GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Fakultät für
Physik

Bachelor’s Thesis

Optimierung der Unterdrückung von Jet-Untergrundprozessen in der Identifizierung von hadronisch zerfallenden Tau-Leptonen am Atlas-Detektor

Optimization of Jet Background Suppression for the Identification of Hadronically Decaying Tau Leptons with the Atlas Detector

prepared by

Nils Gillwald

from Göttingen

at the II. Physikalischen Institut

Thesis number:	II.Physik-UniGö-BSc-2016/03
Thesis period:	4th April 2016 until 11th July 2016
First referee:	Prof. Dr. Stan Lai
Second referee:	Prof. Dr. Arnulf Quadt

Abstract

This thesis deals with the optimization of jet background suppression for the identification of hadronically decaying τ leptons with the ATLAS detector. For this purpose, a boosted decision tree is trained for Monte Carlo simulated 1-prong as well as 3-prong hadronic τ decays. The overtraining and efficiencies of the algorithms are investigated, and also the correlation of the used identification variables is reviewed. Finally, the effect on removing variables from the stable version is discussed.
The best stable configuration reaches a background rejection of 96.72% for the 1-prong and of 99.34% for the 3-prong case, both at a signal efficiency of 50%.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Optimierung der Unterdrückung von Jet-Untergrundprozessen in der Identifizierung von hadronisch zerfallenden τ Leptonen am ATLAS-Detektor. Zu diesem Zweck wird ein verstärkter Entscheidungsbaum auf der Grundlage von Monte Carlo generierten hadronischen 1- und 3-prong τ Zerfällen trainiert. Das Overtraining und die Effizienz des Algorithmus werden untersucht, genauso wie die Korrelation der verwendeten Identifikationsvariablen. Zum Schluss wird diskutiert, welchen Effekt das weitere Entfernen von Identifikationsvariablen aus dem stabilen Algorithmus hat.
Die beste stabile Konfiguration unterdrückt 96.72% der Untergrundereignisse für den 1-prong und 99.34% der Untergrundereignisse für den 3-prong Fall, bei einer Effizienz von 50% für die Signalereignisse.

Contents

Nomenclature

Variables

Variable	Meaning	Unit
m	(invariant) mass	GeV
p_T	transverse momentum	GeV
θ	polar angle	°, rad
ϕ	azimuthal angle	°, rad
η	pseudorapidity	–
ΔR	distance in $\eta - \phi$ space	–
p_μ	four vector	GeV

Abbrevations

Abbrevation	Meaning
τ_{had}	hadronically decaying τ leptons
$\tau_{had-vis}$	visible part of τ_{had}
SM	Standard Model or Standard Model of particle physics
q	quark
ℓ	lepton
g	gluon
QCD	Quantum Chromodynamics
GSW	Glashow, Salam and Weinberg
γ	Photon
NC	Neutral Current
H	Higgs boson
ggF	gluon gluon fusion
VBF	vector boson fusion

Abbrevation	**Meaning**
ISR	initial state radiation
ATLAS	experiment at CERN
CERN	european organization for nuclear research
LHC	Large Hadron Collider
SUSY	supersymmetry
EM	electromagnetic
TRT	transition radiation tracker
LAr	liquid Argon
L1	Level-1 trigger
L2	Level-2 trigger
EF	event filter
HLT	high-level trigger
RoI	Regions-of-Interest
TopoClusters	three dimensional clusters of calorimeter cells
LC	local hadronic calibration
core region	region within $\Delta R < 0.2$ around a direction
TV	τ lepton production vertex
TauID	τ lepton identification
isolation region	ring with radius $0.2 < \Delta R < 0.4$ around a direction
MC	Monte Carlo
PDF	particle distribution function or probability density function
DT	decision tree
BDT	boosted DT
NTree	maximum number of trees in a BDT
MaxDepth	maximum number of cuts in a DT
MinNodeSize	minimal size of a DT node
TMVA	toolkit for multivariate data analysis
ROOT	data analysis program commonly used in particle physics
KS	Kolmogorov-Smirnov
BkgRej@50	background rejection at 50% signal efficiency
ROC curve	receiver operating characteristics curve
intROC	integral over the ROC curve

1 Introduction

The following thesis deals with the discrimination of hadronically decaying τ leptons (τ_{had}) from jet backgrounds at the ATLAS detector. This discrimination is necessary in order to investigate properties of the Higgs boson in the $H \to \tau\tau$ channel. Here, the H-τ Yukawa coupling is of interest, as the $H \to \tau\tau$ channel is the only accessible lepton decay channel so far. This discrimination is especially challenging, as - among other reasons - only the decay products of the τ lepton will be recorded by the detector.

The theoretical backgrounds for this thesis will be introduced in Chapter 2, where the standard model of particle physics (SM) will be discussed, as well as the Higgs mechanism and Higgs boson production and decays. Chapter 3 will explain the setup of the LHC and the ATLAS experiment, including important details of the detector. Properties of the τ lepton and hadronic τ lepton decays will be discussed. Subsequently, the τ lepton reconstruction algorithm in ATLAS and the τ lepton identification variables will be presented in Chapter 4. The used dataset will also be discussed there. The analysis dealing with finding the best performing stable boosted decision tree (BDT) for the discrimination of τ_{had} from jets will be described in Chapter 5, including explanations on boosted decision trees and the tools that were used. A short summary of the results will be given in Chapter 6.

2 Theoretical Concepts

2.1 The SM of particle physics

The Standard Model of Particle Physics is the most accurate and precise theory that describes the known fundamental particles and their interactions with each other. It summarizes the results of over 100 years of research, beginning with the discovery of the electron by J. J. Thomson in 1897 [1, 2].

As illustrated in Figure 2.1, the particles in the SM are divided into two different groups: the bosons, being spin-1 particles, and the fermions, being spin-$\frac{1}{2}$ particles. The bosons are the mediators of the forces and the fermions are the fundamental particles of which matter is made. The latter can again be divided into the quarks (q) and leptons (ℓ), based upon the interactions in which they take part. The coupling of the different forces in the SM or more precisely of the mediating bosons with the particles is shown in Table 2.1. The strong interaction couples to quarks, and not to any leptons. It is reasonable to try to merge the electromagnetic and weak interaction into a single unified electroweak interaction, not least because using only a single $SU(2)_L$ local gauge group to describe the weak interaction does not correctly describe the coupling of the Z^0 boson to particles. This was accomplished by the introduction of a new theory by Glashow, Salam and Weinberg (GSW) in the 1960s [3–5]. In this unified theory, a new quantity, the hypercharge $Y = 2(Q - T_3)$ is introduced, with Q the electric charge and T_3 the third component of weak isospin. The theory also respects the left chiral coupling of the $W^{\pm}$ to particles. The gauge group for this local gauge

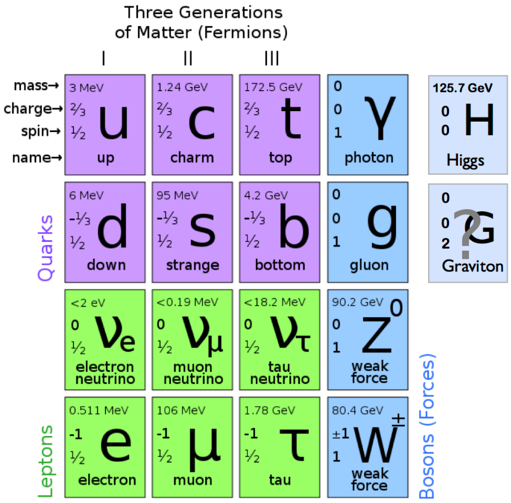


Figure 2.1: The particle content of the Standard Model of Particle Physics.

force	mediating boson	charge	coupling to
electromagnetic	γ	electric	charged leptons, quarks, W
weak	$W^{\pm}$, Z^0	weak isospin	leptons, quarks, Z, W
strong	g	colour	quarks, gluons

Table 2.1: Summary of the forces of the SM, their mediating bosons, charges and couplings.

theory is $SU(2)_L \times U(1)_Y$. Applying local gauge transformations to this theory yields four massless vector bosons, which can be identified as the $W^{\pm}$, the Z^0 and the γ.

The model arranges the elementary particles in left-handed chiral doublets indicated by the first three columns in Figure 2.1 and separated into quark and lepton doublets. Apart from the masses, each doublet has the same properties: The particles in the top row have a weak isospin third component value T_3 of $+\frac{1}{2}$, the ones in the bottom row have a weak isospin third component value of $-\frac{1}{2}$. The right-handed particles are arranged in singlets and have a weak isospin value of 0. As the Z^0- and $W^{\pm}$-bosons couple to the third component of the weak isospin of a particle, they do only couple to the left-handed fermions. The W boson is able to simultaneously couple to quarks from different flavours and thus to change the flavour of quarks. The flavour changes are described with the Cabibbo-Kobayashi-Maskawa (CKM) Matrix [6, 7], which will not be discussed here. For neutral currents (NC), mediated by the Z^0 or γ, a flavour change is not possible.

Photons (γ) couple to the electric charge of a particle. Up-type quarks have an electric charge Q of $+\frac{2}{3}\,\mathrm{e}$ and down-type quarks of $-\frac{1}{3}\,\mathrm{e}$. The neutrinos do not carry an electric charge and the charged leptons have an electric charge of $-1\,\mathrm{e}$.

The gluon (g) is the exchange particle of the strong interaction, coupling to a particle property called colour, coming in red, blue and green, described by a quantum number that follows an SU(3) group structure. Only quarks and the gluons themselves carry this property. There are eight different gluons, distinguished by the combination of colour they carry. In the theory of the strong interaction, called Quantum Chromodynamics (QCD), it is only possible for quarks to form bound states with other quarks because it is not possible to form coloured states and they therefore cannot exist as free particles. This so-called confinement, an artefact of the running coupling constant of the strong force rising for small energies, yields mesons formed out of two and baryons formed out of three quarks. Together with the leptons, the mesons and baryons form the "observable matter" in the universe.

Finally, the Higgs mechanism, which is responsible for the acquisition of mass of the SM particles implies the existence of an additional particle, the Higgs boson (H). The Higgs boson and the Higgs mechanism are essential concepts of the SM and are described in the following sections.

2.1.1 Electroweak symmetry breaking and the Higgs mechanism in the SM of particle physics

One way to include masses for the vector bosons in the SM is to add mass terms of the form

$$\mathcal{L}_{mass} = \frac{1}{2} m_V V_\mu V^\mu \tag{2.1}$$

to the SM Lagrangian. Unfortunately, these terms break the local gauge invariance of the SM Lagrangian for both bosons and fermions, the bosons and fermions are required to be massless. This is in contradiction to experimental observations, for instance the measured masses of the $W^\pm$ boson and the Z^0 boson [8–10].

The solution to the problem of generating masses was provided by P. Higgs [11] as well as R. Brout, F. Englert [12], G. Guralnik, C. Hagen and T. Kibble [13] in 1964 by introducing a new scalar spin-0 field ϕ, today known as the Higgs field.

This complex scalar field ϕ is incorporated in a weak isospin doublet

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + \mathrm{i}\phi_2 \\ \phi_3 + \mathrm{i}\phi_4 \end{pmatrix} \tag{2.2}$$

which is called the Higgs doublet. It provides four additional degrees of freedom to the theory. Adding a potential of the form

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda \left(\phi^\dagger \phi\right)^2 \tag{2.3}$$

to the Lagrangian, the Higgs doublet induces spontaneous symmetry breaking. The parameters of the potential must fulfil $\lambda > 0$ and $\mu^2 < 0$, which is shown in Figure 2.2. The minima of the potential are then found to be on a circle with radius $\frac{v^2}{2} = -\frac{\mu^2}{2\lambda}$ around the origin of the coordinate system (cp. Fig. 2.2) and the physical vacuum state will be located at one arbitrary point on this minima circle. The vacuum expectation value for ϕ can thus be chosen as

$$\langle 0| \phi |0\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} . \tag{2.4}$$

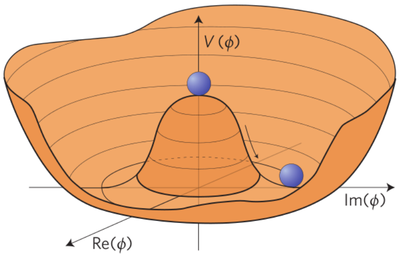


Figure 2.2: Sketch of the Higgs potential for $\lambda > 0$ and $\mu^2 < 0$. The ring of minima is at $\frac{v^2}{2}$.

The scalar fields can then be expanded about the vacuum state, yielding

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1(x) + \mathrm{i}\phi_2(x) \\ v + h(x) + \mathrm{i}\phi_4(x) \end{pmatrix} \tag{2.5}$$

with the Higgs field $h(x)$ and three additional massless Goldstone bosons $\phi_{1,2,4}(x)$. Using the so-called unitary gauge, the latter will disappear, contributing the longitudinal degrees of freedom needed for the massive $W^{\pm}$ and Z^0 bosons. The Higgs doublet is written as

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} . \tag{2.6}$$

in this gauge. Now, the mass terms of the theory need to be identified. Firstly, the Lagrangian must respect the $SU(2)_L \times U(1)$ group symmetry which yields

$$\mathcal{L} = \left(D_\mu\phi\right)^\dagger \left(D^\mu\phi\right) - V\left(\phi\right) \tag{2.7}$$

$$\text{with the covariant derivative} \quad D_\mu\phi = \left(\partial_\mu + \mathrm{i}g_W T_a \cdot W_\mu^a + \mathrm{i}g'\frac{Y}{2}B_\mu\right)\phi\,.$$

Here, $T_a = \frac{1}{2}\sigma_a$ are the Pauli matrices and g_W, g' are the couplings of the $SU(2)_L$ and the $U(1)$, respectively. The W_μ^a and B_μ are the four degrees of freedom which will be superposed to form the mass terms for the $W^{\pm}$ and the Z^0.

To generate the masses for the fermions, the left-handed chiral ones are placed in $SU(2)$ doublets whilst the right-handed chiral ones are placed in $SU(2)$ singlets. It is possible to show that the combination of a left-handed doublet $\bar{L} = L^\dagger\gamma^0$ with the Higgs doublet, $\bar{L}\phi$, is invariant under $SU(2)_L$ gauge transformations, and that this term multiplied with a right-handed singlet from the right, $\bar{L}\phi R$, is invariant under both $SU(2)_L$ and $U(1)$ gauge transformations. Therefore, terms of the form

$$-g_f \left(\bar{L}\phi R + \bar{R}\phi^\dagger L \right) \tag{2.8}$$

where g_f is the Yukawa coupling of the fermions to the Higgs field can be added to the SM Lagrangian, respecting the symmetry of the theory while providing mass and interaction terms for the fermions.
All in all, the theory yields well-defined masses for all particles in the SM

$$m_W = \frac{1}{2} g_W v \ , \quad m_A = 0 \ , \quad m_Z = \frac{1}{2} v \sqrt{g_W^2 + g'^2} \ , \quad m_f = \frac{1}{\sqrt{2}} v g_f \ . \tag{2.9}$$

The remaining fourth degree of freedom of the Higgs doublet corresponds to the physical Higgs boson mass.

2.2 The Higgs boson

The SM Higgs boson is a CP-even spin-0 particle. Its mass is determined by the parameters of the Higgs potential from Equation 2.3 [10],

$$m_H = \sqrt{2\lambda} v \quad \text{with} \quad v \approx 246\,\text{GeV}\,, \tag{2.10}$$

and not predicted by the SM. In experiment, a value of $m_H = 125.09 \pm 0.24\,\text{GeV}$ is measured [14].
The following sections describing the production, decay and discovery of the Higgs boson rely on [10, 15–17] if not stated otherwise.

2.2.1 Production mechanisms

The LHC is a proton-proton collider. Therefore, the Higgs boson will be produced by processes involving gluons or quarks, respectively. The leading order Feynman graphs of the most important production mechanisms are shown in Figure 2.3 and are gluon gluon fusion (ggF), vector boson fusion (VBF), Higgsstrahlung and associated Higgs top production. As shown in Figure 2.4, the ggF ($pp \to H$ in the Figure), where two gluons merge into a Higgs boson via a virtual top loop (and with a much lower probability other quark loops) is the dominant Higgs production channel at all energies. In about half the events, there will be QCD initial state radiation (ISR) seen in the detector for this process due to the high probability of gluons emitting further gluons, in addition to the Higgs

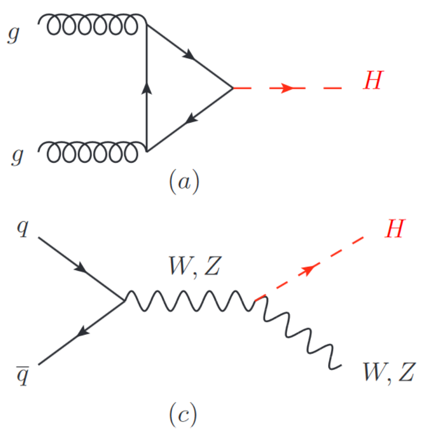


Figure 2.3: Main Higgs production mechanisms at the LHC at leading order. Shown are (a) gluon gluon fusion (b) vector boson fusion (c) Higgs strahlung (d) associated Higgs top production.

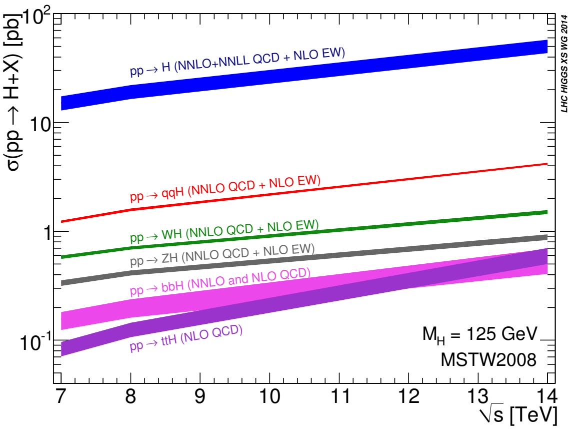


Figure 2.4: Cross section for different Higgs production processes as a function of the center of mass energy $\sqrt{s}$ [18].

boson decay signature.

The process with the second highest production cross section is VBF ($pp \to qqH$ in the Figure), where two quarks each emit a $W^{\pm}$ or Z^0 boson, which then merge into a Higgs boson. The quarks will form forward jets with a rapidity gap.

The production cross sections of Higgsstrahlung ($pp \to H(W/Z)$ in the Figure), where a Higgs boson is produced in association with a $W^{\pm}$ or Z^0 and associated Higgs top production ($pp \to ttH$ in the Figure), where a Higgs boson is produced in association with two top quarks, are even lower.

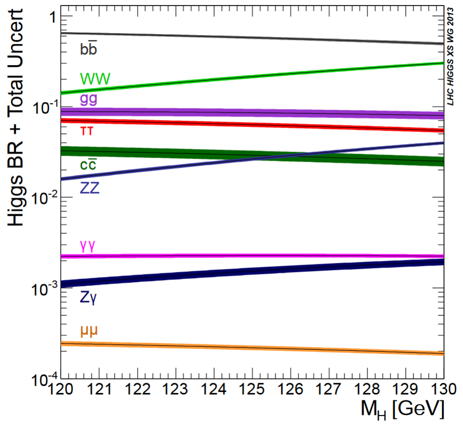


Figure 2.5: Branching ratios of the different Higgs decay channels as a function of the Higgs mass [10]. The theoretical uncertainties are indicated as a band.

Decay channel	Branching ratio
$H \to b\bar{b}$	5.77×10^{-1}
$H \to WW^*$	2.15×10^{-1}
$H \to gg$	8.57×10^{-2}
$H \to \tau\tau$	6.32×10^{-2}
$H \to c\bar{c}$	2.91×10^{-2}
$H \to ZZ^*$	2.64×10^{-2}
$H \to \gamma\gamma$	2.28×10^{-3}
$H \to \gamma Z$	1.54×10^{-3}
$H \to \mu\mu$	2.19×10^{-4}

Table 2.2: Branching ratio of different Higgs boson decay channels for a Higgs mass of 125 GeV, adapted from [10].

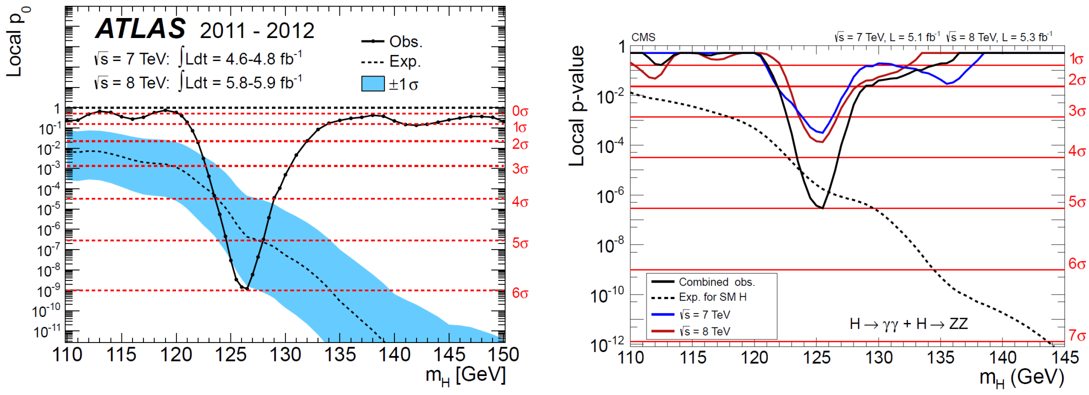
2.2.2 Decay modes

The main Higgs boson decay modes are shown in Figure 2.5, with the corresponding absolute cross section values for a Higgs boson mass of 125 GeV given in Table 2.2. The main decay mode of the $m_H = 125\,\text{GeV}$ Higgs boson is the $H \to b\bar{b}$ mode, because a $b\bar{b}$ pair is the heaviest SM particle pair which has a lower mass than the Higgs boson itself. Thus, it has the highest possible Yukawa coupling to the Higgs boson whilst it is still possible to produce to real b quarks in the decay. The decay with the second highest branching ratio is the decay of the Higgs boson into one real and one virtual W boson, followed by a decay into two gluons. The latter is again moderated by a quark loop which is dominated by top quarks.

The decay into a τ pair is the fourth most common decay mode and will be discussed below in more detail. The Higgs boson can also decay into a pair of Z^0 bosons, γ or a mix of both, where the modes containing photons must also be moderated by a virtual loop.

2.2.3 Discovery

In 2012, both the ATLAS and CMS experiments discovered a new particle [19, 20]. The associated plots are shown in Figure 2.6. They show the observed and expected p_0 values with the resulting significances for a SM Higgs boson signal as a function of the Higgs boson mass m_H. An excess of the local significance of the data of $5.9\,\sigma$ for ATLAS and



(a) ***(b)***

Figure 2.6: Significances of the observation of the Higgs boson for different Higgs masses in the (a) ATLAS experiment [19] (b) CMS experiment [20].

$5\,\sigma$ for CMS is found for a Higgs boson mass hypothesis of $m_H = 126.5\,\mathrm{GeV}$ and $m_H = 125.5\,\mathrm{GeV}$, respectively [19, 20]. Every measured property of this new particle is consistent with the predicted SM Higgs boson this far.

The main discovery channels were $H \to \gamma\gamma$ and $H \to Z^0 Z^{0*} \to \ell\ell\ell\ell$ with $\ell = e, \mu$, as these channels have a very clear signature in the detector which is well distinguishable from background processes. The $H \to b\bar{b}$ and the $H \to \tau\tau$ channel suffer from large backgrounds yielding a low sensitivity. This applies also to the $H \to WW^* \to \ell\nu\ell\nu$ channel, additionally to uncertainties coming from the unknown neutrino energy.

The latest combined measurement of the Higgs mass by ATLAS and CMS yields $m_H = 125.09 \pm 0.21\,(\mathrm{stat}) \pm 0.11\,(\mathrm{sys})\,\mathrm{GeV}$ [14].

2.2.4 $H \to \tau\tau$ Decays

The Higgs boson decays in a pair of τ leptons in $6.3\,\%$ of all cases. The latest measurement of ATLAS of the signal strength μ is $1.43^{+0.43}_{-0.37}$ times the SM expectation, as shown in Figure 2.7 [21]. ATLAS and CMS combined measure a signal strength of $\mu = 1.12^{+0.25}_{-0.23}$ for the $H \to \tau\tau$ decay, and observe the decay with a measured significance of $5.5\,\sigma$ [22]. ATLAS and CMS used the full Run I data set for the analysis, corresponding to an integrated luminosity per experiment of $5\,\mathrm{fb}^{-1}$ for the 2011 data with $\sqrt{s} = 7\,\mathrm{TeV}$ and $20\,\mathrm{fb}^{-1}$ for the 2012 data with $\sqrt{s} = 8\,\mathrm{TeV}$.

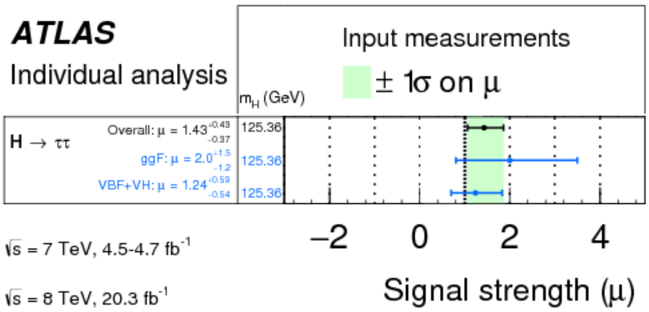


Figure 2.7: Measurement of the $H \to \tau\tau$ signal strength. The overall signal strength (black) is the combination of the single signal strengths (blue) assuming SM values for the cross section ratios. The error bands represent $\pm 1\sigma$ intervals and the green bands are the uncertainty on the overall signal strength obtained by each analysis. Adapted from [21].

3 The LHC and the ATLAS detector

This thesis is written within the context of the Atlas Collaboration which is a collaboration at Cern, Geneva. Cern is the biggest particle research center in Europe with 21 member nations and several thousand people working on different research projects.

3.1 The LHC

With a circumference of 27 km, the Lhc is the main accelerator ring at Cern, colliding protons. As shown in Figure 3.1, the protons are injected in the Lhc after several stages of pre-acceleration: First, they are accelerated in LINAC2. Then, they pass the booster, the proton synchrotron and the super proton synchrotron as the last pre-acceleration stage.

When two protons collide at energy ranges as that used in the Lhc, in fact the different partons of the protons - quarks, antiquarks and gluons - collide, and not point-like protons. The Lhc uses a magnetic field of up to 8.33 T for the single beams to force the protons on their path and reaches a center of mass energy of 14 TeV in Run II.

To reach the design luminosity of $10^{34}\,\text{cm}^{-2}\text{s}^{-1}$, the Lhc uses proton bunches containing up to 10^{11} protons. The bunches are travelling nearly at the speed of light and are colliding every 25 ns.

There are four experiments operating at the Lhc. Alice, Atlas and Lhcb are located in the southern part of the ring, whereas Cms is located in the northern part. Atlas and Cms search amongst other things for physics beyond the SM and have discovered the Higgs boson in 2012 [23], but also measure properties of already known particles like the Higgs boson or top quark. Lhcb studies CP-violation and Alice investigates lead-ion collisions to understand the physics of the early universe. In difference to the protons, the lead-ions are pre-accelerated using LINAC3 and LEIR.

The Atlas experiment will be described in more detail in the following section.

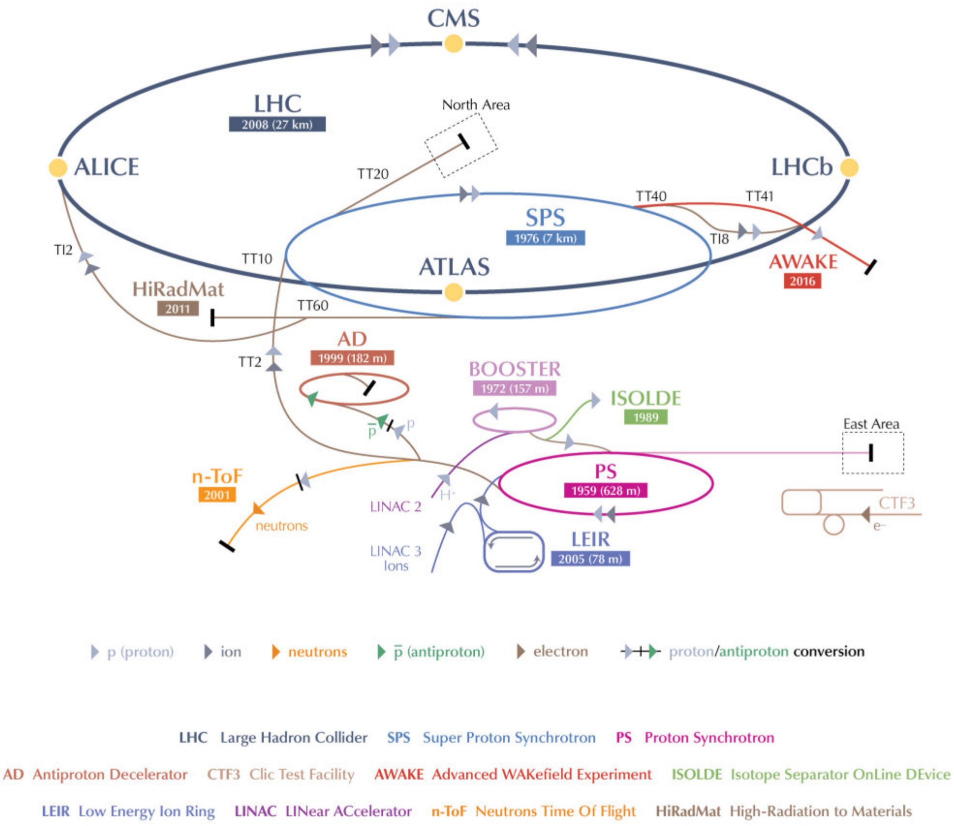


Figure 3.1: Schematic view of the CERN accelerator complex in the year 2013. ©2008-2017 CERN.

3.2 The ATLAS Detector

Amongst other things, ATLAS is designed to find new particles such as the Higgs boson or SUSY particles. As pictured in Figure 3.2, the detector has a cylindrical form and consists of a barrel part in the central region and two end caps, closing the detector on both sides. It has a height of 25 m and a width of 44 m, weighing about 7 kt.

3.2.1 Coordinate system

ATLAS uses a right handed coordinate system, where the origin is placed at the collision point of the bunches. The x-axis points to the center of the LHC ring, the y-axis to the top and the z-axis accordingly in the beam direction. The xy-plane is thus orthogonal to the proton beam and defines the transversal momentum

$$p_T = \sqrt{p_x^2 + p_y^2}\,. \tag{3.1}$$

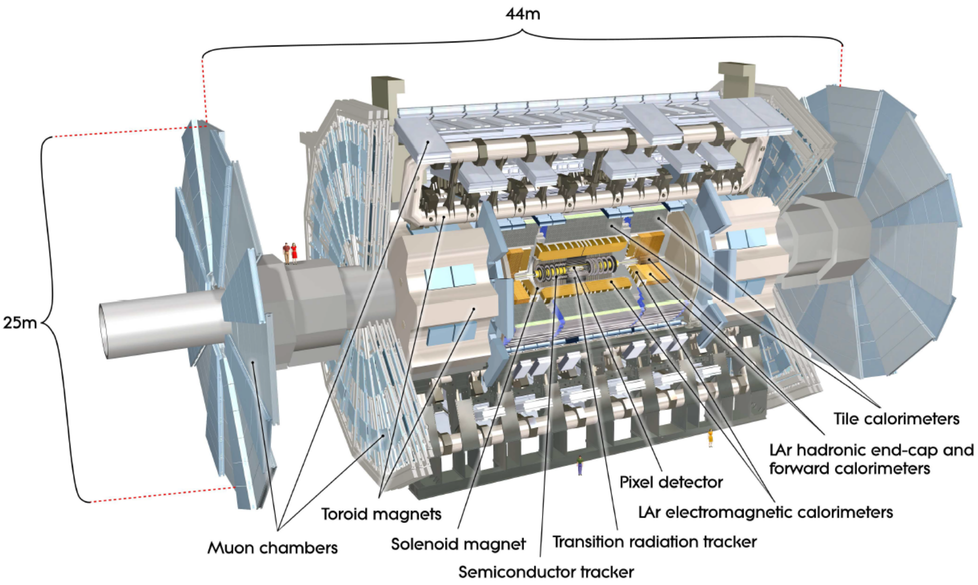

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Transition radiation tracker
Muon chambers
Solenoid magnet
Semiconductor tracker

Figure 3.2: Cut-away view of the ATLAS detector [24].

Using a Cartesian coordinate system is quite impractical in most tasks the ATLAS detector is used for. Therefore, ATLAS uses three dimensional cylindrical coordinates, where θ is the polar angle from the beam axis z and ϕ is the azimuthal angle. Often, the pseudorapidity

$$\eta = -\ln\tan\frac{\theta}{2} \tag{3.2}$$

is used instead of θ, as differences in the pseudorapidity $\Delta\eta$ are invariant under Lorentz transformations in z-direction in the high energy approximation $m \ll E$. This is especially useful as the boost of the center-of-mass system is not a priori known in hadron colliders. Differences in the $\eta - \phi$ space are defined as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{3.3}$$

and are also invariant under boost along the z-axis in this approximation.

3.2.2 Detector components

The detector exhibits an onion-like structure. Starting from the beam pipe, there is a layer of different tracking detectors (inner detector), magnets, the electromagnetic (EM) and the hadronic calorimeter and finally muon chambers.

The inner detector consists of three components. First, there are silicon pixel detectors with sensor sizes of $50 \times 400\,\mu\mathrm{m}^2$ [24]. Using cylindrical coordinates, they are segmented in $R - \phi$ and z, where the z coordinate is the beam axis. The second layer is provided by silicon microstrip trackers with a size of $80\,\mu$m × 12 cm each. The third inner detector layer consists of transition radiation trackers (TRT), consisting of 4 mm diameter straw tubes. On the one hand, the TRT is used as a tracking detector, but on the other hand it also helps in particle identification, discriminating mainly electrons from pions. This is possible as the amount of transition radiation photons is proportional to the γ-factor of the trespassing particles, and particles with a different mass having a different γ-factor at the same energy level.

The whole inner detector is immersed in a 2 T solenoidal magnetic field parallel to the beam axis [25]. Charged particles moving in this field will be forced to move on a circular path in the transverse plane, yielding information about the momentum of the particle from the measurement of the path curvature. The relative uncertainty of the momentum measurement is proportional to the momentum itself, as the curvature of the path gets smaller with higher momentum of the particle [26].

The next stage in the ATLAS detector structure are the EM and hadronic calorimeters. The EM calorimeter is used to measure the energy of incoming particles which interact mainly electromagnetically, except for Muons which will be discussed below. Hadrons just interact lowly with the EM calorimeter and are stopped and measured in the hadronic calorimeter. The EM and hadronic calorimeters of ATLAS are both sampling calorimeters, which means that they consist of alternating layers of an absorbing and an active material. The latter detects and measures the particle showers.

The EM calorimeter uses liquid argon (LAr) as the active and lead as the absorber material [25] arranged in an "accordion structure", which provides a full ϕ coverage of the detector without any gaps. It covers a region of $|\eta| < 1.475$ with the barrel and $1.375 < |\eta| < 3.2$ with the end-cap. The hadronic calorimeter uses steel as the absorbing and scintillating ceramics as the active material [25]. It covers a region of $|\eta| < 1.7$ with the barrel and $1.5 < |\eta| < 3.2$ with the end-caps. The end-caps of both the EM and hadronic calorimeter are made of LAr, copper and tungsten. An additional region of $3.1 < |\eta| < 8.2$ is covered by LAr forward calorimeters on both sides of the detector.

As Muons travel through all other detector layers without having a significant energy loss, the muons must be detected in the outermost layer of the detector, the muon chambers. These are high-precision tracking chambers using drift tubes inside a large air-core toroid magnet.

The resolutions of the different detector parts are summarized in Table 3.1. All particles which cannot be detected in one of the mentioned detector parts, such as neutrinos or eventual neutral exotic particles, will leave the detector and can only be seen as missing

Detector part	Resolution
Inner detector	$\sigma_{p_T}/p_T = 0.05\,\% \times p_T \oplus 1\,\%$
EM Calorimeter	$\sigma_E/E = 10\,\%/\sqrt{E} \oplus 1\,\%$
Hadronic Calorimeter	$\sigma_E/E = 50\,\%/\sqrt{E} \oplus 0.03$
Muon Chambers	$\sigma_{p_T}/p_T = 10\,\%$ for $p_T = 1\,\text{TeV}$

Table 3.1: Resolutions of the different Atlas detector parts [24]. The transverse momentum p_T and the energy E are given in GeV.

transverse momentum $p_{T,miss} = \sqrt{p^2_{x,miss} + p^2_{y,miss}}$ due to momentum conservation in the transverse plane.

3.2.3 The trigger system

Every 25 ns, or at a rate of 40 MHz, there is a proton collision in the Atlas detector. An event, as it is used in the analysis, is a snapshot of such a collision. But just a small fraction of all collisions are of interest for further investigation. Therefore, a trigger system is applied in order to distinguish between interesting and uninteresting events.

The Run I Atlas event trigger consisted of three levels of event selection: The Level-1 (L1), Level-2 (L2) and event filter (EF). The L2 and the EF were software-based, whilst the L1 was hardware-based.

First, the L1 searched for signatures of high-p_T muons, electrons, photons, jets and τ_{had}, as well as missing transverse energy $\not{E}_T$ and large total transverse energy E_T. Special trigger chambers were used for muons and all calorimeter sub-systems with a reduced granularity for the other particles. The L1 reduced the event rate from 40 MHz to 75 kHz and indicated Regions-of-Interest (RoI) to the L2 trigger.

The L2 trigger further investigated these RoI, where the L1 trigger had identified possible trigger objects in the event. In contrast to the L1, L2 had access to the whole detector information. It reduced the event rate to below 3.5 kHz. The EF used offline analysis methods on the whole event. It reduced the event rate to about 200 Hz, which could be recorded for further offline analysis.

In Run II, the L1 trigger remains unchanged, but the L2 and the EF are combined to the high level trigger (HLT). It analyses the event using either the RoI or the full detector information to refine the event selection. Threshold cuts, which decide if an event gets selected for further investigation, are improved by better information on energy deposits, while the particle identification is enhanced by track reconstruction. The event rate is reduced from about 100 kHz to about 1 kHz.

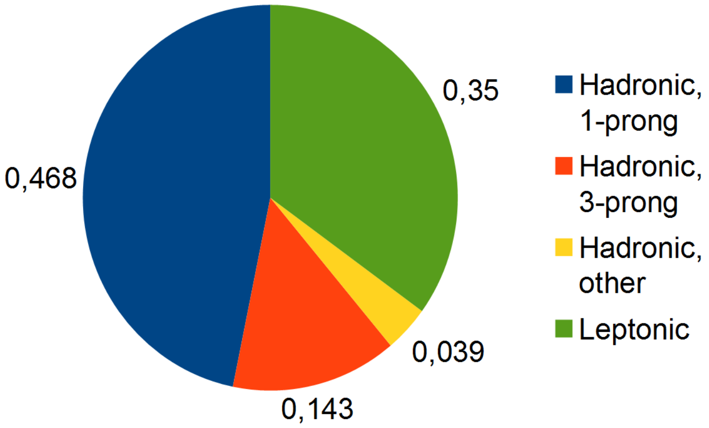


Figure 3.3: Branching ratio of τ lepton decays.

3.2.4 τ leptons in the ATLAS detector

The τ lepton is the heaviest of the three known leptons with a mass of $1,776.82$ MeV [10]. It has a mean lifetime of only 2.9×10^{-13} s, yielding a typical decay length of $87.03\,\mu$m. Therefore, the vast majority of all τ leptons produced at the ATLAS experiment will decay before having reached the innermost detector layer, and only the decay products will be measurable.

As shown in Figure 3.3, the τ lepton will decay hadronically in 65 % of all cases, where the decay products are one or three charged pions (1-prong and 3-prong events) in 72 % and 22 % of all cases, respectively, and mostly Kaons in the rest of the cases. In the remaining 35 % of all cases, the τ lepton will decay leptonically. This case will not be considered in this thesis.

The main background to τ_{had} are high-energetic jets resulting from hadronization of gluons and quarks. These may be produced by ISR in an actual process containing τ leptons in the final state, but can also be processes where the jet is misidentified as a τ lepton. To distinguish such background events from proper τ lepton decays, discriminating variables must be introduced. They will be discussed in the next chapter.

4 Reconstruction and Identification of τ leptons in ATLAS for LHC Run II at 13 TeV

4.1 τ lepton reconstruction algorithm

To reconstruct the visible part of hadronically decaying τ leptons ($\tau_{had-vis}$), the anti-k_t algorithm is used with the distance parameter R set to 0.4. Three dimensional clusters of calorimeter cells (TopoClusters) calibrated at local hadronic calibration (LC) are used as input to the algorithm. They are used as seeds of the $\tau_{had-vis}$ reconstruction algorithm if they satisfy $p_T > 10\,\mathrm{GeV}$ and $|\eta| < 2.5$. The p_T of the $\tau_{had-vis}$ candidate is set to the total energy of the TopoCluster within $\Delta R < 0.2$ (core region) around the $\tau_{had-vis}$ direction.
Tracks are associated to the $\tau_{had-vis}$ candidate, when they are in the core region around $\tau_{had-vis}$ direction, have a $p_T > 1\,\mathrm{GeV}$, have at least two hits in the pixel detector and at least seven hits in the pixel and strip detectors combined. Also, the track must not be further away from the τ lepton production vertex (TV) than 1.0 mm in the transverse plane and should satisfy $|\Delta z_0 \sin\theta| < 1.5\,\mathrm{mm}$ longitudinally.
The TV is identified among the previously reconstructed primary vertex candidates in the event to reduce pile-up effects and to increase the reconstruction efficiency. The TV association algorithm takes all τ lepton candidate tracks in the region $\Delta R < 0.2$ around the $\tau_{had-vis}$ jet seed direction into account, sums the p_T of these tracks and then defines the TV as the one with the largest fraction of the p_T sum matched to it. The TV is used to determine the $\tau_{had-vis}$ direction, calculate new impact parameters, associate tracks and build the coordinate system in which the τ lepton identification (TauID) variables are calculated.
Investigations of the reconstruction efficiency, defined as the fraction of $\tau_{had-vis}$ decays which are reconstructed, show that the detector is almost fully efficient for finding a jet-seed for a $\tau_{had-vis}$ lepton in the acceptance region of $p_T > 20\,\mathrm{GeV}$, $|\eta| < 2.5$ and $|\eta|$ outside [1.37, 1.52] [27].

Sample	case	# Events	Parton Shower Model	Matrix Element	Detector model
$Z^0 \to \tau\tau$	1-prong	69,051	Pythia 8.186	CT10 PDF	Geant 4
	3-prong	28,514	Pythia 8.186	CT10 PDF	Geant 4
$Z^0 \to ee$	1-prong	64,483	Pythia 8.186	CT10 PDF	Geant 4
$+jets$	3-prong	137,786	Pythia 8.186	CT10 PDF	Geant 4

Table 4.1: Properties and details of the used MC samples for the discrimination of τ_{had} from jets. The number of events is given after applying the cuts.

4.2 Monte Carlo Samples

In this thesis, a boosted decision tree (BDT) is used as the τ lepton identification algorithm. For the training of the BDT, a Monte Carlo (MC) simulated sample is used. In particle physics, MC samples are used to generate multiple random events or four vectors p_μ of final states, respectively, at a certain energy regime. They interpret the matrix element squared $|\mathcal{M}|^2$ of the regarded process as a probability function and take the PDF of protons into account to simulate single proton-proton collisions. Then, initial and final state soft QCD radiation (parton shower) and low-energy jets from underlying events (pile-up) are added. Finally, the interaction with the Atlas detector is simulated.
In this thesis, the MC samples contain $Z^0 \to \tau\tau$ signal events generated using Powheg-Box v2 interfaced to the Pythia 8.186 parton shower model. For the matrix element, the CT10 PDF set is used, as well as AZNLO tune, with the PDF set Cteq 6L1 for the modeling of non-perturbative effects. Bottom and charm quark decays are simulated using the Evtgen v1.2.0 program and Photos++ v3.52 for QED emissions [27].
For the background, $Z^0 \to e^+e^- + jets$ events simulated with the same generator settings as for the signal events are used [27].
The pileup is simulated by overlaying event-by-event some minimum-bias interactions extracted from a Poisson distribution, with a number of interactions per bunch crossing μ with a mean value of 25. The interaction of particles with the Atlas detector is simulated by Geant 4 using the FTFP_BERT hadronic shower model [27].
For an event to be considered as a signal event in the analysis, the leading τ lepton must be truth-matched to a simulated τ lepton. Additionally, the leading τ lepton must fulfil $p_T \geq 20\,\text{GeV}$ as well as $|\eta| < 2.5$ and $|\eta|$ outside [1.37, 1.52] on reconstruction level. The number of charged tracks associated with the leading τ lepton is required to be one for 1-prong events and three for 3-prong events, respectively. For an event to be considered as a background event in the analysis, it must not be truth-matched to an electron or photon and should not be filled with an dummy event with a pdgId of 0, as this pdgId is not connected to any particle. The leading τ lepton should also fulfil the same conditions

Variable	1-prong	3-prong
f_{cent}	●	●
$f^{-1}_{leadtrack}$	●	●
$R^{0.2}_{track}$	●	●
$\lvert S_{leadtrack} \rvert$	●	
f^{track}_{iso}	●	
ΔR_{Max}		●
S^{flight}_{T}		●
m_{track}		●
$f^{track-HAD}_{EM}$	●	●
f^{EM}_{track}	●	●
$m_{EM+track}$	●	●

Table 4.2: TauID variables used as input to the TauID algorithm for 1- and 3-prong events. The bullets indicate if the particular variable is used for the selection. Slightly modified from [27].

for p_T and $|\eta|$ as a signal event.
The properties of the simulated signal and background samples using these cuts are summarized in Table 4.1.

4.3 τ lepton identification variables

The τ lepton identification algorithm is trained against jet backgrounds. This is done in order to be able to investigate e.g. the $H - \tau$ Yukawa coupling, but the algorithm could also be used in identification of τ leptons which do not come from a Higgs decay.
The challenge in identifying τ leptons in the Atlas detector is the short lifetime of the τ lepton. Decaying in average in $2.903(5) \times 10^{-13}\,\mathrm{s}$, the typical τ lepton decays before having reached the innermost detector layer and only the products of the τ lepton decay are detected. This implies an uncertainty on whether a signature in the detector which looks like it comes from a τ lepton decay really comes from such a decay or was faked by background events. To distinguish these two possibilities, different TauID variables were defined. They are based on characteristic properties of τ_{had}, namely the narrower shower shape than gluon or quark jets, the number of charged particle tracks (1 track, 3 tracks, or other) and the displaced τ lepton decay vertex in respect to the τ lepton production vertex. Translating into detector properties, the variables exploit information from the tracks in the tracking detector and the TopoClusters in the core and isolation region around the $\tau_{had-vis}$ candidate direction. The core region is defined as the cone between radius $R < 0.2$ around the $\tau_{had-vis}$ candidate direction and the isolation region

as the ring with radius $0.2 < R < 0.4$ around the same direction.
The ATLAS TauID algorithm uses twelve TauID variables which are discussed below. Eleven of them will be investigated in this thesis and are summarized in Table 4.2. A separation of the use of variables between 1- and 3-prong events is necessary as not all TauID variable definitions make sense for either of these cases. The twelfth variable was not accessible by the end of the thesis for technical reasons. The definitions are taken from [27]:
Central energy fraction (f_{cent}): Fraction of the calorimeter transverse energy deposited in the region $\Delta R < 0.1$ with respect to all energy deposited in the region $\Delta R < 0.2$ around the $\tau_{had-vis}$ candidate. It is calculated by summing the energy deposited in all cells belonging to TopoClusters with a barycentre in these regions, calibrated at the EM energy scale.
Leading track momentum fraction ($f^{-1}_{leadtrack}$): The transverse energy sum, calibrated at the EM energy scale, deposited in all cells belonging to TopoClusters in the core region of the $\tau_{had-vis}$ candidate, divided by the transverse momentum of the highest-p_T charged particle in the core region.
Track radius ($R^{0.2}_{track}$): p_T-weighted ΔR distance of the associated tracks to the $\tau_{had-vis}$ direction, using only tracks in the core region.
Leading track IP significance ($|S_{leadtrack}|$): Absolute value of transverse impact parameter of the highest-p_T track in the core region, calculated with respect to the TV, divided by its estimated uncertainty.
Fraction of tracks p_T in the isolation region (f^{track}_{iso}): Scalar sum of the p_T of tracks associated with the $\tau_{had-vis}$ candidate in the region $0.2 < \Delta R < 0.4$ divided by the sum of the p_T of all tracks associated with the $\tau_{had-vis}$ candidate.
Maximum ΔR (ΔR_{Max}): The maximum ΔR between a track associated with the $\tau_{had-vis}$ candidate and the $\tau_{had-vis}$ direction. Only tracks in the core region are considered.
Transverse flight path significance (S^{flight}_T): The decay length of the secondary vertex (vertex reconstructed from the tracks associated with the core region of the $\tau_{had-vis}$ candidate) in the transverse plane, calculated with respect to the TV, divided by its estimated uncertainty. It is defined only for multi-track $\tau_{had-vis}$ candidates.
Track mass (m_{track}): Invariant mass calculated from the sum of the four-momentum of all tracks in the core and isolation regions, assuming a pion mass for each track.
Fraction of EM energy from charged pions ($f^{track-HAD}_{EM}$): Fraction of the electromagnetic energy of tracks associated with the $\tau_{had-vis}$ candidate in the core region. The numerator is defined as difference between the sum of the momentum of tracks in the core

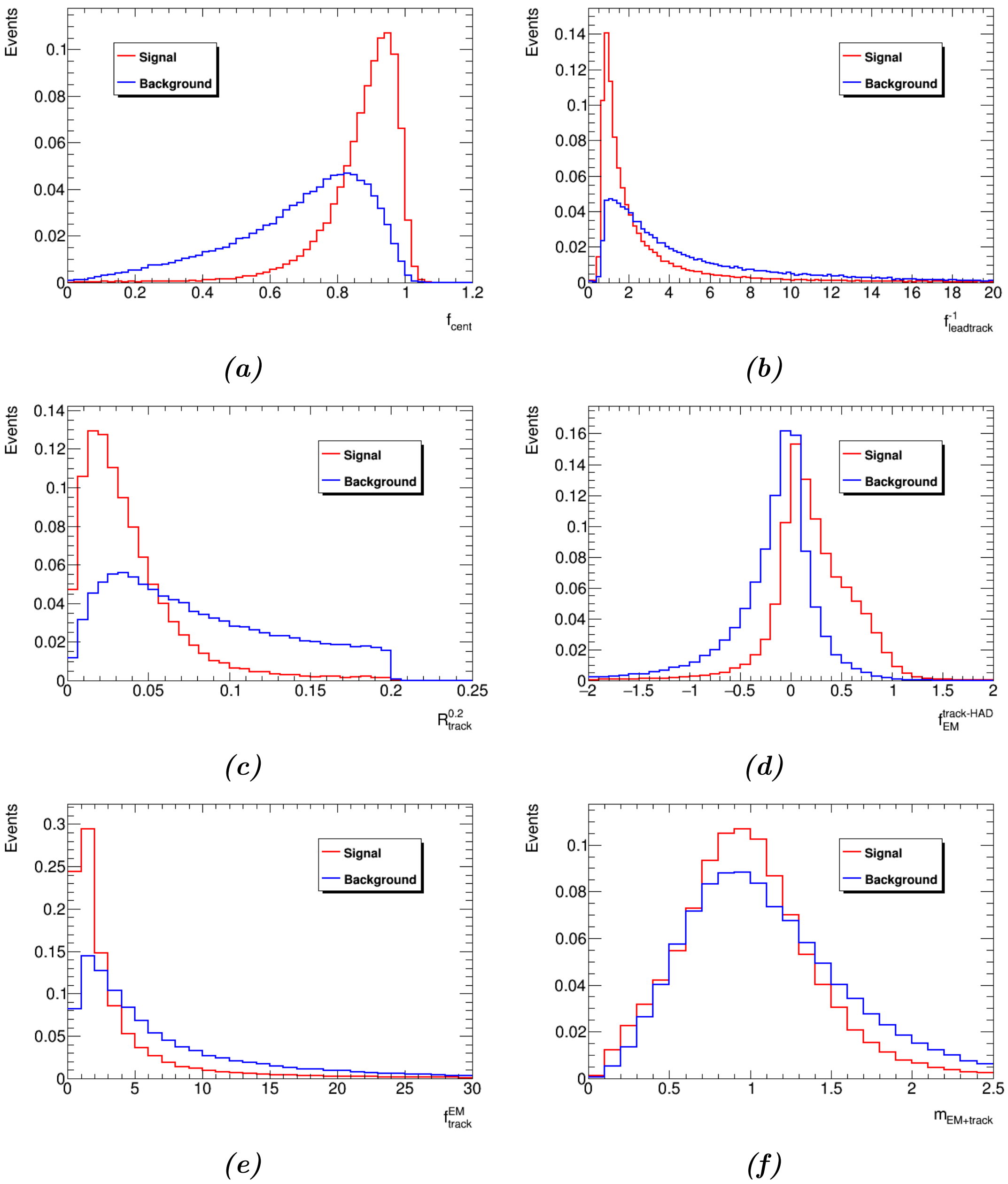
region and the sum of cluster energy deposited in the hadronic part of each TopoCluster (including the third layer of the EM calorimeter) associated with the $\tau_{had-vis}$ candidate. The denominator is the sum of cluster energy deposited in the electromagnetic part of each TopoCluster (presampler and first two layers of the EM calorimeter) associated with the $\tau_{had-vis}$ candidate. All clusters are calibrated at the LC energy scale.

Ratio of EM energy to track momentum (f_{track}^{EM}): Ratio of the sum of cluster energy deposited in the electromagnetic part of each TopoCluster associated with the $\tau_{had-vis}$ candidate to the sum of the momentum of tracks in the core region. All clusters are calibrated at the LC energy scale.

Track-plus-EM-system mass ($m_{EM+track}$): Invariant mass of the system composed of the tracks and up to two most energetic EM clusters in the core region, where EM cluster energy is the part of TopoCluster energy deposited in the presampler and first two layers of the EM calorimeter, and the four-momentum of an EM cluster is calculated assuming zero mass and using TopoCluster seed direction.

Ratio of track-plus-EM-system to p_T ($p_T^{EM+track}/p_T$): Ratio of the $\tau_{had-vis}$ p_T, estimated using the vector sum of track momenta and up to two most energetic EM clusters in the core region to the calorimeter-only measurement of $\tau_{had-vis}$ p_T (not used in this thesis).

After calculating the variables for each event, a correction depending linearly on the average number of pile-up events at the instantaneous luminosity, μ, is applied. The variable distributions plotted from the MC sample presented in Chapter 4.2 are shown in Figures 4.1, 4.2 and 4.3. Comparisons of these distributions can be made with [27] which includes the official Atlas Run II 13 TeV sample distributions, although not all variable distributions are shown there. A comparison to the definitions and distributions of the Atlas Run I identification variables can be made with [28].



(a)

(b)

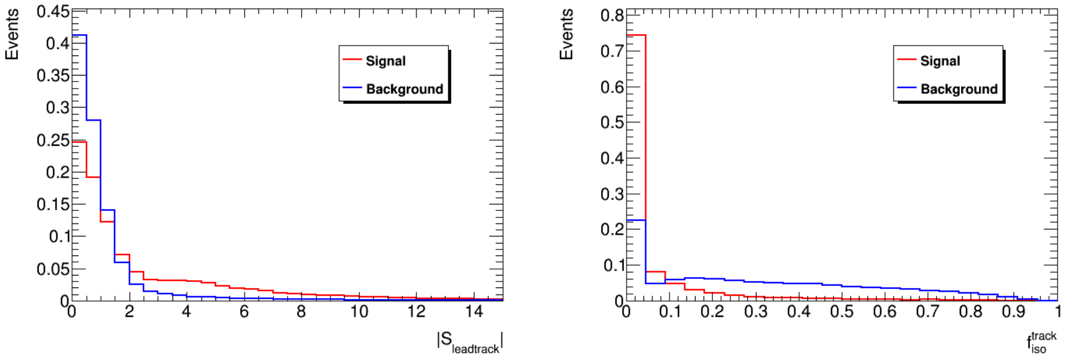
(c)

(d)

(e)

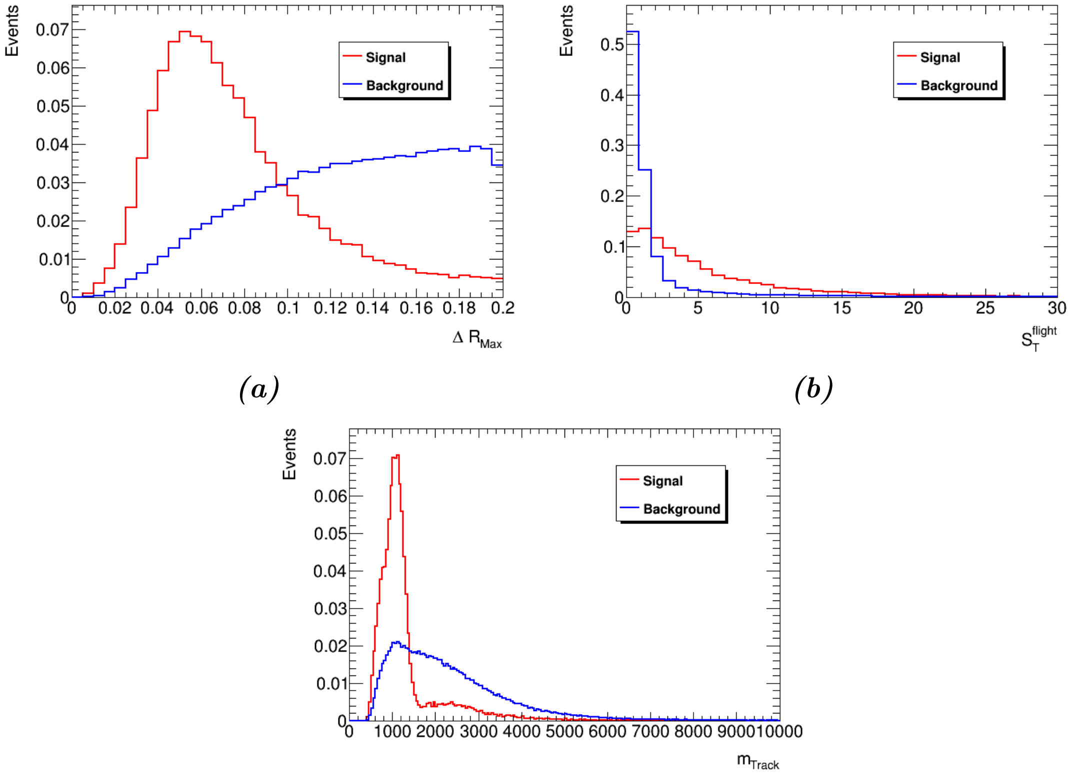
(f)

Figure 4.1: Distributions of the used 1- and 3-prong TauID variables after applying pile-up corrections, plotted from 1-prong events. The values bigger than 1 in e.g. (a) are a result of the pile-up correction.



(a) *(b)*

Figure 4.2: Distributions of the used 1-prong only TauID variables after applying pile-up corrections.



(a) *(b)*

(c)

Figure 4.3: Distributions of the used 3-prong only TauID variables after applying pile-up corrections.

5 Analysis

In the following, the research to find a well-performing stable algorithm which discriminates τ_{had} from jet background is presented. The concept of a BDT and the different analysis steps are discussed.

5.1 Boosted decision trees

A decision tree (DT) must be trained on a data sample in which each of the events is known to be signal or background. As shown in Figure 5.1, a DT is a simple cut-based sequence of criteria to classify an event as either signal or background. In each decision, the variable with the highest separation power is used and every variable can be used as often as needed. In this analysis, the Gini index defined by $p \cdot (1 - p)$ with the purity $p = \frac{\#\text{signal events}}{\#\text{total events}}$ is used to define the separation power. The algorithm selects the variable and cut value that maximizes the increase in the Gini index between the parent node and the sum of indices of the two daughter nodes, weighted by their relative fraction of events [29]. At the end of the DT, signal-like and background-like nodes are obtained, where the class of the majority of events in the final nodes is the defining property.

The single DTs have different stopping parameters, and if specified, the DT can be prevented from adding nodes past a certain depth. There are also other criteria which can be optimized in the analysis in order to find the best stable version of the BDT for separating hadronically decaying τ leptons from jet background. The criteria used for optimizing the BDT in this thesis are the maximum number of trees (NTree), which limits the number of subsequent boosts and thus the number of trees evaluated in the boosting process, the maximal depth (MaxDepth) which limits the number of subsequent cuts in one DT and the minimal node size (MinNodeSize) which limits the minimal number of events in one node as a fraction of the total number of events used in the training sample.

"Boosting" is a procedure to obtain new DTs by reweighting misclassified events in the previous DT, so that they will be more important in the training and application of cuts of the new DT. One thus obtains a so-called forest out of several subsequent DTs and each event obtains a BDT output value which is the sum of outputs of each tree over the

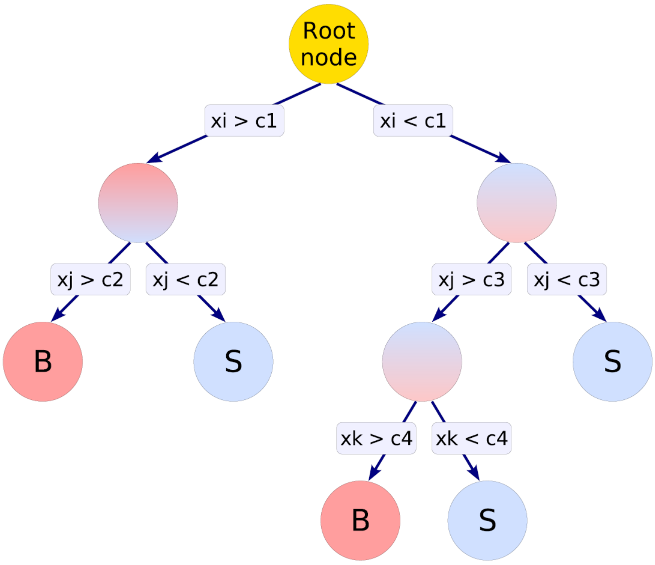


Figure 5.1: Example of a simple decision tree with depth 3, using the cut variables xi, xj and xk. S stands for signal-like and B for background-like nodes.

number of trees in the forest. Assigning each background-like node with a value -1 and each signal-like node with a value +1 in the single DT, the BDT score of an event must also be between -1 and +1. Events which have a BDT score close to +1 are more likely to be signal events and events which have a BDT score close to -1 are more likely to be background events.

5.2 Training a BDT discriminant for tau identification

The analysis aims at finding the best possible separation of hadronically decaying τ leptons (signal) and jet background. For this purpose, the Toolkit for Multivariate Data Analysis (Tmva) [29] is used. It provides an environment to process, evaluate and apply various multivariate analysis techniques integrated in Root [30], a data analysis program commonly used by the Atlas Collaboration and elsewhere in particle physics. In this thesis, the BDT method will be used with AdaBoost [31] as the boosting algorithm. Tmva is given half of the data sample to train the algorithm to separate signal and background events. The other half of events form a statistically independent sample used to assess the performance of the trained BDT. Checks on overtraining can be performed by comparing the BDT output distribution on the training and testing sample through a

Parameter	NTree	MinNodeSize	MaxDepth
Used values	50	2.0	1
	250	4.0	2
	500	6.0	3
	750	8.0	4
	850	10.0	5

Table 5.1: Used values for the different stopping parameters in the BDT training. Each possible combination of the different values is trained and investigated.

Kolmogorov-Smirnov (KS) test. Overtraining of the algorithm occurs when the algorithm regards the training sample in too much detail and becomes sensitive to statistical fluctuations in the training sample. These will be different in the test sample and therefore the output distributions of the BDT will be different if overtraining is present. In other words, the algorithm will become less effective. The KS Test is a measure of consistency of e.g. two data samples, which calculates the probability that these two samples come from the same probability distribution. Thus, the output of the KS Test is uniformly distributed between 0 and 1 for two statistically independent samples that result from the same PDF.

5.3 Tau identification performance

The set of variables investigated for the 1- and 3-prong hadronic τ lepton decays are not the same (cp. Table 4.2). Thus, the actual analysis will distinguish between the 1- and 3-prong τ_{had} decays. This will also yield more effective algorithms for both cases.

5.3.1 1-prong events

To find the best stable BDT configuration, different sets of the stopping parameters NTree, MaxDepth and MinNodeSize are applied in the training of the BDT. The used parameters for each variable are listed in Table 5.1. Every possible combination of the parameters is used in the training. To be able to compare the different training outcomes, three different measures are investigated. First, the KS test number of the signal and background BDT output distribution is considered. If it is below 0.1, the BDT is discarded due to overtraining of the algorithm. Second, the remaining BDTs with the highest background rejection at 50% signal efficiency (BkgRej@50) are selected, where the background rejection is defined as 1 minus the background efficiency. Also, the error on this quantity is regarded. The properties of the best performing BDTs are given in Table 5.2. Overall,

NTree	MinNodeSize	MaxDepth	KSTest Sig	KSTest Bkg	intROC	BkgRej@50
500	8.0	2	0.51	0.19	0.916	0.97 ± 0.03
750	4.0	2	0.25	0.27	0.916	0.97 ± 0.03
750	6.0	2	0.30	0.35	0.915	0.97 ± 0.03
750	8.0	2	0.42	0.39	0.916	0.97 ± 0.03
850	4.0	2	0.17	0.22	0.916	0.97 ± 0.03
850	8.0	2	0.24	0.33	0.916	0.97 ± 0.03

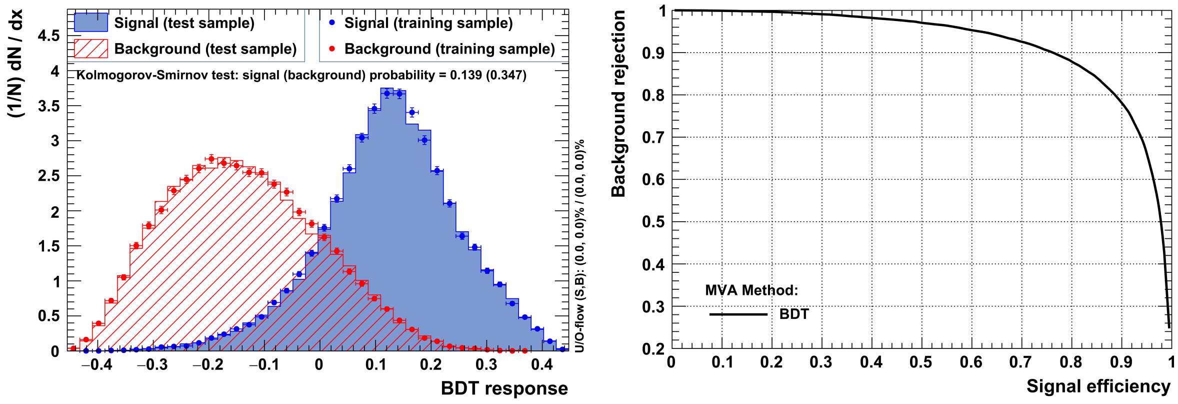
Table 5.2: Properties of the best performing BDTs for the 1prong events.

configurations with a higher number of trees seem to yield a better performance, as well as medium-valued minimal node sizes.

Third, the integral over the receiver operating characteristic (ROC) curve (intROC) is evaluated, which has a theoretical maximum of 1. The ROC curve for the stable 1-prong configuration is shown in Figure 5.2b. In a ROC curve, the background rejection is plotted against the signal efficiency which both should be high for an efficient algorithm. Thus, both the background rejection and the intROC should be as high as possible for the BDT to yield a good performance. For this reason, the BDT trained using NTree = 850, MinNodeSize = 4.0 and MaxDepth = 2 is chosen to be the best performing one, as it has the highest background rejection with the smallest uncertainty and the highest intROC. This configuration will be called stable 1-prong configuration and is used in the following analysis. The BDT score output distributions for the stable 1-prong configuration is shown in Figure 5.2a. The separation of the signal and background distributions is clearly visible. As expected from the KS Test number, the training and testing sample distributions only differ within the statistical uncertainties, which does not hint at overtraining of the algorithm. The ROC curve is near the top right corner of the coordinate system at (1,1) which indicates an efficient algorithm.

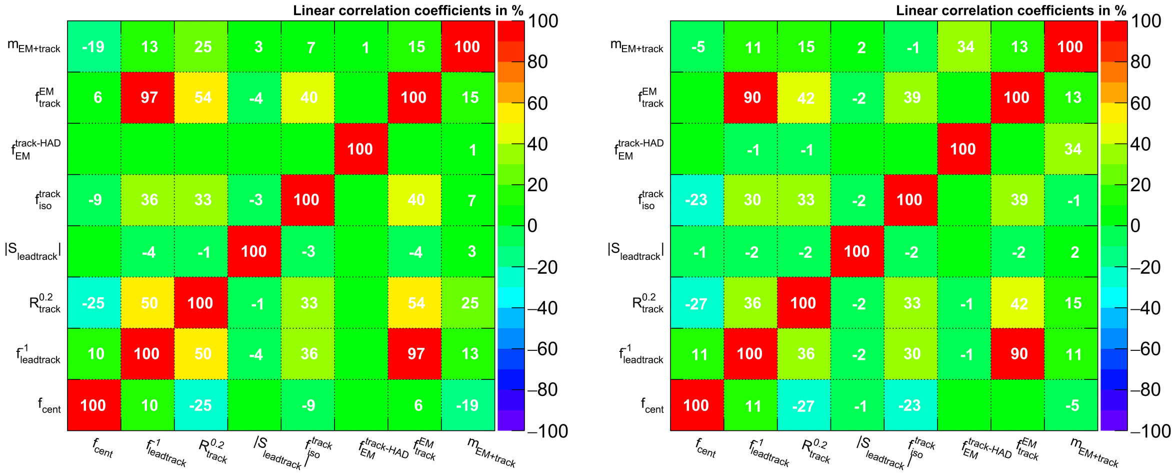
The correlation matrices of the TauID variables for signal and background are given in Figure 5.3. The most correlated variables are the Leading track momentum fraction $f_{leadtrack}^{-1}$ and the Ratio of EM energy to track momentum f_{track}^{EM} with a correlation of 97% for the signal events and 90% for the background events. The 2D scatter plot of these two variables is shown in Figure 5.4.

Given this high degree of correlation, the performance of the BDT was also investigated with one of these two variables omitted. There are also other correlated variables, but not to such an extent that it would be necessary to investigate the effects on the algorithm if one of them is left out (max 54% (42%) correlation). Running the stable 1-prong configuration again but leaving out either of the highly correlated TauID variables once yields performance values which are shown in Table 5.3. The BDT is more efficient when



(a) *(b)*

Figure 5.2: BDT score output distribution for signal and background events in both the training and testing sample (a) and ROC curve (b) for the stable 1-prong configuration. The KS Test numbers in (a) do not correspond to the ones given in Table 5.2 as TMVA uses a different implementation to calculate the KS Test number in (a). The background rejection is defined as (1 - background efficiency).



	f_{cent}	$f^{-1}_{leadtrack}$	$R^{0.2}_{track}$	$\lvert S_{leadtrack}\rvert$	f^{track}_{iso}	$f^{track\text{-}HAD}_{EM}$	f^{EM}_{track}	$m_{EM+track}$
$m_{EM+track}$	-19	13	25	3	7	1	15	100
f^{EM}_{track}	6	97	54	-4	40		100	15
$f^{track\text{-}HAD}_{EM}$						100		1
f^{track}_{iso}	-9	36	33	-3	100		40	7
$\lvert S_{leadtrack}\rvert$		-4	-1	100	-3		-4	3
$R^{0.2}_{track}$	-25	50	100	-1	33		54	25
$f^{1}_{leadtrack}$	10	100	50	-4	36		97	13
f_{cent}	100	10	-25		-9		6	-19

	f_{cent}	$f^{-1}_{leadtrack}$	$R^{0.2}_{track}$	$\lvert S_{leadtrack}\rvert$	f^{track}_{iso}	$f^{track\text{-}HAD}_{EM}$	f^{EM}_{track}	$m_{EM+track}$
$m_{EM+track}$	-5	11	15	2	-1	34	13	100
f^{EM}_{track}		90	42	-2	39		100	13
$f^{track\text{-}HAD}_{EM}$		-1	-1			100		34
f^{track}_{iso}	-23	30	33	-2	100		39	-1
$\lvert S_{leadtrack}\rvert$	-1	-2	-2	100	-2		-2	2
$R^{0.2}_{track}$	-27	36	100	-2	33	-1	42	15
$f^{1}_{leadtrack}$	11	100	36	-2	30	-1	90	11
f_{cent}	100	11	-27	-1	-23			-5

Figure 5.3: Linear correlation matrices for the TauID variables for 1-prong events for the signal and background events. The boxes without a number represent a linear correlation of 0.

KS Test Sig	KSTestBkg	intROC	BkgRej@50
0.69	0.87	0.920	0.97 ±0.03
0.70	0.53	0.920	0.97 ±0.03

Table 5.3: Properties of the chosen stable 1-prong configuration leaving out either the Ratio of EM energy to track momentum f^{EM}_{track} (first row) or the Leading track momentum fraction $f^{-1}_{leadtrack}$ (second row).

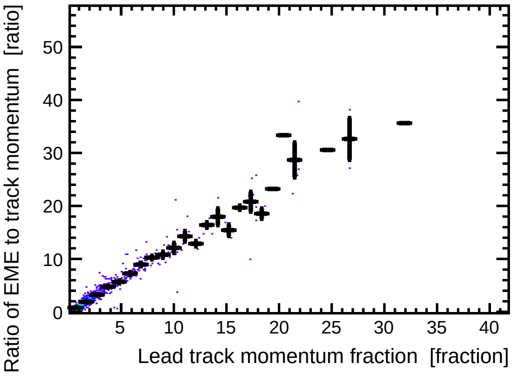
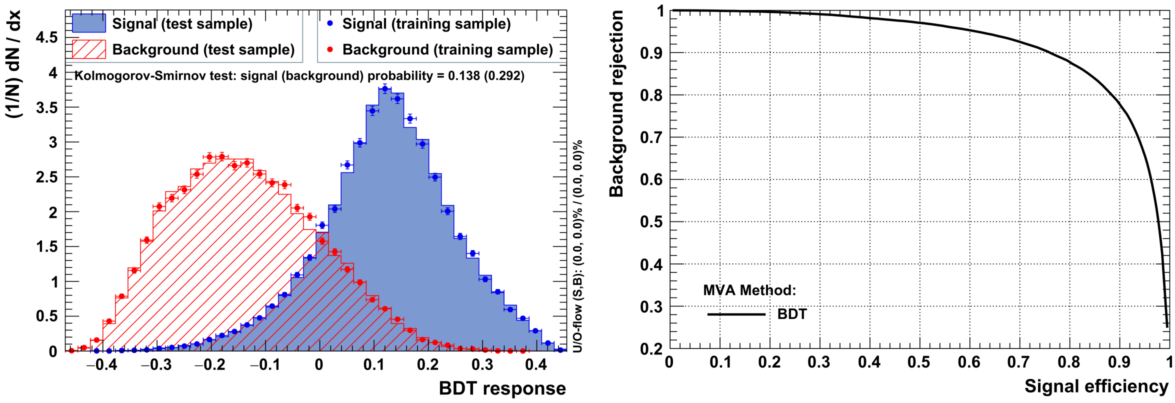


Figure 5.4: 2D scatter plot of the Leading track momentum fraction $f_{leadtrack}^{-1}$ and the Ratio of EM energy to track momentum f_{track}^{EM} for the 1-prong events. The correlation of the variables is clearly visible.



(a) ***(b)***

Figure 5.5: BDT score output distribution for signal and background events in both the training and testing sample (a) and ROC curve (b) for the stable 1-prong configuration after removing f_{track}^{EM}. The KS Test numbers in (a) do not correspond to the ones given in Table 5.3 as TMVA uses a different implementation to calculate the KS Test number in (a). The background rejection is defined as (1 - background efficiency).

Rank	Variable	Var. Importance	Rank	Variable	Var. Importance
1	f_{cent}	0.1890	1	f_{cent}	0.1907
2	f_{iso}^{track}	0.1738	2	f_{iso}^{track}	0.1787
3	$R_{track}^{0.2}$	0.1650	3	$R_{track}^{0.2}$	0.1642
4	f_{track}^{EM}	0.1309	4	$f_{leadtrack}^{-1}$	0.1383
5	$\|S_{leadtrack}\|$	0.1281	5	$\|S_{leadtrack}\|$	0.1347
6	$m_{EM+track}$	0.1115	6	$m_{EM+track}$	0.1114
7	$f_{EM}^{track-HAD}$	0.1018	7	$f_{EM}^{track-HAD}$	0.0820

Table 5.4: Ranking of the TauID variables for the stable 1-prong configuration when either the Ratio of EM energy to track momentum f_{track}^{EM} (right) or the Leading track momentum fraction $f_{leadtrack}^{-1}$ (left) is left out in the BDT training. The Variable Importance is the fraction of the variable usage for cuts in the whole forest.

Left out variable	KSTestSig	KSTestBkg	intROC	BkgRej@50
None	0.69	0.87	0.920	0.97 ±0.03
f_{cent}	0.96	0.73	0.893	0.95 ±0.04
$f_{leadtrack}^{-1}$	0.44	0.59	0.915	0.97 ±0.03
$R_{track}^{0.2}$	0.48	0.98	0.917	0.97 ±0.03
$\|S_{leadtrack}\|$	0.58	0.996	0.905	0.96 ±0.04
f_{iso}^{track}	0.31	0.9996	0.896	0.96 ±0.04
$f_{EM}^{track-HAD}$	0.80	0.73	0.918	0.97 ±0.03
$m_{EM+track}$	0.48	0.97	0.917	0.97 ±0.03

Table 5.5: Properties of the BDT if either of the remaining TauID variables is removed from the stable 1-prong configuration in training, applying the same stopping parameters.

the Ratio of EM energy to track momentum f_{track}^{EM} is left out. The BkgRej@50 and also the intROC is slightly larger for this case. The background rejection rises by an absolute value of 0.0027 in comparison to the BDT using all TauID variables. The ROC curve and the BDT score output distributions for this case are given in Figure 5.5. As shown in Table 5.4, there are no huge impacts on the variable importance being the percentage of usage in all cuts in the forest, implying that the removal of f_{track}^{EM} does not significantly change the algorithm. Thus, the Ratio of EM energy to track momentum f_{track}^{EM} can be safely removed from the algorithm.

Leaving out one of the remaining TauID variables does not have a great impact on the BDT in most cases and the algorithm remains stable. However, the BkgRej@50 as well as the intROC drops in every case. This is expected as the variables are not highly correlated. Also, removing the variables with the highest separation power, f_{cent}

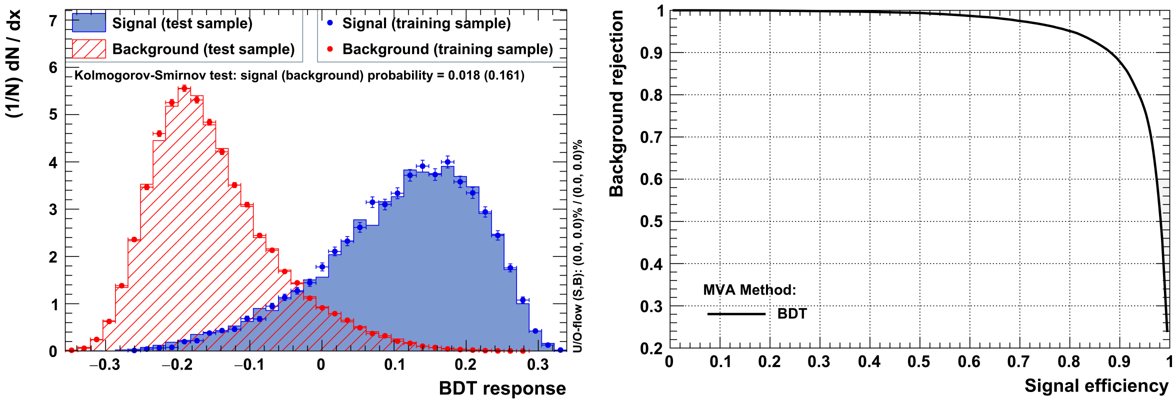
NTree	MinNodeSize	MaxDepth	KSTest Sig	KSTest Bkg	intROC	BkgRej@50
500	2.0	4	0.77	0.89	0.953	0.99 ±0.01
500	2.0	5	0.51	0.50	0.953	0.99 ±0.01
750	2.0	4	0.35	0.49	0.953	0.99 ±0.01
750	2.0	5	0.25	0.50	0.954	0.99 ±0.01
850	2.0	3	0.88	0.71	0.954	0.99 ±0.01
850	2.0	4	0.35	0.75	0.954	0.99 ±0.01
850	2.0	5	0.22	0.59	0.954	0.99 ±0.01
850	4.0	3	0.75	0.17	0.953	0.99 ±0.01
850	4.0	4	0.69	0.44	0.953	0.99 ±0.01

Table 5.6: Properties of the best performing BDTs for the 3-prong events.

and f_{iso}^{track}, has the highest impact on the efficiency of the algorithm. The above results confirm the choice of NTree = 850, MinNodeSize = 4.0 and MaxDepth = 2 for the BDT training parameters using the TauID variables f_{cent}, $f_{leadtrack}^{-1}$, $R_{track}^{0.2}$, $|S_{leadtrack}|$, f_{iso}^{track}, $f_{EM}^{track-HAD}$ and $m_{EM+track}$ as the best stable BDT to separate hadronically decaying τ leptons from jet background for 1-prong events.

5.3.2 3-prong events

Like in the analysis for the 1-prong events, the stopping parameters NTree, MaxDepth and MinNodeSize are applied in the training of the BDT. The same combination of parameters as before is used (cp. Table 5.1). As in the 1-prong analysis, the KS test number of the signal and background BDT output distribution, the background rejection at 50% signal efficiency including its error and the intROC are investigated. The properties of the best performing BDTs are given in Table 5.6. The algorithm seems to be more efficient when a higher number of trees is used in the forest and medium to higher values of the MaxDepth are used. Also, effective algorithms tend to have small values of MinNodeSize. The BDT trained using NTree = 850, MinNodeSize = 2.0 and MaxDepth = 4 is chosen to be the best performing one, as it has the highest background rejection with the smallest uncertainty whilst the intROC values of all trained BDTs are nearly the same. This configuration will be called stable 3-prong configuration and is used in the following analysis. The BDT score output distribution and the ROC curve are given in Figure 5.6a and 5.6b, respectively. The output distributions are well-separated and the training and testing sample outputs lie within their statistical uncertainties, which does not hint at overtraining of the algorithm. Also, the ROC curve hints to an efficient algorithm rejecting major parts of the background events while keeping most of the signal events, as it is near the top right corner of 100% background rejection and signal efficiency.



(a) *(b)*

Figure 5.6: BDT score output distribution for signal and background events in both the training and testing sample (a) and ROC curve (b) for the stable 3-prong configuration. The KS Test numbers in (a) do not correspond to the ones given in Table 5.6 as TMVA uses a different implementation to calculate the KS Test number in (a). The background rejection is defined as (1 - background efficiency).

KS Test Sig	KSTestBkg	intROC	BkgRej@50
0.65	0.42	0.952	0.99 ±0.01
0.54	0.39	0.953	0.99 ±0.01

Table 5.7: Properties of the chosen stable 3-prong configuration leaving out either the Ratio of EM energy to track momentum f_{track}^{EM} (first row) or the Leading track momentum fraction $f_{leadtrack}^{-1}$ (second row).

The correlation matrices of the TauID variables for signal and background are given in Figure 5.7. As in the 1-prong analysis, the most correlated variables are the Leading track momentum fraction $f_{leadtrack}^{-1}$ and the Ratio of EM energy to track momentum f_{track}^{EM} with a correlation of 84% for both signal and background events. The 2D scatter plot of these two variables is shown in Figure 5.8. Given this high degree of correlation, the performance of the BDT was also investigated with one of these two variables omitted. Other variables are also correlated, but not to such an extent that it would be necessary to investigate the performance of the BDT omitting one of them (max 63% (59%) correlation). Running the stable 3-prong configuration again but leaving out either of the highly correlated TauID variables once yields performance values which are shown in Table 5.7. In contrast to the 1-prong case, the BDT is more efficient with respect to the BkgRej@50 when the Leading track momentum fraction $f_{leadtrack}^{-1}$ is left out, and also the error on the background rejection is smaller. The intROC is also larger for this case. The background rejection drops by an absolute value of 0.00012 in comparison to the BDT

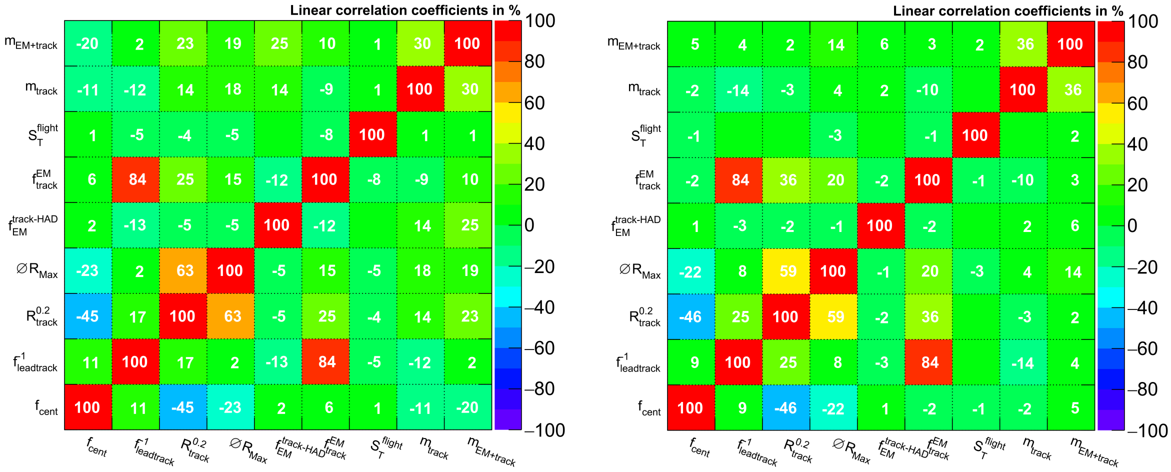


Figure 5.7: Linear correlation matrices for the TauID variables for 3-prong events for the signal and background events. The boxes without a number represent a linear correlation of 0.

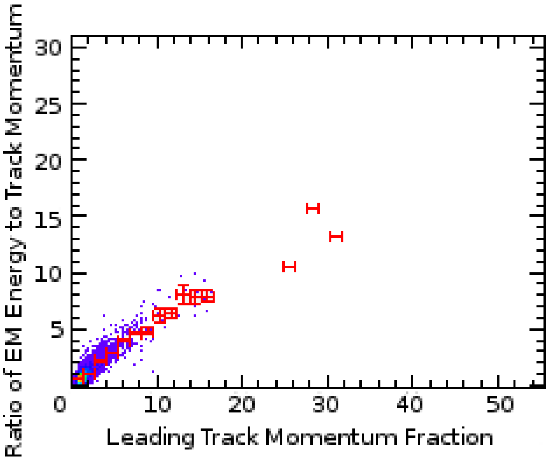
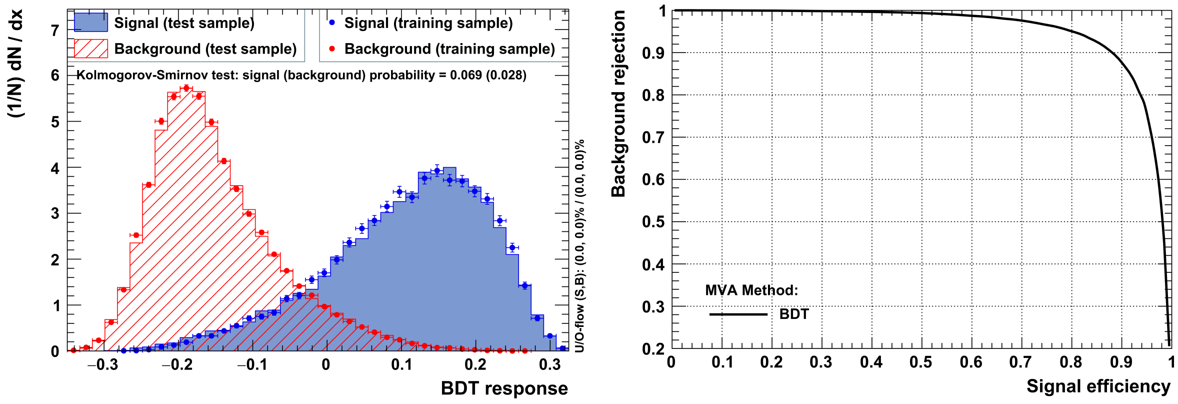


Figure 5.8: 2D scatter plot of the Leading track momentum fraction $f_{leadtrack}^{-1}$ and the Ratio of EM energy to track momentum f_{track}^{EM} for the 3-prong events. The correlation of the variables is clearly visible.



(a) *(b)*

Figure 5.9: BDT score output distribution for signal and background events in both the training and testing sample (a) and ROC curve (b) for the stable 3-prong configuration after removing $f^{-1}_{leadtrack}$. The KS Test numbers in (a) do not correspond to the ones given in Table 5.7 as TMVA uses a different implementation to calculate the KS Test number in (a). The background rejection is defined as (1 - background efficiency).

using all TauID variables, which is negligible. The BDT output distribution and the ROC curve are shown in Figure 5.9. Table 5.8 shows that a removal of $f^{-1}_{leadtrack}$ has no huge impacts on the variable importance being the percentage of usage in all cuts in the forest, implying no significant changes of the algorithm if $f^{-1}_{leadtrack}$ is left out. Thus, the Ratio of EM energy to track momentum $f^{-1}_{leadtrack}$ can be safely removed from the algorithm.

Leaving out one of the remaining TauID variables does not have a great impact on the BDT in most cases. An exception is the omission of S_T^{flight} which yields a drop in the background rejection of two percentage points. For the latter, also the intROC gets significantly smaller. The same applies for m_{track}, but the effects are not as high as for S_T^{flight}. The results are summarized in Table 5.9. These effects were expected, as S_T^{flight} and m_{track} are two of the four variables with the highest separation power. However, it is strange that leaving out the variables with the highest separation power, $R^{0.2}_{track}$ and ΔR_{Max}, seems to have no effect on the BDT at all. Also, no strange behaviour in the frequency of usage of TauID variables is observed when either of them is left out. A reason for this could be the relatively high correlation between both variables of 63% for signal and 59% for background.

The above results confirm the choice of NTree = 850, MinNodeSize = 2.0 and MaxDepth = 4 for the BDT training parameters using the TauID variables f_{cent}, f^{EM}_{track}, $R^{0.2}_{track}$, $f^{track-HAD}_{EM}$, $m_{EM+track}$, ΔR_{Max}, S_T^{flight} and m_{track} as the best stable BDT to separate hadronically decaying τ leptons from jet background for 3-prong events.

Rank	Variable	Var. Importance	Rank	Variable	Var. Importance
1	ΔR_{Max}	0.1658	1	ΔR_{Max}	0.1704
2	$R^{0.2}_{track}$	0.1488	2	$R^{0.2}_{track}$	0.1509
3	f_{cent}	0.1410	3	f_{cent}	0.1458
4	m_{track}	0.1239	4	m_{track}	0.1290
5	f^{EM}_{track}	0.1169	5	$m_{EM+track}$	0.1154
6	$m_{EM+track}$	0.1104	6	$f^{-1}_{leadtrack}$	0.1076
7	S^{flight}_{T}	0.1054	7	S^{flight}_{T}	0.1044
8	$f^{track-HAD}_{EM}$	0.0878	8	$f^{track-HAD}_{EM}$	0.0766

Table 5.8: Ranking of the TauID variables when either the Ratio of EM energy to track momentum f^{EM}_{track} (right) or the Leading track momentum fraction $f^{-1}_{leadtrack}$ (left) is left out in the BDT training for the stable 3-prong configuration. The Variable Importance is the fraction of the variable usage for cuts in the whole forest.

Left out variable	KSTestSig	KSTestBkg	intROC	BkgRej@50
None	0.54	0.39	0.953	0.99 ±0.01
f_{cent}	0.56	0.63	0.950	0.99 ±0.01
$R^{0.2}_{track}$	0.70	0.21	0.952	0.99 ±0.01
$f^{track-HAD}_{EM}$	0.38	0.44	0.947	0.99 ±0.01
f^{EM}_{track}	0.47	0.49	0.948	0.99 ±0.01
$m_{EM+track}$	0.45	0.10	0.949	0.99 ±0.01
ΔR_{Max}	0.57	0.21	0.952	0.99 ±0.01
S^{flight}_{T}	0.46	0.89	0.928	0.98 ±0.01
m_{track}	0.35	0.34	0.939	0.99 ±0.01

Table 5.9: Properties of the BDT if either of the remaining TauID variables is removed from the stable 3-prong configuration in training, applying the same stopping parameters.

6 Conclusion

In this thesis, two BDTs separating 1- and 3-prong τ_{had} leptons from jet backgrounds were trained. The analysis used simulated $Z^0 \to \tau\tau$ and $Z^0 \to e^+e^- + jets$ as signal and background samples, respectively. The analysis was done using TMVA with AdaBoost as the boosting algorithm.

The properties of the best performing stable algorithm for both the 1- and 3-prong events are shown in Table 6.1. For the 1- prong events, the Ratio of EM energy to track momentum f_{track}^{EM} could safely be omitted, as it was found to be highly correlated to the Leading track momentum fraction $f_{leadtrack}^{-1}$. Thus, the variable did not provide any new information to the algorithm. The performance of the BDT for 1-prong events was higher omitting f_{track}^{EM}, both regarding the BkgRej@50 and the intROC.

For the 3-prong events, it turned out that the algorithm performed better omitting $f_{leadtrack}^{-1}$ regarding both the BkgRej@50 and the intROC.

Omitting more than just one of the two highly correlated variables from the BDT training does not yield any further insights. The algorithm remains stable for both the 1- and 3-prong case and no unexpected effects are observed, as the performance of the algorithm goes down for most of the cases. The variable ranking is also not affected. However, removing $R_{track}^{0.2}$ or ΔR_{Max} from the 3-prong BDT training did not have any effect on the algorithm at all, which is strange as these are the variables most often used to do cuts in this analysis. A further investigation would be interesting here.

A significant rise in the performance of the BDTs could have been achieved if new TauID variables were defined. However, this would have exceeded the limits of this thesis.

case	NTree	MinNodeSize	MaxDepth	intROC	BkgRej@50
1-prong	850	4.0	2	0.920	0.97 ±0.03
3-prong	850	2.0	4	0.953	0.99 ±0.01

Table 6.1: Performance of the stable configurations for the 1- and 3-prong BDTs after omitting the Ratio of EM energy to track momentum f_{track}^{EM}.

Bibliography

[1] J. J. Thomson, *Cathode Rays*, The Electrician **39**, 104 (1897)

[2] J. J. Thomson, *Cathode Rays*, Philosophical Magazine **5. 44(293)** (1897)

[3] S. L. Glashow, *Partial Symmetries of Weak Interactions*, Nucl. Phys. **22**, 579 (1961)

[4] A. Salam, *Weak and Electromagnetic Interactions*, Conf. Proc. **C680519**, 367 (1968)

[5] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19**, 1264 (1967)

[6] N. Cabibbo, *Unitary Symmetry and Leptonic Decays*, Phys. Rev. Lett. **10**, 531 (1963)

[7] M. Kobayashi, T. Maskawa, *CP Violation in the Renormalizable Theory of Weak Interaction*, Prog. Theor. Phys. **49**, 652 (1973)

[8] DØ Collaboration (D0), *Measurement of the W boson mass with the D0 detector*, Nucl. Part. Phys. Proc. **273-275**, 2237 (2016)

[9] *Measurement of the W Boson Mass in Proton-Proton Collisions at $s = 7$ TeV with the ATLAS Detector*, Technical Report ATLAS-CONF-2016-113, CERN, Geneva (2016)

[10] K.A. Olive et al. (Particle Data Group), Chin. Phys. C **38, 090001** (2014)

[11] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13**, 508 (1964)

[12] F. Englert, R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13**, 321 (1964)

[13] G. S. Guralnik, C. R. Hagen, T. W. B. Kibble, *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13**, 585 (1964)

[14] Atlas, Cms Collaborations, *Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments*, Phys. Rev. Lett. **114**, 191803 (2015)

[15] LHC Higgs Cross Section Working Group, *Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables* (2011), `arXiv:1101.0593`

[16] LHC Higgs Cross Section Working Group, *Handbook of LHC Higgs Cross Sections: 2. Differential Distributions* (2012), `arXiv:1201.3084`

[17] LHC Higgs Cross Section Working Group, *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties* (2013), `arXiv:1307.1347`

[18] LHC Higgs Cross Section Working Group, as of June 27, 2016, URL `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGCrossSectionsFigures`

[19] 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. **B716**, 1 (2012)

[20] Cms Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. **B716**, 30 (2012)

[21] Atlas Collaboration, *Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment*, Eur. Phys. J. **C76(1)**, 6 (2016)

[22] Atlas and Cms Collaborations, *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV* (2015)

[23] Atlas Collaboration, *A particle consistent with the Higgs Boson observed with the ATLAS Detector at the Large Hadron Collider*, Science **338**, 1576 (2012)

[24] Atlas Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3**, S08003 (2008)

[25] Atlas Collaboration, *ATLAS detector and physics performance: Technical Design Report, 1*, Technical Design Report ATLAS, CERN, Geneva (1999), URL `https://cds.cern.ch/record/391176`

[26] C. Berger, *Elementarteilchenphysik: Von den Grundlagen zu den modernen Experimenten*, Springer-Lehrbuch, Springer (2006)

[27] Atlas Collaboration, *Reconstruction, Energy Calibration, and Identification of Hadronically Decaying Tau Leptons in the Atlas Experiment for Run II of the Lhc* (2015)

[28] ATLAS Collaboration, *Identification and energy calibration of hadronically decaying tau leptons with the ATLAS experiment in pp collisions at $\sqrt{s}$=8 TeV*, Eur. Phys. J. **C75(7)**, 303 (2015)

[29] A. Hoecker, et al., *TMVA - Toolkit for Multivariate Data Analysis* (2009)

[30] R. Brun, F. Rademakers, *ROOT: An object oriented data analysis framework*, Nucl. Instrum. Meth. **A389**, 81 (1997)

[31] Y. Freund, R. E. Schapire, *A desicion-theoretic generalization of on-line learning and an application to boosting*, pages 23–37, Springer Berlin Heidelberg, Berlin, Heidelberg (1995)

Acknowledgements

Being part of the II. Institute of Physics in Göttingen was a great experience which I would not like to miss. I loved getting an inside view on research in particle physiscs, with all its ups and downs.
A first big thank you goes to Prof. Dr. Stan Lai, who provided an excellent supervision. The weekly meetings were a great support which could not have been replaced by anything, and were always enlightening.
The second big thank you goes to Dr. Michel Janus, who did an extraordinary great job in supervising Alejandra and me in both our theses. It was always a huge enrichment when you just did not stop talking to provide us even more explanations and information around technical and scientific details, regardless of what you had to do at the moment! Two hours can pass really quickly this way.
I also want to thank Alejandra for always being able to discuss the problems we encountered on the way to finishing our theses and for showing me that I am not the only one getting weird error messages in my analysis.
A really huge thank you goes to my parents and to my brother. Thank you for always being there and supporting me not only through this 14 weeks of the Bachelors thesis, but the whole last four years of my studies! Without you, I would have hardly made it here. DANKE!
Finally, I want to thank all the kind, lovely and crazy persons and friends on my way up till now. Thank you for being there and supporting me :-) (especially Leo!)!

Erklärung nach §13(9) der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen:

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe.
Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, im Rahmen einer nichtbestandenen Prüfung an dieser oder einer anderen Hochschule eingereicht wurde.

Göttingen, den 19. April 2017

(Nils Gillwald)