kad metodas atpažintų signalo ir fono bruožus ir nepriskirtų jiems fliuktuacijų.

Procesams $Wh, h \to \mu^+\mu^-$ ir $WZ, Z \to \mu^+\mu^-$ pasirinktas miško dydis yra 150 medžių, medžių gylis - 3 šakojimosi lygiai, kiekviename suklasifikuotame įvykių krepšelyje mažiausiai 5% visų įvykių. $Wh, h \to \mu^+\mu^-$ ir $WZ, Z \to \mu^+\mu^-$ įvykių klasifikacija geresnė nei antrojo signalo atvejo, tačiau šiame kanale fono įvykių daugiau (13167) ir labai mažas tikimasis signalo įvykių skaičius (3) $L_{\text{int}} = 100$ fb^{-1} protonų susidūrimų intensyvumui (angl. luminosity). Pasirinktas miško dydis $Wh, h \to \tau^+\tau^-$ ir $WZ, Z \to \tau^+\tau^-$ atveju yra 30 medžių, medžių gylis - 3 šakojimosi lygiai, kiekviename suklasifikuotame įvykių krepšelyje mažiausiai 5% visų įvykių. Medžių skaičius daug mažesnis nei pirmuoju atveju, kad būtų išvengta apmokymo fliuktuacijomis. Signalo ir fono klasifikacija prastesnė nei pirmuoju atveju. Dėl didesnio neutrinų skaičiaus galutinėje būsenoje $Wh, h \to \tau^+\tau^-$ ir $WZ, Z \to \tau^+\tau^-$ procesus atskirti sudėtingiau, tačiau šiuo atveju tikimasi signalo įvykių daugiau (28) ir fono įvykių skaičius mažesnis (400) $L_{\text{int}} = 100$ fb^{-1} protonų susidūrimų intensyvumui.

Sprendimų medžių apmokymas, kad būtų išvengta apmokymo fliuktuacijomis tikrinamas dviem būdais: tikrinant pasiekiamas signalo ir fono efektyvumo vertes nepriklausomiems apmokymo ir

tikrinimo įvykių rinkiniams - jos turi sutapti paklaidų ribose. Taip pat tikrinama Kolmogorovo-Smirnovo tikimybė signalo ir fono procesams abiejų signalų atvejais.

Signalo ir fono procesų atskyrimo sprendimų medžių išvesties parametro vertė pasirenkama taip, kad būtų pasiekiamas maksimalus reikšmingumas, arba, jei reikšmingumas yra tam tikrame intervale yra stabilus - jo vertė, esanti kuo arčiau maksimalios su didesniu signalo įvykių skaičiumi (signalo efektyvumu). Šiame darbe pasirinktos parametrų pjūvių vertės yra: 0.35 $Wh, h \to \mu^+\mu^-$ ir $WZ, Z \to \mu^+\mu^-$ procesams atitinkanti 40% signalo efektyvumą esant mažesniam nei 5% fono efektyvumui bei -0.1 $Wh, h \to \tau^+\tau^-$ ir $WZ, Z \to \tau^+\tau^-$ procesams atitinkanti 75% signalo efektyvumą esant 50% fono efektyvumui.

Sprendimų medžių klasifikavimo veikimas buvo patikrintas su nepriklausomais Monte Carlo simuliacijų duomenimis, kuriuose įvykių skaičius buvo parinktas pagal protonų susidūrimų intensyvumą L_{int}, susidarymo skerspjūvius ir skilimo santykius. Apmokytas daugiadimensinės analizės metodas bus naudojamas analizuojant protonų susidūrimus antrajame LHC veikimo etape esant 13 TeV protonų susidūrimų energijai.

References

[1] Particle Data Group, *Review of Particle Physics*, Chin. Phys. C. **38**, 090001 (2014)
http://pdg.lbl.gov/

[2] S. F. Novaes, *Standard Model: An Introduction*, (2000)
https://arxiv.org/abs/hep-ph/0001283

[3] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716**, 1, (2012)

[4] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B **716**, 30, (2012)

[5] L. Evans and P. Bryant, *LHC Machine*, J. Instrum. **3** S08001, (2008)

[6] F. Englert and R. Brout, Broken Symmetry and the Mass of Gauge Vector Mesons, Phys. Rev. Lett. **13**, (1964)
http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.13.321

[7] P. W. Higgs, Broken symmetries, massless particles and gauge fields, Physics Letters **12**, (1964)
http://www.sciencedirect.com/science/article/pii/0031916364911369

[8] ATLAS Collaboration and CMS Collaboration, Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments, Phys. Rev. Lett. **114**, 191803 (2015)
http://journals.aps.org/prl/pdf/10.1103/PhysRevLett.114.191803

[9] M. Venčkauskaitė, *Prospects for the W-boson associated Higgs boson production with three leptons in the final state at 13 TeV with CMS*, (2016)

[10] ATLAS Collaboration, The ATLAS experiment at the CERN Large Hadron Collider, J. Instrum. **3**, (2008).

[11] CMS collaboration, The CMS experiment at the CERN LHC, J. Instrum. **3**, S08004, (2008)

[12] C. De Melis, *The CERN accelerator complex. Complexe des accélérateurs du CERN*, OPEN-PHO-ACCEL-2016-001, (2016)
http://cds.cern.ch/record/2119882

[13] T. Sakuma and T. McCauley, Detector and Event Visualization with SketchUp at the CMS Experiment, J. Phys.: Conf. Ser. **513**
http://iopscience.iop.org/article/10.1088/1742-6596/513/2/022032

[14] D. Barney and E. Quigg, *Interactive Slice of the CMS detector*, CMS-doc-4172-v2, (2010) https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=4172

[15] CMS Collaboration, CMS reconstruction improvement for the muon tracking by the RPC chambers, J. Instrum. **3**, (2013) https://arxiv.org/pdf/1306.6905v2.pdf

[16] G. S. Guralnik, C. R. Hagen and T. W. B. Kibble, Global Conservation Laws and Massless Particles, Phys. Rev. Lett. **13**, (1964) http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.13.585

[17] M. Carena *et. al*, *Status of Higgs boson physics*, (2014). http://pdg.lbl.gov/2015/reviews/rpp2015-rev-higgs-boson.pdf

[18] LHCPhysics, *SM Higgs production cross sections at* $\sqrt{s} = 13 - 14$ *TeV (CERN Report 3)*, (2016). https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageAt1314TeV2014

[19] LHC Higgs Cross Section Working Group, *Standard Model Higgs-Boson Branching Ratios with Uncertainties*, (2011). http://arxiv.org/pdf/1107.5909v2.pdf

[20] CMS Collaboration, *CMS Physics : Technical Design Report Volume 1: Detector Performance and Software*, CERN-LHCC-2006-001, CMS-TDR-8-1, (2006) http://cds.cern.ch/record/922757/files/lhcc-2006-001.pdf

[21] T. Sjöstrand, S. Mrenna and P. Skands, A Brief Introduction to PYTHIA 8.1, Comput. Phys. Comm. **178** 852, (2008). http://arxiv.org/abs/0710.3820

[22] T. Sjöstrand, S. Mrenna and P. Skands, PYTHIA 6.4 Physics and Manual, JHEP05 **026**, (2006). http://arxiv.org/abs/hep-ph/0603175

[23] N. Davidson *et al.*, *Universal Interface of TAUOLA*, IFJPAN-IV-2009-10, (2010).

[24] J. Allison *et al.*, Geant4 developments and applications, IEEE Transactions on Nuclear Science **53**, (2006)

[25] CMS collaboration, *Performance of CMS muon reconstruction in pp collision events at* $\sqrt{s} = 7$ *TeV, 2013*, CMS-MUO-10-004, (2013). http://arxiv.org/pdf/1206.4071v2.pdf

[26] J. F. Owens *et al.*, Parton Distribution Functions of Hadrons, Annu. Rev. Nucl. Part. Sci. **42**, (1992)

[27] H. L. Lai *et al.*, Global QCD analysis of parton structure of the nucleon: CTEQ5 parton distributions, Eur. Phys. J. C *12*, (2000)

[28] CMS collaboration, Measurement of the top quark pair production cross section in proton-proton collisions at $\sqrt{s_{NN}} = 13$ TeV, Phys. Rev. Lett. **116**, 052002, (2015).
http://arxiv.org/pdf/1510.05302.pdf

[29] CMS Collaboration, *Measurement of the WZ production cross section in pp collisions at* $\sqrt{s_{NN}} = 13$ *TeV*, CMS-PAS-SMP-15-006, (2015).
http://inspirehep.net/record/1409819/

[30] ROOT a Data analysis Framework.
http://root.cern.ch/

[31] A. Hoecker *et al.*, *Toolkit for Multivariate Data Analysis with ROOT Users Guide*, (2007)
http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf

[32] B. P. Roe *et al.*, Boosted Decision Trees as an Alternative to Artificial Neural Networks for Particle Identification, Nucl. Instrum. Meth. **A543** (2005)
http://arXiv:physics/0408124

[33] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Second Edition, Springer, Stanford, California, (2008).

[34] I. Narsky, *StatPatternRecognition: A C++ Package for Statistical Analysis of High Energy Physics Data*, (2005).
http://arXiv:physics/0507143

[35] Kolmogorov-Smirnov test in ROOT,
https://root.cern.ch/doc/master/classTH1.html#TH1:KolmogorovTest

VENČKAUSKAITĖ, Monika. *Identification of the W-boson associated Higgs boson production events with the CMS detector at 13 TeV proton collision energy.* Vad. dr. Andrius Juodagalvis. Vilnius: Vilniaus universitetas, Teorinės fizikos ir astronomijos institutas, 2016, 58 p.