Technische Universität München
Physik-Department

Masterarbeit

Search for top squark Production at the LHC at $\sqrt{s} = 13$ TeV with the ATLAS Detector Using Multivariate Analysis Techniques

von
Jonas Graw

München
Freitag, den 18. August 2017

Max-Planck-Institut für Physik
(Werner-Heisenberg-Institut)

Technische Universität München
Physik-Department
Erstgutachter (Themensteller): Prof. Dr. H. Kroha
Zweitgutachter: Prof. Dr. L. Oberauer

I assure the single handed composition of this master's thesis is only supported by declared resources.

Ich versichere, dass ich diese Masterarbeit selbstständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 18. August 2017

Unterschrift

Abstract

Supersymmetry is a very promising extension of the Standard Model. It predicts new heavy particles, which are currently searched for in the ATLAS experiment at the Large Hadron Collider at a center-of-mass energy of 13 TeV. So far, all searches for supersymmetric particles use a cut-based signal selection. In this thesis, the use of multivariate selection techniques, Boosted Decision Trees and Artificial Neural Networks, is explored for the search for top squarks, the supersymmetric partner of the top quark. The multivariate methods increase the expected lower limit in the mass of top squarks by approximately 90 GeV from currently 990 GeV for small neutralino masses.

Acknowledgments

First of all, I would like to thank Professor Hubert Kroha for giving me the opportunity to study and research in such a fantastic group. I am also very grateful for the proof-reading of my thesis. I greatly benefited from his keen insight.

Secondly, I must express my profound gratitude to Nicolas Köhler and Johannes Junggeburth, who always provided help when I needed some – even at night or on weekends. In addition, I am much obliged to them for proof-reading the thesis. I really enjoyed our time together and I greatly appreciate their unconditional efforts.

Finally, I would also like to express my gratitude to all my friends and my family, particularly to my mother Regina Graw, for always supporting me.

Contents

Chapter 1

Introduction

The Standard Model of particle physics has successfully passed all tests over the last several decades. The last missing particle of the Standard Model, the Higgs boson, was discovered in the ATLAS [1] and CMS [2] experiments at the Large Hadron Collider (LHC) in 2012. There are, however, many experimental and theoretical reasons to believe that the Standard Model is not the final theory of elementary particles. Dark Matter and Dark Energy, established by astrophysical observations, dominate the energy content of the universe and cannot be explained by the Standard Model [3, 4]. The smallness of the Higgs boson mass compared to the energy scale of the unification of all interaction strengths requires an explanation, the so-called naturalism or fine-tuning problem of the Higgs mass in the Standard Model.

Supersymmetry is a very promising extension of the Standard Model, in which every Standard Model particle is accompanied by a supersymmetric partner particle differing by spin $\frac{1}{2}$. The searches for supersymmetric particles like the top squark – the partner of the top quark – have been performed with cut-based signal selection methods in the ATLAS experiment up to now. In this thesis, a search for top squarks with ATLAS at 13 TeV centre of mass energy has been performed using machine learning or multivariate methods which allow for the efficient combination of many partially correlated discriminating variables.

In Chapter 2, the Standard Model and supersymmetry are introduced, while in Chapter 3 the Large Hadron Collider and the ATLAS Experiment are described. Chapter 4 explains the signatures of top squark production in ATLAS and the cut-based search as a reference. Chapter 5 introduces the machine learning methods used, Boosted Decision Trees (BDT) and Artificial Neural Network (ANN). Chapter 6 finally describes the results of the multivariate signal selection techniques for the top squark search. The summary of the analysis and results are given in Chapter 7.

Chapter 2

Theoretical Framework

2.1 The Standard Model of Particle Physics

The particle content of the Standard Model is shown in Figure 2.1. Besides the three generations of quarks and leptons, the matter particles with spin $\frac{1}{2}$, there are force mediators with spin 1 and the scalar Higgs boson.

Four fundamental forces are known – the strong, the weak, the electromagnetic and the gravitational force. All fundamental forces, except gravitational, are described by the Standard Model (SM based on the gauge symmetry groups [5, 6]

$$SU(3)_C \otimes SU(2)_L \otimes U(1)_Y. \tag{2.1}$$

The gauge symmetries predict mediators of the interactions with spin 1, called gauge bosons. $SU(3)_C$ is the gauge group describing the strong force with eight gluons as force carriers, whereas $SU(2)_L \otimes U(1)_Y$ is the gauge group of the unified electroweak interaction. $SU(2)_L$ is the gauge group of the weak force with three force carriers W^i, $i = 1, 2, 3$, belonging to the weak isospin outer components T_i as charge operators. $U(1)_Y$ is the symmetry group of the weak hypercharge Y, which is related to the electric charge Q and T_3 via

$$Q = T_3 + \frac{Y}{2} \tag{2.2}$$

and is mediated by a gauge boson B.

The electroweak gauge bosons mix to the electrically charged mass eigenstates.

$$W^{\pm} = \frac{1}{\sqrt{2}} \cdot \left(W^1 \mp iW^2\right), \tag{2.3}$$

and to neutral mass eigenstates

$$\begin{pmatrix} Z \\ A \end{pmatrix} = \begin{pmatrix} \cos\theta_W & -\sin\theta_W \\ \sin\theta_W & \cos\theta_W \end{pmatrix} \cdot \begin{pmatrix} W^3 \\ B \end{pmatrix}, \tag{2.4}$$

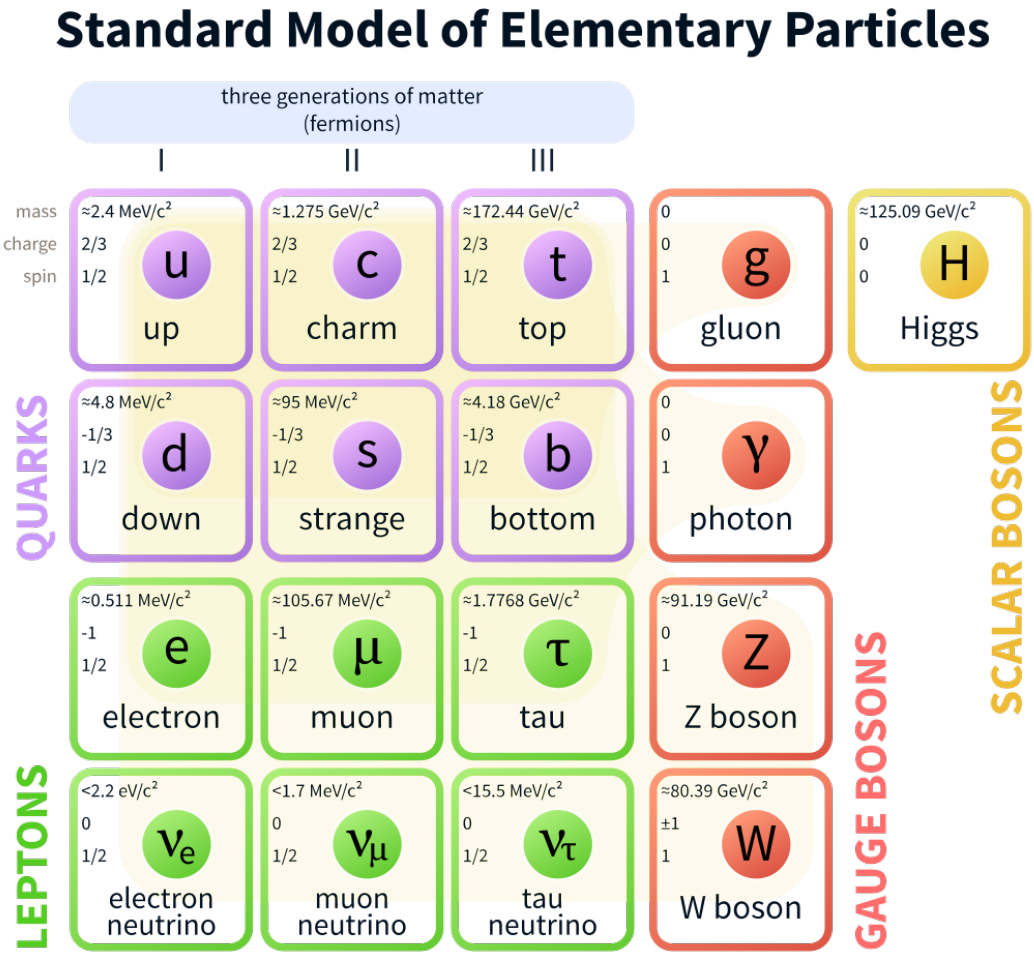


Figure 2.1: Particles of the Standard Model [7].

where Z^0 is the mediator of the neutral weak interaction and heavy partner of the photon field A. The Weinberg mixing angle θ_W is related to the gauge coupling constants corresponding to the hypercharge, g_Y and g_W of the weak hypercharge and the weak isospin interaction via

$$\tan\theta_W = \frac{g_Y}{g_W}. \tag{2.5}$$

The $SU(3)_C$ color gauge interaction predicts the existence of eight gluons as carriers of the strong force coupling to the quarks which all appear in three color states [8].

The gauge symmetries predict invariance of the theory under local phase transformations of the matter fields and the massless spin 1 gauge bosons. The problem that the gauge bosons of the weak interaction are massless in contradiction to experimental observations was solved by introducing spontaneous weak interaction breaking of the electroweak gauge symmetry [9–11], which, in its minimal form requires a complex scalar weak isospin doublet Φ with $Y = 1$. The potential of the scalar field

$$V(\phi) = \mu|\phi|^2 + \lambda|\phi|^4. \tag{2.6}$$

acquires a non-zero vacuum expectation value

$$\Phi_0 = \sqrt{\frac{-\mu}{2\lambda}} \tag{2.7}$$

for negative parameter μ such that the ground vacuum state violates the $SU(2)_L \otimes U(1)_Y$ gauge symmetry. The spontaneous symmetry breaking leads to masses for the weak $W^\pm$ and the Z^0, while the photon field A remains massless corresponding to the unbroken $U(1)$ gauge group of the electromagnetic interaction. A further consequence is the existence of a new scalar particle, the Higgs boson. It was discovered by ATLAS and CMS experiments of the LHC in 2012 [1, 2] with a mass of about 125 GeV. All measured properties so far fit the Standard Model predictions [12–16].

The spin $\frac{1}{2}$ particles in the SM comprise six quarks with electric charges $\frac{2}{3}e$ or $-\frac{1}{3}e$ and six leptons which obtain their mass via Yukawa interactions with the Higgs field [8]. The Yukawa couplings are proportional to the fermion masses. Neutrinos may have additional mass terms.

The left-handed states of the quark and the leptons are doublets of the $SU(2)_L$ weak interaction gauge symmetry,

$$\begin{pmatrix} u \\ d \end{pmatrix}_L, \begin{pmatrix} c \\ s \end{pmatrix}_L \begin{pmatrix} t \\ b \end{pmatrix}_L, \tag{2.8}$$

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L, \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L, \tag{2.9}$$

and participate in the charged weak interaction via the exchange of $W^{\pm}$ bosons and transition from up to down states and vice versa.

Right-handed quarks and right-handed leptons have isospin $T = 0$ and are singlets of the $SU(2)_L$ symmetry, i. e. they do not participate in the charged weak interaction.

2.2 Limitations of the Standard model

Despite the greatness of the SM passing all experimental tests with high precision, there is a large number of open questions which show that the Standard Model (SM) cannot be complete.

The SM does not allow for neutrino oscillations [17–19] as neutrinos are massless in the SM.

From astrophysical and cosmological observations it is known that only 4.9% of the universe consists of SM particles, whereas 26.8% is so-called Dark Matter and the remaining fraction Dark Energy, neither of which is described in the SM [3, 4]. It also appears that the matter-anti-matter asymmetry in the universe cannot be explained by the Standard Model.

Another problem – and one of the key arguments for supersymmetry – is the naturalness- and, related, hierarchy-problem. The question is, why the electroweak scale (≈ 100 GeV) is so much smaller than the Planck-scale (10^{19} GeV) [20]. Since the quantum loop corrections to the Higgs boson mass diverge quadratically with the cutoff scale, it remains unnatural fine-tuning of the SM parameters to keep the Higgs boson mass as low as 125 GeV.

Finally, the SM does not describe (quantum) gravity. One very promising extension of SM which solves several problems is supersymmetry.

2.3 Supersymmetry

Supersymmetry (SUSY) transforms boson into fermion states and vice versa.

As a consequence, in supersymmetric extensions of the SM, there is a supersymmetric partner, each SM particle which differs by spin $\frac{1}{2}$, but otherwise has the same mass as its super-partner. In this case, fermionic and bosonic loop corrections to the Higgs mass would exactly cancel, solving the naturalness problem of to the Higgs mass.

SUSY, however, has to be a broken symmetry, since no super-partners of the SM particles with the same mass have been observed. In order to solve the naturalness

problem, the SUSY breaking scale must not be higher than a few TeV, which could be within the reach of the LHC, in particular, the masses of the partners of the top quark.

2.4 The Minimal Supersymmetric Standard Model (MSSM)

The MSSM is the minimal supersymmetric extension of the SM with the smallest number of additional particles of the MSSM. The particles of the MSSM are shown in Figure 2.2. Instead of one Higgs doublet, two Higgs doublets are required, one for up-type and one for down-type fermions in order to avoid gauge anomalies. Spontaneous electroweak symmetry breaking in this case leads to five Higgs bosons: a Standard Model-like h, another heavier neutral scalar boson H^0, a pseudo-scalar neutral boson A and two charged Higgs bosons $H^{\pm}$.

Each of these bosons is provided with its own superpartner. The nomenclature is as follows: The supersymmetric partners of bosons are denoted by an -ino ending, the five Higgsinos and the gauginos, the three Winos, the Bino, the eight gluinos and the photino. The supersymmetric partners of the SM fermions are denoted by an s- at the beginning, squarks and sleptons as the partners of the quarks and the leptons. The symbols of the supersymmetric particles are marked with a tilde on the SM particles symbols, e. g. $\tilde{t}$ for a top squark. The spin-$\frac{1}{2}$ Higgsinos, Binos and Winos mix to four neutral mass eigenstates $\chi_i^0, i = 1, ..., 4$, the neutralinos, and four charged mass eigenstates $\chi_i^{\pm}, i{+}1, 2$, the charginos. Since right-handed and left-handed SM fermions couple differently to the weak interaction, the two charginos are distinguished for each SM fermion with mass coupling scalar superpartner states $\widetilde{l}_{1,2}$ and $\widetilde{q}_{1,2}$ for sleptons and squarks, respectively.

2.5 R-Parity

As the lepton and baryon number are not conserved in SUSY theories, the proton is allowed to decay rapidly without an extra symmetry, which prevents it.

In the MSSM, a new conserved quantum number is introduced, the so-called R-parity

$$P_R = (-1)^{3B+L+2S} \tag{2.10}$$

where B is the baryon number, L the lepton number and S the spin of the particle. The R-parity is 1 for SM particles and -1 for supersymmetric particles. In this

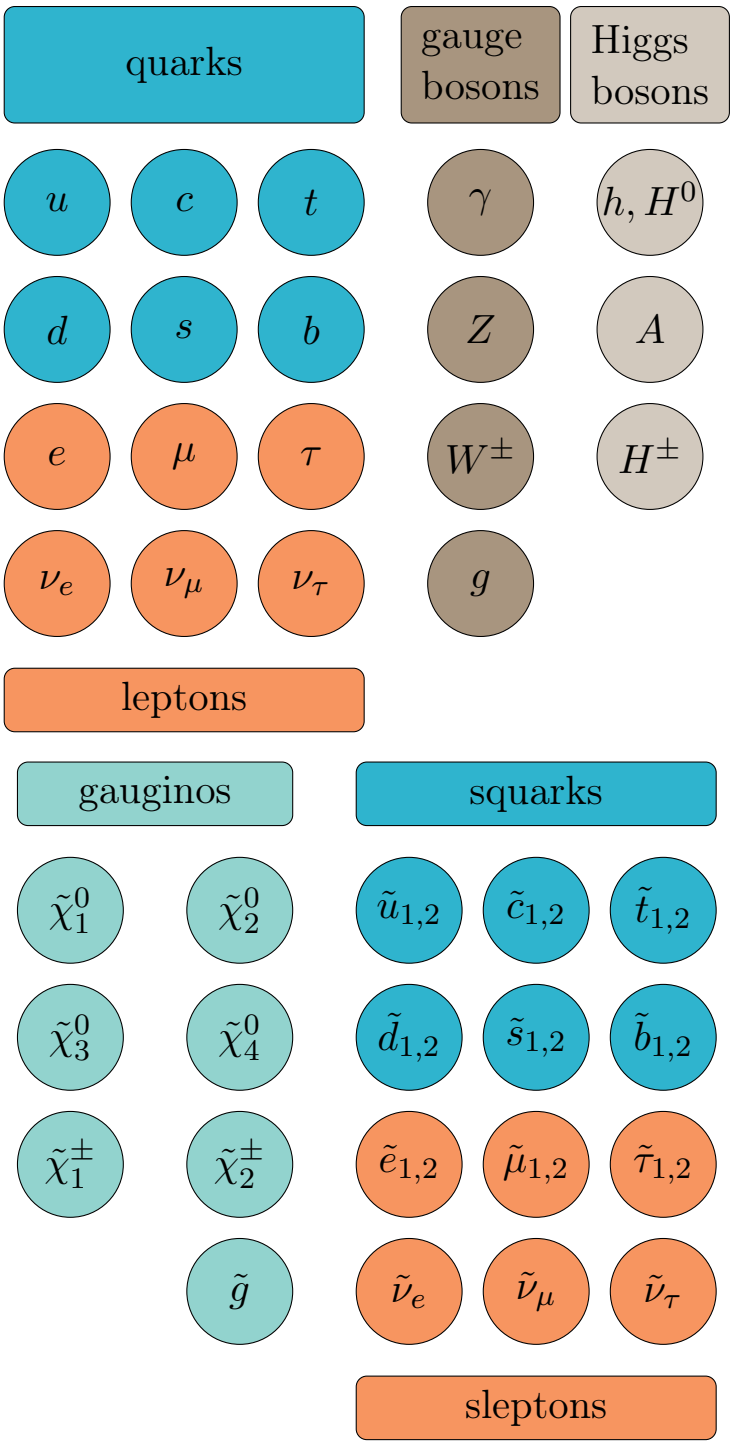


Figure 2.2: Particles of the Minimal Supersymmetric Standard Model [21].

thesis, any SUSY model with R-parity conservation is considered. In this case, the lightest supersymmetric particle (LSP) is stable, and super-partners are produced in pairs from SM particle collisions. If the LSP is neutral and only weakly-interacting, it proves to be an excellent candidate for a Dark Matter particle. The lightest neutralino, $\tilde{\chi}_1^0$, is often assumed to be the LSP. In SUSY models with R-parity violation, other symmetries are introduced to prevent rapid proton decay.

Chapter 3

The ATLAS Experiment

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) at CERN, the European centre for particle physics near Geneva, collides protons cylindrically at a centre-of-mass energy of 13 TeV and a luminosity of $\mathcal{L} = 1.5 \cdot 10^{34}\text{cm}^{-2}\text{s}^{-1}$.

It is located in a tunnel at a depth between 50m and 175m underground and has a circumference of about 26.6 km.

The LHC utilizes 1232 superconducting dipole magnets to keep the protons on their circular path. They are operated at a maximum current of 12 kA resulting in a magnetic field of 8.3 T. The proton beams consist of 2808 bunches of 10^{11} protons each. The bunch spacing time is 25 ns.

Before injection into the LHC, the protons are pre-accelerated in several other accelerators: First a linear accelerator, then the proton synchrotron (PS) and finally the Super Proton Synchrotron (SPS), which increases the proton energy to the injection energy of 450 GeV. Figure 3.1 shows the CERN accelerator complex schematically.

There are four detectors at the LHC: ATLAS (**A** **T**oroidal **L**HC **A**pparatu**S**) and CMS (**C**ompact **M**uon **S**olenoid), which are designed to search for production and decays of the Higgs boson and for signatures of physics beyond SM, LHCb (**L**arge **H**adron **C**ollider **b**eauty detector), a bottom quark physics experiment and ALICE (**A** **L**arge **I**on **C**ollider **E**xperiment), built for the investigation of heavy ion collisions.

The main goal and greatest success of the LHC so far was the discovery of the Higgs boson in 2012 by the ATLAS [1] and CMS [2] experiments.

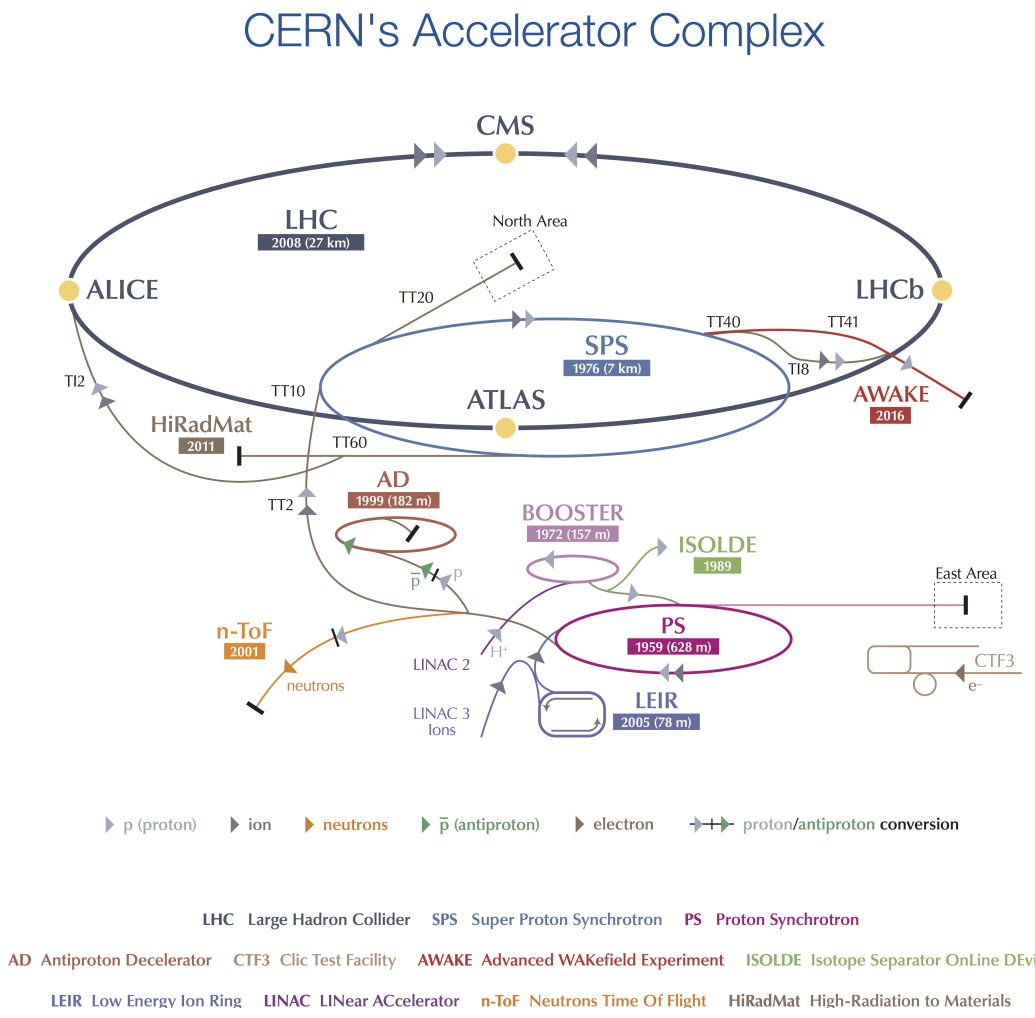


Figure 3.1: Schematic layout of the CERN accelerator complex [22].

3.2 The ATLAS Detector

ATLAS [23] is one of the two main experiments at the LHC, along with the CMS detector. It has a length of 46 m, a diameter of 25 m and a mass of 7000 tons. It is located about 100 m underground.

Figure 3.2 shows the ATLAS Detector and its main components. Like other collider detectors, it consists of a barrel region containing co-axial cylindric detector layers and endcaps with two disk-shaped detector layers. Figure 3.3 explains the functionality of the sub-detectors for the identification and measurement of various particle types.

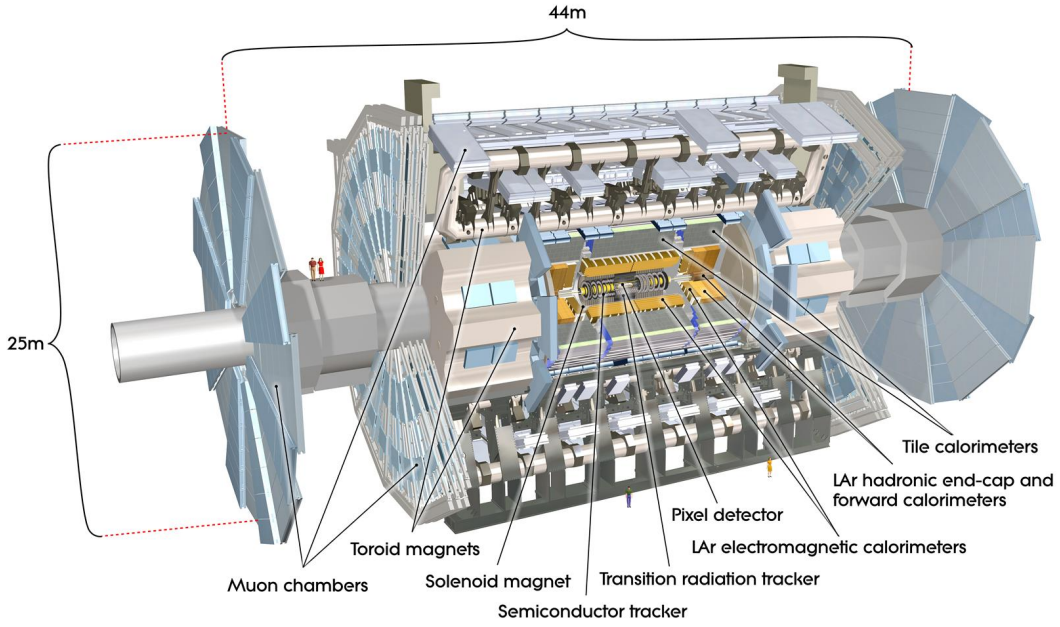


Figure 3.2: Schematic views of the ATLAS detector [23].

ATLAS uses a right-handed coordinate system of which the origin cuts the nominal interaction point. The z-axis is the proton beam axis, whereas the x-axis points towards the centre of the LHC ring. The x-y-plane is the transverse plane with respect to the proton beam. To describe points in the detector, spherical coordinates (r, θ, ϕ) are used, where r is the distance from the origin, θ the polar angle and Φ the azimuthal angle in the x-y-plane.

Alternatively, the pseudo-rapidity $\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right)$ serves to describe particle production in the r-z-plane. It is related to the transverse momentum p_T and the transver-

sal energy E_T of particles relative to the beam axis by the expressions

$$p_T = \sqrt{p_x^2 + p_y^2} = \frac{|p|}{\cosh(\eta)}, \tag{3.1}$$

$$E_T = \frac{E}{\cosh(\eta)}. \tag{3.2}$$

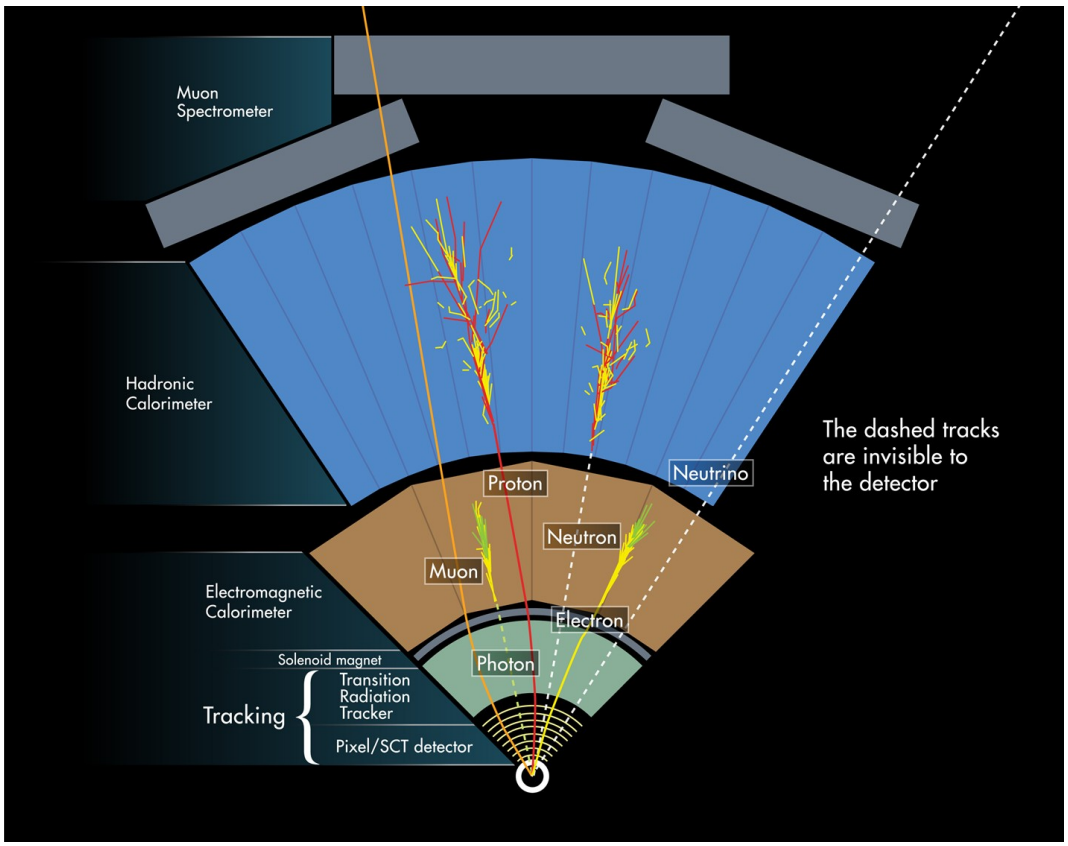


Figure 3.3: Schematic illustration of the detection of particles in the ATLAS detector components [24].

3.2.1 The Inner Detector

The Inner Detector, depicted in Figure 3.4, is the innermost detector component, which consists of three layers: the Pixel Detector, the Semiconductor Tracker and

the Transition Radiation Tracker. The Inner Detector is surrounded by a 2T superconducting solenoid magnet in order to measure the deflection and momentum of charged particles up to a pseudo-rapidity of $|\eta|$ =2.5.

The Pixel Detector is the detector closest to the beamline. Its purpose is to reconstruct interaction and decay vertices. It consists of silicon pixel sensors with 80 million pixels in total and a spatial resolution of 14 μm in the transverse plane and 115 μm in the direction of the beam axis which are arranged in three layers in the barrel region and three disks in the end-caps. The innermost layer is called B-layer, since it is particularly important for the tagging of b-quark jets making use of the long lifetime and decay length of B hadrons.

The Semiconductor Tracker is built around the Pixel Detector, consisting of four layers and nine disks of silicon microstrip detectors with a resolution of 17 μm and 580 μm in transverse and in longitudinal direction respectively.

The outermost layer of the Inner Detector is the Transition Radiation Tracker, which also helps to identify electrons. It consists of straw drift tubes with a diameter of 4 mm, which are filled with a gas mixture of Xenon (70%) and carbon dioxide (27%). It measures the particle tracks in the transverse plane with about 30 layers of straw tubes with a resolution of 130 μm. Electrons are distinguished by measuring the transition radiation emitted when they traverse carbon foils inserted between the straw tube layers.

3.2.2 The Calorimeter System

The Calorimeter System (illustrated in Figure 3.5) surrounds the Inner Detector and measures the energy of particles interacting electromagnetically or hadronically. There are two main parts – the electromagnetic calorimeter and the hadronic calorimeter. Both consist of metal absorber plates interleaved between active detector material, liquid argon in the electromagnetic and hadronic endcap calorimeters and scintillating plastic tiles in the barrel hadronic calorimeter.

The electromagnetic calorimeter, consisting of a barrel and an endcap part, is located in the front and measures the energy of electrons and photons. The barrel part covers $|\eta| < 1.475$, whereas the end-caps cover $1.375 < |\eta| < 2.5$. Copper plates are used as absorber material. Electromagnetic showers start in the copper plates. The shower particles ionize the Argon atoms in the active groups. The ionization electrons are collected on electrode boards in the liquid Argon gaps. The thickness of the electromagnetic calorimeter is more than 20 radiation lengths, such that photons and electrons are completely absorbed.

The hadronic calorimeter surrounds the electromagnetic calorimeter and captures

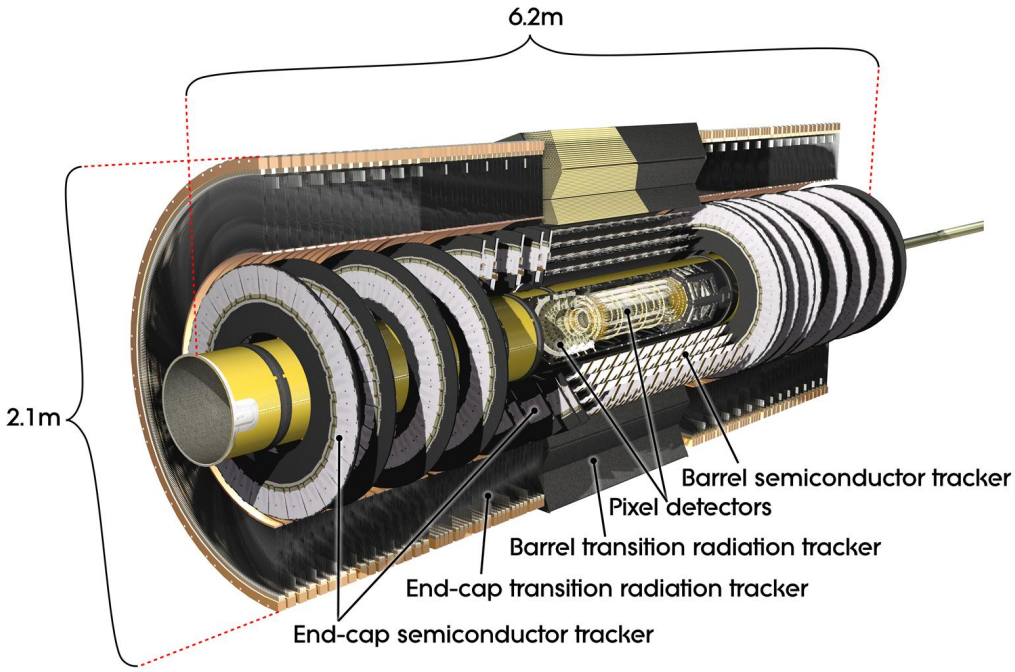


Figure 3.4: The ATLAS Inner Detector [23].

hadrons traversing the electromagnetic calorimeter. It covers the range $|\eta| < 4.9$ and consists of three different components: The Tile Calorimeter utilizes steel absorber plates and scintillating tiles as active material covering $|\eta| < 1.7$, whereas the hadronic endcap calorimeter uses copper absorber plates and liquid argon as active material in the range $1.5 < |\eta| < 3.2$. The forward liquid-argon calorimeter consists of three layers with tungsten absorber plates and covers $3.1 < |\eta| < 4.5$. The thickness of the hadronic calorimeter is greater than 10 radiation lengths in order to absorb all hadrons.

3.2.3 The Muon Spectrometer

Apart from neutrinos, muons are the only Standard Model particles, that pass the calorimeters. The Muon Spectrometer (shown in Figure 3.6) is the outermost and by far the largest component of the detector. It uses a toroidal magnetic field of 0.5 to 1 T to bend the muon tracks. Its pseudo-rapidity range is divided into three regions: The barrel part covers the region $|\eta| \leq 1.0$, while the regions $1.4 \leq |\eta| \leq 2.7$ are

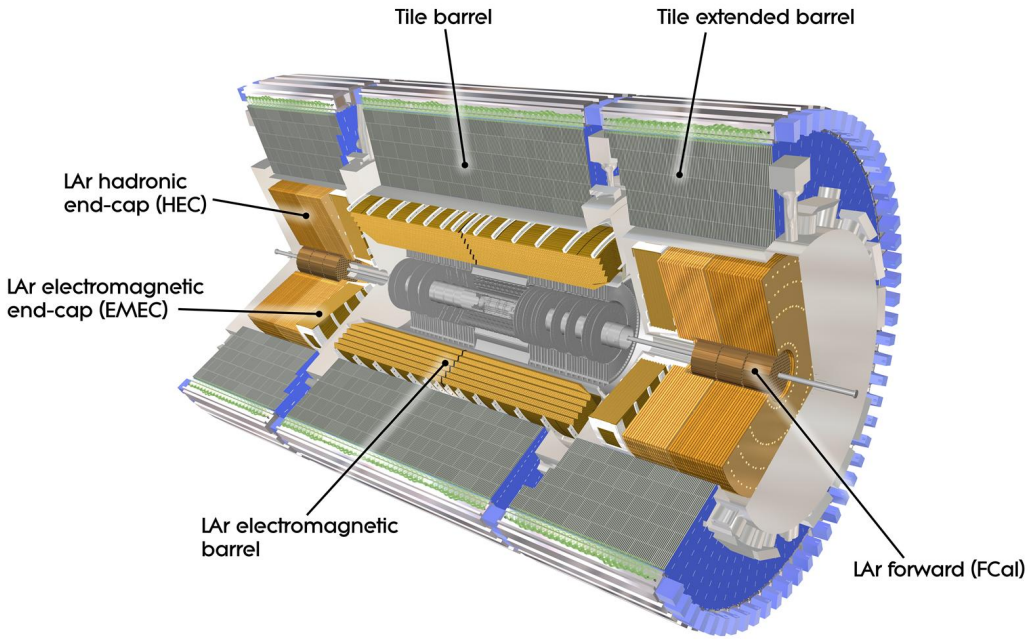


Figure 3.5: The ATLAS Calorimeter System [23].

covered by two separate endcap magnets. The η regions in between are covered by both magnet systems.

Two different kinds of muon chambers are used: Precision tracking chambers and trigger chambers. The precision muon tracking chambers allow for very precise measurement of the momenta of muons. The momentum resolution for 1 TeV muons is 10%, dominated by the chamber position resolution of 35 μm and the chamber alignment accuracy of 30 μm, while it is 3-4% for 100 GeV muons. Multiple scattering starts to become relevant. Two types of tracking chambers are used:

The Monitored Drift Tube chambers (MDT) consist of aluminium drift tubes of 30 mm diameter filled with a gas mixture of Argon (93%) and carbon dioxide (7%) at 3 bar pressure. Since there are higher background rates of neutrons and γ-rays close to the beamline, Cathode Strip Chambers (CSC), operated with Argon (30%), carbon dioxide (50%) and tetrafluoromethane CF_4 (20%), are used in the very forward regions of the inner end-caps wheels.

Two types of muon trigger chambers are made use of: Resistive Plate Chambers (RPC), in the barrel region ($|\eta|$ <1.05) and Thin Gap Chambers (TGC) in the end-

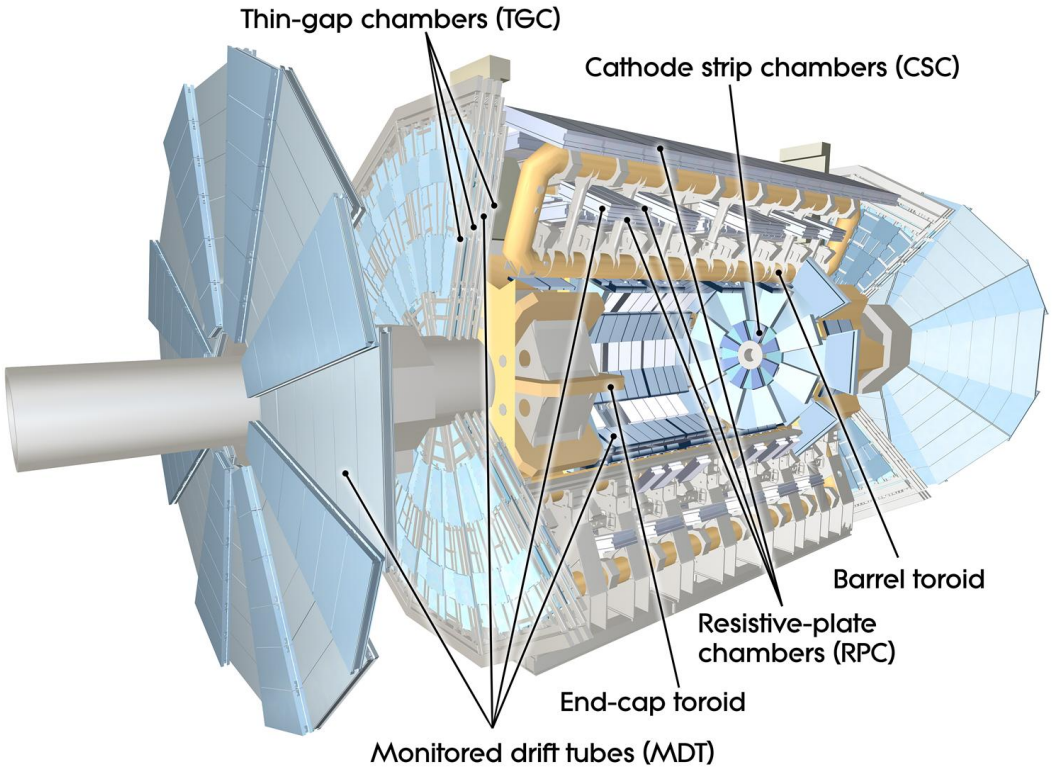


Figure 3.6: The ATLAS Muon Spectrometer with eight barrel toroid coils and two endcap toroid magnets in separate cryostats. The muon chambers are installed on the barrel toroid coils or on separate wheel-shaped support structures in the end-cap [23].

cap regions ($1.05 < |\eta| < 2.4$). The trigger chambers have a poorer spatial resolution than the tracking chambers but provide fast signals for the muon trigger and for the determination of the proton bunch collision time that the muon belongs to.

3.2.4 The Trigger System

At the LHC, proton bunch collisions take place at a frequency of 40 MHz. A collision event contains about 1 MB of information. The resulting data rate would be much too high to be recorded. Therefore, the data rate is drastically reduced by a two-stage trigger system down to a maximum recording rate of 1 kHz [25]. The trigger scheme is presented in Figure 3.7.

The first trigger stage (level-1 trigger) is completely hardware-based and uses the transverse momentum p_T, the missing transverse energy E_T^{miss} and the energy of

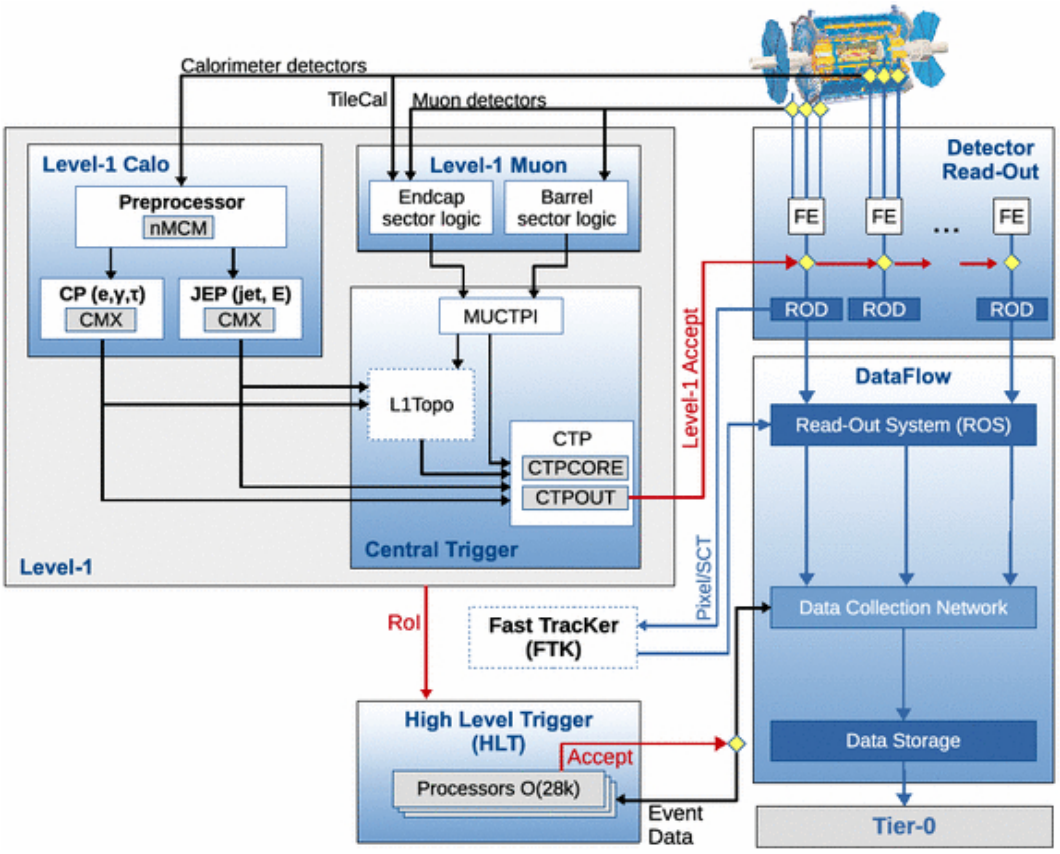


Figure 3.7: Schematics of the ATLAS trigger system [25].

hadron jets to reject events and define regions of interest (RoI) for evaluation by the next trigger stage.

The second trigger stage is called High-Level Trigger (HLT) and is software-based. It utilizes simplified versions of the reconstruction algorithms used by ATLAS for the measurement of particle and event properties.

Chapter 4

The search for top squarks

In this thesis, the search for $\tilde{t}$-quarks is examined by using multivariate analysis techniques. The considered model together with the final state is presented in this chapter. This is followed by a short introduction of physical objects, by analyzing the different background events and by describing the Monte-Carlo event generation.

Figure 4.1 depicts a sketch of the simplified $\tilde{t}_1$-model used throughout this thesis. The $\tilde{t}_1$ is assumed to be produced pairwise in the pp-collisions. Each $\tilde{t}_1$ then decays into a fully hadronically decaying top and in $\tilde{\chi}_1^0$, which is assumed to be stable and to leave the detector system undetected. So the final state of interest consists of six jets, two of which are b-tagged, and a large amount of $E_{\mathrm{T}}^{\mathrm{miss}}$. The masses of the two sparticles are considered to be the only free parameters of the model, while all remaining sparticles are decoupled.

4.1 Reconstruction of physics objects

The physics objects considered here are jets, electrons, muons and the missing transverse energy. Hadronic τ lepton are identified from jets using substructure information [27].

The primary interaction vertex reconstruction uses charged particle tracks reconstruction in the Inner Detector [28]. Events are required to obtain at least one vertex with at least two well reconstructed tracks [29]. Well-reconstructed tracks are required to have a transverse momentum p_T >400 MeV and a pseudo-rapidity $|\eta|$ <2.5. The primary interaction vertex is defined as the vertex with a maximum $\sum p_T^2$ of the associated tracks. The remaining reconstructed vertices are called pileup vertices.

The anti-k_t algorithm [30] with a radius parameter of $R = 0.4$ is employed to reconstruct jets starting from topological clusters [31]. Area-based corrections using calorimeter information are performed for contributions from pileup interactions [32].

Vertex reconstruction of b-hadron decays is an important ingredient for b-tagging,

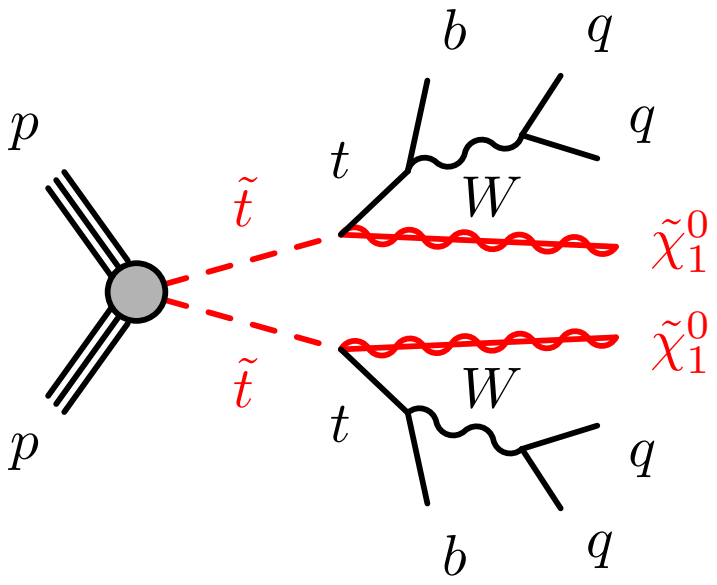


Figure 4.1: Sketch of pair production of top squarks with fully-hadronic decay into final states with b-quark jets of which two are from b-quarks and missing energy due to the lightest super-symmetric particle $\tilde{\chi}_1^0$ leaving the detector undetected [26].

the identification of b-quark jets. Since b-hadrons have a long lifetime (some more than 1 ps [33]), it is possible to distinguish between b-jets and jets from lighter quarks or gluons. Multivariate methods, neural networks and Boosted Decision Trees (BDT), are used in order to collect the maximum information for b-tagging [34]. A jet is b-tagged if it satisfies the 77% selection efficiency working point at a misclassification rate of less than 1% for light jets [35]. Both numbers are extracted from simulated $t\bar{t}$-events.

Electrons are reconstructed in the range $|\eta| < 2.47$ and $p_T > 7$ GeV using energy deposit in the electromagnetic calorimeter matching a track in the Inner Detector in energy and direction [36]. Likelihood-based multivariate identification techniques are applied to discriminate electrons from jets. Electron candidates that satisfy the Loose LH selection criteria are selected if there is not any jet in the cone $0.2 < \Delta R < 0.4$ or a b-jet within $\Delta R < 0.2$. Light jets within $\Delta R < 0.2$ are discarded as electrons. The electromagnetic showers of electrons are distinguished from the ones of hadron jets by a likelihood-based method. A reconstructed jet is taken as an electron if

$$\Delta R = \sqrt{(\eta_2 - \eta_1)^2 + (\phi_2 - \phi_1)^2} < 0.2. \tag{4.1}$$

If the jet is b-tagged, it is always kept, and the electron is discarded.

Both the Inner Detector and the Muon Spectrometer are used to reconstruct muons in the range $|\eta| < 2.7$ and $p_T > 6$ GeV. The Inner Detector and Muon Spectrometer tracks are combined into one single track forming the muon. The ΔR separation between a muon and a nearby jet must be greater than 0.4, in order to keep the muon. Otherwise, the jet is kept.

The missing transverse momentum $p_{\mathrm{T}}^{\mathrm{miss}}$ is the negative sum of the transverse momenta of all reconstructed physics objects in an event. Since neutrinos and other only weakly interacting particles are not detected by any elements of the detector, they lead to missing transverse momentum. The magnitude of $p_{\mathrm{T}}^{\mathrm{miss}}$ is the missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$. High $E_{\mathrm{T}}^{\mathrm{miss}}$ is an important criterium in the search for processes beyond the Standard Model.

4.2 Background Processes

Most of the SM background in the 0ℓ channel arises from $t\bar{t}$ where one top quark decays hadronically and the other one leptonically. The produced lepton from the decay is not reconstructed, and the neutrino leads to an amount of $E_{\mathrm{T}}^{\mathrm{miss}}$ after preselection comes from $t\bar{t}$, where top quarks are produced, which decay into a W-boson and a b-quark each, where the W-boson decays either into two quarks, or into a lepton and a neutrino with the lepton either being falsely classified as a jet through a hadronic decay (τ) or not being reconstructed (e, μ) leading to $E_{\mathrm{T}}^{\mathrm{miss}}$. The second largest background arises from the production of a Z-boson with heavy jets, where the Z-Boson decays into two neutrinos $Z \to \nu\bar{\nu}$, leading to very high $E_{\mathrm{T}}^{\mathrm{miss}}$. Another important background process is the production of a W-boson in combination with heavy jets, where the W-boson mostly decays via $W \to \ell\nu$, and the lepton escapes directly. Further background processes are single top quark production in combination with at least one jet. In comparison to $t\bar{t}$, single top quarks are produced weakly and were discovered at the Fermilab Tevatron Collider in 2009 [37, 38]. The production of $t\bar{t} + Z$ where the Z-boson decays into neutrinos also contributes to the final states. Processes involving a pair of two bosons, also known as diboson, are rare, as well as $t\bar{t} + W$ processes. Single Photon and $tt\gamma$ processes are negligible [26].

4.3 Event Selection

In order to search for top squark $\left(\tilde{t}_1\right)$ decays, signal regions (SR) are defined [26] by cuts which suppress the SM background while keeping the signal events with high efficiency. Two signal regions are used to search for high top squarks masses: signal

region A (SRA) for the case of large mass differences $\Delta m(\tilde{t}_1, \tilde{\chi}_1^0)$ between top squarks and the neutralinos $\tilde{\chi}_1^0$ and signal region B (SRB) for smaller, intermediate mass differences $\Delta m(\tilde{t}_1, \tilde{\chi}_1^0)$. SRA is optimized for $m_{\tilde{t}_1}$ =1 TeV and $m_{\tilde{\chi}_1^0} = 1$ GeV, SRB for $m_{\tilde{t}_1} = 600$ GeV and $m_{\tilde{\chi}_1^0} = 300$ GeV.

Both signal regions have the following preselection in common: A missing transverse momentum trigger is required which is fully efficient for offline calibrated missing transverse energies $E_{\mathrm{T}}^{\mathrm{miss}} > 250$ GeV. Thus, a cut on $E_{\mathrm{T}}^{\mathrm{miss}} > 250$ GeV is applied. In addition, at least four jets are required, where the transverse momenta of the two most energetic ones must be at least 80 GeV, while the third and forth jet have to have transverse momenta greater than 40 GeV. In addition, two of the b-tagged jets are required to have a b-tag. This preselection is also the starting point for all studies presented in the thesis.

Events with wrongly measured E_T^{miss}, due to multijet, arising from mismeasured jets, will be rejected by requiring the difference between the azimuthal angles $\left|\Delta\phi\left(jet^{0,1,2}, \boldsymbol{p}_T^{miss}\right)\right|$ of the three highest p_T jets and the missing transverse momentum vector $\mathbf{p}_T^{\mathrm{miss}}$ to be greater than 0.4.

The track transverse momentum p_T^{track} is defined as the sum of the transverse momenta of all ID-tracks belonging to the primary vertex. The corresponding missing track transverse momentum and energy are denoted by $\mathbf{p}_T^{\mathrm{miss,track}}$ and $E_T^{\mathrm{miss,track}}$. It is required that $E_T^{miss,track} > 30$ GeV and

$$\left|\Delta\phi\left(\boldsymbol{p}_T^{miss}, \boldsymbol{p}_T^{miss,track}\right)\right| < \frac{\pi}{3}. \tag{4.2}$$

Depending on the boost of the decay products, sometimes the six jets with a radius parameter $R = 0.4$ are either all reconstructed individually or if two jets are a distance R of less than $2R$ apart from each, some of the jets merge. In the latter case, hadronic top decays must be reconstructed by the anti-k_t-clustering algorithm with $R = 0.8$ or $R = 1.2$ reclustering. For high top squark masses, two $R = 1.2$ jets (with mass $m^0_{\mathrm{jet},R=1.2}$ and $m^1_{\mathrm{jet},R=1.2}$) and at least one top-candidate, i. e. $m^0_{jet,R=1.2} > 120$ GeV, are required. Three different regions are defined: one (TT) with a second top candidate ($m^1_{jet,R=1.2} >$120 GeV), one (TW) without a second top candidate but with a W candidate (60 GeV$< m^1_{jet,R=1.2} <$ 120 GeV), and one (T0) where the second jet is not even a W candidate ($m^1_{jet,R=1.2} <$ 60 GeV). Each region has different optimized selection cuts and signal-to-background-ratios in $t\bar{t}$-events.

Since the neutralinos $\tilde{\chi}_1^0$ are undetectable, the missing transverse energy is a highly important discriminating variable. Therefore, background events with semi-leptonic $t\bar{t}$ decays, where the lepton is not reconstructed, are rejected by requiring the transverse mass $m_T = \sqrt{E_T^2 - p_T^2}$ of $\mathbf{p}_T^{miss}$ and the b-tagged jet with minimal azimuthal

distance to $\mathbf{p}_T^{miss}$ to be

$$m_T^{b,min} = \sqrt{2p_T^b E_T^{miss} \left(1 - cos\left[\Delta\phi\left(\boldsymbol{p}_T^b, \boldsymbol{p}_T^{miss}\right)\right]\right)} > 200 \text{ GeV}. \tag{4.3}$$

As shown in Figure 4.2, the majority of $t\bar{t}$ events is rejected by this requirement while minor proportions of the two signal models are rejected.

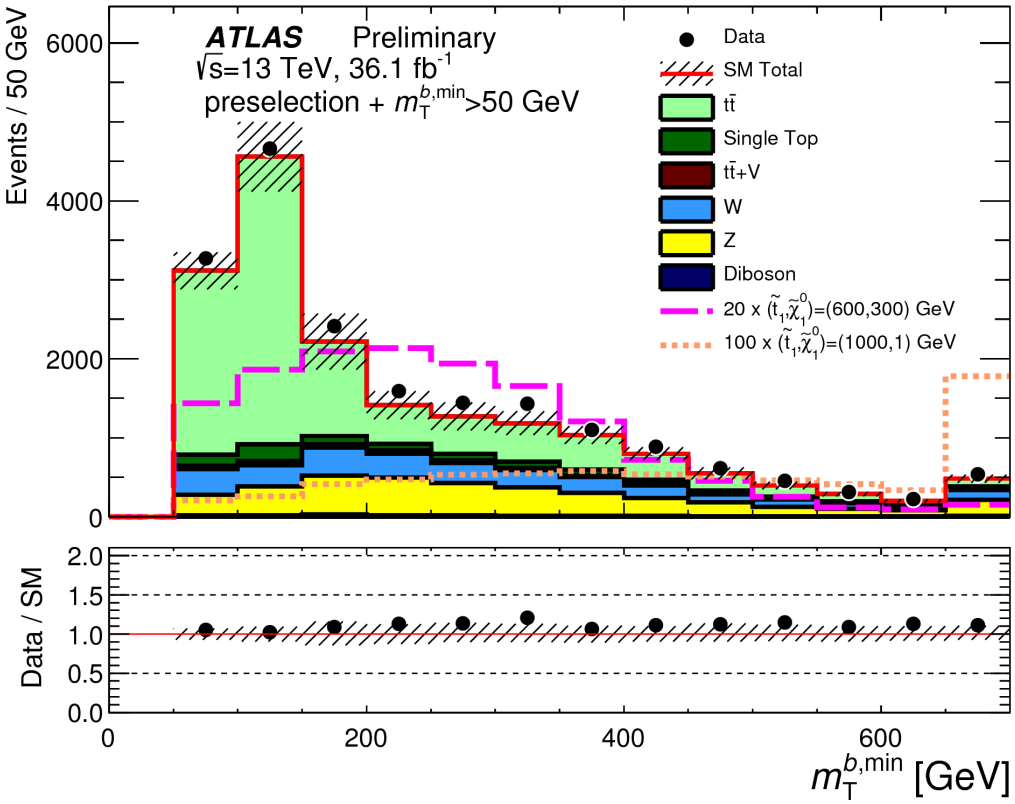


Figure 4.2: $m_T^{b,min}$ distribution for signal (purple and orange dashed and dotted lines) and background (histograms) events after preselection and one further cut $m_{\mathrm{T}}^{b,\mathrm{min}} > 50$ GeV. $t\bar{t}$ will be surpressed [26].

The following selection criteria are used for signal region A only: The mass $m^0_{\mathrm{jet},R=0.8}$ of the leading $R = 0.8$ reclustered jet must be greater than 60 GeV. It can be interpreted as a W candidate.

The transverse momentum of an individual neutralino cannot be evaluated, since

there are two of them. The stransverse mass m_{T2} is defined by [39]:

$$m_{T2}^2 = \min_{\mathbf{q}_1+\mathbf{q}_2=\mathbf{p}_T^{\text{miss}}} \left[\max\{m_T^2\left(\boldsymbol{p}_T^{\bar{t}}, \boldsymbol{q}_1\right), m_T^2\left(\boldsymbol{p}_T^{t}, \boldsymbol{q}_2\right)\}\right], \tag{4.4}$$

where two the transverse momenta of top-candidates are denoted by $\mathbf{p}_T^t$ and $\mathbf{p}_T^{\bar{t}}$. Two methods have been used to form a top quark candidate, a χ^2 and a ΔR_{min} based method. In both cases, the two b-jets with the highest b-tagging BDT-score are taken as b-jets b_1 and b_2 when forming the top candidates. The remaining jets are regarded as light quark jets. In the case of the χ^2 method, $\chi^2 = \frac{(m_{cand}-m_{true})^2}{m_{true}}$ is minimized by using all pairs of light quark jets and m_{true} =80.4 GeV (W-boson mass). The two W candidates with minimal χ^2 are then combined with b_1 and b_2 jets to form top candidates by minimizing $\chi^2 = \frac{(m_{cand}-m_{true})^2}{m_{true}}$ with m_{true} = 172 GeV (top quark mass). The ΔR_{min} method employs $\Delta\phi$ in the transverse plane instead of χ^2 to find the W candidates. The masses of the top candidates obtained by the ΔR_{min} method are denoted by m_{jjj}^0 and m_{jjj}^1. The top candidates going into the stransverse mass calculation are the ones from the χ^2 method. Events with a transverse mass of $m_{\text{T2}}^{\chi^2}$ <400 (500) GeV respectively are discarded. The ΔR_{min} candidates are used in another signal region not described.

For signal region B, $m_T^{b,max}$ is constructed additionally from the b-jet farest away from $\mathbf{p}_T^{\text{miss}}$. Table 4.1 summarizes all cuts.

Figure 4.3 shows the 95% CL upper limits on the $\tilde{t}_1\overline{\tilde{t}_1}$ production cross section decay branching as a function of the top squark and neutralino mass for the cut-based signal selection described above. The solid red line depicts the observed limits. The area below it is excluded at 95% confidence level. The question analyzed in this thesis is whether these limits can be improved with the help of multivariate analysis methods.

4.4 Monte-Carlo Event Simulation

In order to search for supersymmetric particles, it is necessary to simulate them. Signal models are generated with MadGraph5_aMC@NLO [40], PYTHIA8 [41] and EvtGen [42] v.1.2.0. The matrix elements are calculated at tree level and matched with the parton showering via the CKKW-L prescription [43]. In order to generate the signal samples, the parton distribution function set NNPDF2.3LO [44] is used together with the A14 fine-tuning parameter set [45]. Cross-sections calculations are based on next-to-leading order logarithmic accuracy, matched to next-to-leading order supersymmetric QCD [46–48], which might also serve for possible future accelerators [49].

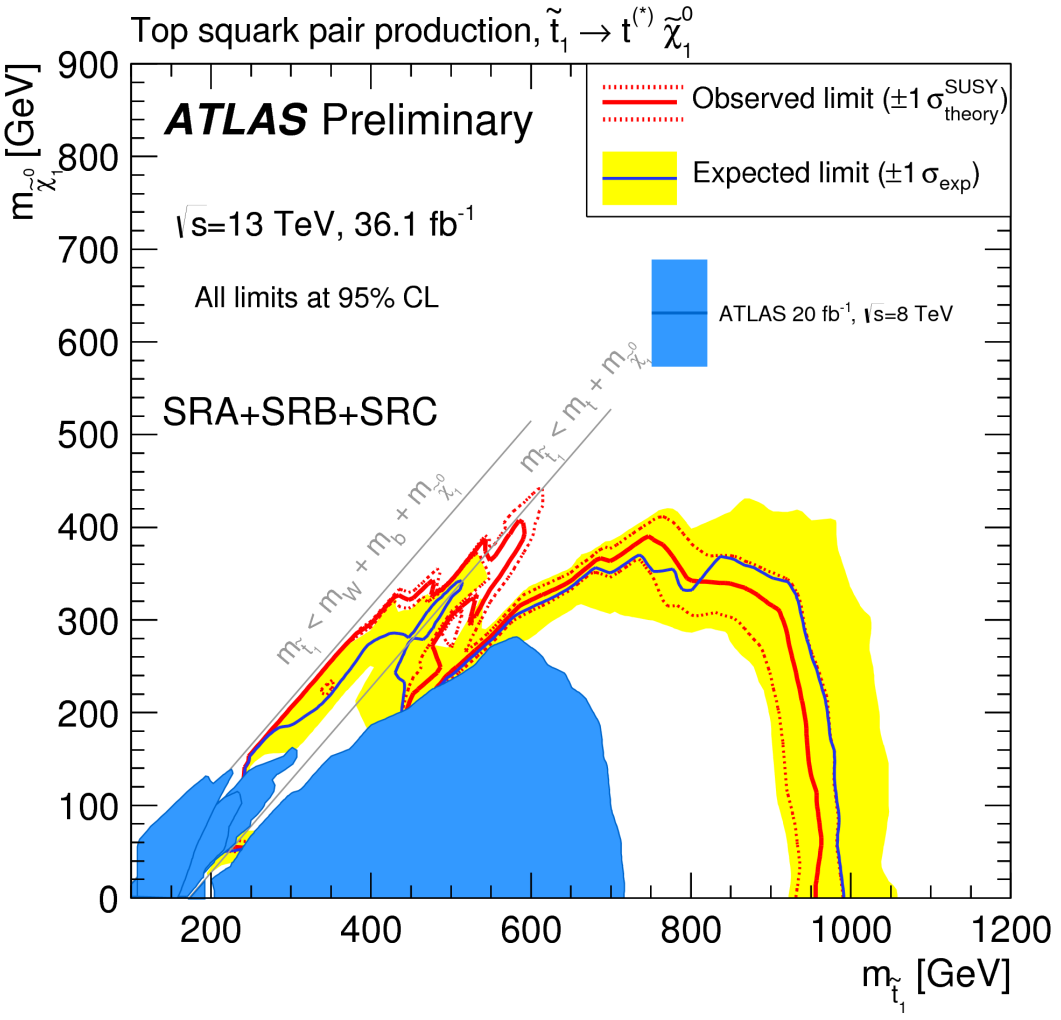


Figure 4.3: Exclusion limits in the $\tilde{t}_1 - \tilde{\chi}_1^0$ mass plane with the cut method for signal selection [26].

Table 4.1: Selection criteria for SRA and SRB [26]

Signal Region		**TT**	**TW**	**T0**
	$m^0_{\text{jet},R=1.2}$	$> 120\,\text{GeV}$		
	$m^1_{\text{jet},R=1.2}$	$> 120\,\text{GeV}$	$[60, 120]\,\text{GeV}$	$< 60\,\text{GeV}$
	$m_{\text{T}}^{b,\text{min}}$	$> 200\,\text{GeV}$		
	$N_{b-\text{jet}}$	≥ 2		
	τ-veto	yes		
	$\left\|\Delta\phi\left(jet^{0,1,2}, \boldsymbol{p}_T^{miss}\right)\right\|$	> 0.4		
A	$m^0_{\text{jet},R=0.8}$	$> 60\,\text{GeV}$		
	$\Delta R\,(b, b)$	> 1	-	
	$m^{\chi^2}_{\text{T2}}$	> 400 GeV	> 400 GeV	> 500 GeV
	$E_{\text{T}}^{\text{miss}}$	$> 400\,\text{GeV}$	$> 500\,\text{GeV}$	$> 550\,\text{GeV}$
B	$m_{\text{T}}^{b,\text{max}}$	$> 200\,\text{GeV}$		
	$\Delta R\,(b, b)$	> 1.2		

Likewise, simulated events are used for the description of the SM background processes. Top-quark pair production, where at least one of the top quarks decays into a lepton and the single top production are simulated with POWHEG-Boxv.2 [50] and PYTHIA6 [51] with the CT10 PDF set [52] and the Perugia set [53]. $Z+$ jets and $W+$ jets are generated with SHERPAv2.2.1 [54] and the NNPDF3.ONLNLO [42] set. MadGraph 5_a MC@NLO [40] and PYTHIA8 [41] generate $t\bar{t} + V$ and $t\bar{t} + \gamma$ samples with NNPDF3.0NLO [55], for which NNPDF2.3LO [44] with the A14-set [45] is used. SHERPAv2.2.1 [54] and the CT10 [52] set generate diboson samples, whereas SHERPAv2.1 [54] and the CT10 set [52] generate $V\gamma$ samples.

The detector simulation [56] is performed using GEANT4 [57]. The simulation of the calorimeter response can be achieved with a full or a fast simulation [58]. Events involving the interaction of more than one proton pair are generated with the MSTW 2008 [59] and the A2 set [60].

Chapter 5

Machine Learning

In this thesis, the possibility of increasing the signal significance is investigated by using machine learning algorithms for the search for top squarks in the full hadronic decay channel using simulated signal and background events. The Monte Carlos (MC) events are split into two halves: One half is used to train the algorithms (training sample) and the other half serves to apply the event selection to (test sample). After training with MC simulated events, the algorithm can also be applied to the data.

For the analysis, the TMVA (Toolkit for Multivariate Data Analysis with ROOT) program package [61] was employed.

5.1 Boosted Decision Trees

One widely-used machine learning method is the *Boosted Decision Trees* (BDT) [61]. A decision tree is illustrated in Figure 5.1. It is a binary tree consisting of nodes at which the data are divided into two subsets, a signal-like and a background-like subset. Variables are used, which separate best at the node. The purity $p = \frac{S}{S+B}$ gives the fraction of signal events at the node.

Before determining the optimum cuts for the top squark search, N_{cuts} grid points are chosen for each of the N_{var} input variable, picked at equal distance: A variable var^j $(j = 1, ..., N_{var})$ is selected for which the maximum value of the training sample is var^j_{max}, whereas the corresponding minimum is denoted by var^j_{min}. Then the grid points var^j_i with $i = 1, ..., N_{Cuts}$ are determined by

$$var^j_i = var^j_{min} + \left(var^j_{max} - var^j_{min}\right) \cdot \frac{i}{n_{Cuts} + 1}. \tag{5.1}$$

For illustration, a single variable is assumed, e. g. the missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$ with a maximum value of 1000 GeV and a minimum value of 200 GeV. For three grid points, $E^{\mathrm{miss}}_{T,1}$ =400 GeV, $E^{\mathrm{miss}}_{T,2}$ =600 GeV and $E^{\mathrm{miss}}_{T,3}$ =800 GeV define the possible cut values.

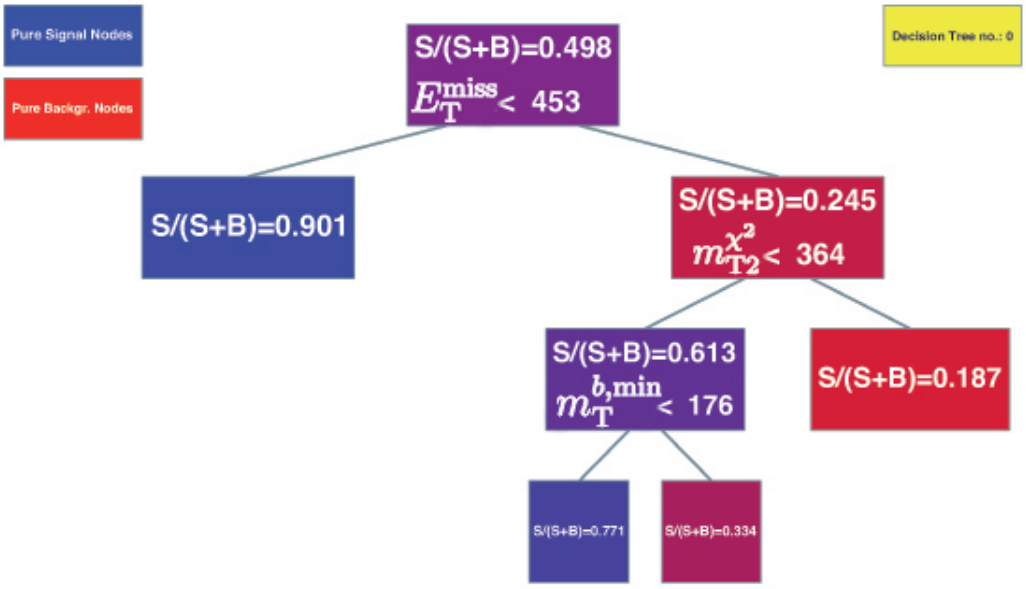


Figure 5.1: Example of the training of an initial decision tree for the top squark search chosen. The colors visualize the signal purity $p = \frac{S}{S+B}$. The blue color stands for a pure signal sample with $p = 1$ and red for a pure background $p = 0$. The colors illustrate the local purity of the node. The bluer, the larger the signal fraction.

For each variable n_{Cuts} many possible cuts are received. The question arises which variables and which cut values are optimal. As a criterion, the best signal purity p

$$p = \frac{S}{S+B} \tag{5.2}$$

is used along with the so-called Gini-Index $p \cdot (1-p)$. The index has its maximum at $p = 0.5$ and is symmetric with respect to it, meaning that it does not distinguish between the case of M signal and N background events and the case of N signal and M background events. For $p = 0$ and $p = 1$ the Gini index equals zero. The smaller the Gini index, the greater the excess of either signal or background. In order to decide which variable and which cut value to use, the sum of the two Gini Indices for the signal-like region and the complementary background-like region of each variable and of each cut value is taken, being weighted by the relative fraction of events (in the following just called Gini index)

$$\mathcal{G} = w_1 \cdot p_1 \cdot (1-p_1) + w_2 \cdot p_2 \cdot (1-p_2) = \frac{N_1}{N_{tot}} \cdot \frac{S_1 \cdot B_1}{N_1^2} + \frac{N_2}{N_{tot}} \cdot \frac{S_2 \cdot B_2}{N_2^2}, \tag{5.3}$$

which is calculated for each variable and each cut value. Here $N_i = S_i + B_i$ is the

total number of (weighted) events in node i, and $N_{tot} = N_1 + N_2$ is the total number of (weighted) events in the mother node. *Equation* 5.3 can be simplified:

$$\mathcal{G} = w_1 \cdot p_1 \cdot (1 - p_1) + w_2 \cdot p_2 \cdot (1 - p_2) = \frac{1}{N_{tot}} \cdot \left(\frac{S_1 \cdot B_1}{N_1} + \frac{S_2 \cdot B_2}{N_2} \right) \tag{5.4}$$

By minimizing Equation 5.4, the so-called separation gain $\mathcal{S}$ is maximized:

$$\mathcal{S} = p_0 \cdot (1 - p_0) - \mathcal{G} \tag{5.5}$$

Here, p_0 is the purity of the mother node. In order to determine the Gini indices for $E_{\mathrm{T}}^{\mathrm{miss}}$ at a cut value of 600 GeV, the Gini Indices for $E_{\mathrm{T}}^{\mathrm{miss}}$ <600 GeV and for $E_{\mathrm{T}}^{\mathrm{miss}}$ ≥600 GeV are calculated and added together. The variable and the cut value are chosen for which the Gini index is minimal compared to all other choices.

The cut giving best separation in the first node at the initial tree is $E_{\mathrm{T}}^{\mathrm{miss}} \leq 453$ GeV for the exemplary training. Events satisfying the labeled condition $E_{\mathrm{T}}^{\mathrm{miss}}$ <453 GeV in Figure 5.1 are piped to the right node in the chosen scheme of this thesis. The right node, fulfilling the cut condition, is also called "true" branch. The right branch is the more background-like event branch and the left one is the more signal-like branch, which means that the purity of the right branch p_r is smaller than the purity of the left branch: $p_r \leq p_l$.

More than 90% of the events in the left branch are signal events (purity $p = 0.901$). This node does not split further due to the hypothetical new background-like node being smaller than the minimal node size allows. The minimal node size will be discussed in the next paragraph. The right node has a signal contamination of 25%. These events are further separated using the next best discriminating variable, in this case $m_{\mathrm{T2}}^{\chi^2}$ with a separation edge of 364 GeV. The best discriminating variable cut of the remaining events is $m_{\mathrm{T2}}^{\chi^2} < 364$ GeV. The subsequent right side node has a purity $p = 0.187$ and does not split further. The left side node splits once more for the cut $m_{\mathrm{T}}^{b,\mathrm{min}}$ <176 GeV, resulting in purities of $p = 0.772$ and $p = 0.334$ respectively. When the events at the final split are looked at, then events located on the left side are classified as signal events (and should have a purity $p > 0.5$, whereas the events on the right side are classified as background events (and should have a purity $p \leq 0.5$).

If the number of nodes were not constrained, a purity of either 0 or 1 would mostly be the case for the final nodes. The problem of such training is then that the tree essentially memorizes each event in the training sample individually. This situation is called overtraining. The events, not being trained with (test samples) have a much worse score. Therefore, the amount of branchings needs to be limited. Several parameters control when a node will not be split further. The parameter maximal depth defines the maximal number of branchings. The default value is three, which

means that the events can be split three times at most until reaching the final node, as depicted in Figure 5.1. Another criterion is the minimal node size which determines the minimum fraction of events at each node. The default value is 2.5%.

A single decision tree, however, gives poorer results than the Cut and Count method and is highly prone to overtraining. Therefore, boosting [62] is introduced: Multiple trees are trained, which return poor results as individual trees, but good ones as a collective. The set of all decision trees used for the boosting is called decision forest. The most commonly applied Boosting Algorithm is Adaptive Boosting (AdaBoost) [63]. The training of a tree is carried out with weighted events, such that training events with a larger weight have a greater impact on the determination of the cut variables and the cut values. Initially, all signal training events (in total $N_{sig,train}$ events) are weighted with $\frac{1}{N_{sig,train}}$, such that the total number of the weighted events is normalized to one. Equivalently, the initial weights for the background training samples are given by $\frac{1}{N_{bkg,train}}$. After training one tree, the weight of the events which were classified falsely is increased, and the weight of the events that were truly classified is decreased, whereby falsely classified events are those which were categorized by this particular tree as signal even though they were background and vice versa. Then training is repeated with the newly weighted events in order to obtain a new decision tree (see Fig. 5.2). The maximum number of trees is defined by the parameter N_{trees}. After the training of all trees, the score of each event is determined from all trees as will be explained later.

The first tree is trained as explained above with every event having the same weight. Then, the misclassification error err

$$err = \frac{\text{number of misclassified events}}{\text{number of total events}}. \tag{5.6}$$

is determined.

The error fraction is, per construction, $err \leq 0.5$ (otherwise more misclassified than truly classified events, meaning that switching classification would provide better results).

The error fraction is applied to determine the boost factor α:

$$\alpha = \beta \cdot ln\left(\frac{1 - err}{err}\right), \tag{5.7}$$

where β is called the learning rate of the boost algorithm. It ranges typically between 0 and 1 and $\alpha = 0$ for $err = 0.5$. In the entire analysis, the default value 0.5 will be

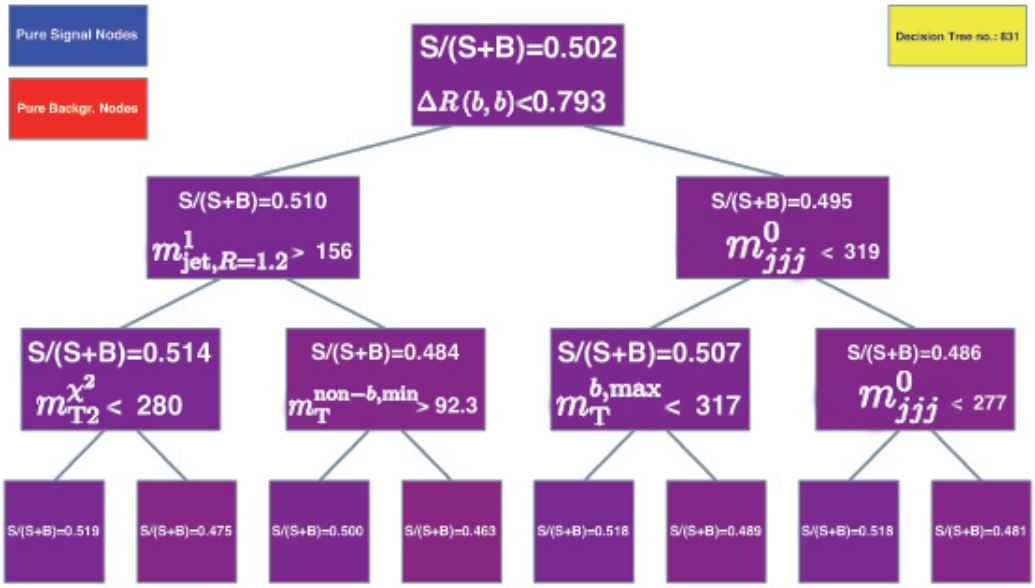


Figure 5.2: Example of an 831st reweighted decision tree in a forest.

used. The weight of a misclassified event i, denoted by w_m^i, is multiplied by

$$exp\,(\alpha) = \left(\frac{1-err}{err}\right)^{\beta}, \tag{5.8}$$

while the weights of correctly classified (true) events j, denoted by w_t^j, are not modified.

Subsequently, the weights are renormalised such that the sum of all weights is one:

$$w^i_{m,new} = \frac{w^i_{m,old} \cdot exp\,(\alpha)}{\sum_{k\in I_m} w^k_{m,old} \cdot exp\,(\alpha) + \sum_{l\in I_t} w^l_{t,old}} \tag{5.9}$$

$$w^j_{t,new} = \frac{w^j_{t,old}}{\sum_{k\in I_m} w^k_{m,old} \cdot exp\,(\alpha) + \sum_{l\in I_t} w^l_{t,old}} \tag{5.10}$$

where I_m and I_t denote the sets of misclassified and truly classified events respectively. This results in $w^i_{m,new} \geq w^i_{m,old}$ and $w^i_{t,new} \leq w^i_{t,old}$. There hardly exists, if at all, any event which is classified truly by all events or falsely. Thus, most of the events are truly classified by a certain number of trees and falsely classified by the rest, meaning that the weight of an event is almost unique.

The higher the level of decision trees, the larger will be the error fraction since it becomes increasingly hard to classify the weighted events. Figure 5.2 shows an example of a decision tree, constructed at a very late stage of training (tree number 831). All purities are very close to $p = 0.5$, resulting in an error fraction close to $err = 0.5$, which indicates that a separation of signal and background events is hardly possible. This means that the boost weight will be equally close to one. Figure 5.3 shows the error fraction as a function of the tree number. For trees with tree numbers above 100, the error fraction is close to $err = 0.5$ and increases only slightly.

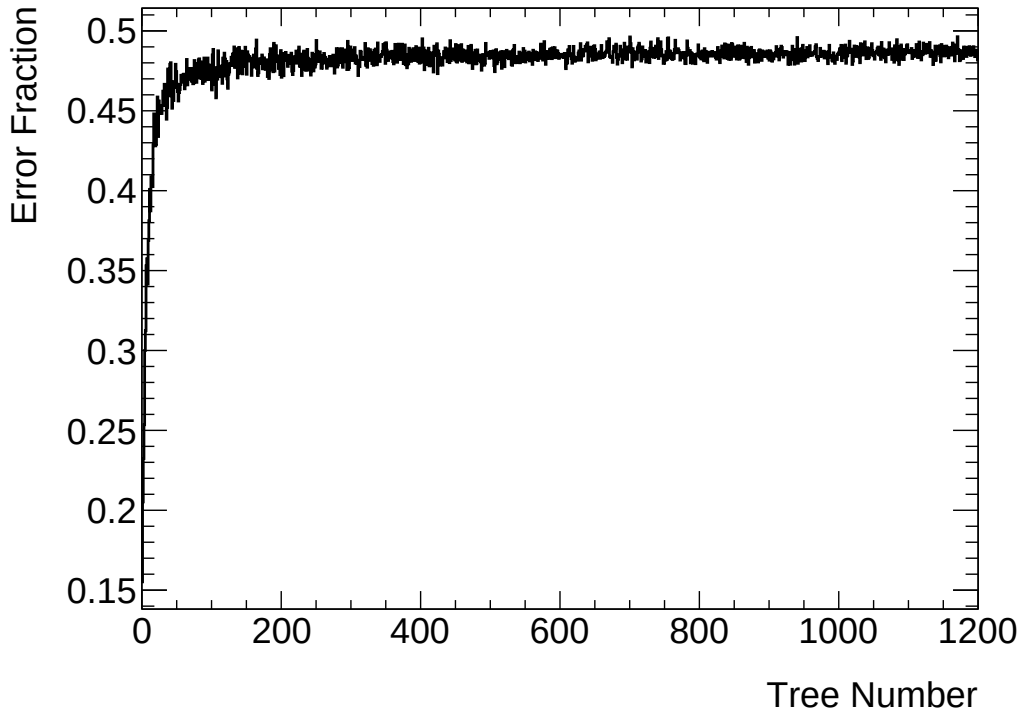


Figure 5.3: The error fraction of each decision tree in the decision forest is depicted for the top squark search. The error fraction takes values close to the poorest value 0.5 for any but the first trees.

The total BDT-score s^i constructed from all decision trees of the event i is calculated by all trees:

$$s^i = \frac{\sum_{m=0}^{NTrees} \alpha_m \cdot q_m^i}{\sum_{m=0}^{NTrees} .\alpha_m}. \tag{5.11}$$

The final boost factor of tree m is denoted by α_m, while the decision of tree m whether event i is signal-like or background-like is denoted by

$$q_m^i = \begin{cases} 1 & \text{if tree } m \text{ classifies event } i \text{ as signal} \\ -1 & \text{else} \end{cases} . \tag{5.12}$$

The smaller the error fraction, the more it affects the final score. The BDT-score is limited inside the interval between -1 and 1 (cf. Equation 5.11). The larger the BDT-score, the more likely the event is a signal event. A BDT-score $s^i = 1$ means that all trees in the forest with boost factor greater zero classify event i as signal. A BDT-score $s^i = -1$ means the opposite. There is a strong reason to believe that the events have the expected properties.

The problems looked at were classification problems in a sense that the events were classified either signal or background. Another type of problems are regression problems, having – instead of a boolean variable as an output – a real-valued variable, i. e. the energy deposit of a particle in a calorimeter. Regression trees are trained such that the error function is minimized.

The main focus of the thesis, however, is the classification problems.

5.2 Neural Networks

Artificial neuronal networks (ANN) are inspired by the structure and working principles of the brains. An input signal activates a first set of neurons. Depending on the internal structure and on possible other input pulses, each neuron generates an output propagated to another set of neurons. Finally, the signal is propagated to the last layer of the ANN, returning the net's output. In this thesis, the so-called Multi-Layer Perceptron approach is applied [64, 65].

The basic layout of an ANN is sketched in Figure 5.4. The information of the four input variables is distributed to the first layer of the network. This layer consists of as many neurons as variables plus an extra, so-called bias node, which generates a constant output signal of $y_0^{(1)} = 1$. Each neuron i of a given layer l is connected to each neuron j of the subsequent layers. For simplicity and also to save computing resources, neurons are only considered to have a direct connection from one layer to the following neighbouring layer (MLP). The strength of these connections is described by weights, $w_{i,j}^{(l)}$, whose values will be optimized during the training of the network after its layout has been established. In general, the signal can be propagated to an arbitrary number of middle layers (also called hidden layers) – each with a

different number of neurons – before reaching the final output-layer. The latter usually consists of one single neuron returning one single score.

The determination which value $y_j^{(l)}$ a given neuron j in the l^{th} layer returns to the input is illustrated in Figure 5.5 and will be discussed in the following. In a first step, the response of all neurons of the previous layer is summarized in the so-called synapse function κ, given by

$$\kappa\left(y_1^{(l-1)}, y_2^{(l-1)}, ..., y_n^{(l-1)}, w_{1,j}^{(l-1)}, ..., w_{n,j}^{(l-1)}, w_{0,j}^{(l-1)}\right) = w_{0,j}^{(l-1)} + \sum_{i=1}^{n} y_i^{(l-1)} w_{i,j}^{(l-1)} \tag{5.13}$$

There are also other possibilities, commonly considered to define κ, whose details can be found in Reference [61]. The synapse function is then connected to the activation function, which is generally any differentiable function mapping $\mathbb{R} \to \mathbb{R}$. The activation function is supposed to detect potential (non-)linear correlations between the input variables of the network which is justified by the Stone-Weierstrass theorem [66, 67]. In this thesis three different activation functions are examined:

- the sigmoid function: $\alpha(x) = \frac{1}{1+exp(-x)}$
- tangens hyperpolicus: $\alpha(x) = \tanh(x)$
- radial function: $\alpha(x) = \exp\left(-\frac{x^2}{2}\right)$.

The output y_{ANN} of a neural network as depicted in Figure 5.4 is given for tangens hyperbolicus as the neuron activation function:

$$y_{ANN}\left(x_1, x_2, x_3, x_4 \mid \mathbf{w}\right) = w_{0,1}^{(2)} + \sum_{j=1}^{5} w_{j,1}^{(2)} \cdot \tanh\left(\sum_{i=1}^{4} x^{(i)} \cdot w_{i,j}^{(1)} + w_{0,j}^{(1)}\right) \tag{5.14}$$

Training of a neural network

As in the case of the training of a BDT, N labeled simulated events are applied to train the network. Each event is represented by an M_{input} dimensional vector z^i containing all considered training variables. The goal of the training is to adjust all weights $w_{i,j}^{(l)}$ such that the so-called error function of all training events is minimized, which is given by

$$E\left(\mathbf{z}^{(1)}, ..., \mathbf{z}^{(N)} \mid \mathbf{w}\right) = \sum_{i=1}^{N} E_i\left(\mathbf{z}^i \mid \mathbf{w}\right) = \frac{1}{2}\sum_{i=1}^{N}\left(y_{ANN}^{(i)} - \hat{y}^{(i)}\right)^2. \tag{5.15}$$

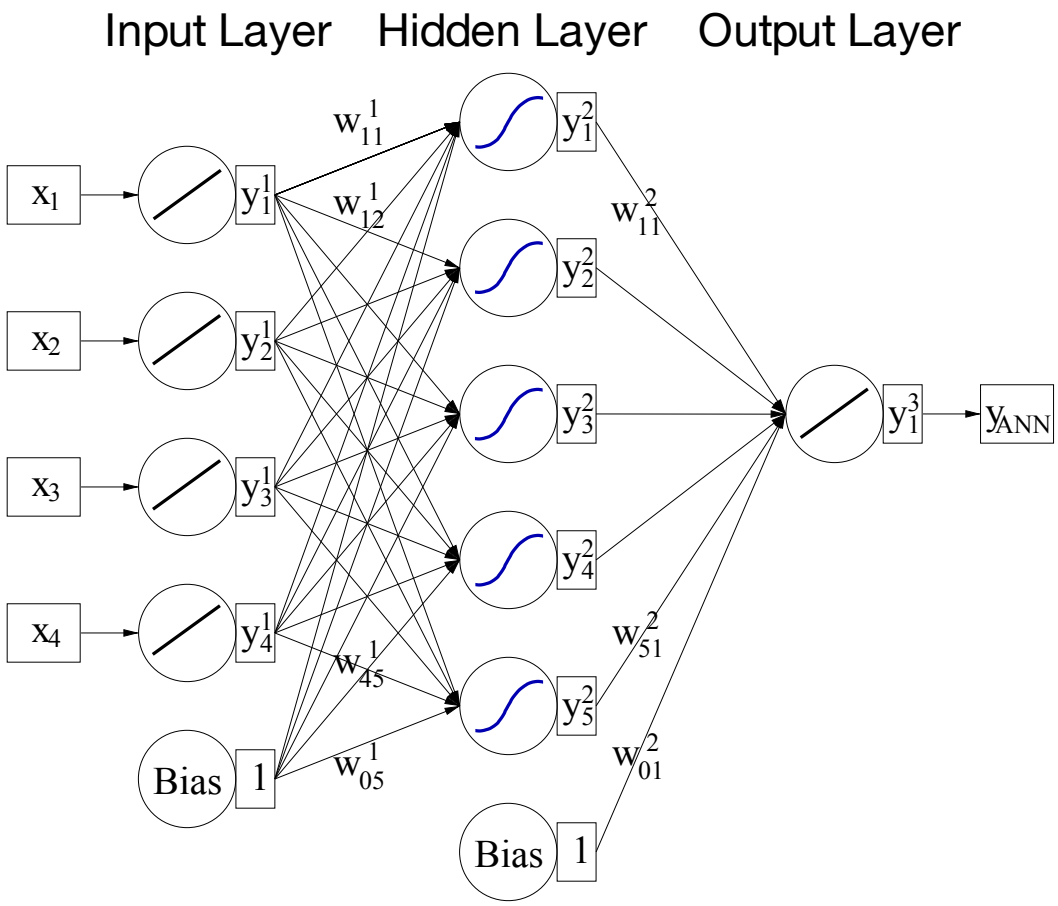


Figure 5.4: Schematic layout of an artificial neural network [61].

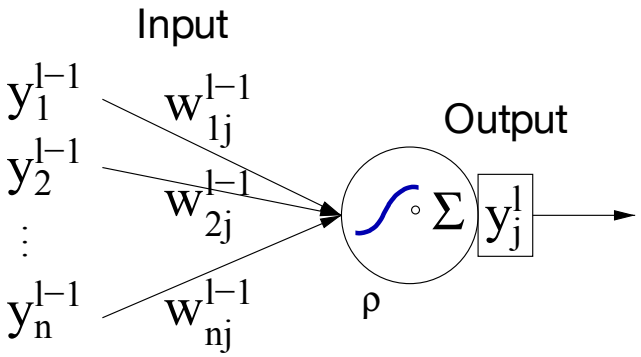


Figure 5.5: Schematic layout of a single neuron [61].

The $\hat{y}_i$ denote the truth-label of each event, whereas signal (background) events are labeled by $y_i = 1(0)$. Due to the complexity, this minimization problem has to be solved by using numerical methods.

Two ways of obtaining the weights of the neural network are considered: the Back-Propagation (BP) and the Broyden-Fletcher-Goldfarb-Shannon (BFGS) method.

The error function reaches its extremum when

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = 0 \tag{5.16}$$

is satisfied. The Back-Propagation method is based on the steepest descent approach. Iteratively, the $\mathbf{w}^{[q+1]}$ weight is adjusted from the $\mathbf{w}^q$ weight via

$$\mathbf{w}^{[q+1]} = \mathbf{w}^{[q]} - \eta \nabla_{\mathbf{w}^{[q]}} E\left(\mathbf{z}^{(1)}, ..., \mathbf{z}^{(N)} \mid \mathbf{w}^{[q]}\right), \tag{5.17}$$

where η is the learning rate fixed to $\eta = 0.02$, which determines the difference between $\mathbf{w}^{[q]}$. The larger the learning rate, the faster the convergence, but the greater the risk of taking too big steps, which often results in divergences. The number of cycles determines how many times this optimization is carried out, and when the algorithm is stopped. The procedure is terminated either if the minimum is reached or after a maximum number of iterations – also called cycles – is met. This limit is an external training parameter, which is optimized in Appendix 3.2

The other algorithm, the BFGS-method, is explained in Appendix 3.1.

An example of a neural network can be found in Figure 5.6. In this case, the network was trained with five input variables. One output variable provides the final results. The weight can be evaluated from the thickness and the color of the lines [61]. The thicker the line, the larger the absolute value of the weight. Blue colors indicate a large negative weight, orange-yellow colors represent a large positive weight, green color indicates a small weight. No line is an indication of very small weights.

5.3 Overtraining

One of the main important aspects in machine learning is to avoid overtraining. The algorithm must not memorize the training events. It is rather supposed to learn the main features of signal and background, in order to filter a potential signal from real data. This section introduces the main concepts of how overtraining is detected and quantified based on arbitrary BDT training as an example. All concepts can be transferred straight forward to neural networks.

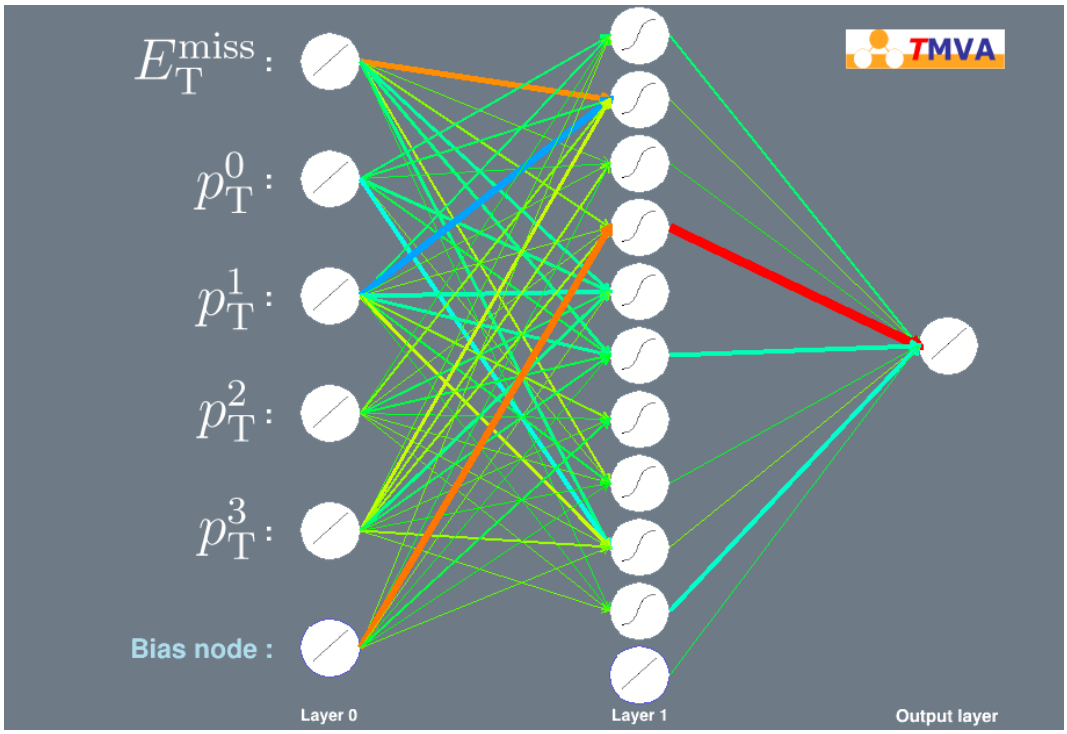


Figure 5.6: An example of a neural network trained with five input variables. The lines indicate the weights of the network. The thickness corresponds to the absolute value of the weights. The colors indicate the sign. Blue colors show the weight being negative, orange-yellow colors indicate a positive weight, whereas green lines correspond to small weights. When no line is plotted, this means that the weight is smaller than the weights marked with a green color.

Figure 5.7 shows the resulting information from a BDT. The red histogram is the distribution of the BDT-score s^i for background test events, whereas the blue histogram is the s^i distribution for signal test events. The dots indicate the distributions for training events. The two histograms are normalized to one.

A BDT is overtrained if the testing and the training samples for signal or background do not share the same shape. Since the differences in the histograms in Figure 5.7 are small, little overtraining can be expected. Overtraining, however, must be quantified for the final analysis.

As a measure of overtraining, i. e. of the level of disagreement between the BDT score distributions for test and training events, a Kolmogorov-Smirnov-Test [68] can be used[1]. With the cumulative, normalized BDT-score probability distribution of the

[1]This is not the Kolmogorov-Smirnov-Score, used by TMVA.

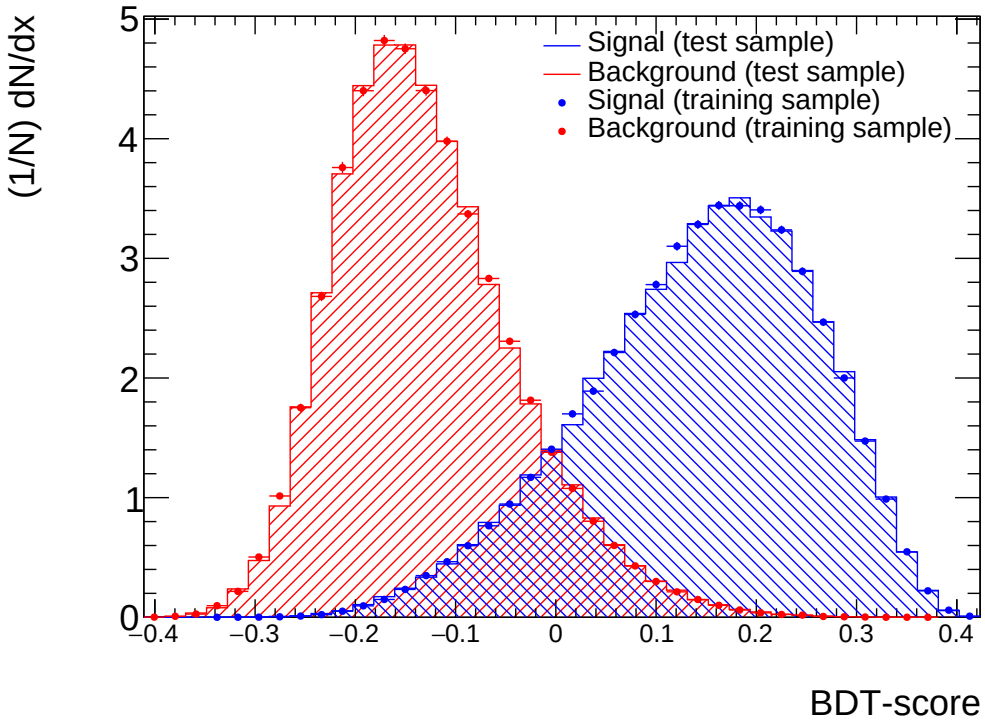


Figure 5.7: BDT-score distribution of signal (blue) and background (red) for training (dots) and testing (lines). A good separation between signal and background can be observed.

background training events

$$F^b_{training}(x) = \frac{\text{number of background training events } i \text{ with BDT-score } s^i \leq x}{\text{number of background training events } i} \tag{5.18}$$

and the corresponding distribution $F^b_{test}(x)$ for background test events, the Kolmogorov-Smirnov-Score k_B of the background events is defined as:

$$k^b = \max_{x \in [-1,1]} \left| F^b_{train} - F^b_{test} \right|. \tag{5.19}$$

There are only finitely many bins to check in the range $[-1, 1]$ in order to find k^b.

For large enough numbers of events N, the relevant quantity is:

$$d^b = k^b \cdot \sqrt{\frac{N^b_{train} \cdot N^b_{test}}{N^b_{train} + N^b_{test}}} \tag{5.20}$$

N^b_{train} is the number of training background events given. d^b can be translated into the significance $1 - \alpha$ to reject H_0. In this thesis, the output of a machine learning algorithm is said to be not overtrained if d^b is $d^b < 1.07$ [68].

5.4 Performance Evaluation

An important characteristic of the performance of machine learning techniques, in fact of any signal-selection method, is the so-called ROC (Receiver Operating Characteristic) Curve [69]. It is based on the test samples. An example of a ROC-Curve is illustrated in Figure 5.8. When a cut in the BDT-score distributions is applied, also part of the signal events is rejected in addition to background events. Then the signal efficiency e_{sig} and background rejection r_{bkg} are defined as

$$e_{sig} = \frac{\text{number of selected signal events in the test sample}}{\text{number of signal events in the test sample}}, \tag{5.21}$$

$$r_{bkg} = \frac{\text{number of rejected background events in the test sample}}{\text{number of background events in the test sample}} \tag{5.22}$$

respectively. The plot provides the essential information on the performance of the selection. The ROC-Curve shown in Figure 5.8 encodes information on the performance of the training. Generally, an algorithm distinguishes background better from signal than another one if its associated ROC-Curve approaches closer the point of full background rejection while keeping all signals ($e_{sig} = r_{bkg} = 1$). On the other hand, the ROC-Curve is bounded by the diagonal from $(1, 0)$ to $(0, 1)$, since it is possible to exchange the labels of background and signal if the curve crosses the diagonal. Thus, the area under the ROC-Curve is a suitable measure to compare various configurations of the same machine-learning algorithm or entirely different algorithms.

The larger the area under the ROC-Curve, the better the performance of the algorithm, and the results of the training. The poorest performance region is the diagonal with an area of 0.5, while the best performance can be found in rectangles with one corner at $(1, 1)$.

Another measure used in this analysis is significance: It is a standard measure in particle physics, which is a basis for the decision whether particles are detected or if certain mass ranges are excluded. More details can be found in [70]. If particles are detected with a significance of more than 5σ, it is assumed that the particle exist.

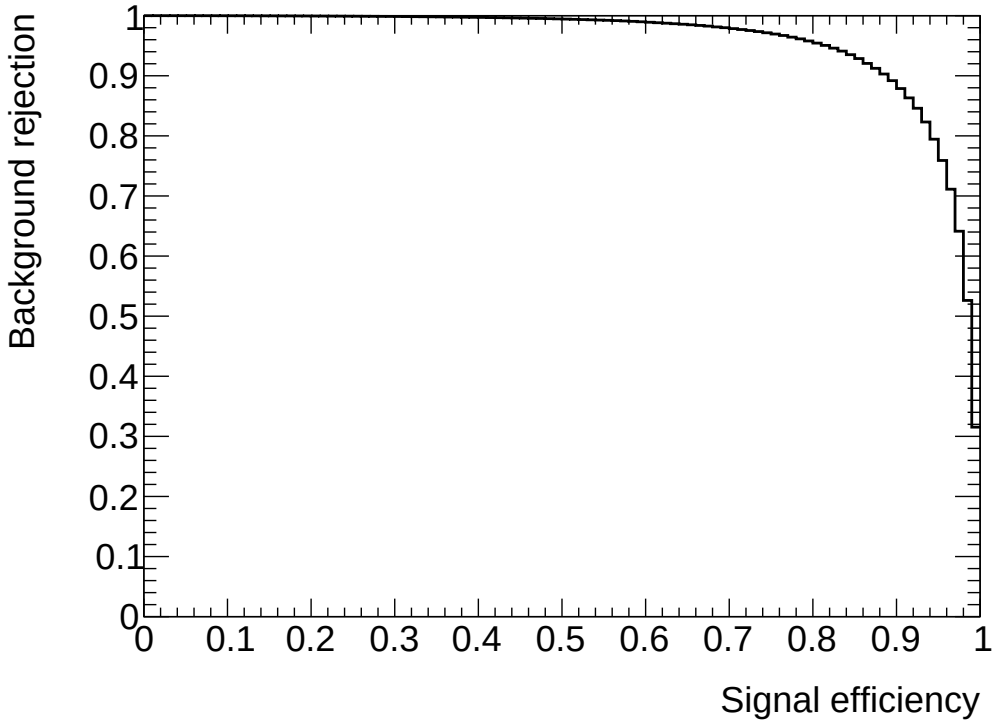


Figure 5.8: ROC-Curve of a BDT. It can be seen that the area under the ROC-Curve is distinctly closer to 1 than to 0.5, showing that the training was successful.

Chapter 6

Analysis

6.1 Input Variables

For multivariate methods, the selection of the input variables is essential. All discriminating variables of the cut-based method were made use of (see Table 6.1).

The exemplary distributions of four input variables after the preselection are shown in Figure 6.1 - 6.4 for ATLAS data taken at $\sqrt{s} = 13$ TeV in 2015 and 2016 by the ATLAS experiment. Figure 6.2 shows the distribution of the number of jets with $R = 0.4$. The preselection requires at least four jets. Events with more than ten jets are very rare. The data are described reasonably well in the SM background MC-simulation. The data are slightly above the SM prediction. Further, no shape of the ratio is recognizable in a sense that the ratio increases or decreases as a function of the number of jets.

Figure 6.1 shows the distribution of the transverse mass $m_{\mathrm{T}}^{\mathrm{non-}b,\mathrm{min}}$ of the non-b-tagged jet closest in φ to the missing transverse momentum. Again, the data are described reasonably by the background MC modulation.

The missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution is shown in Figure 6.3. The preselection requires $E_{\mathrm{T}}^{\mathrm{miss}} > 200$ GeV. Data and MC show good agreement.

Figure 6.4 shows that the distribution of $m^{1}_{\mathrm{jet},R=1.2}$ also has good agreement between data and MC. A cut at 100 GeV would discriminate a lot of background events.

The distribution of the other input variables used in the signal selection can be found in Appendix 1. The Monte-Carlo-simulations are in good agreement with the data.

Table 6.1: List of discriminating variables used in the BDT-based event selection. b-tagged jets are referred to as b-jets. *Candidate* is abbreviated to cand.

Variable	Description
$m^0_{\text{jet},R=0.8}$	Mass of the 1$^{\text{st}}$ leading (in p_T) reclustered $R = 0.8$ jet
$m^0_{\text{jet},R=1.2}$	Mass of the 1$^{\text{st}}$ leading reclustered $R = 1.2$ jet
$m^1_{\text{jet},R=1.2}$	Mass of the 2$^{\text{nd}}$ leading reclustered $R = 1.2$ jet
$\Delta R\,(b, b)$	Distance between the two leading $R = 0.4$-jets
$E^{\text{miss}}_{\text{T}}$	Missing transverse energy
$H^{\text{sig}}_{\text{T}}$	$H^{\text{sig}}_{\text{T}} = E^{\text{miss}}_{\text{T}}/\sqrt{H_{\text{T}}}$ with $H_{\text{T}} = \sum_{jets} \lvert\mathbf{p}_{\text{T}}\rvert$
p^0_{T}	p_T of the of the 1$^{\text{st}}$ $R = 0.4$ jet
p^1_{T}	p_T of the of the 2$^{\text{nd}}$ leading $R = 0.4$ jet
p^2_{T}	p_T of the of the 3$^{\text{rd}}$ leading $R = 0.4$ jet
p^3_{T}	p_T of the of the 4$^{\text{th}}$ leading $R = 0.4$ jet
m_{bb}	Invariant mass of the two leading b-jets
m_{eff}	$m_{\text{eff}} = E^{\text{miss}}_{\text{T}} + H_{\text{T}}$
$m^{\chi^2}_{\text{T2}}$	$m^{\chi^2}_{\text{T2}} = \min_{\mathbf{q}_1+\mathbf{q}_2=\mathbf{p}_M} \left[\max\{m^2_T\left(\boldsymbol{p}^{\bar{t}}_T, \boldsymbol{q}_1\right), m^2_T\left(\boldsymbol{p}^{t}_T, \boldsymbol{q}_2\right)\}\right]$ (cf. Eq. 4.4)
$m^{b,\text{max}}_{\text{T}}$	Transverse mass of the b-jet, which is furthest away in φ from $\mathbf{p}^{\text{miss}}_{\text{T}}$
$m^{b,\text{min}}_{\text{T}}$	Transverse mass of the b-jet, which is closest in φ to $\mathbf{p}^{\text{miss}}_{\text{T}}$
$m^{\text{non}-b,\text{max}}_{\text{T}}$	Transverse mass of the non-b-jet, which is furthest away in φ from $\mathbf{p}^{\text{miss}}_{\text{T}}$
$m^{\text{non}-b,\text{min}}_{\text{T}}$	Transverse mass of the non-b-jet, which is closest in φ to $\mathbf{p}^{\text{miss}}_{\text{T}}$
N_{jet}	Number of $R = 0.4$ jets
$N_{b-\text{jet}}$	Number of b-jets
m^0_{jjj}	Mass of top cand. with smallest ΔR between b-jet and W ($\rightarrow q\bar{q}'$)
m^1_{jjj}	Mass of top cand. with 2$^{\text{nd}}$ smallest ΔR between b-jet and W ($\rightarrow q\bar{q}'$)
$p^0_{\text{T},jjj}$	p_{T} of top cand. with smallest ΔR between b-jet and W ($\rightarrow q\bar{q}'$)
$p^1_{\text{T},jjj}$	p_{T} of top cand. with 2$^{\text{nd}}$ smallest ΔR between b-jet and W ($\rightarrow q\bar{q}'$)

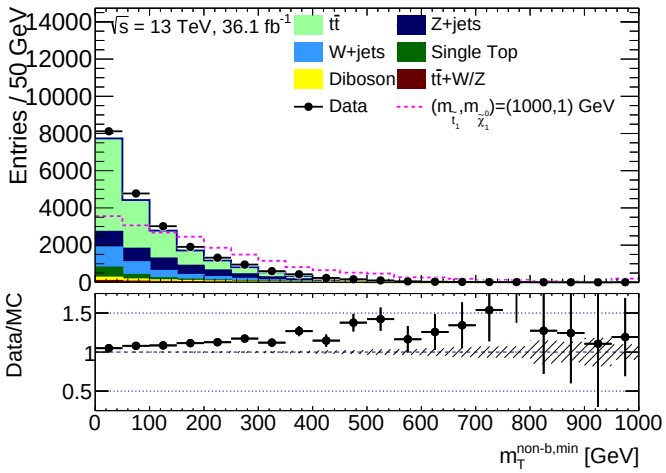


Figure 6.1: Distributions of $m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{min}}$ after the preselection requirement. The stacked histograms show the SM expectation, whereas the dashed purple line shows the distribution of the reference model normed to the integral of the background events. The panel beneath illustrates the ratio of data events to the total SM expectation. The hatched uncertainty band around the SM expectation and in the ratio plot illustrates statistical uncertainties. The rightmost bin includes all overflows.

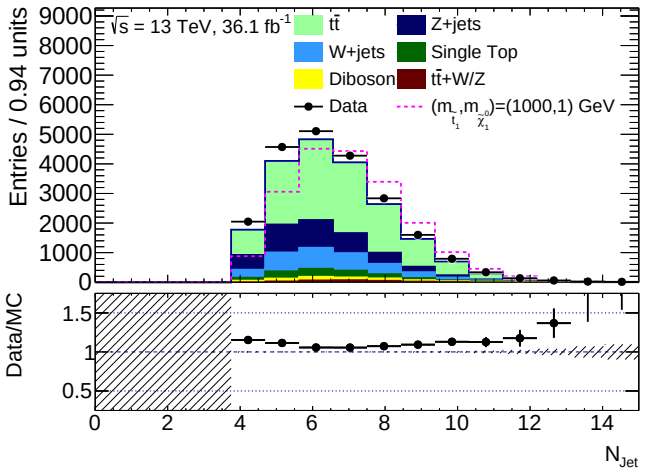


Figure 6.2: Distributions of N_{Jet} after the preselection requirement. The stacked histograms show the SM expectation, whereas the dashed purple line indicates the distribution of the reference model normed to the integral of the background events. The panel beneath shows the ratio of data events to the total SM expectation. The hatched uncertainty band around the SM expectation and in the ratio plot illustrates statistical uncertainties. The rightmost bin includes all overflows.

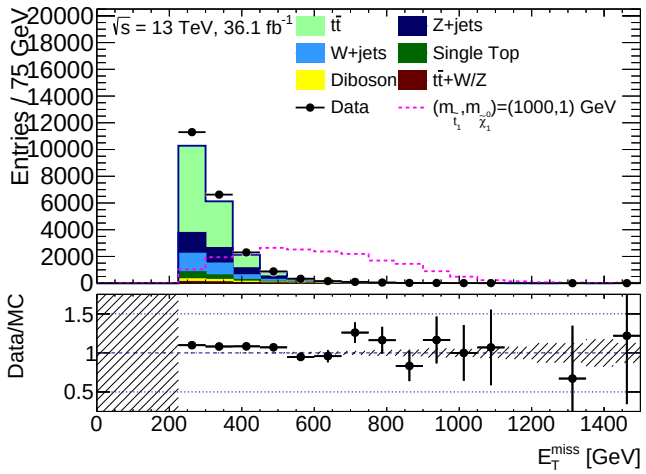


Figure 6.3: Distributions of $E_{\mathrm{T}}^{\mathrm{miss}}$ after the preselection requirement. The stacked histograms show the SM expectation, whereas the dashed purple line indicates the distribution of the reference model normed to the integral of the background events. The panel beneath shows the ratio of data events to the total SM expectation. The hatched uncertainty band around the SM expectation and in the ratio plot illustrates statistical uncertainties. The rightmost bin includes all overflows.

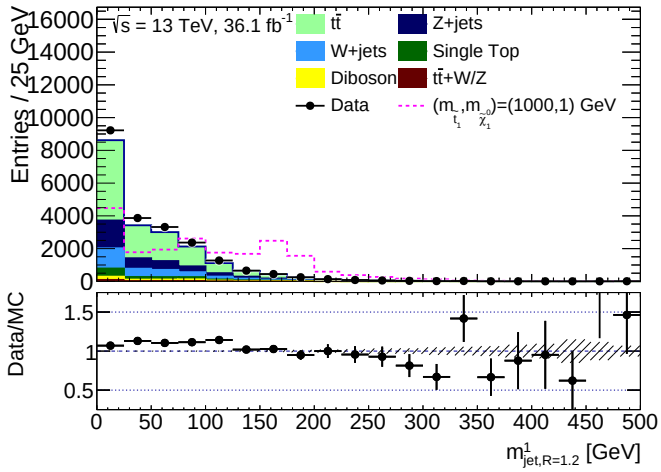


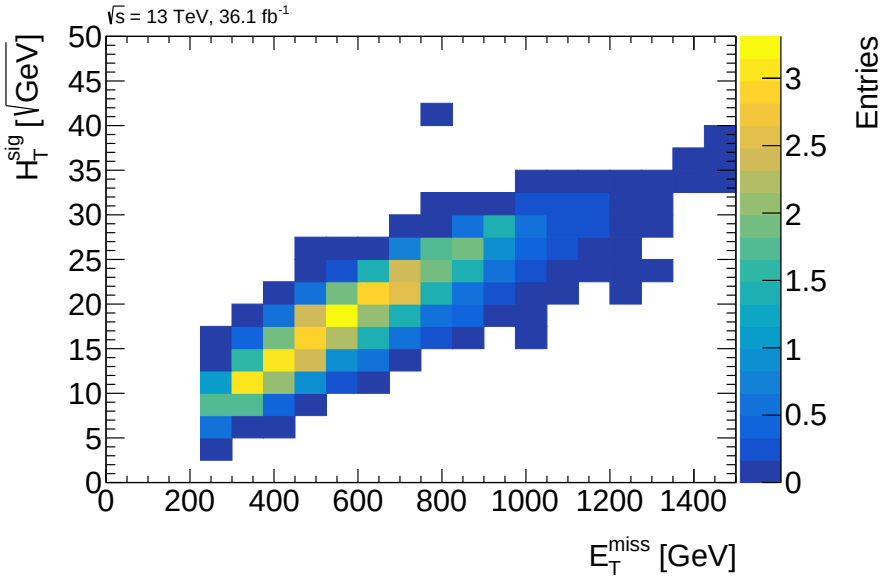
Figure 6.4: Distributions of $m^{1}_{\mathrm{jet},R=1.2}$ after the preselection requirement. The stacked histograms show the SM expectation, whereas the dashed purple line shows the distribution of the reference model normed to the integral of the background events. The panel beneath shows the ratio of data events to the total SM expectation. The hatched uncertainty band around the SM expectation and in the ratio plot illustrates statistical uncertainties. The rightmost bin includes all overflows.

6.2 Correlations between Input Variables

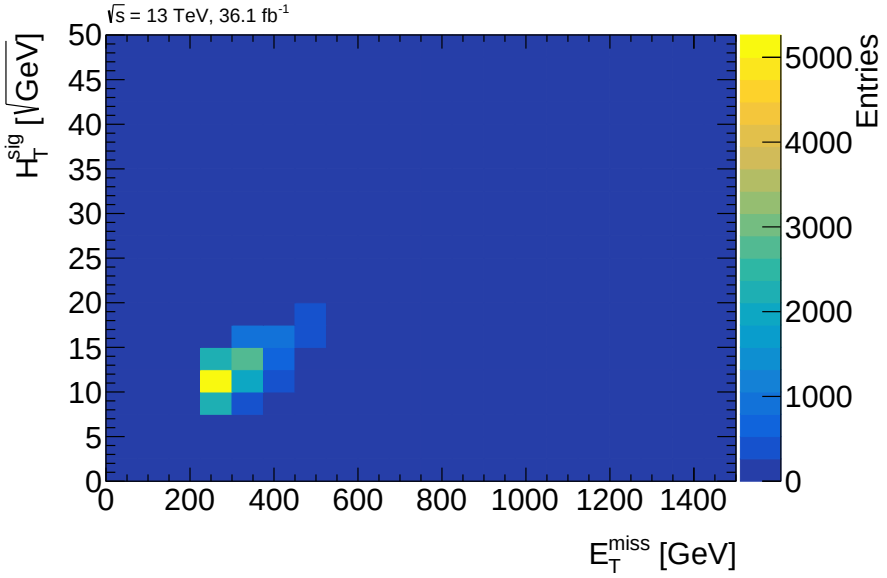
In the Cut and Count method, the discrimination of signal from SM backgrounds is achieved by just cutting on certain variables. The improvement of multivariate methods is to make use of correlations between different input variables in signal and background to discriminate them accordingly. Figure 6.5 shows the distribution of the signal events with $m_{\tilde{t}_1} = 1$ TeV and $m_{\tilde{\chi}_1^0} = 1$ GeV (Figure 6.5a) and of all background events (Figure 6.5b) with respect to $E_{\mathrm{T}}^{\mathrm{miss}}$ and $H_{\mathrm{T}}^{\mathrm{sig}}$. Particularly for signal, a strong linear positive correlation can be observed, meaning that an increase of $E_{\mathrm{T}}^{\mathrm{miss}}$ leads to an increase in $H_{\mathrm{T}}^{\mathrm{sig}}$. This is reasonable since the later variable is defined by $H_{\mathrm{T}}^{\mathrm{sig}} = E_{\mathrm{T}}^{\mathrm{miss}}/\sqrt{H_{\mathrm{T}}}$. When comparing signal and background, it can be seen that a cut on $E_{\mathrm{T}}^{\mathrm{miss}} > 400$ GeV suppresses most of the backgrounds while keeping the majority of the signal. For this reason, this cut was done for Signal Region SRA_TT (cf. Table 4.1). It can be seen that similar results could also be achieved with a cut on $H_{\mathrm{T}}^{\mathrm{sig}} > 15$ GeV due to the high linear correlation.

Figure 6.6 shows the distribution of p_{T}^1 and $m_{\mathrm{T}}^{b,\mathrm{min}}$. A bulk can be seen, both for signal and background, showing that there is no linear correlation between these two variables. There are, nevertheless, two different shapes recognizable, however, with quite a lot of overlap. A cut on $m_{\mathrm{T}}^{b,\mathrm{min}} = 200$ GeV still discriminates lots of signal from background, which therefore is done in the Cut and Count analysis (cf. Table 4.1).

Figure 6.7 depicts the signal and background distribution of $m^0_{\mathrm{jet},R=1.2}$ and $m^1_{\mathrm{jet},R=1.2}$. It is noticeable that signal and background are distributed completely differently. This is justified by the fact that a top squark decays into a neutralino and a top quark, which decays into a W-boson and a bottom quark. This means that the biggest jet $m^0_{\mathrm{jet},R=1.2}$ is almost always a reconstructed top-quark with a mass of 174 GeV [71]. The second largest jet $m^1_{\mathrm{jet},R=1.2}$ can either be the second reconstructed top jet from the other top squark, which would have an equal mass, or it can also be a reconstructed W-boson, being the heaviest decay product of the top quark with a mass of 80 GeV [72]. It can also be the jet of the b-quark, which is the other decay product of the top quark and has a mass below 4.2 GeV [73]. Other event possibilities are scarce and may also result from false jet reconstruction. This explains the three major signal peaks. Background processes, on the other hand, do not involve a top or even two. Therefore, these peaks cannot be observed. These peaks once more explain the cuts on $m^0_{\mathrm{jet},R=1.2}$ and $m^1_{\mathrm{jet},R=1.2}$ for the Cut and Count method (cf. Table 4.1). Multivariate techniques are supposed to recognize and exploit various distributions like these (perhaps more complex distributions) in order to discriminate background from signal even better.

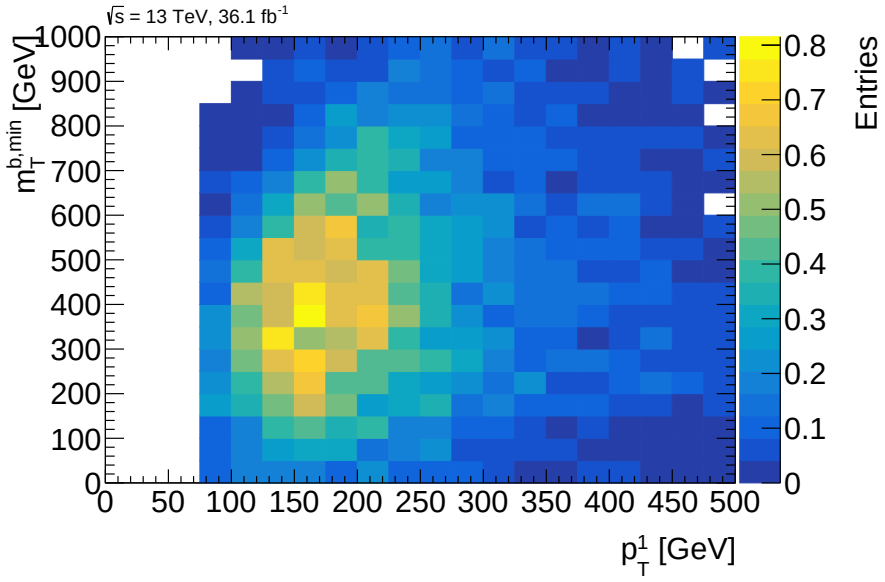


(a) Distribution of $m_{\tilde{t}_1} = 1$ TeV and $m_{\tilde{\chi}_1^0} = 1$ GeV signal.

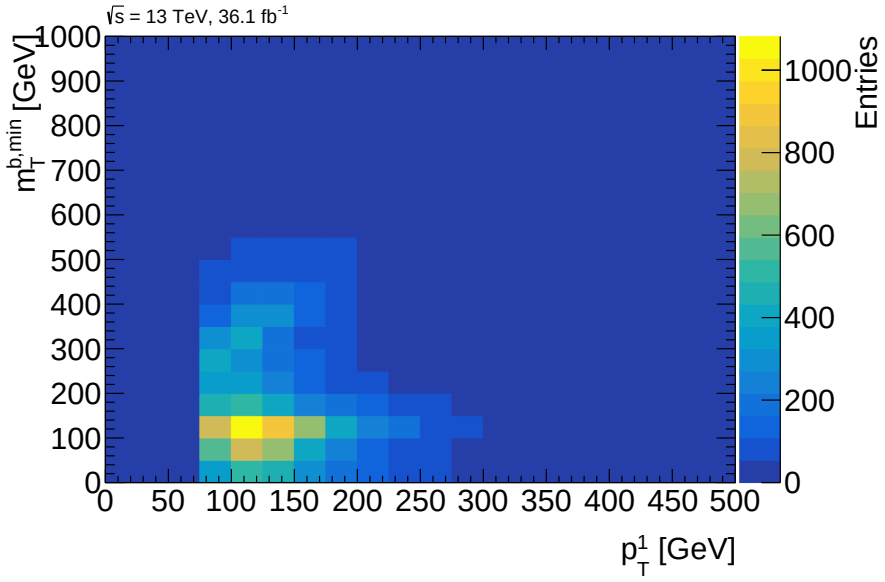


(b) Distribution of SM backgrounds.

Figure 6.5: Distribution of $E_{\mathrm{T}}^{\mathrm{miss}}$ and $H_{\mathrm{T}}^{\mathrm{sig}}$. Positive linear correlation between the input variables can be seen.

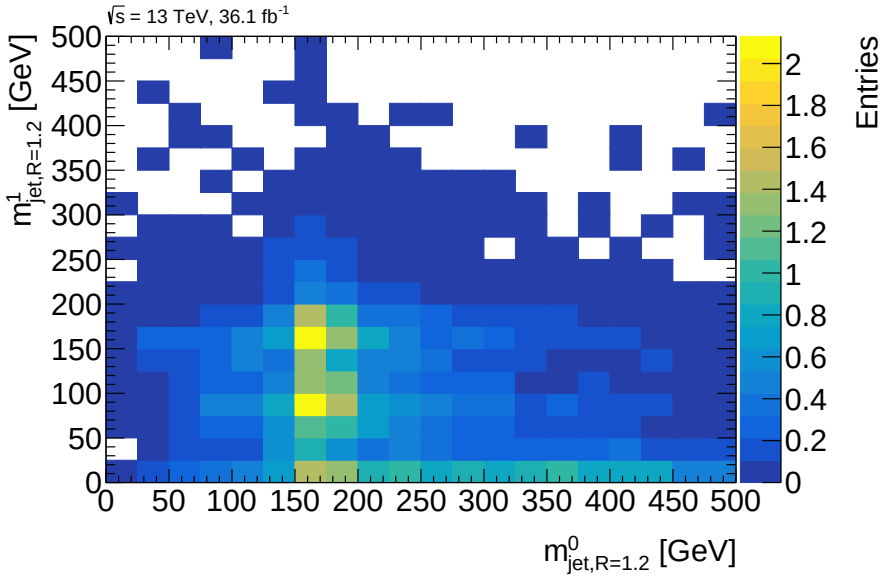


(a) Distribution of $m_{\tilde{t}_1} = 1$ TeV and $m_{\tilde{\chi}_1^0} = 1$ GeV signal.

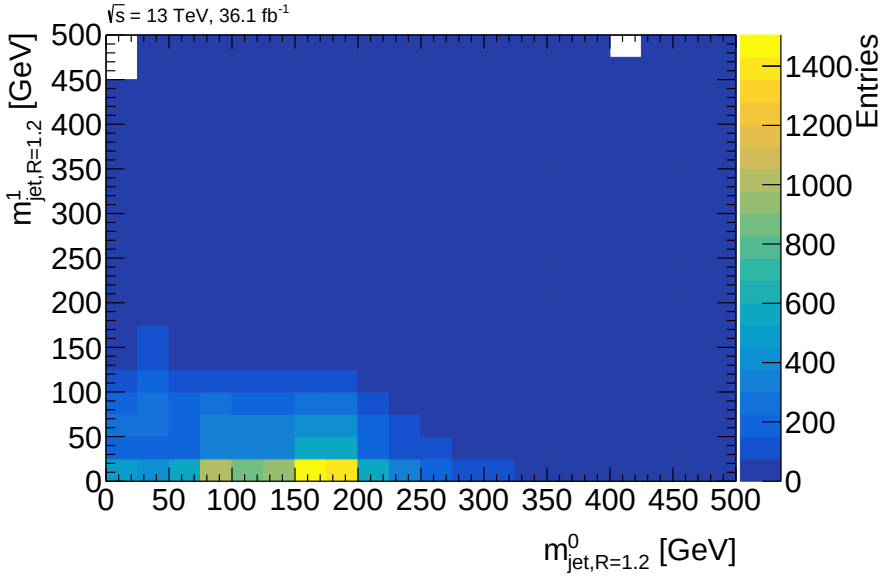


(b) Distribution of SM backgrounds.

Figure 6.6: Distribution of p_{T}^{1} and $m_{\mathrm{T}}^{b,\mathrm{min}}$. Hardly any correlation can be observed.



(a) Distribution of $m_{\tilde{t}_1} = 1$ TeV and $m_{\tilde{\chi}_1^0} = 1$ GeV signal.



(b) Distribution of SM backgrounds.

Figure 6.7: Distribution of $m^0_{\text{jet},R=1.2}$ and $m^1_{\text{jet},R=1.2}$. Different shapes between signal and background can be observed.

6.3 Boosted Decision Trees

6.3.1 Selection of Signal Samples

Figure 6.8a depicts the expected significance as a function of the top squark mass $m_{\tilde{t}_1}$ and the neutralino mass $m_{\tilde{\chi}_1^0}$ achieved with the Cut and Count method. On the other hand, Figure 6.8b depicts the expected exclusion limit. It can be seen that the reference point cannot be excluded, even though the exclusion line is very close. It will be analyzed whether better results can be achieved with multivariate analysis methods.

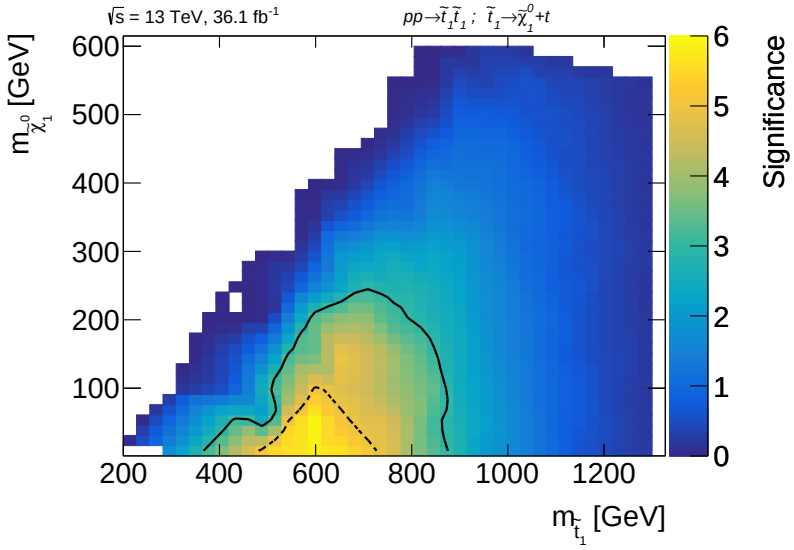
The choice of signal samples used for machine learning is essential for the success of the algorithms, e. g. enough statistics must be available and the event kinematics have to be similar in case that different samples are used simultaneously. In order to maximize the significance to the reference model, which is the model Signal Region A (Section 4.3) was optimized to and in the following called reference model, three different choices to pick the signal MC samples were considered: First, using the reference signal point only, second, using all signal models and third using all signal models with $m_{\tilde{t}_1} \geq 1$ TeV. The standard BDT settings were used in order to compare the significances after applying the BDT. In addition, all input variables were used (cf. Table 6.1).

Figure 6.9 presents the expected significances as a function of the top squark mass $m_{\tilde{t}_1}$ and the neutralino mass $m_{\tilde{\chi}_1^0}$ (Figure 6.9a) of a BDT training using only the reference signal point and the corresponding BDT-score (Figure 6.9b). The solid line in the significance plot is the 3σ-contour, whereas the dashed line, which does not exist in this setting, is the 5σ-line. It can be seen that the 3σ line gets close to $m_{\tilde{t}_1} = 1$ TeV with low neutralino masses. Figure 6.9b shows the different background events and the data depending on the BDT-score. For events with enough statistics, data and MC fit very well.

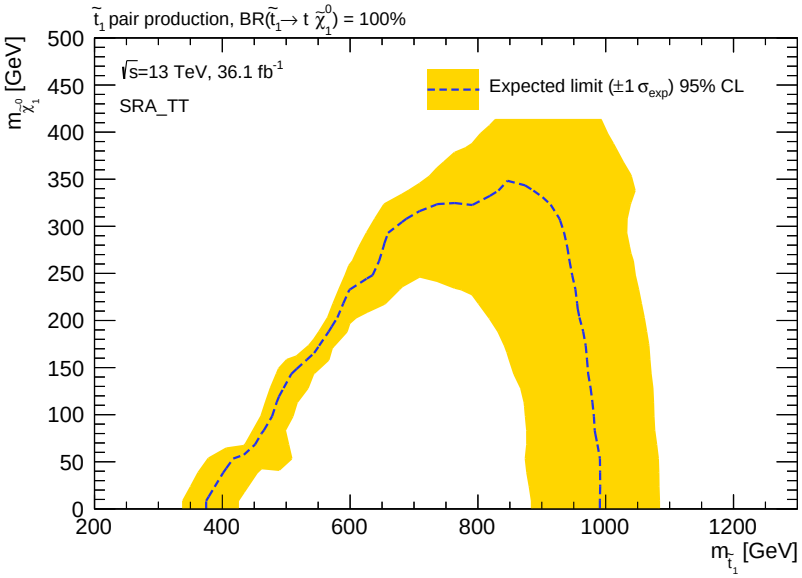
Figure 6.10 depicts the significances which are obtained by using a BDT training with all $\tilde{t}_1 \to t + \tilde{\chi}_1^0$-models and the corresponding Data-MC-plot. Very high significances are achieved in the region 500 GeV $< m_{\tilde{t}_1} <$ 700 GeV. Significances of 5σ can be observed from roughly $m_{\tilde{t}_1} = 400$ GeV until $m_{\tilde{t}_1} = 800$ GeV. 5σ are also obtained for high neutralino masses until approximately $m_{\tilde{\chi}_1^0} = 200$ GeV at certain $m_{\tilde{t}_1}$. Furthermore, the 3σ line is very close to the kinematic border, which is defined via $m_{\tilde{t}_1} = m_t + m_{\tilde{\chi}_1^0}$. This seems to be in fact a very successful training, however, it is not all that sensitive towards samples with high $m_{\tilde{t}_1}$. Again, the MC-events show good agreement with data.

Figure 6.11 shows the results of a training when all signal samples with $m_{\tilde{t}_1} \geq 1$ TeV

are used. This training does not result in any discovery sensitivity for the bulk region of the $m_{\tilde{t}_1} - m_{\tilde{\chi}_1^0}$ plane, which for the most part has already been excluded in the current analysis [26]. It rather achieves an increase in the sensitivity to models with $\tilde{t}_1$-masses beyond 1 TeV. This model has been found to be the most appropriate configuration for the selected test sample. Therefore, it can be concluded that in order to achieve the highest possible sensitivity to high $\tilde{t}$-mass scenarios, it is crucial that the selected signal points for the machine learning training have also high $\tilde{t}_1$-masses, as low-mass signals strongly influence the training due to their much higher production cross-section. Thus, they would pull back the sensitivity as the algorithm recognizes their characteristics. Analogous consideration can be given to the kinematic border of the phase space where $m_{\tilde{t}_1} \approx m_{\tilde{\chi}_1^0} + m_t$. A dedicated MVA training using signals close to that line in combination with advanced high-level variables, like the ones constructed from the recursive Jigsaw technique [74], will result in large sensitivities to this particular regime. This scenario has not been exploited in this thesis and is thus left open for future studies.

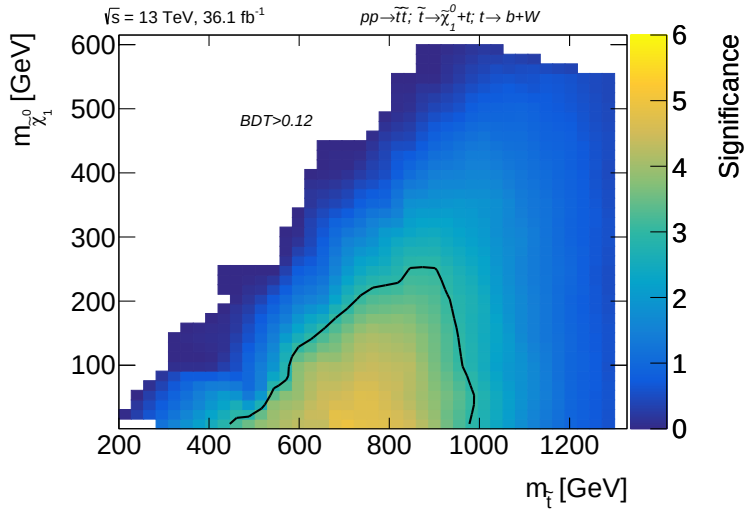


(a) Expected significance for the observation of $\tilde{t}_1$-quarks in depency of $m_{\tilde{t}_1}$ and $m_{\tilde{\chi}_1^0}$.

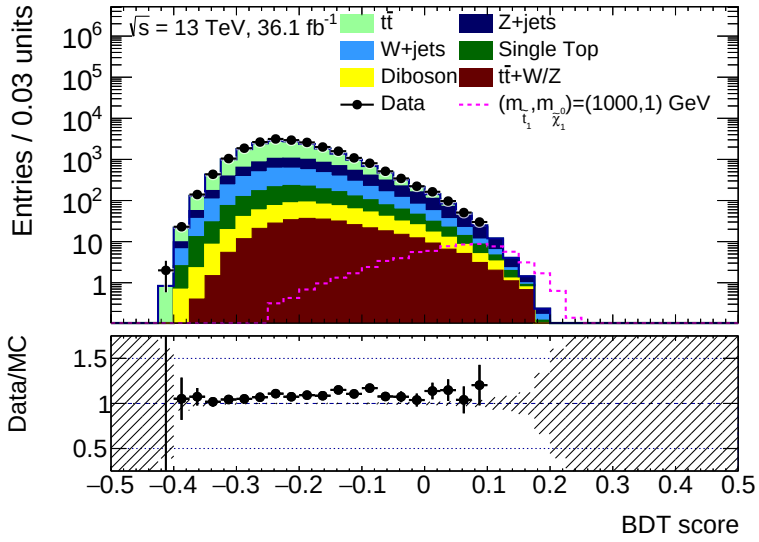


(b) Expected exclusion plot for the observation of $\tilde{t}_1$-quarks in depency of $m_{\tilde{t}_1}$ and $m_{\tilde{\chi}_1^0}$.

Figure 6.8: Expected significances and expected exclusion as a function of $m_{\tilde{t}_1}$ and $m_{\tilde{\chi}_1^0}$ achieved with the Cut and Count method. The reference point cannot be excluded.

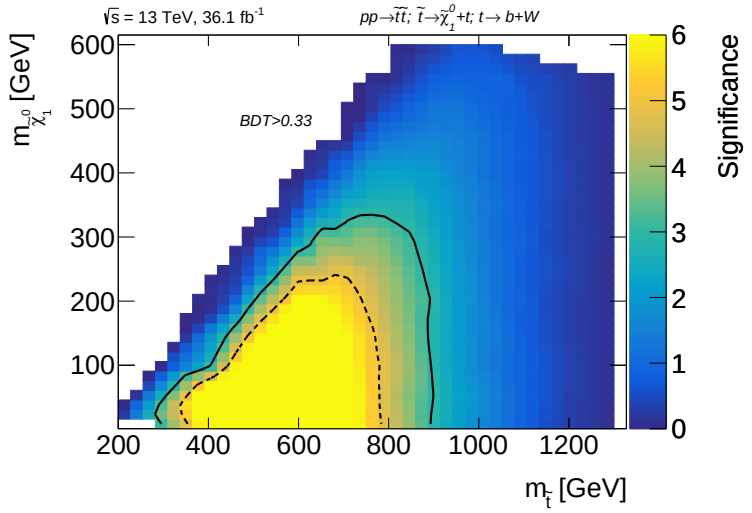


(a) Expected significance for the observation of $\tilde{t}_1$-quarks in depency of $m_{\tilde{t}_1}$ and $m_{\tilde{\chi}_1^0}$.

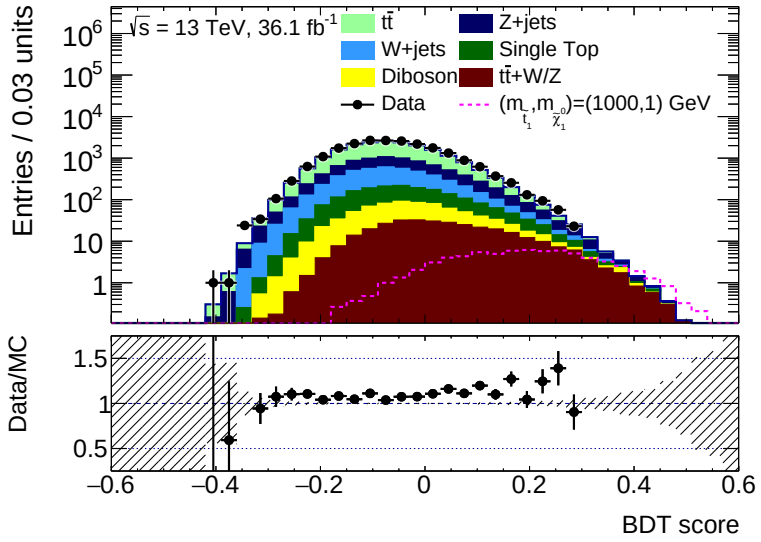


(b) Data-MC plot as a function of the BDT-score.

Figure 6.9: Events from models with $m_{\tilde{t}} = 1$ TeV and $m_{\tilde{\chi}^0} = 1$ GeV were used to develop the BDT training. Expected significances (a) are shown as a function of the top squark mass and the neutralino mass for various training models. The solid (dashed) line indicates the 3 (5)σ contour. The BDT cut was done in a way that highest sensitivities are achieved for large top squark masses. The corresponding Data-MC-plot is depicted in (b), showing the different background events logarithmically in color, as well as data events (black dots) and the signal events of the reference model in the purple dashed line.

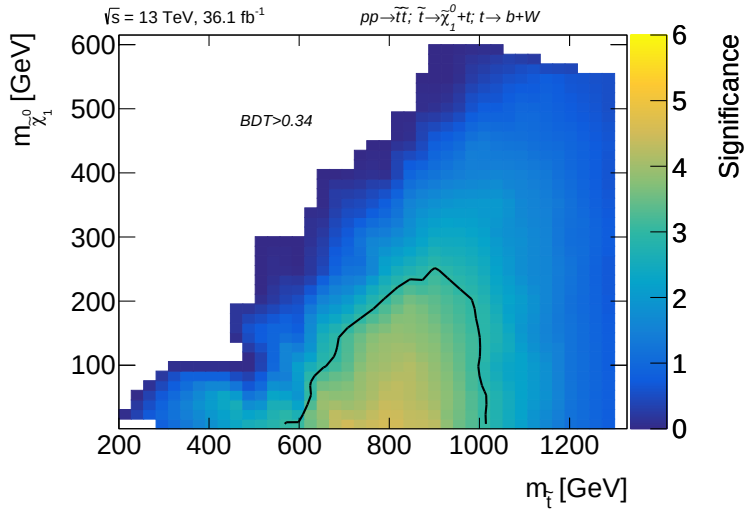


(a) Expected significance for the observation of $\tilde{t}_1$-quarks in depency of $m_{\tilde{t}_1}$ and $m_{\tilde{\chi}_1^0}$.

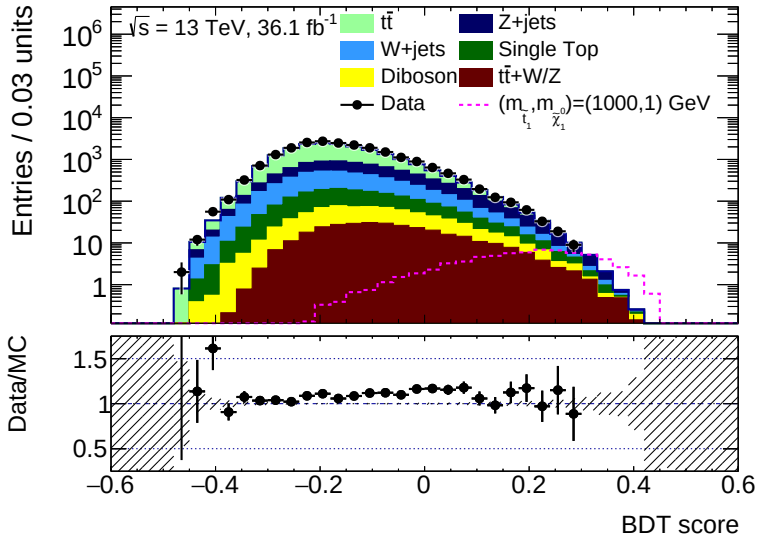


(b) Data-MC plot as a function of the BDT-score.

Figure 6.10: All considered mass parameters were used to develop the BDT training. Expected significances (a) are shown as a function of the top squark mass and the neutralino mass for various training models. The solid (dashed) line indicates the 3 (5)σ contour. The BDT cut was done in a way that highest sensitivities are achieved for large top squark masses. The corresponding Data-MC-plot is depicted in (b), showing the different background events logarithmically in color, as well as data events (black dots) and the signal events of the reference model in the purple dashed line.



(a)



(b) Data-MC plot as a function of the BDT-score.

Figure 6.11: Events from models with $m_{\tilde{t}} \geq 1$ TeV were used to develop the BDT training. Expected significances (a) are shown as a function of the top squark mass and the neutralino mass for various training models. The solid (dashed) line indicates the 3 (5)σ contour. The BDT cut was done in a way that highest sensitivities are achieved for large top squark masses. The corresponding Data-MC-plot is depicted in (b), showing the different background events logarithmically in color, as well as data events (black dots) and the signal events of the reference model in the purple dashed line.

Table 6.2: Standard Settings of the BDT-method.

Parameter	Parameter Value
Number of Trees	850
Minimal Node Size	2.5%
Maximal Depth of the Trees	3
Boost Type	AdaBoost
Separation Type	Gini Index
Number of Cuts	20
AdaBoost-β	0.5
Use Bagged Boost	True
Bagged Sample Fraction	0.5

6.3.2 Optimization of the Training Configurations

In this subchapter, it will be discussed which is the best set of parameters to achieve the highest sensitivity and avoid overtraining. A list of the standard parameters used for training is depicted in Table 6.2. Details can be found in [61]. All input variables, listed in Table 6.1, are used for training.

Separation Type
The first relevant setting is the selection of the separation algorithm, referred to as *separation type*. TMVA [61] provides four different types of classification problems: The *Gini Index*, explained in Section 5.1, the *Cross Entropy*, the *Gini Index with Laplace* and the *Misclassification Error* [61].

If the separation type *Gini Index* is selected, the discriminating variable and the cut value are chosen according to the minimal sum (Equation 5.4) of the Gini Indexes (p as in Equation 5.2):

$$p \cdot (1 - p) = \frac{S \cdot B}{(S + B)^2} \tag{6.1}$$

The *Gini Index with Laplace* is very similar to the Gini Index and also bounded from the top to be 0.25 for S=B. In contrast to the ordinary separation index, this formula has no absolute minimum. By using the substitution $p = \frac{S+1}{S+B+2}$, in this case the

expression

$$\mathcal{G}^L = p \cdot (1-p) = \frac{(S+1)\cdot(B+1)}{(S+B+2)^2} \tag{6.2}$$

is minimized in order to determine the best choice of a cut on the input variable.

For the remaining separation types discussed in this section, the purity p, defined in Equation 5.2, is used.

The *Cross Entropy* is defined via

$$\mathcal{S} = -p \cdot \ln(p) - (1-p) \cdot \ln(1-p) \tag{6.3}$$

This separation value maps the purity to the interval $[0, \ln(2)]$.

The maximum value will be reached at $p = 0.5$ with Cross Entropy of $-\ln(0.5) \approx 0.69$. The smallest values will also be at $p = 0$ or $p = 1$, the corresponding values are set to be 0.

The final separation type is the Misclassification Error, given by

$$\mathcal{E}^M = 1 - \max(p, 1-p). \tag{6.4}$$

This separation type obtains its extrema for $\mathcal{E}^M(0.5) = 1$ and $\mathcal{E}^M(0) = \mathcal{E}^M(1) = 0$.

All separation types have in common that the maximum is at $p = 0.5$ and that it is fully symmetric with the axis $p = 0.5$.

Figure 6.12 depicts the Area under the ROC-Curves of four BDT trainings, each performed with a different separation type for training and testing data. The variation of the separation algorithm impacts the area under the ROC-Curve of testing data negligibly. The tested algorithms, however, tend differently towards overtraining as demonstrated in Figure 6.13. They show the Kolmogorov-Smirnov-Score, described in Section 5.3 for signal and background. The Misclassification Error separation type has the largest score with a value of 0.0034. Large fluctuations between signal and background are observed for the Gini Index with Laplace. The reason for the latter has not been fully understood yet. Figure 6.14 shows the best possible significance of the reference model obtained by applying a minimum requirement on the BDT response of each of the four training settings. All configurations result in a significance of at least $\approx 3\sigma$. The Misclassification Error and the Gini Index achieve the highest significances of 3.22σ and 3.2σ respectively. Considering the observed trend to overtraining of the misclassification, the Gini Index has been found to provide the best results in order to search for high-mass $\tilde{t}_1$.

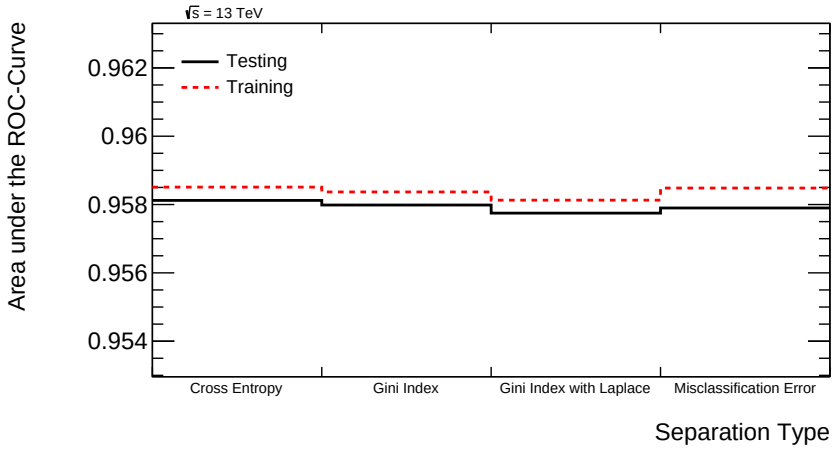


Figure 6.12: Area under ROC-Curves obtained from performing a $\tilde{t}_1$-BDT training with high-mass $\tilde{t}_1$ models with varied training parameter separation type for MVA testing and training events.

Boost Type

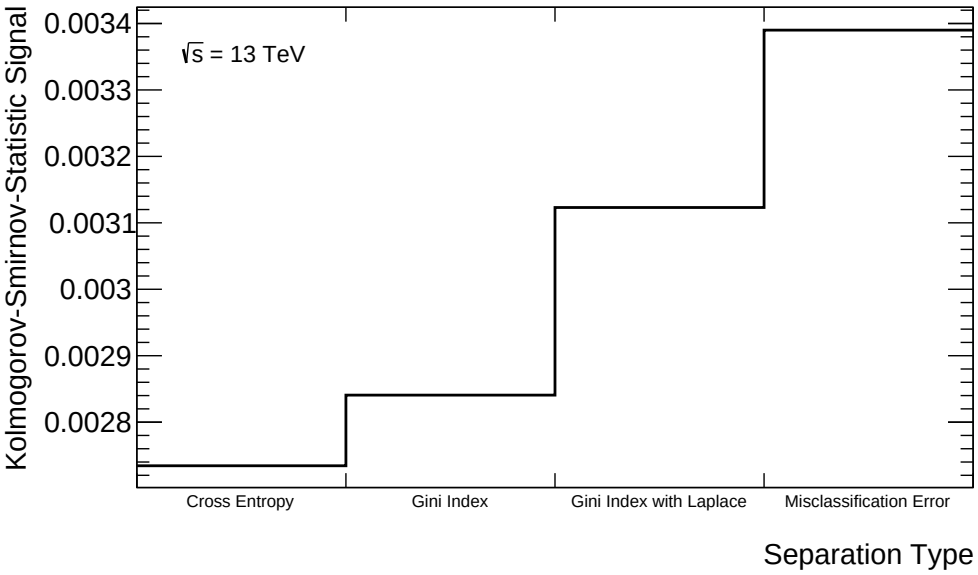
In a further step, the best boosting algorithm for reweighing falsely classified events for the next decision tree in the forest is determined. The algorithms studied are *AdaBoost*, *Real AdaBoost*, *Gradient* and *Bagging*. The AdaBoost has been described in detail in Section 5.1. The Real AdaBoost procedure is a modification of Ada Boost, the final BDT score chosen as

$$s^i = \frac{\sum_{m=0}^{N_{Trees}} \alpha_m \cdot p_m^i}{\sum_{m=0}^{N_{Trees}} \alpha_m}, \tag{6.5}$$

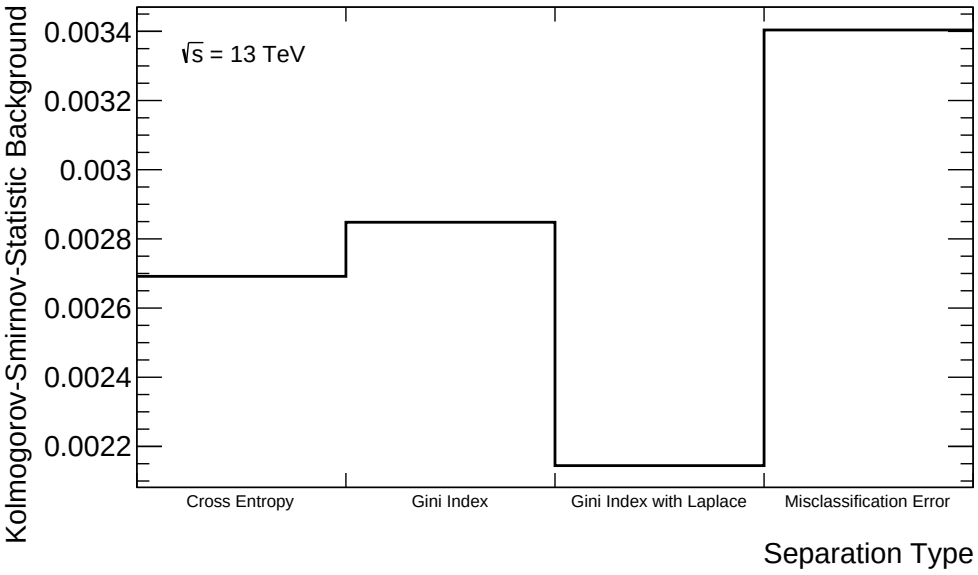
where p_m^i is the purity $p = \frac{S}{S+B}$ of the leaf node concerning event i. Since the purity is a number between 0 and 1, the Real AdaBoost BDT score s^i ranges equally between 0 and 1, unlike the AdaBoost BDT score, which assumes between -1 and 1 (see Equation 5.12). An event i is assumed signal if the Real AdaBoost score s^i is greater than 0.5.

The Gradient Boost algorithm, in contrast to AdaBoost, does not reweight misclassified events in order to train a new tree, but instead builds new trees to reduce the error. More details can be found in 2.2.

The final Boost Type, which was investigated, is the Bagging algorithm. A certain number of events is randomly chosen to create a decision tree. The remaining trees are



(a) Signal Kolmogorov-Smirnov-Score.



(b) Background Kolmogorov-Smirnov-Score.

Figure 6.13: Kolmogorov-Smirnov-Score of the different Separation Types.

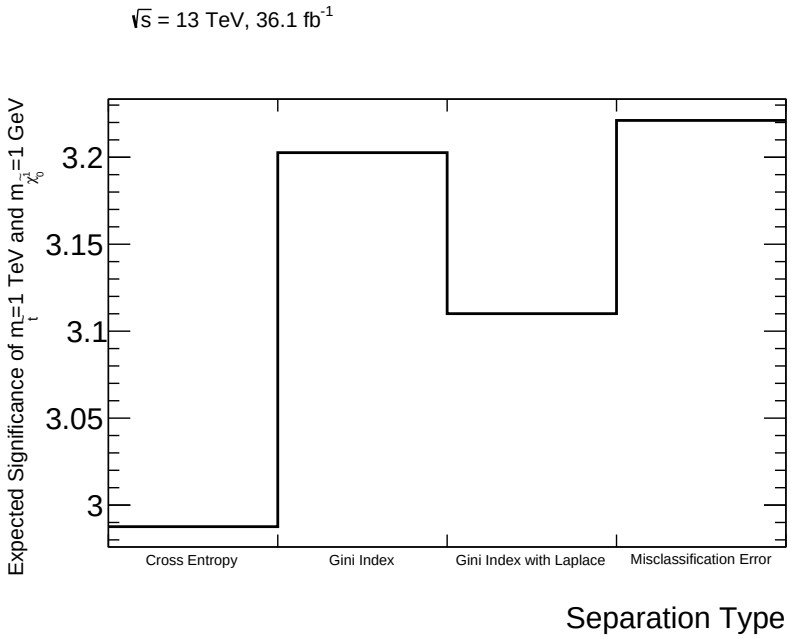


Figure 6.14: Significance of the test signal model for BDT trainings with four choices for the separation type. In each case, the best possible value is shown for a dataset of $L = 36.1$ fb^{-1} at $\sqrt{s} = 13.1$ TeV.

also trained by picking a randomly chosen ensemble of events. The final BDT-score is then given by the average classification outcome of each tree in the forest.

The Area under the ROC-Curves of each boosting type is illustrated in Figure 6.15 for testing and training events of the training model. The bagging algorithm results in the smallest Area under the ROC-Curves. This can be expected in a way, because all trees are based on smaller training statistics, in which all events are weighted equally, regardless of whether they have been misclassified or not. Thus, there is no real learning of the main characteristics of a potential signal. This assumption is supported by the fact that the Bagging-Boost training does not provide any sensitivity to observe a $\tilde{t}_1$-signal with $m_{\tilde{t}_1} = 1$ TeV in 36.1 fb^{-1}.

The Real Ada Boost shows slightly poorer results than its original pandon although both construct the same decision forest. The difference arises by comparing the evaluation of the final BDT-score given in Equation 5.11 and Equation 6.5. In the case of the AdaBoost, background-classified events contribute destructively to the score while in the case of Real AdaBoost always positiv numbers are added. This leads to a smaller separation between background and signal BDT responses. The corresponding plots can be found in the appendix in Figure 5.

Almost equal values for the Area under the ROC-Curve are obtained from BDT trainings with AdaBoost and Gradient. The Kolmogorov-Smirnov-Score, however, shows that the Gradient Boost algorithm tends to two times more overtraining than AdaBoost, as depicted in Figure 6.16.

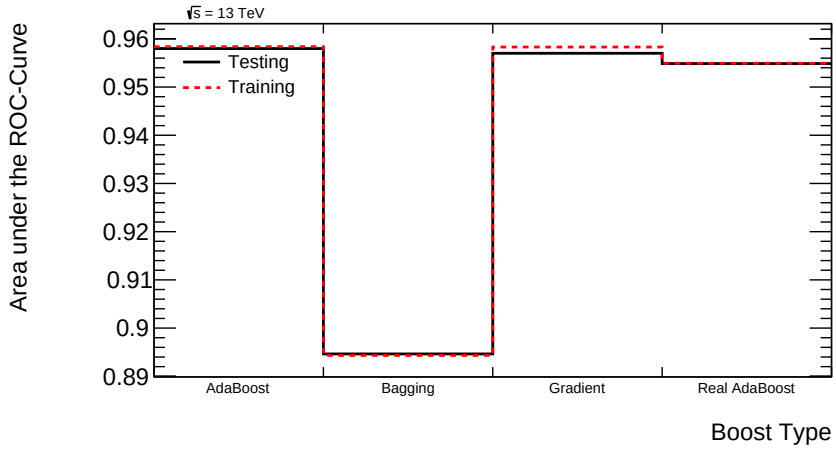
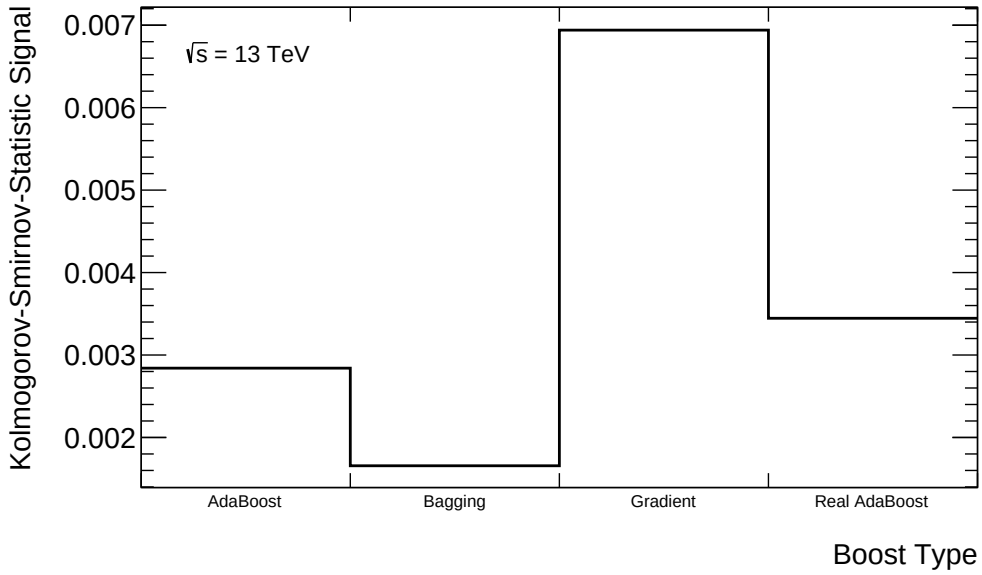


Figure 6.15: Area under ROC-Curves obtained from performing a $\tilde{t}_1$-BDT training with high-mass $\tilde{t}_1$ models with varied training parameter boost type.

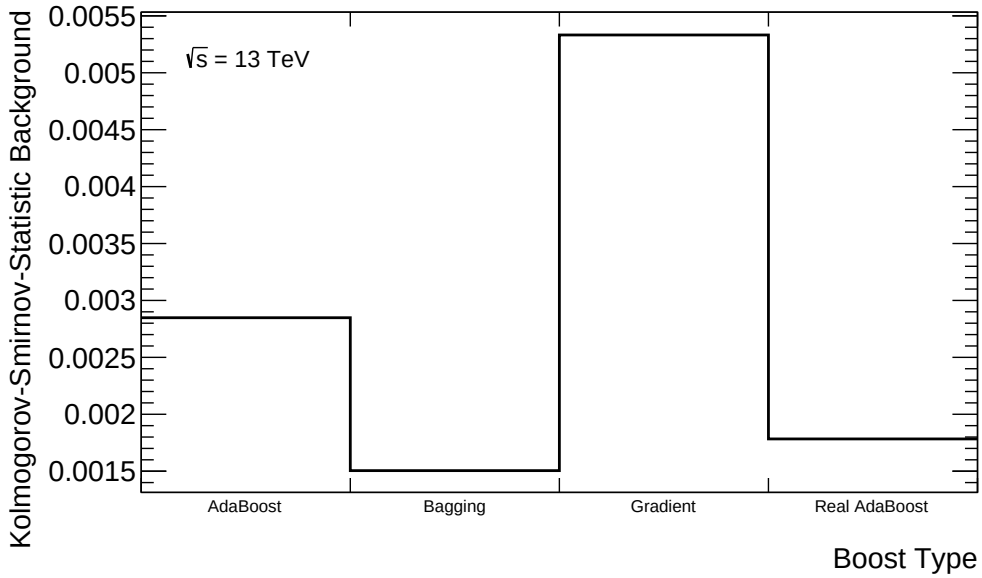
Two-Dimensional Parameter Scans

After determining the optimal algorithms to develop the BDT, the four driving configuration parameters, which determine the size and the branching of each tree, are optimized in the following. These are the maximal depth, the minimal node size, the number of cuts and the number of trees. Since it was observed in preliminary studies that the best choice of a given parameter depends on the choices of the other parameters, two parameters are varied at the same time while keeping the others at the standard values, listed in Table 6.2.

As an example, the Area under the ROC-Curves obtained from BDT trainings with a simultaneous variation of the maximum tree depth and the minimal node size is shown for signal and background events in Figure 6.18. The scan shows that for an increasing maximum depth and shrinking minimal node size, the highest values for the area under the ROC-Curves are obtained for training events, whereas in the case of testing events, the best values seem to be a maximal depth of four and a minimal node size smaller than 2%. The observed deviation is reflected in the associated Kolmogorov-Smirnov-Score, shown in Figure 6.19. Highest testing scores are observed for trainings with small node sizes and high tree branching. The significance of the



(a)



(b)

Figure 6.16: Kolmogorov-Smirnov-Score of the diferent boosting types.

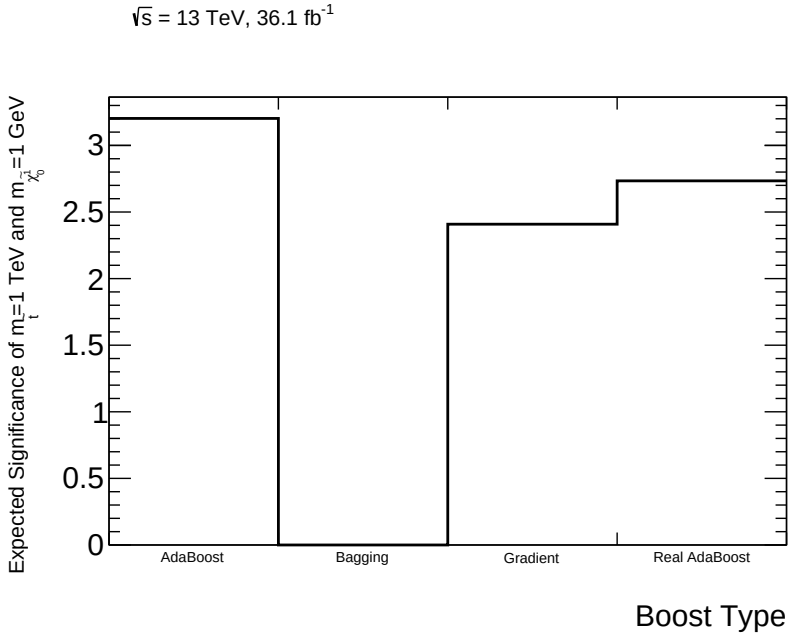


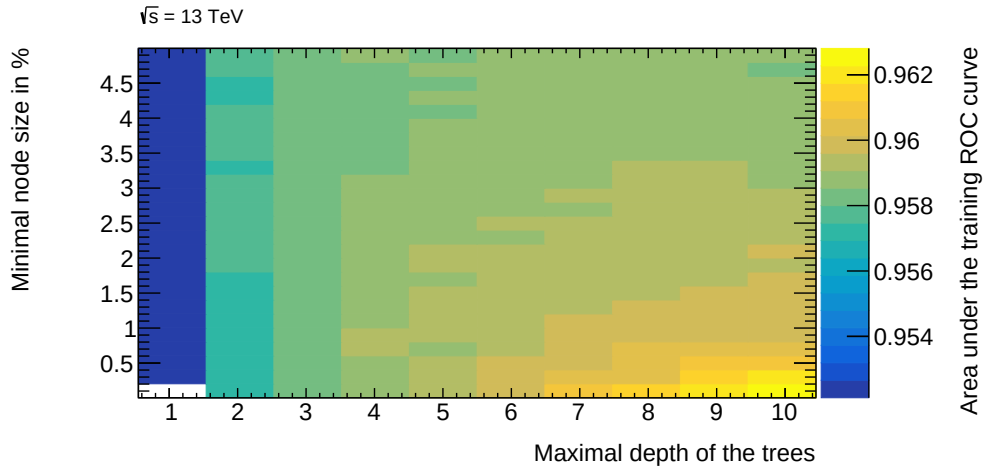
Figure 6.17: Significance of the test signal model for BDT trainings with four choices for the boost type. In each case the best possible value is shown for a dataset of $L = 36.1$ fb^{-1} at $\sqrt{s} = 13.1$ TeV.

test signal model, depicted in Figure 6.20, shows similar results. Greatest significances are achieved with a high maximal depth and a small minimal node size.

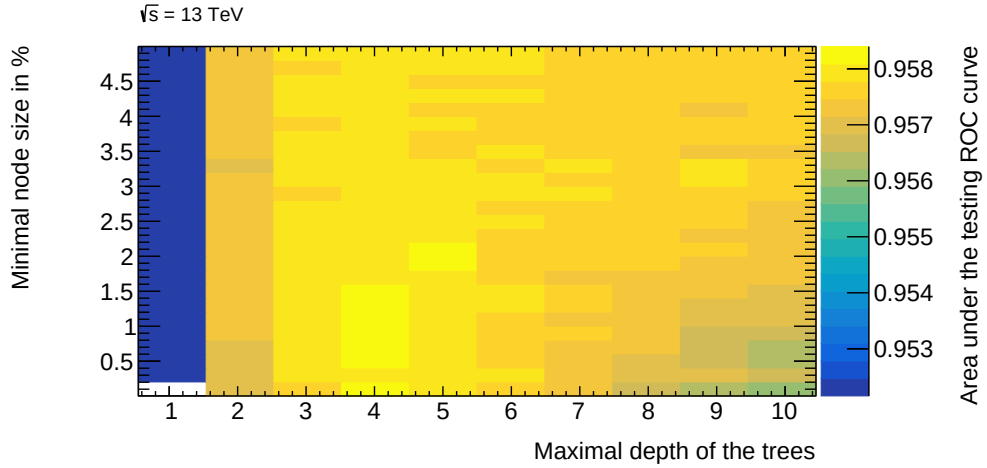
Both, the Kolmogorov-Smirnov-Test and the area under the ROC-Curves of training events in comparison with the testing areas hint at strong overtraining for values with high maximal depth and small minimal node size, which is the reason for the high significance in this region. A greater maximal depth redounds to overtraining, since with a higher maximal depth it will be easier to memorize the events ”by heart”. Therefore, the minimal node size is a parameter to prevent the algorithm from getting overtrained: The larger the minimal node size, the less the possibility to memorize the events since more events need to be clustered. The maximal depth shall therefore be four at most, whereas a minimal node size smaller than 2% and greater than 0.5% is reasonable. Overtraining has to be investigated carefully, before looking at final physics results (cf. to the high significances shown in Figure 6.20).

In Appendix 2.3, the other parameter plots are depicted.

As already discussed, the maximal depth should at most be four. From the Kolmogorov-Smirnov-Score of the maximal depth – number of cuts and of the max-

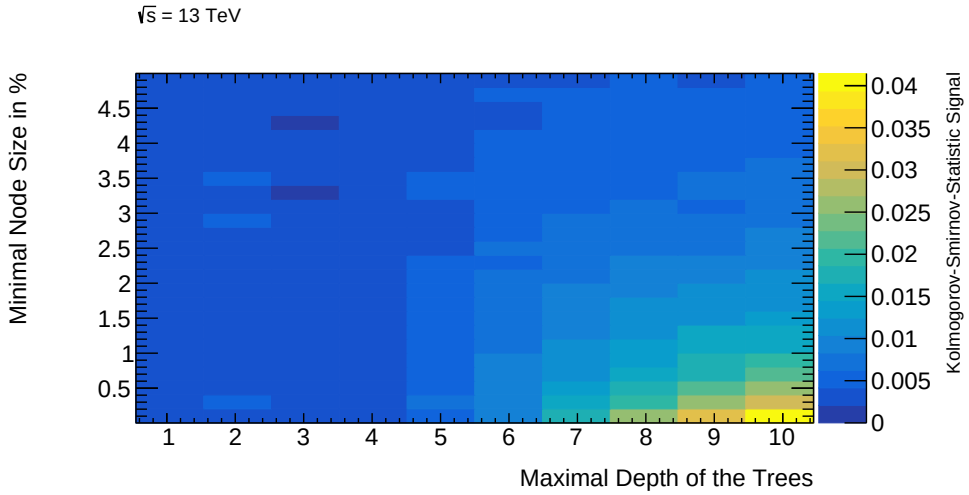


(a) Area under the Training-ROC-Curves.

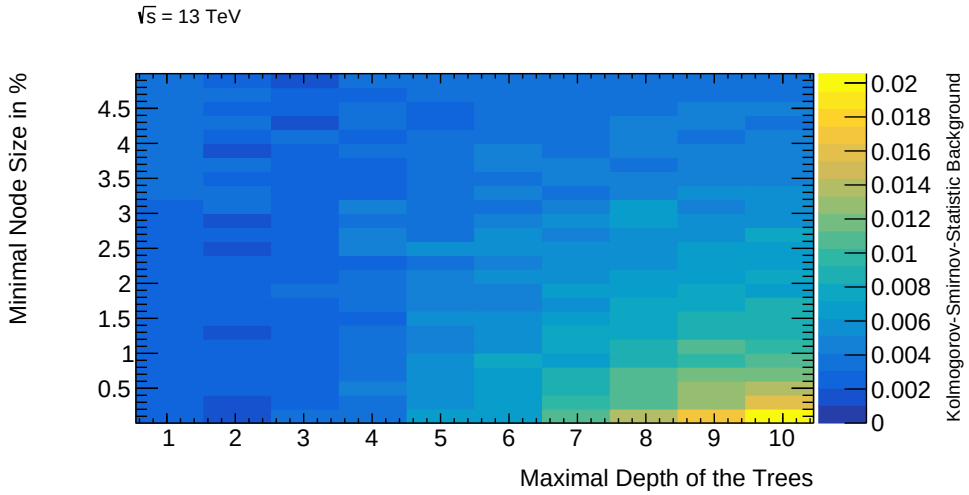


(b) Area under the Testing-ROC-Curves.

Figure 6.18: Area under ROC-Curves obtained from performing a $\tilde{t}_1$-BDT training with high-mass $\tilde{t}_1$ models with varied training parameters maximal depth and minimal node size for MVA testing and training events.



(a)



(b)

Figure 6.19: Kolmogorov-Smirnov-Score dependent on the maximal depth of the trees and the minimal node size.

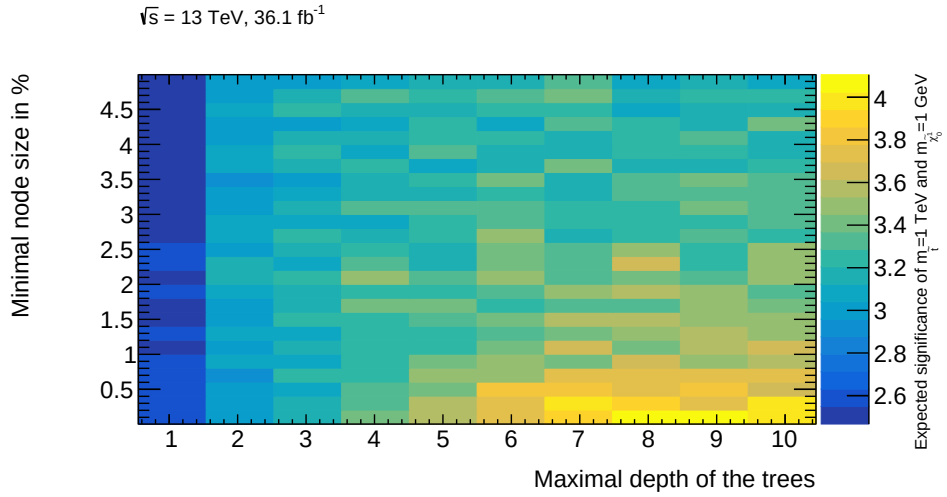


Figure 6.20: Significance of the test signal model for BDT trainings dependent on the maximal depth and the minimal node size. In each case, the best possible value is shown for a dataset of $L = 36.1$ fb^{-1} at $\sqrt{s} = 13.1$ TeV.

imal depth – number of trees two-dimensional parameter plots, it can additionally be seen that a maximal depth of three reduces overtraining notably. Therefore, a maximal depth of three was chosen for the optimal BDT settings.

The final minimal node size value shall be greater than 0.5% and smaller than 2%, according to the previous section. The plots in the appendix show that a change in the minimal node size hardly changes anything and the deviations are rather comparable to statistical fluctuations. Overall, a minimal node size of 1.5% was selected.

A high number of cuts tends to lead to overtraining (cf. Trees-Cuts-plots in Appendix). Since there is hardly any improvement of the Area under the ROC-Curves for more than 20 cuts, it was determined to use 20 cuts for the final analysis.

Furthermore, the trees – cuts plots (area under ROC-Curve) suggest taking more than 1000 trees, whereas the Background Kolmogorov-Smirnov-Score of the trees – cuts plots shows that high numbers of trees leads to overtraining. Therefore, it was determined to train with 1200 trees.

6.3.3 Input Variable Ranking

In this section, the separation power of the input variables used in the training is analyzed. Multiple Boosted Decision Forests have been trained for the analysis with

the final parameter settings. In each case, one variable has been dropped from the training.

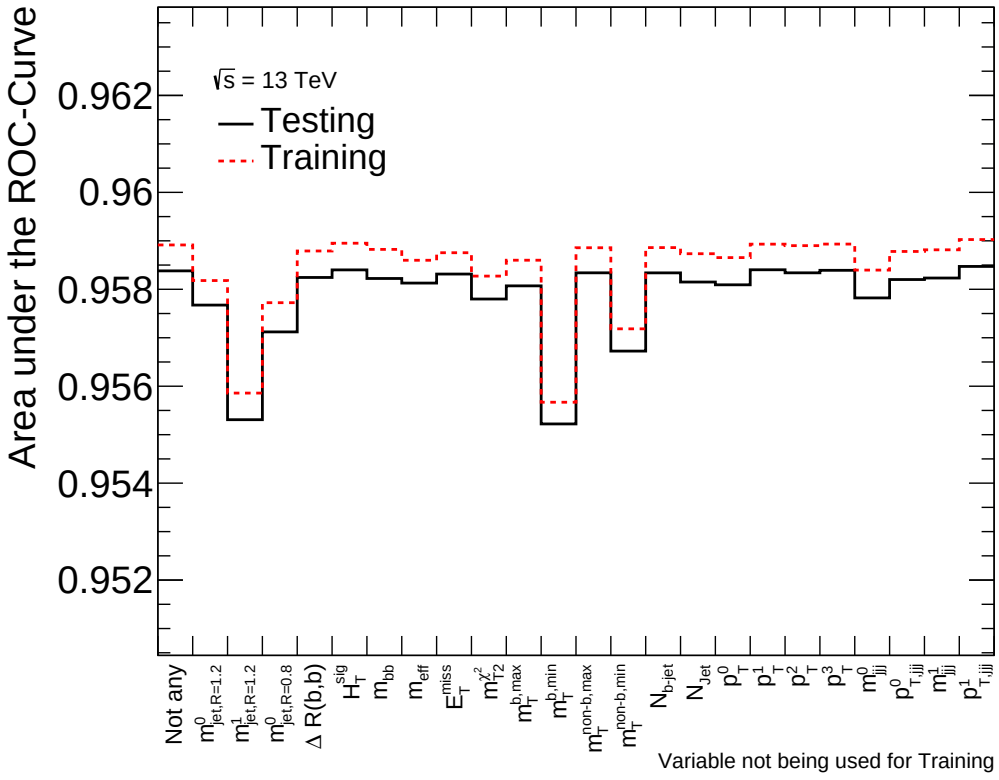


Figure 6.21: Area under the ROC-Curves of the BDT with the final parameter settings, for which all but one variables were used for training. The bin labels denote which variable from Table 6.1 was not used.

Figure 6.21 depicts the area under the ROC-Curves. In comparison to the area of the BDT trained with all input variables, there does not exist a BDT trained with all but one variable (subsequently called N-1 BDTs) with a significantly higher area. There do exist N-1 BDTs though, which do not have a significantly different area under the ROC-Curve. This means that these variables – not all simultaneously – may have no significant effect on the performance of the training. On the other hand, this means that the variables, whose N-1 BDT ROC-Curves have a significantly smaller area, are of great importance to the BDT.

In particular, it can be seen that the mass of the second highest p_{T} reclustered ($R = 1.2$) jet mass $m^1_{\mathrm{jet},R=1.2}$ is essential for the BDT. The other two input variables

$m^0_{\text{jet},R=1.2}$ and $m^0_{\text{jet},R=0.8}$ do not seem to be similarly indispensable. One reason might be that in case of removing either $m^0_{\text{jet},R=1.2}$ or $m^0_{\text{jet},R=0.8}$, the other still remains in the training, thus the information of the leading reclustered fat jet is not completely lost whereas this is the case when removing $m^1_{\text{jet},R=1.2}$. This could mean that the absence of $m^0_{\text{jet},R=1.2}$ will result in an increase of the importance of $m^0_{\text{jet},R=0.8}$ and vice versa.

Another very important variable is $m_{\text{T}}^{b,\text{min}}$. It has already been shown in Figure 4.2 that this variable discriminates signal from $t\bar{t}$ very well. In addition, $m_{\text{T}}^{\text{non}-b,\text{min}}$ is also a valuable variable. The other variables do not seem to be equally important. Essentially, one of the reasons for their minor importance is that they are covered by other variables in a manner that removing one variable will increase the importance of another or several other variable(s).

Figure 6.22 and Figure 6.23 show the corresponding Kolmogorov-Smirnov-Scores. The BDT, trained with all input variables, is less overtrained compared to the N-1 BDTs for both signal and background. Interestingly, the differences are partly very high, in particular the missing of top quark candidates leads to an increase of overtraining. This can be explained by the fact that the properties of top candidates prove to be a better measure for the event topology than simpler variables and thus it is more difficult for the BDT to memorize them. Another reason might be that the parameter settings are not ideal for the N-1 BDT plots since they were optimized merely for a BDT training with all input variables.

Figure 6.24 shows the corresponding significances of the test signal model. The best variables, according to the area under the ROC-Curves, also lead to the lowest expected significances when not trained with, which means that they are very important variables to train with. A training without $m_{\text{T}}^{b,\text{max}}$ leads to an increase of 0.3σ. This does not fit the observations of the Testing ROC-Curves. In fact, the Kolmogorov-Smirnov-Score suggests that this setting is distinctly more overtrained.

Table 6.3 shows the importance measure performed by TMVA [61]. The importance is defined by how many times the variable was used as a node splitting variable, weighted by the number of events in the node and by the square of the separation gain. Since the missing transversal energy $E_{\text{T}}^{\text{miss}}$ is the variable measuring undetected particles and/or jets in an indirect manner, this variable is best-ranked. On the second and third rank are the stransversal mass $m_{\text{T2}}^{\chi^2}$ and the effective mass m_{eff}, followed by the reclustered jet masses and $m_{\text{T}}^{b,\text{min}}$.

There are various ways of defining the importance of a variable, each with different results. While $E_{\text{T}}^{\text{miss}}$ might be replaceable due to other similar variables, it is the variable providing most separation. On the other hand, $m^1_{\text{jet},R=1.2}$ gives only the fifth

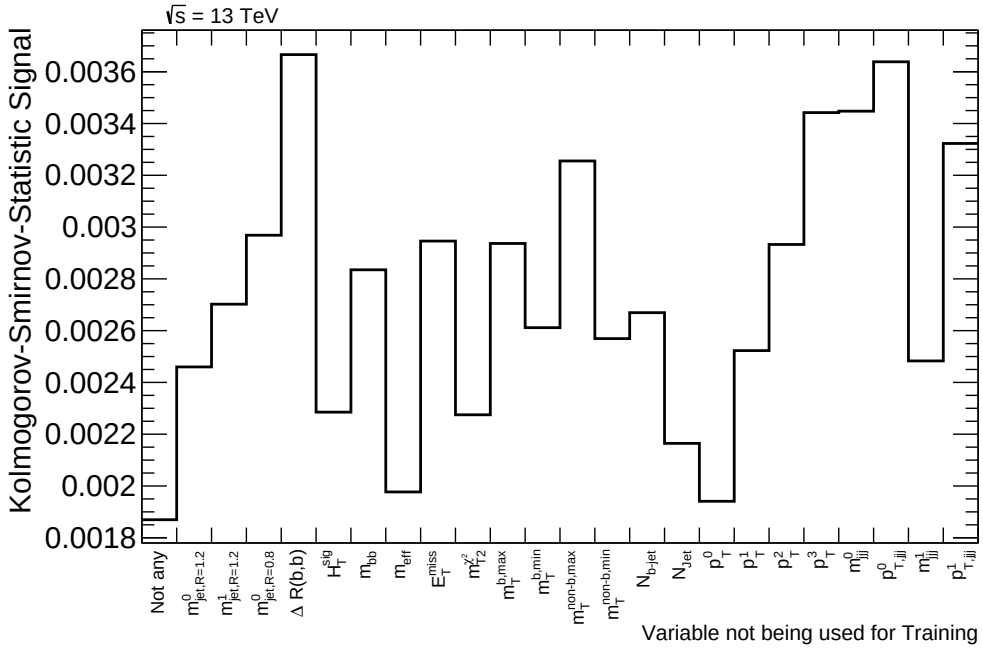


Figure 6.22: Signal Kolmogorov-Smirnov-Score of the BDT with the final parameter settings for which all but one variable used for training. The bin labels denote which variable from Table 6.1 was not used.

best separation while being one of the two most important variables according to the N-1 plots.

6.3.4 Final Results

The final BDT significance plot can be found in Figure 6.25. In comparison to the initial parameter settings, hardly any improvement in the significances is achieved. This shows that the prior settings were already very satisfactory. The 3σ-line is beyond the reference model with $m_{\tilde{t}_1} = 1$ TeV and $m_{\tilde{\chi}_1^0} = 1$ GeV. A 5σ-line cannot be observed. Data fits Monte Carlo very well, showing the accurate approximation of the background. The cut was applied on BDT-score > 0.27. This shows that most of the background was cut, however, also most of the signal. When comparing the results with the significances of the signal region SRA_TT, considerable improvements are noticeable. Whereas a signal with $m_{\tilde{t}_1} =$ TeV and $m_{\tilde{\chi}_1^0} = 1$ GeV would be seen with an expected significance of 1.7σ in the cut and count based signal region, an expected significance of almost 3.2σ can be achieved. This shows, that the Boosted Decision Tree scores very successfully. The signal Kolmogorov-Smirnov-Score is $k^b_{Sig} = 1.9{\cdot}10^{-3}$

Table 6.3: Variable ranking of the BDT with final settings. The separation is determined by the amount a variable was used at a splitting node.

rank	variable	separation
1	$E_{\mathrm{T}}^{\mathrm{miss}}$	$7.503 \cdot 10^{-2}$
2	$m_{\mathrm{T2}}^{\chi^2}$	$6.414 \cdot 10^{-2}$
3	m_{eff}	$6.101 \cdot 10^{-2}$
4	$m^0_{\mathrm{jet},R=0.8}$	$5.687 \cdot 10^{-2}$
5	$m^0_{\mathrm{jet},R=1.2}$	$5.191 \cdot 10^{-2}$
6	$m_{\mathrm{T}}^{b,\mathrm{min}}$	$5.142 \cdot 10^{-2}$
7	$m^1_{\mathrm{jet},R=1.2}$	$5.134 \cdot 10^{-2}$
8	$\Delta R\,(b,b)$	$4.498 \cdot 10^{-2}$
9	$m_{\mathrm{T}}^{b,\mathrm{max}}$	$4.407 \cdot 10^{-2}$
10	$p^0_{\mathrm{T},jjj}$	$4.332 \cdot 10^{-2}$
11	N_{jet}	$4.313 \cdot 10^{-2}$
12	$m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{min}}$	$4.312 \cdot 10^{-2}$
13	m^0_{jjj}	$4.180 \cdot 10^{-2}$
14	$H_{\mathrm{T}}^{\mathrm{sig}}$	$4.121 \cdot 10^{-2}$
15	p^0_{T}	$4.031 \cdot 10^{-2}$
16	$m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{max}}$	$3.982 \cdot 10^{-2}$
17	p^2_{T}	$3.377 \cdot 10^{-2}$
18	p^3_{T}	$3.362 \cdot 10^{-2}$
19	p^1_{T}	$3.351 \cdot 10^{-2}$
20	$p^1_{\mathrm{T},jjj}$	$3.129 \cdot 10^{-2}$
21	m^1_{jjj}	$2.914 \cdot 10^{-2}$
22	m_{bb}	$2.806 \cdot 10^{-2}$
23	$N_{b-\mathrm{jet}}$	$1.713 \cdot 10^{-2}$

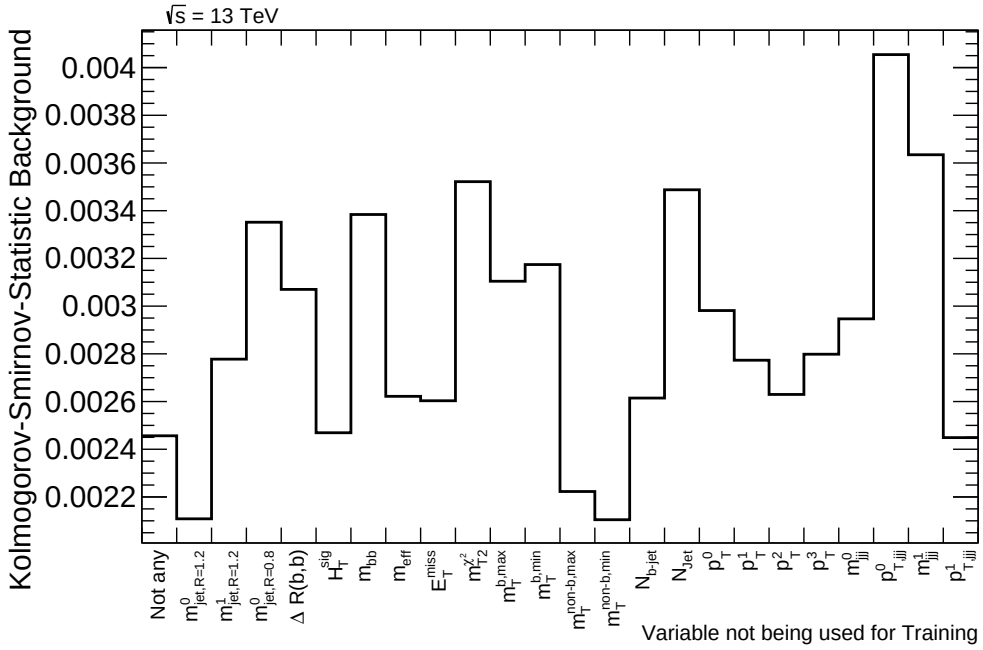


Figure 6.23: Background Kolmogorov-Smirnov-Score of the BDT with the final parameter settings for which all but one variables were used for training. The bin labels denote which variable from Table 6.1 was not used.

whereas the background Kolmogorov-Smirnov-Score is about $k^b_{Bkg} = 2.5 \cdot 10^{-3}$. From $d^b < 1.07$, it can be inferred that $k^b_{Bkg} < 2.6 \cdot 10^{-3}$ and $k^b_{Sig} < 3.7 \cdot 10^{-3}$ is required in this analysis (cf. Equation 5.20). This shows that the overtraining of the BDT is acceptably small.

The exclusion plot is shown in Figure 6.26. It can be seen that the exclusion line cuts the top squark mass axis at about 1.08 TeV, showing that the limit for small neutralino masses was pushed about 9% (cf. Figure 6.8b). This illustrates that the training was in fact very successful.

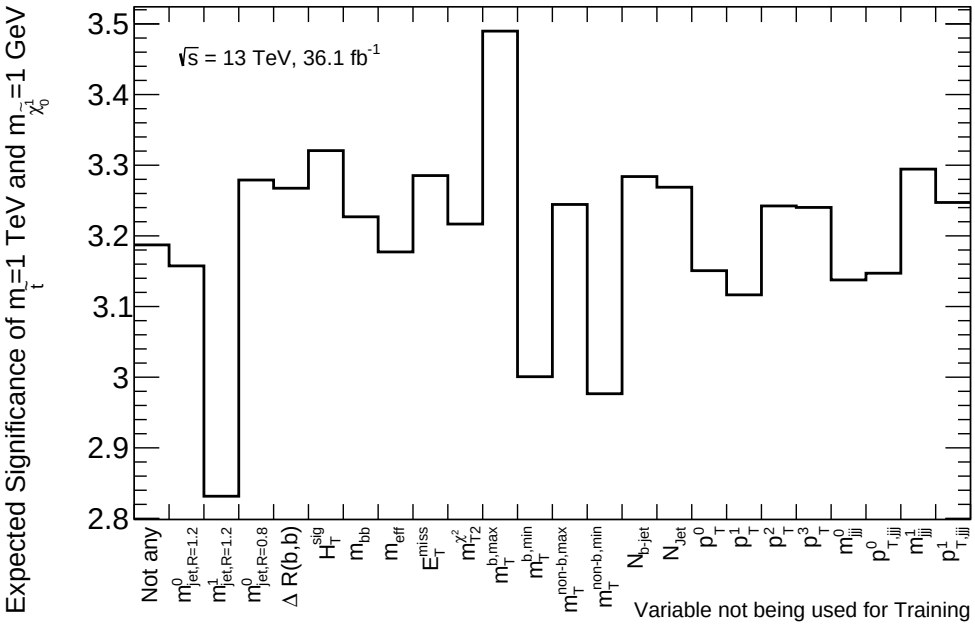
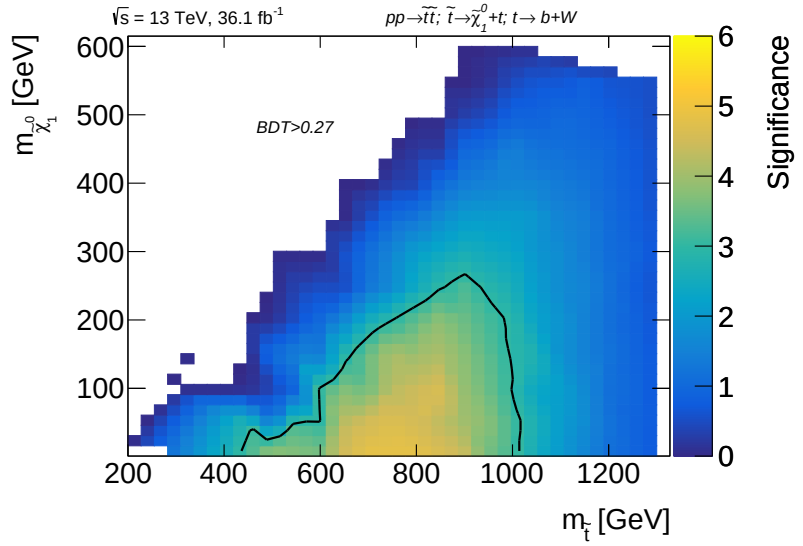
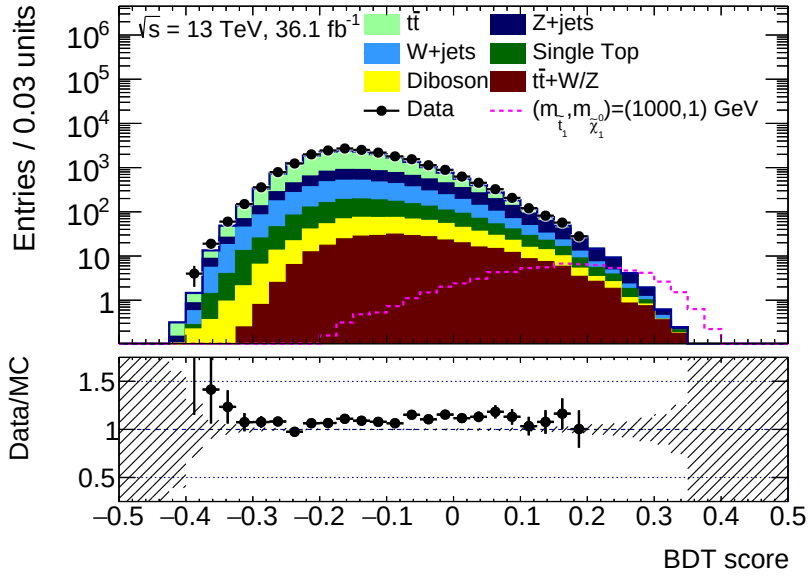


Figure 6.24: Significance of the test signal model for the BDT with the final parameter settings for which all but one variables were used for training. The bin labels denote which variable from Table 6.1 was not used.



(a) Expected significance for the observation of $\tilde{t}_1$-quarks in depency of $m_{\tilde{t}_1}$ and $m_{\tilde{\chi}_1^0}$.



(b) Data-MC plot as a function of the BDT-score.

Figure 6.25: Significance and Data-MC plots of the final BDT training.

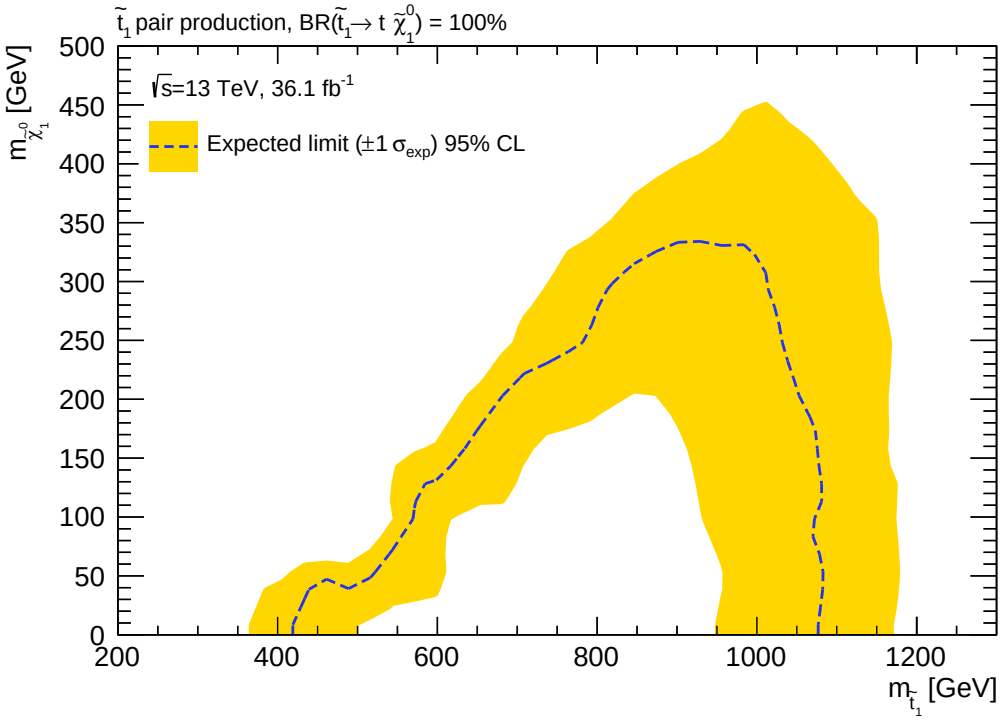


Figure 6.26: Exclusion plot obtained with the results of the BDT-Training. It can be seen, that the exclusion line is beyond 1 TeV, showing that the training achieved excellent results.

6.4 Neural Networks

6.4.1 Setup

As for the final BDT studies, all samples with $m_{\tilde{t}_1} \geq 1$ TeV were used as signal samples and all variables, listed in Table 6.1, were used as input variables.

The parameter settings were selected on criteria similar to those for the BDT in Section 6.3.2. The corresponding plots can be seen in Appendix 3.2. The Back-Propagation algorithm is the most successful training method, tangens hyperbolicus is the best neuron type. 600 Cycles are run for optimization. In addition, having a high number of nodes diminishes the gain of a second node. Therefore, it was decided to take one hidden layer with 30 nodes.

Figure 6.27 shows the exact layout of the neural network.

In analogy to Section 6.3.3, several neuronal network were trained by using the training configuration described in the previous section with the exception that exactly one variable is dropped from the training. This approach indicates the impact of each training variable on the separation power of the neuronal network.

The areas under the ROC-Curves for each of the 23 trainings are summarized in Figure 6.28. The first bin quantifies the area for the training using all variables, whereas the other bin labels denote which variables has been dropped from the training. All tested configurations deviate from the standard setting by a maximum of 1%. The largest declines in the areas are achieved for training without $m_{\mathrm{T}}^{b,\mathrm{min}}$, $m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{min}}$ and $m^{1}_{\mathrm{jet},R=1.2}$. The discrimination power of $m_{\mathrm{T}}^{b,\mathrm{min}}$ against $t\bar{t}$ is discussed in Section 4.3. The relative shapes of the SM-background and of the reference signal are shown in Figure 6.1 and 6.4 after the $\tilde{t}_1$ preselection has been applied. Both distributions demonstrate that large fractions of the background are rejected by applying a minimum requirement of ≈ 150 GeV and 100 GeV respectively.

There is also a small gain in the area under the ROC-Curve, when $E_{\mathrm{T}}^{\mathrm{miss}}$ is not used for training. However, this result is likely to be attributed to an increase in the amount of overtraining for signal events as it is shown by the Kolmogorov-Smrinov-Score in Figure 6.29. Another reason might be that the large correlations between the $E_{\mathrm{T}}^{\mathrm{miss}}$ and other variables e.g. $H_{\mathrm{T}}^{\mathrm{sig}}$ (c. f. Table 6.1) partially encode the information of this variable and thus give the network more freedom to make use of it. All remaining

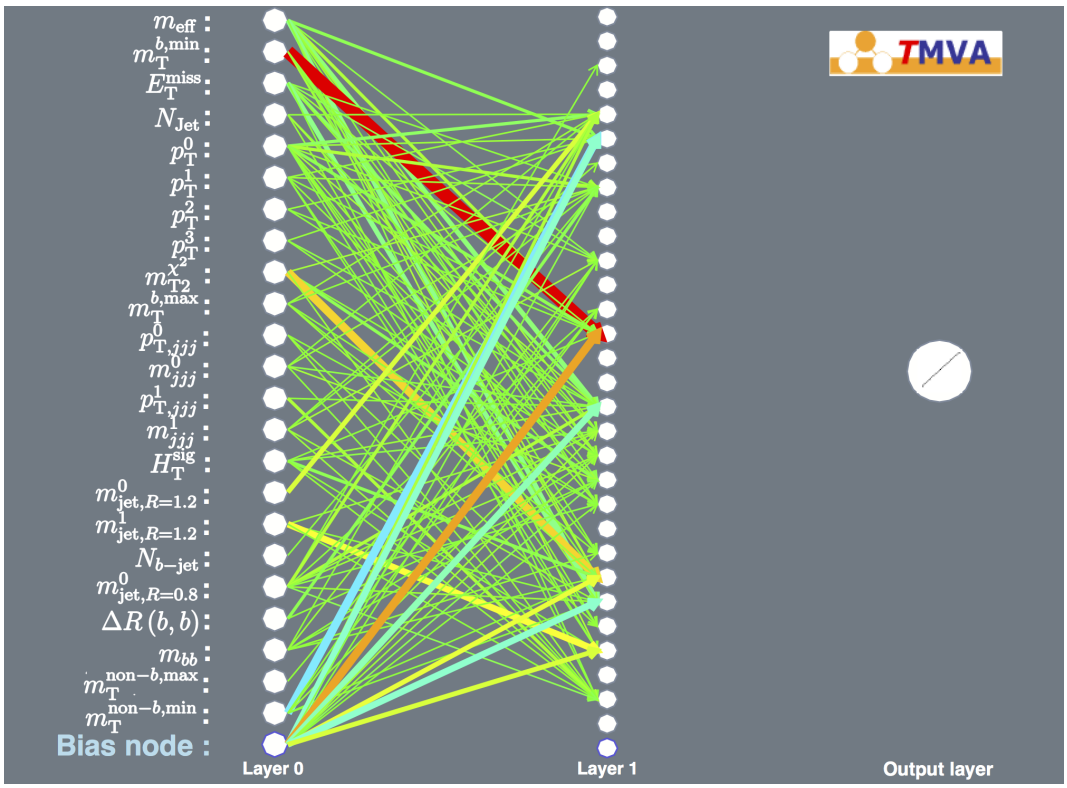


Figure 6.27: Final neural network, the training of which was done with all input variables. The lines indicate the weights of the network. The thickness corresponds to the absolute value of the weights. The colors also indicate the sign. Blue colors show that the weight is negative, orange-yellow colors indicate a positive weight, whereas green lines correspond to small weights. The absence of a plotted line means that the weight is smaller than the weights marked green.

variables do not show a particularly large impact on the area covered by the ROC-Curve. However, missing variables like p_{T}^{0} or m_{bb} increase overtraining up to 30%.

$$I\left(x\right) = \bar{x}^2 \sum_{j=1}^{n} \left(w_{ij}^{(1)}\right)^2 . \tag{6.6}$$

Table 6.4 shows the results at this ranking. The three best-ranked variables are $m_{\mathrm{T}}^{b,\mathrm{min}}$, $m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{min}}$ and m_{eff}. This observation coincides with the results from trainings with $N-1$ variables. $m_{\mathrm{T}}^{b,\mathrm{min}}$ and $m_{\mathrm{jet},R=1.2}^{0}$ gain the largest drops in the area under the ROC-Curves, while training without m_{eff} achieves the lowest sensitivity. The third

one is $m^1_{\text{jet},R=1.2}$ in Table 6.4 fifth-best ranked, showing that the best (ROC-Curve) variables are also among the most important neural network variables.

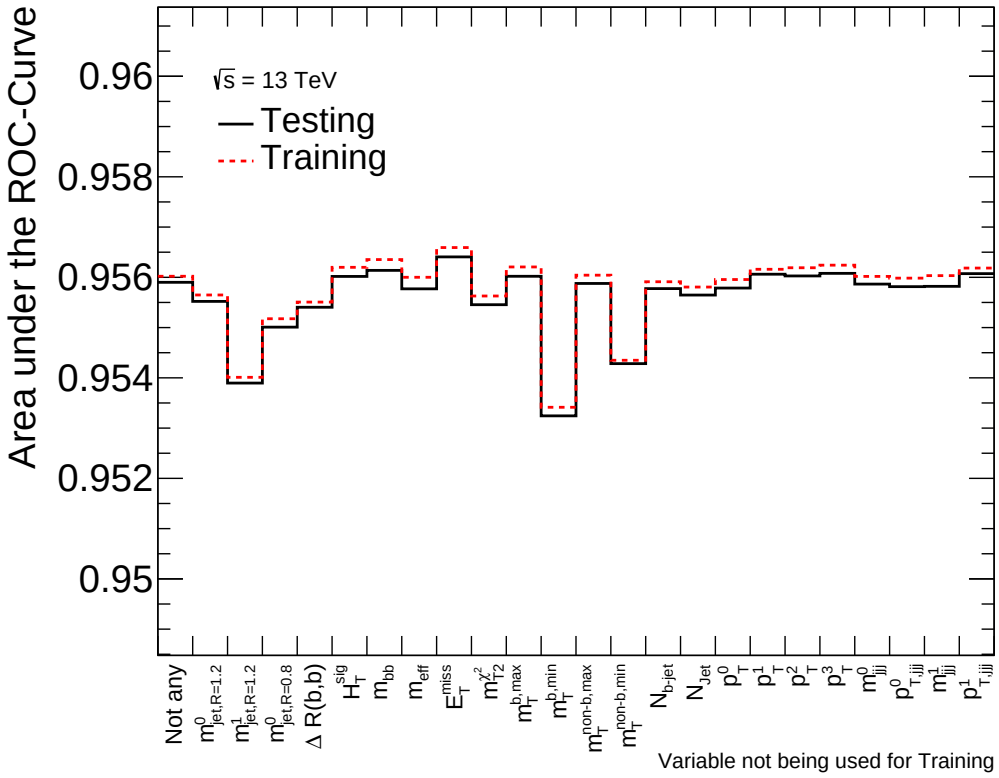


Figure 6.28: Area under the ROC-Curves of the BDT with the final parameter settings for which all but one variables were used for training. The bin labels denote which variable from Table 6.1 was not used.

Table 6.4: Variable ranking of the Neural Network with final parameter settings. The importance is determined by the weights of the connection between the input variable and the neurons of the first (and only) hidden layer.

Rank	Variable	Importance
1	$m_{\mathrm{T}}^{b,\mathrm{min}}$	1202
2	$m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{min}}$	674.5
3	m_{eff}	548.1
4	$m_{\mathrm{T2}}^{\chi^2}$	382.2
5	$m_{\mathrm{jet},R=1.2}^{1}$	349.8
6	$E_{\mathrm{T}}^{\mathrm{miss}}$	321.4
7	p_{T}^{0}	301.5
8	$m_{\mathrm{jet},R=0.8}^{0}$	199.4
9	m_{jjj}^{0}	139.4
10	$m_{\mathrm{jet},R=1.2}^{0}$	136.7
11	p_{T}^{1}	96.64
12	m_{jjj}^{1}	94.16
13	m_{bb}	93.03
14	$p_{\mathrm{T},jjj}^{1}$	77.35
15	p_{T}^{2}	70.18
16	$m_{\mathrm{T}}^{b,\mathrm{max}}$	67.56
17	p_{T}^{3}	58.76
18	$H_{\mathrm{T}}^{\mathrm{sig}}$	50.64
19	$p_{\mathrm{T},jjj}^{0}$	48.33
20	$\Delta R\,(b,b)$	36.28
21	N_{jet}	31.18
22	$m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{max}}$	28.89
23	$N_{b-\mathrm{jet}}$	20.01

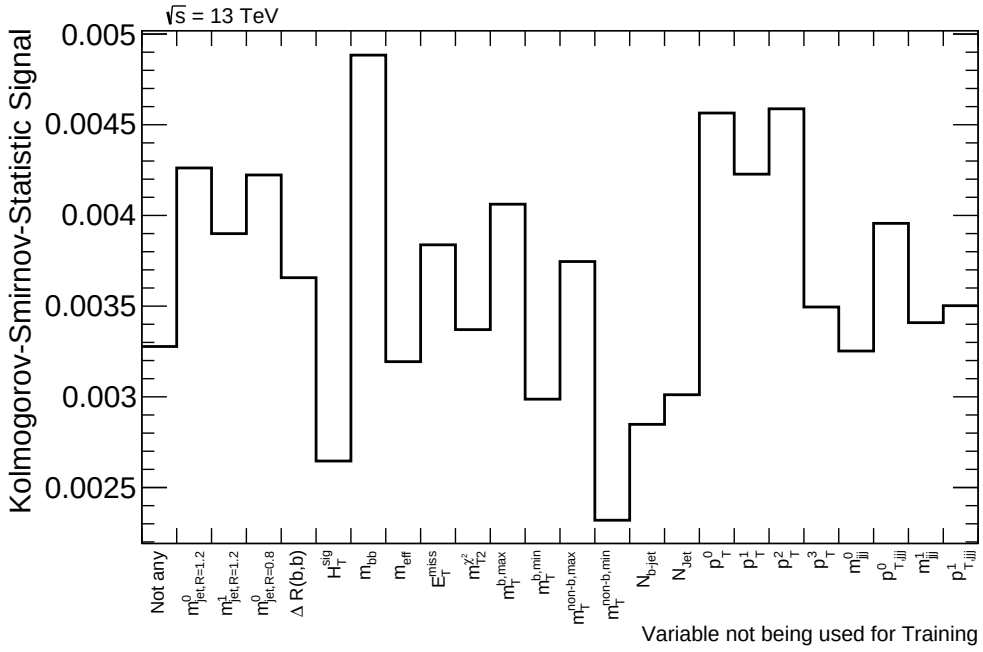


Figure 6.29: Signal Kolmogorov-Smirnov-Score of the neural network with the final parameter settings for which all but one variables were used for training. The bin labels denote which variable from Table 6.1 was not used.

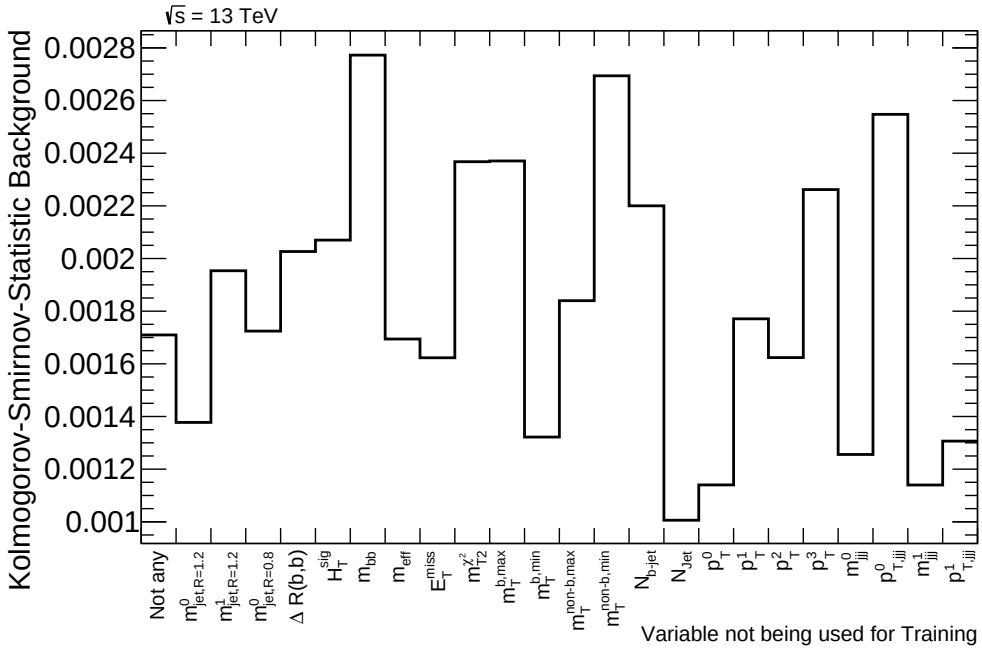


Figure 6.30: Background Kolmogorov-Smirnov-Score of the neural network with the final parameter settings for which all but one variables were used for training. The bin labels denote which variable from Table 6.1 was not used.

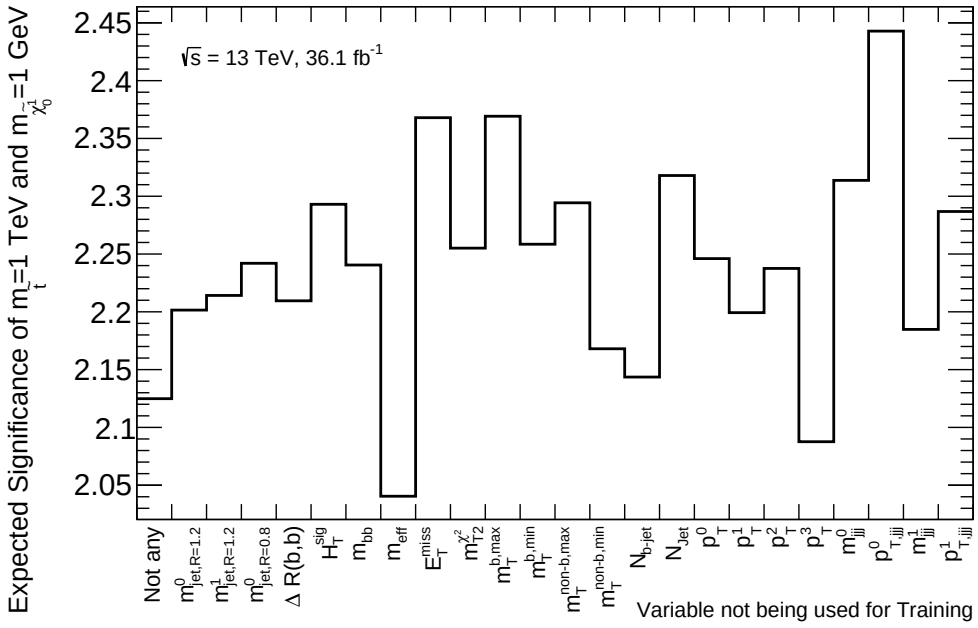


Figure 6.31: Significance of the test signal model for the neural network with the final parameter settings for which all but one variables were used for training. The bin labels denote which variable from Table 6.1 was not used.

6.4.2 Final Results

The maximum expected significance for applying a cut on the ANN-score is shown in Figure 6.32 as a function of the top squark and the neutralino mass. The result does not show the same performance as the significances obtained from using a BDT. The 3σ contour falls back to stop-masses of 900 GeV. They do not reveal a significant improvement to the expected sensitivities of SRA_TT of the cut-based analysis. This is mainly caused by the constant socket of background events for ANN-score > 0.99.

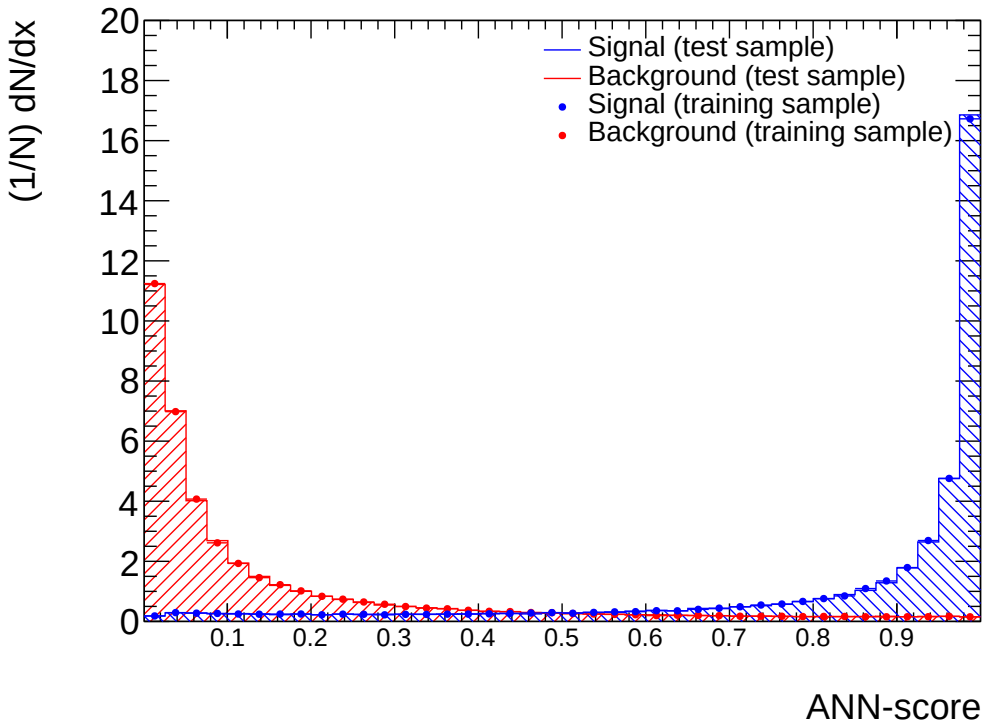


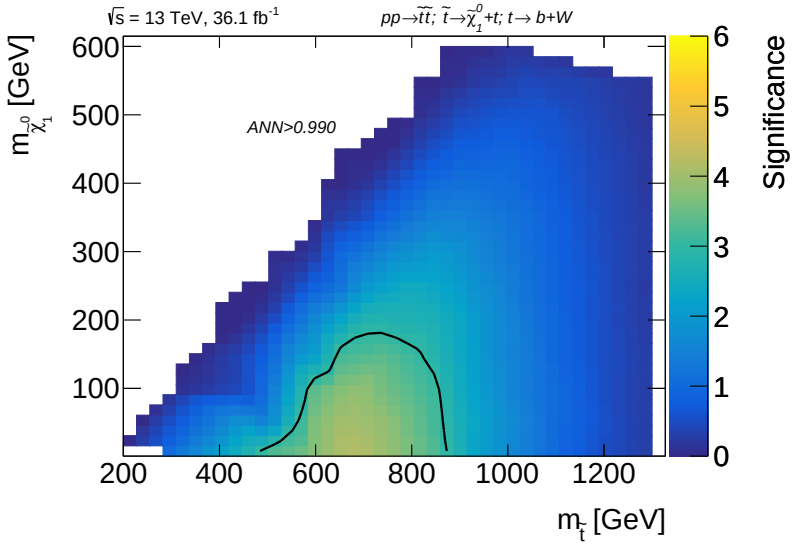
Figure 6.32: The ANN distribution for signal and background training and testing events. The neural network separates both distributions very accurately as the majority of the events is located at the edges of the figure. Throughout the entire range of the distribution, there is a constant socket of background and signal events, which is not observed for the BDT (cf. 5.7).

The significances of the reference model achieved by a neural network, for which all but one variable were used, makes training with fewer variables appear superior. Therefore, a neural net was also trained excluding $E_{\mathrm{T}}^{\mathrm{miss}}$, N_{jet}, $m_{\mathrm{T}}^{b,\mathrm{max}}$, $p_{\mathrm{T},jjj}^{0}$ and m_{jjj}^{0}, the results of which are shown in Figure 6.34. Hardly any differences can be seen here compared to the final neural network in which all variables are used for

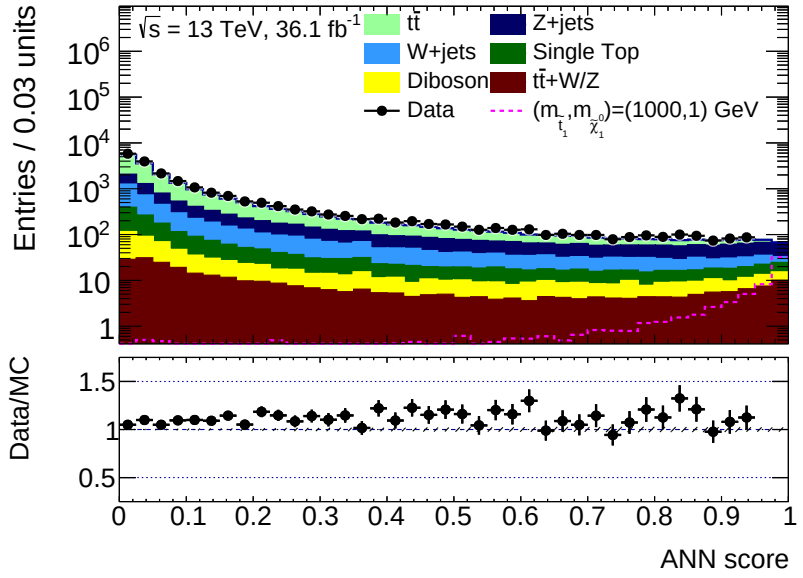
training. The training, performed with all input variables but $E_{\mathrm{T}}^{\mathrm{miss}}$, is also plotted (Figure 6.35). In fact, slight enlargement of the significances can be seen with respect to the other two training settings, but the results are still much poorer compared to the BDT. Since the overtraining of signal is significantly larger, it was decided to select the neural network trained with all input variables.

One reason for the poor results compared to BDT might be that Boosted Decision Trees and cut-based methods simply work better in order to achieve high sensitivity. Additionally, the Decision Trees also make use of cuts since at each node a cut on a variable is performed. This means that well-discriminating variables, such as $E_{\mathrm{T}}^{\mathrm{miss}}$, might have a better usage for BDT and even for the Cut and Count analysis.

Figure 6.36 shows the expected limit in the $\tilde{t}_1 - \tilde{\chi}_1^0$ mass plane for 36.1 fb^{-1}. In the case of low-mass $\tilde{\chi}_1^0$, it is possible to exclude $\tilde{t}_1$ with $m_{\tilde{t}_1} = 1040$ GeV, which is also an improvement with respect to the cut-based analysis in 36.1 fb^{-1} of data.

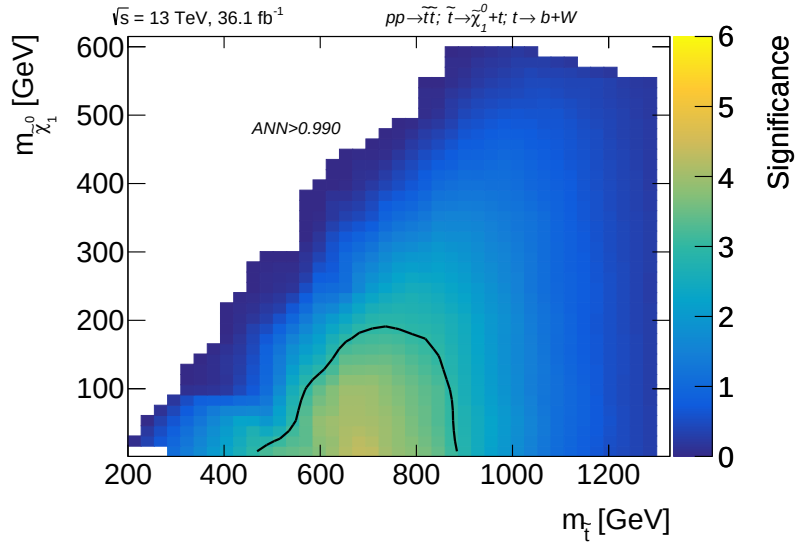


(a) Expected significance for the observation of $\tilde{t}_1$-quarks in depency of $m_{\tilde{t}_1}$ and $m_{\tilde{\chi}_1^0}$.

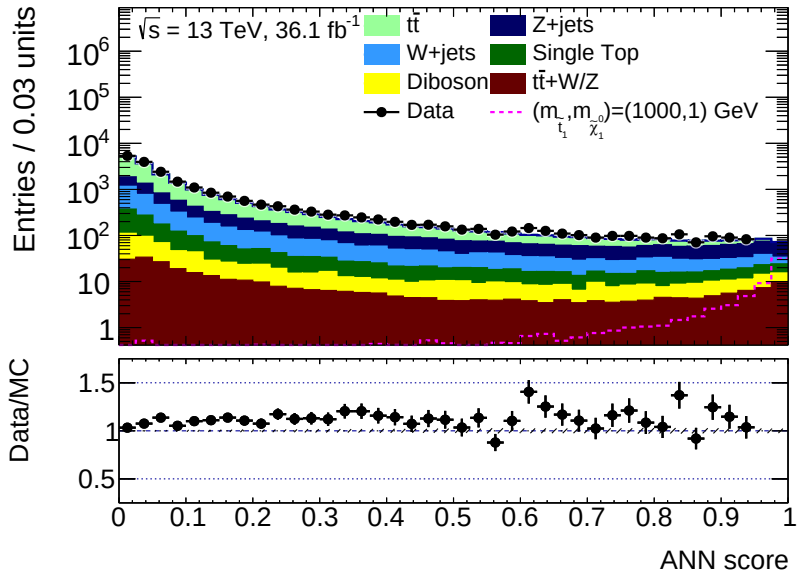


(b) Data-MC plot as a function of the ANN-score.

Figure 6.33: Significance and Data-MC plots of an MLP-training with the final settings for which all input variables were used.

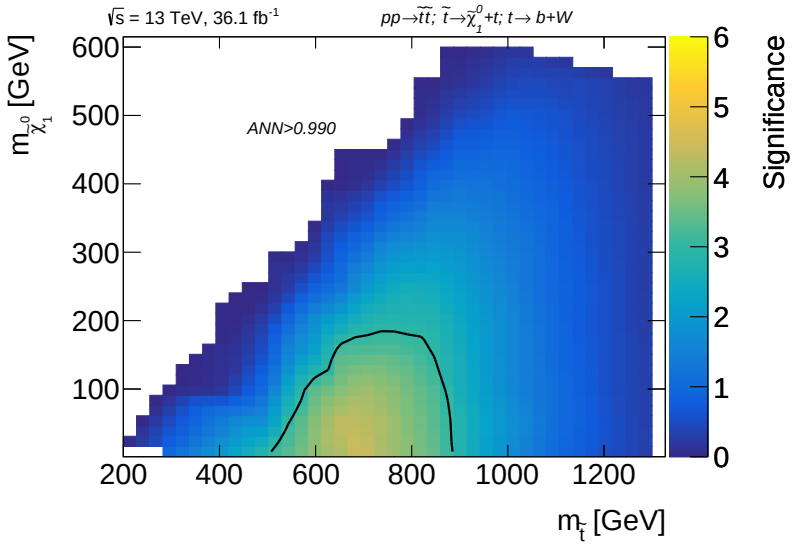


(a) Expected significance for the observation of $\tilde{t}_1$-quarks in depency of $m_{\tilde{t}_1}$ and $m_{\tilde{\chi}_1^0}$.

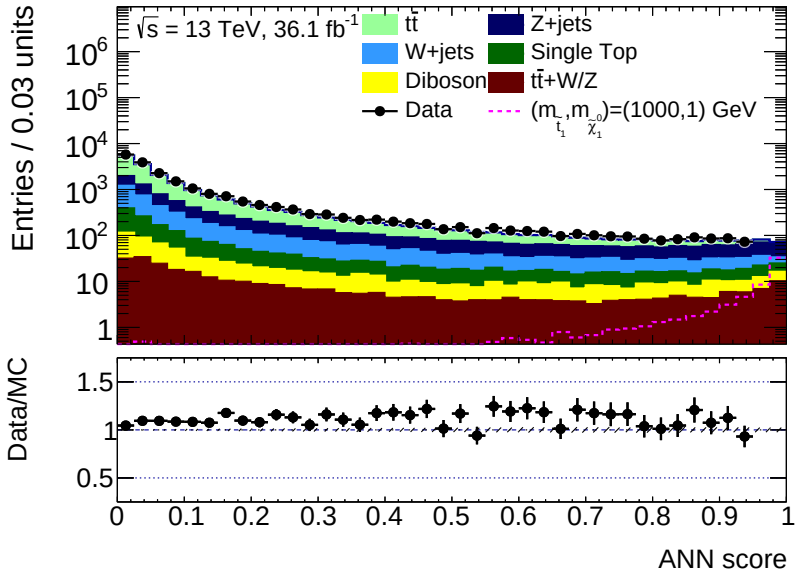


(b) Data-MC plot as a function of the ANN-score.

Figure 6.34: Significance and Data-MC plots of an ANN-training with the final settings for which all input variables but $E_{\mathrm{T}}^{\mathrm{miss}}$, N_{jet}, $m_{\mathrm{T}}^{b,\mathrm{max}}$, $p_{\mathrm{T},jjj}^{0}$ and m_{jjj}^{0} were used.



(a) Expected significance as a function of $m_{\tilde{t}_1}$ and $m_{\tilde{\chi}_1^0}$.



(b) Data-MC plot as a function of the ANN-score.

Figure 6.35: Significance and Data-MC plots of the final MLP-training trained with all variables but $E_{\mathrm{T}}^{\mathrm{miss}}$.

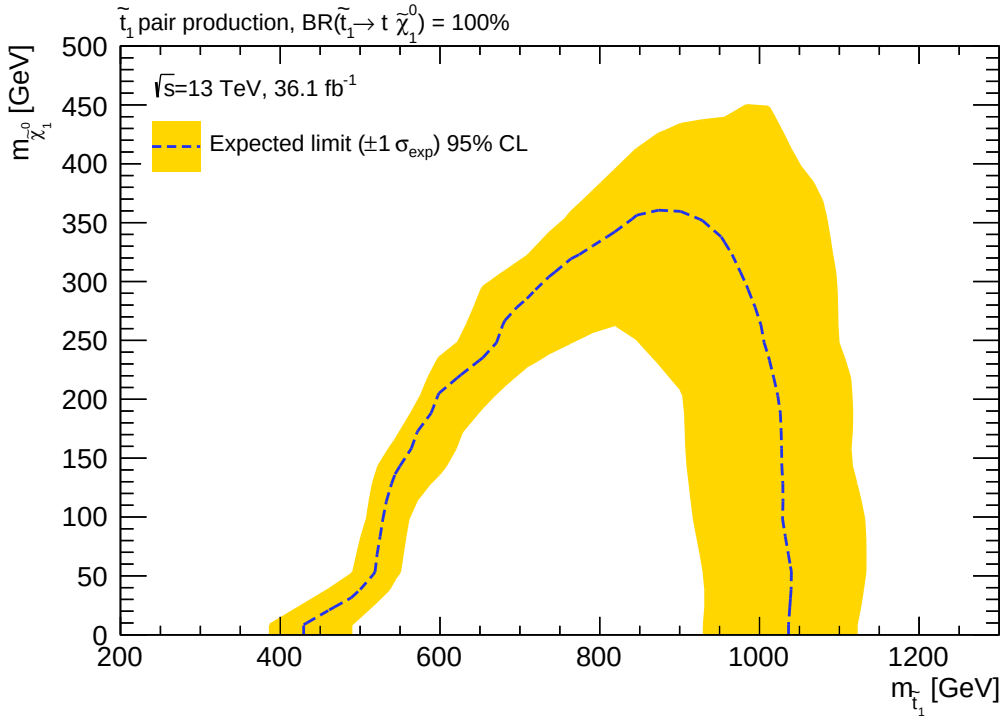


Figure 6.36: Exclusion Plot obtained with the results of the ANN-Training. It can be seen that the exclusion line is beyond 1 TeV, showing that the training achieved very good results.

Chapter 7

Summary and Conclusion

Supersymmetry (SUSY) is the extension of the Standard Model most extensively studied at the LHC. If SUSY were indeed to provide a solution to the naturalness problem of of having a light Higgs boson, at least the lighter top squark $\tilde{t}_1$ is expected to have a mass of around few TeV, which is accessible by the LHC with its centre-of-mass energy of $\sqrt{s} = 13$ TeV. A search for pair production of $\tilde{t}_1$ decaying into top quarks and neutralinos was performed by using 36.1 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV. Only fully-hadronic top quark decays are investigated in. Since neutralinos only interact weakly and gravitationally, they cannot be detected and have to be reconstructed from the missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$.

Up to now, the search for top squarks has been performed by applying cut-based event selection. In this thesis, the possible improvement of the signal sensitivity has been examined, in particular for top squark masses of above ≈ 1 TeV and small neutralino masses (≈ 1 GeV) by using multivariate analysis methods. Two such machine learning methods have been investigated: Boosted Decision Trees (BDT) and artificial neural networks (ANN).

As a first step, the optimal training model was investigated. It has been proved that a training model containing stop-signal models with stop masses beyond 1 TeV leads to the largest increase in the sensitivity of the analysis. This enhancement in the sensitivity was further improved by tuning the training-parameters of the algorithms under the condition of a minimal overtraining.

In a further step, the parameters of the multivariate algorithms were optimized such that the best signal sensitivities could be achieved. The relative weights of the partially correlated discriminating variables in the final score of the multivariate algorithms were determined. $E_{\mathrm{T}}^{\mathrm{miss}}$ plays a more important role in the BDT than in the ANN algorithm.

Finally, the improvement in the expected signal significances and exclusion limits was evaluated. When comparing BDT and ANN results with the original cut-based selection method, the BDT algorithm performs best for high top squark masses with an

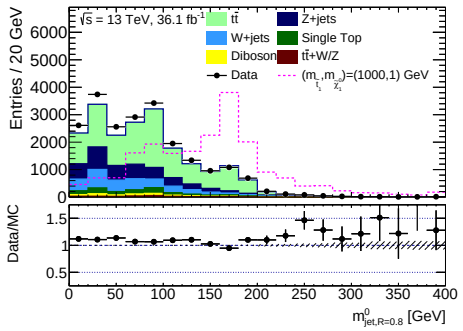
expected exclusion improvement of 9% from 990 GeV to 1080 GeV (for low neutralino masses, i. e. $m_{\tilde{\chi}_1^0} = 1$ GeV) with respect to the cut-based methods. Furthermore, the expected significances increase a lot when using the BDT. While cut-based methods achieve an expected significance of 1.7σ for $m_{\tilde{t}_1} = 1$ TeV and $m_{\tilde{\chi}_1^0} = 1$ GeV, the BDT achieves 3.2σ. By contrast, the ANN algorithm leads to little or no improvement in significance and exclusion reach compared to the cut-based methods over most of the $m_{\tilde{t}_1} - m_{\tilde{\chi}_1^0}$ mass range. Only for high $m_{\tilde{t}_1}$, an improvement in the $m_{\tilde{t}_1}$-limit by 5% is expected. The corresponding expected significance 2.1σ shows that ANN scores better than the cut-based methods, however, the BDT scores significantly better.

Overall, 23 discriminating variables were used for training. About seven of the variables are highly important, whereas about nine variables are ranked low. It is difficult, however, to fix a limit since there are different rankings, in which the variables perform differently well. An ANN, in which five discriminating variables are not used for the training, shows little changes with regard to a neural network trained with all input variables.

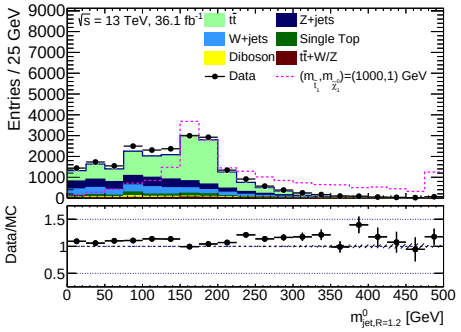
Many scientists, however, are skeptical of machine learning techniques. Therefore, this method is rather suitable as a hint for excesses in the present situation rather than for exclusion limits.

Appendix

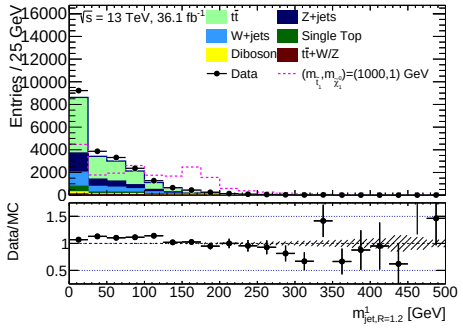
1 Input Variables



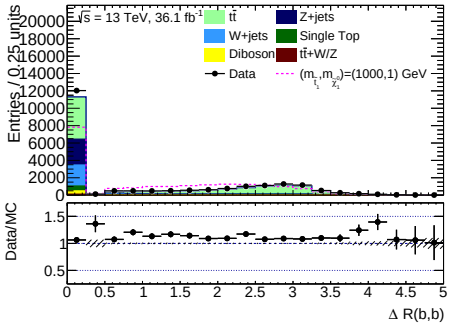
(a) $m^0_{\text{jet},R=0.8}$



(b) $m^0_{\text{jet},R=1.2}$

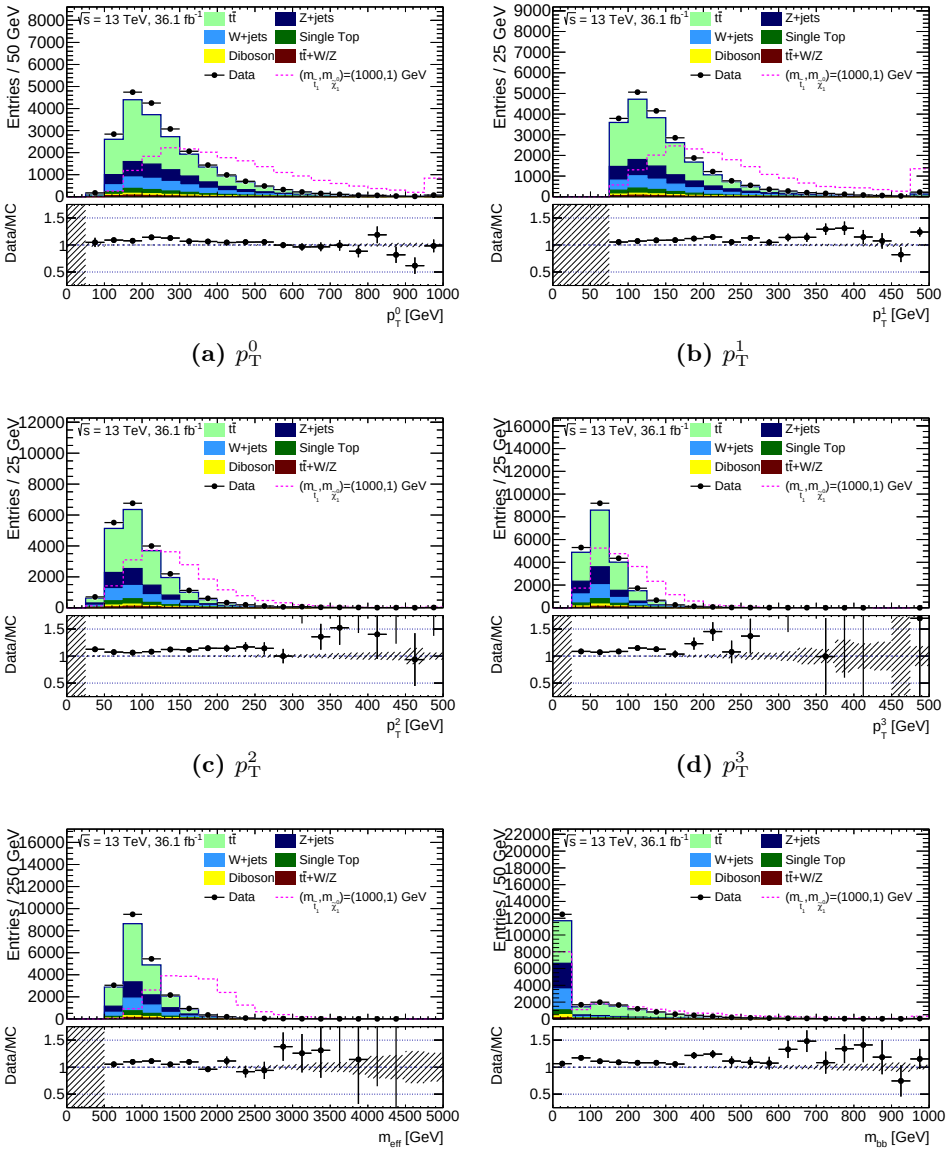


(c) $m^1_{\text{jet},R=1.2}$



(d) $\Delta R\,(b,b)$

Figure 1: Distributions of $m^0_{\text{jet},R=0.8}, m^0_{\text{jet},R=1.2}, m^1_{\text{jet},R=1.2}$ and $\Delta R\,(b,b)$ after the preselection requirement. The stacked histograms show the SM expectation, whereas the dashed purple line shows the distribution of the reference model normed to the integral of the background events. The panel beneath shows the ratio of data events to the total SM expectation. The hatched uncertainty band around the SM expectation and in the ratio plot illustrates statistical uncertainties. The rightmost bin includes all overflows.

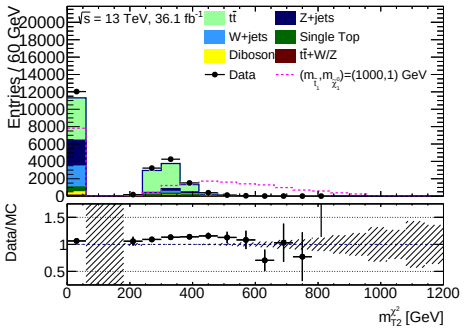


(a) p_{T}^{0} **(b)** p_{T}^{1}

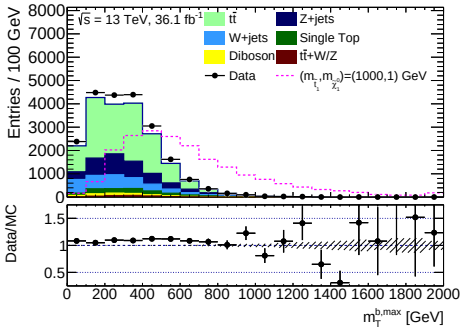
(c) p_{T}^{2} **(d)** p_{T}^{3}

(e) m_{eff} **(f)** m_{bb}

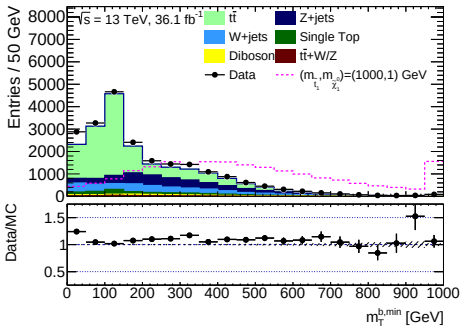
Figure 2: Distributions of $p_{\mathrm{T}}^{0}, p_{\mathrm{T}}^{1}, p_{\mathrm{T}}^{2}, p_{\mathrm{T}}^{3}, m_{\mathrm{eff}}$ and m_{bb} after the preselection requirement. The stacked histograms show the SM expectation, whereas the dashed purple line shows the distribution of the reference model normed to the integral of the background events. The panel beneath shows the ratio of data events to the total SM expectation. The hatched uncertainty band around the SM expectation and in the ratio plot illustrates statistical uncertainties. The rightmost bin includes all overflows.



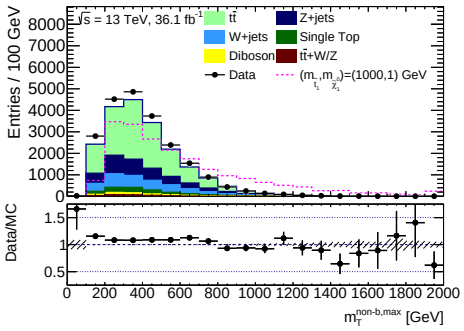
(a) $m_{\mathrm{T2}}^{\chi^2}$



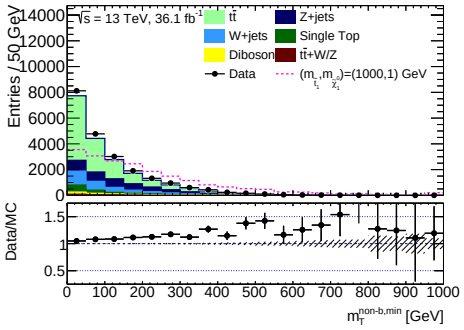
(b) $m_{\mathrm{T}}^{b,\mathrm{max}}$



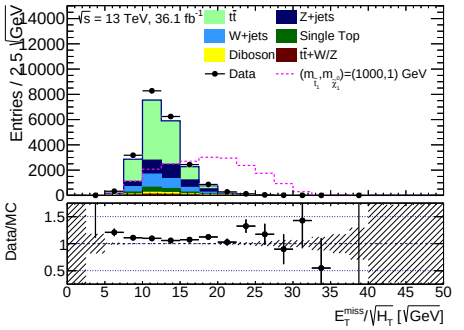
(c) $m_{\mathrm{T}}^{b,\mathrm{min}}$



(d) $m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{max}}$

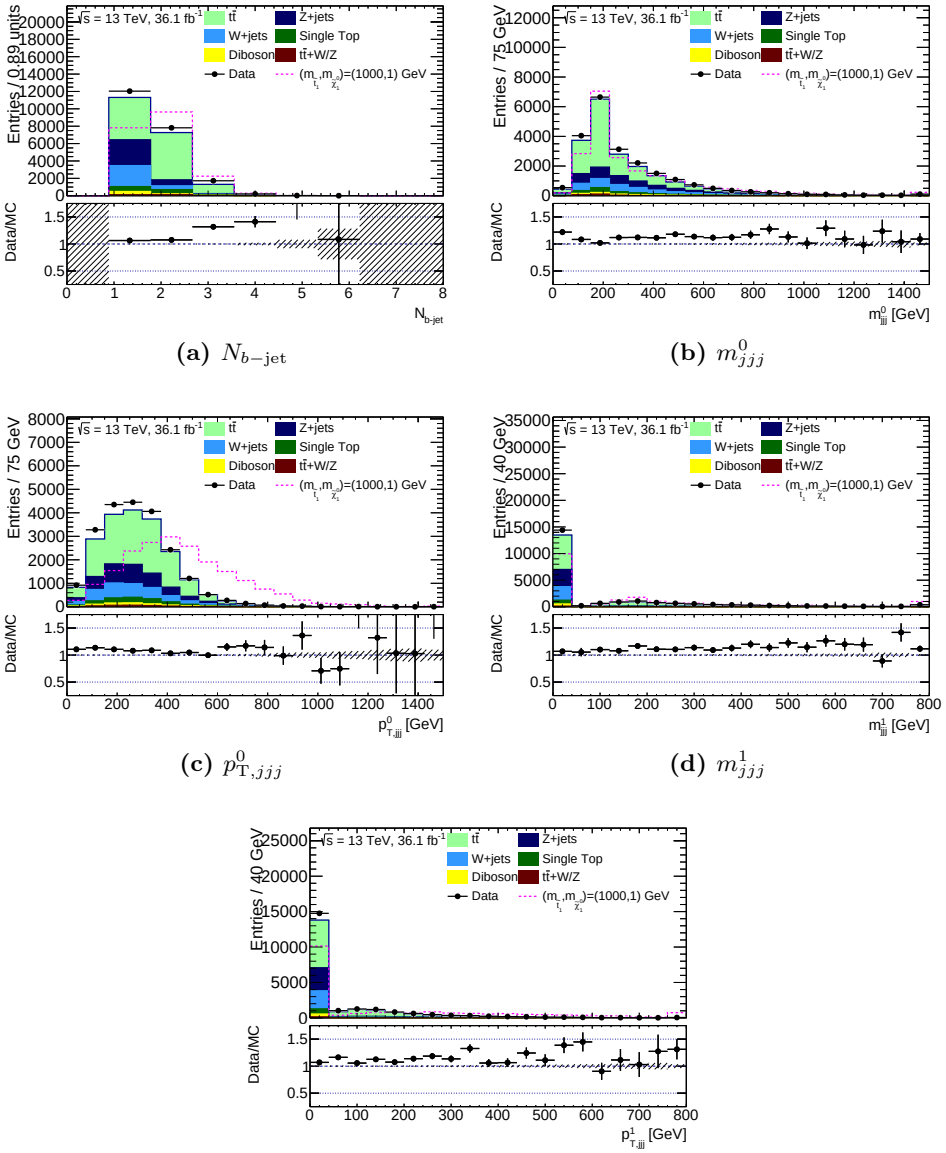


(e) $m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{min}}$



(f) $H_{\mathrm{T}}^{\mathrm{sig}}$

Figure 3: Distributions of $m_{\mathrm{T2}}^{\chi^2}, m_{\mathrm{T}}^{b,\mathrm{max}}, m_{\mathrm{T}}^{b,\mathrm{min}}, m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{max}}, m_{\mathrm{T}}^{\mathrm{non}-b,\mathrm{min}}$ and $H_{\mathrm{T}}^{\mathrm{sig}}$ after the preselection requirement. The stacked histograms show the SM expectation. The panel beneath shows the ratio of data events to the total SM expectation. The hatched uncertainty band around the SM expectation and in the ratio plot illustrates statistical uncertainties. The rightmost bin includes all overflows.



(a) $N_{b-\mathrm{jet}}$ **(b)** m^0_{jjj}

(c) $p^0_{\mathrm{T},jjj}$ **(d)** m^1_{jjj}

(e) $p^1_{\mathrm{T},jjj}$

Figure 4: Distributions of $N_{b-\mathrm{jet}}, m^0_{jjj}, p^0_{\mathrm{T},jjj}, m^1_{jjj}$ and $p^1_{\mathrm{T},jjj}$ after the preselection requirement. The stacked histograms show the SM expectation, whereas the dashed purple line shows the distribution of the reference model normed to the integral of the background events. The panel beneath shows the ratio of data events to the total SM expectation. The hatched uncertainty band around the SM expectation and in the ratio plot illustrates statistical uncertainties. The rightmost bin includes all overflows.

2 Boosted Decision Trees

2.1 Real AdaBoost

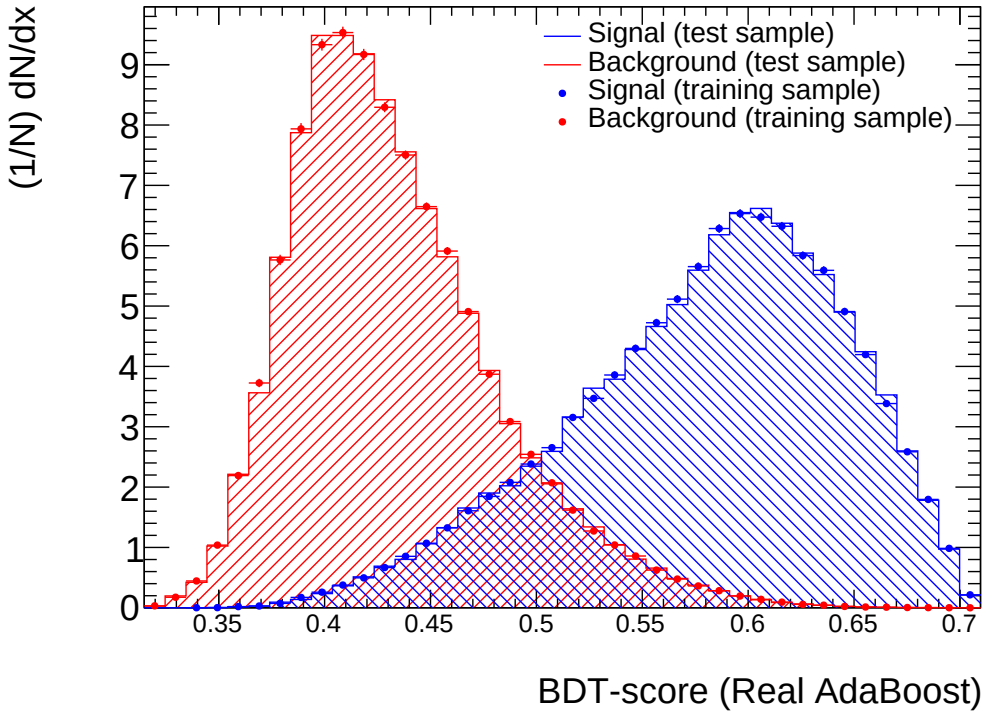


Figure 5: The distribution of signal (blue) and background (red) for training (dots) and testing (lines) for a BDT-training with boost type real AdaBoost.

2.2 Gradient Boost

Training data shall be labeled with $(\mathbf{x}^i, y^i)$, decision trees with f_m and its output of an event with $f_m(\mathbf{x}^i)$. The output function of a decision forest F, shall be a linear combination of all M trees in the forest, with $\beta_m \in \mathbb{R}$ being the weighting coefficients:

$$F(\mathbf{x}^i) = \sum_{j=0}^{M} \beta_j f_j(\mathbf{x}^i). \tag{1}$$

A loss function L, comparing the output of the decision forest $F(\mathbf{x}^i)$ with the true classification y^i, is given by:[1]

$$L(F(\mathbf{x}^i), y^i) = \ln\left(1 + \exp\left(-2F(\mathbf{x}^i)y^i\right)\right). \tag{2}$$

The loss is supposed to be minimized. Since this problem, like most minimization problems, is in general not mathematically exactly solvable, approximations need to be done, in this case, by the steepest descent. The sum of the first n (weighted) trees is defined to be

$$F_m(\mathbf{x}^i) = \sum_{j=0}^{m} \beta_j f_j(\mathbf{x}^i). \tag{3}$$

The first tree $F_0(\mathbf{x}^i) = f_0(\mathbf{x}^i)$ with $\beta_0 = 1$ is merely set to be the value $\rho \in \mathbb{R}$, such that the sum of the loss function is minimized [75]:

$$F_0 = \arg\min_{\rho} \sum_{i=1}^{N} L\left(\rho, y^i\right). \tag{4}$$

N depicts the number of training events. Having determined F_m, the next tree f_{m+1} will be achieved the following way: For each event, the gradient z^i is taken:

$$z^i = \frac{\partial L(F_m(\mathbf{x}^i), y^i)}{\partial F_m(\mathbf{x}^i)} = \frac{-2y^i}{\exp\left(2F_m\left(\mathbf{x}^i\right)y^i\right) + 1}. \tag{5}$$

A regression tree is built (c. f. Session 5.1) with training samples $\left(\mathbf{x}^i, z^i\right)$, meaning trees are built, such that the the difference between expected value $f_{m+1}\left(\mathbf{x}^i\right)$ and true value z^i is as low as possible. The separation type, used for node split decision,

[1] Note, it can be shown that AdaBoost is a similar problem with loss function $L(F(x^i), y^i) = \exp\left(-F(x^i)y^i\right)$ [61].

is the *average squared error*:

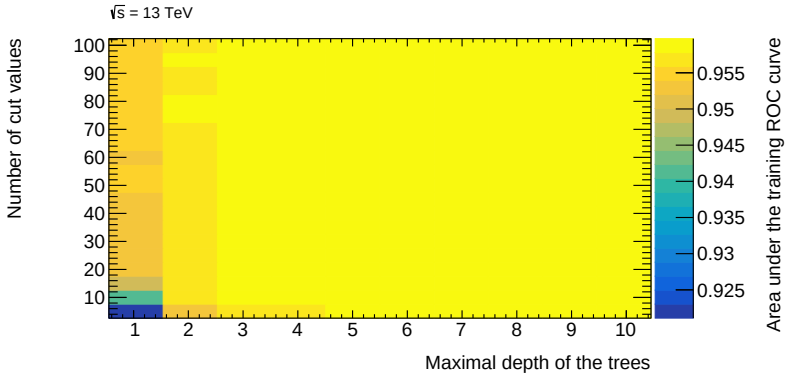
$$\frac{1}{N} \cdot \sum_{i=1}^{N} \left(z^i - \hat{z}\right)^2 . \tag{6}$$

Hereby, $\hat{z}$ is the mean value of all events z_i in the node. Besides this difference, a regression tree is built very similarly to a decision tree. The leaf values equal the mean value of the gradient of all events in this leaf. The weight factor β_{m+1} will be determined by the following equation [75] (approximated):

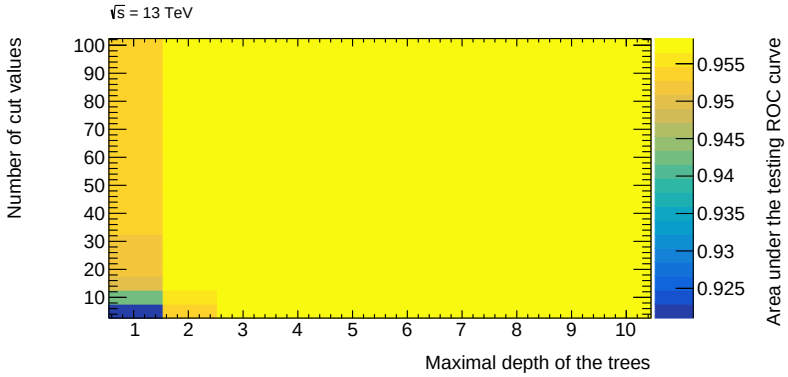
$$\beta_{m+1} = \arg\min_{\gamma \in \mathbb{R}} \sum_{i=1}^{N} L\left(F_m\left(\mathbf{x}^i\right) + \gamma \cdot f_{m+1}\left(\mathbf{x}^i\right), y^i\right). \tag{7}$$

Finally, F_{m+1} is given by Equation 3.

2.3 BDT Parameter Plots

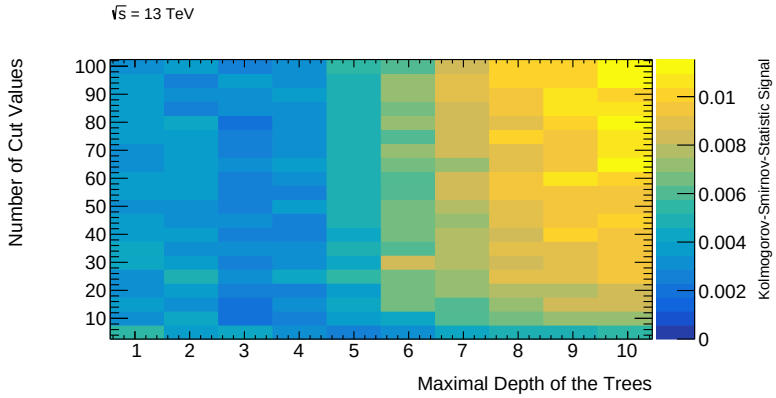


(a) Area under the Training-ROC-Curves.

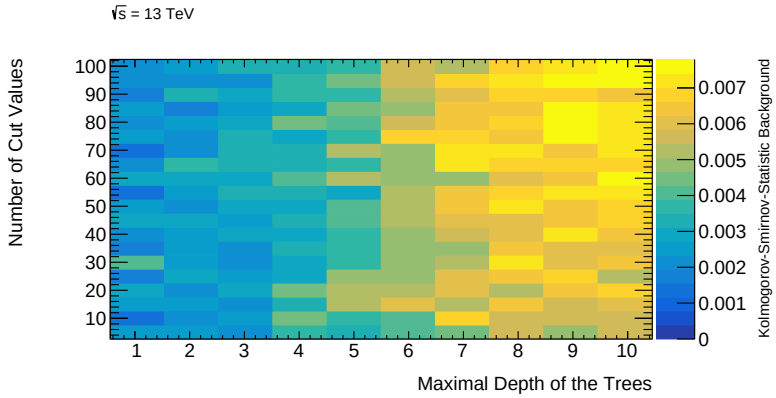


(b) Area under the Testing-ROC-Curves.

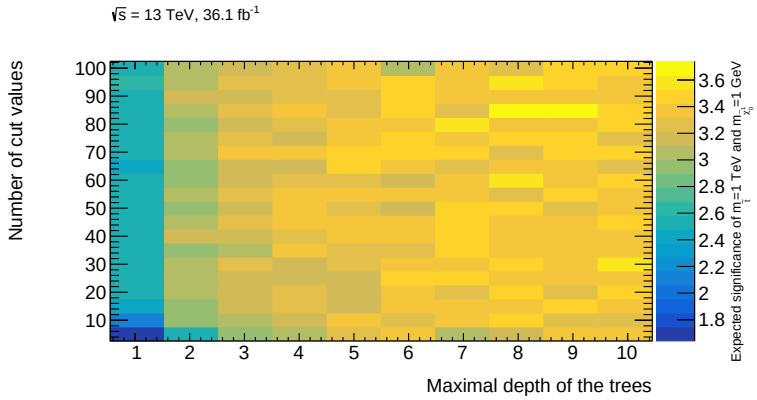
Figure 6: Area under ROC-Curves obtained from performing a $\tilde{t}_1$-BDT training with high-mass $\tilde{t}_1$ models with varied training parameters maximal depth and number of cuts for MVA testing and training events.



(a) Signal Kolmogorov-Smirnov-Score.

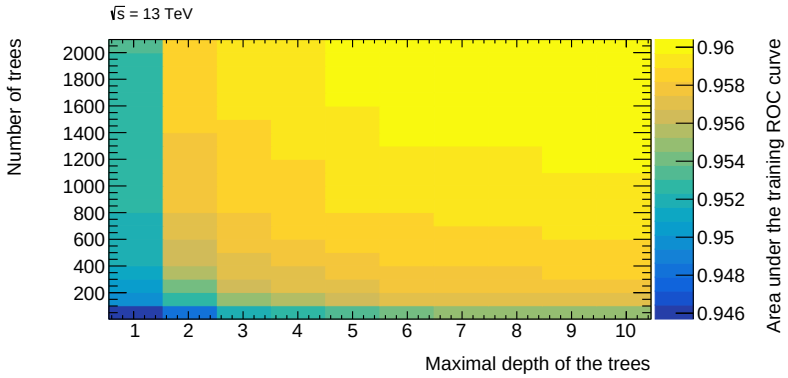


(b) Background Kolmogorov-Smirnov-Score.

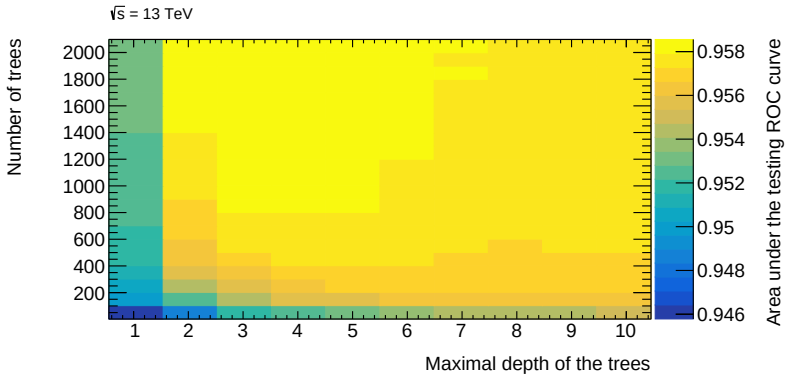


(c) Significance of the reference model for BDT trainings.

Figure 7: Kolmogorov-Smirnov-Score and significance of the reference model as a function of the maximal depth and the number of cut values.

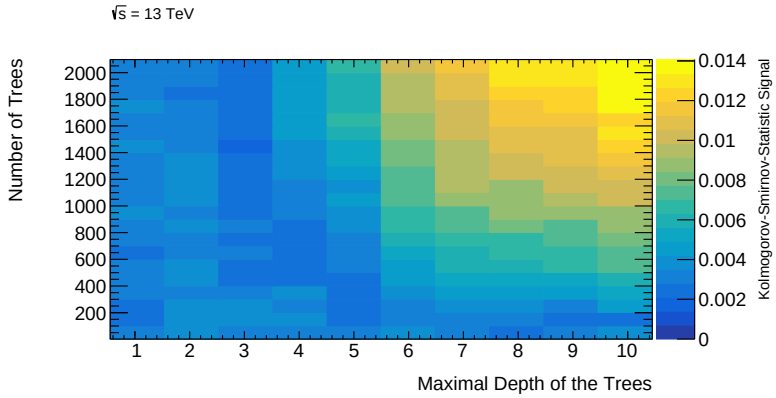


(a) Area under the Training-ROC-Curves.

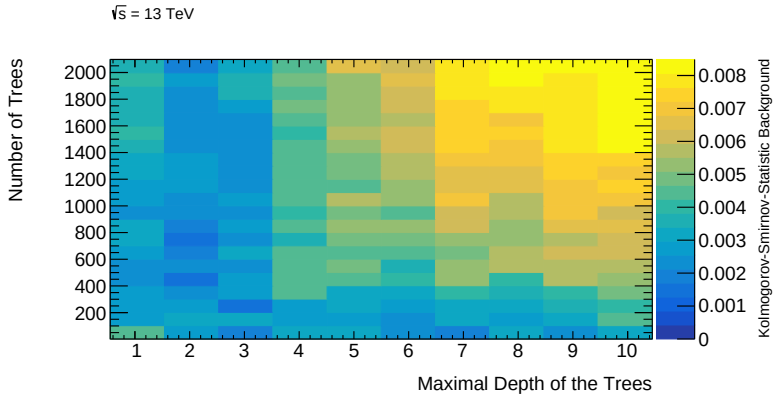


(b) Area under the Testing-ROC-Curves.

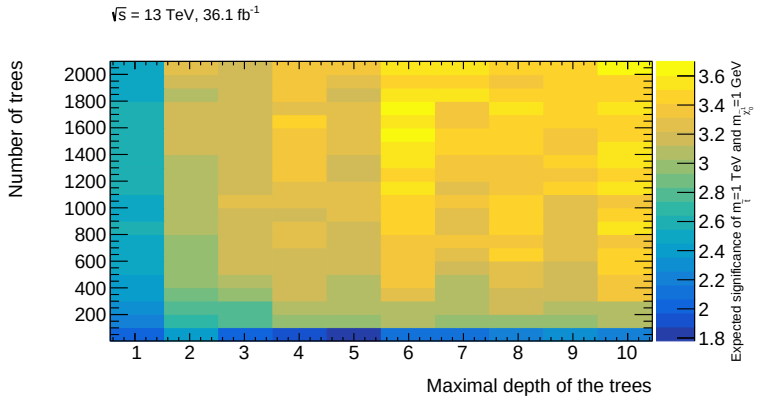
Figure 8: Area under ROC-Curves obtained from performing a $\tilde{t}_1$-BDT training with high-mass $\tilde{t}_1$ models with varied training parameters maximal depth and number of trees for MVA testing and training events.



(a) Signal Kolmogorov-Smirnov-Score.

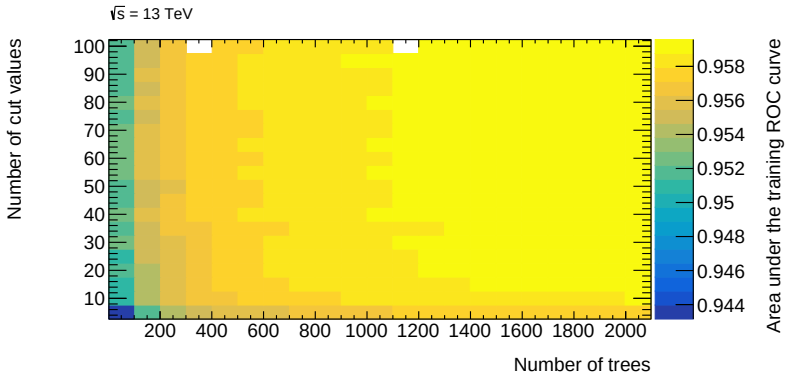


(b) Background Kolmogorov-Smirnov-Score.

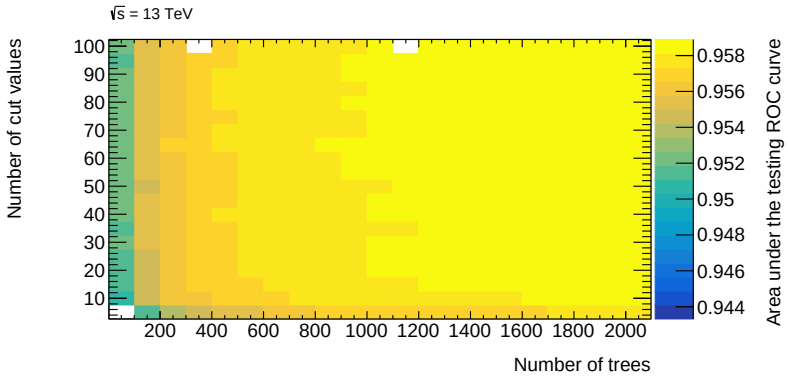


(c) Significance of the reference model for BDT trainings.

Figure 9: Kolmogorov-Smirnov-Score and significance of the reference model as a function of the maximal depth and the number of trees.

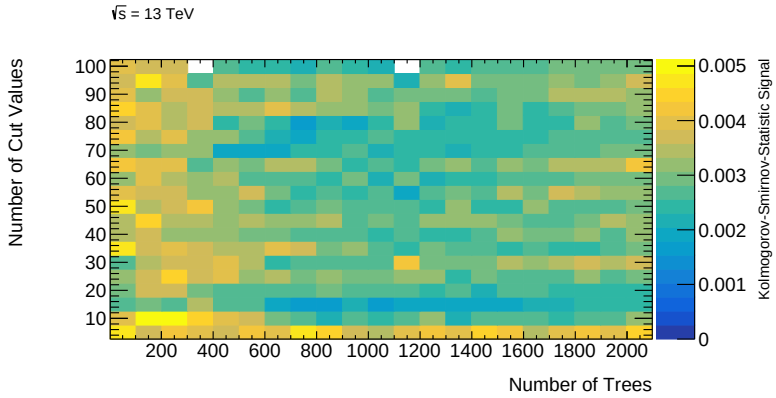


(a) Area under the Training-ROC-Curves.

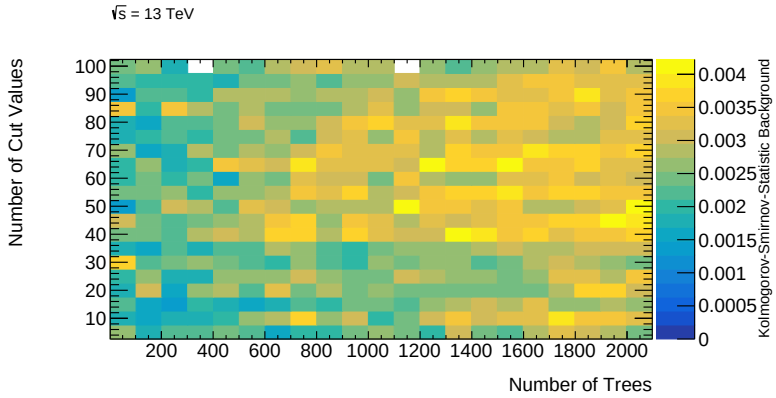


(b) Area under the Testing-ROC-Curves.

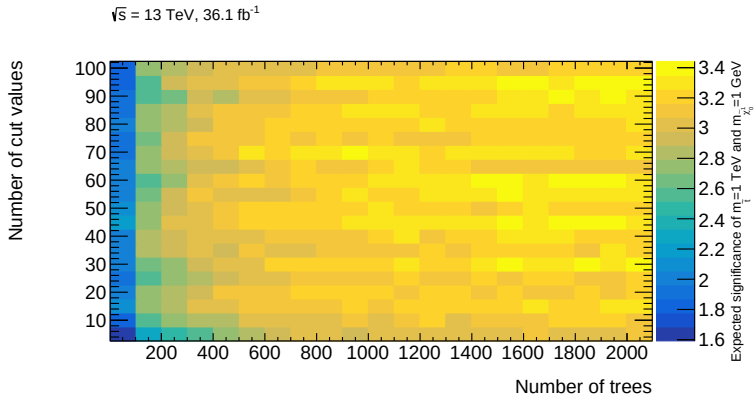
Figure 10: Area under ROC-Curves obtained from performing a $\tilde{t}_1$-BDT training with high-mass $\tilde{t}_1$ models with varied training parameters number of trees and number of for MVA testing and training events.



(a) Signal Kolmogorov-Smirnov-Score.

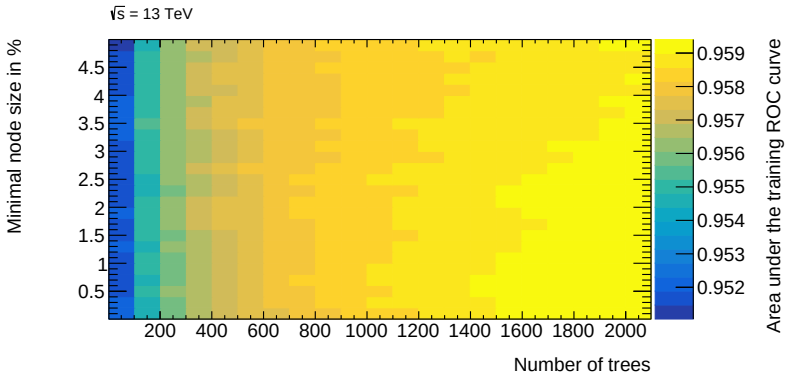


(b) Background Kolmogorov-Smirnov-Score.

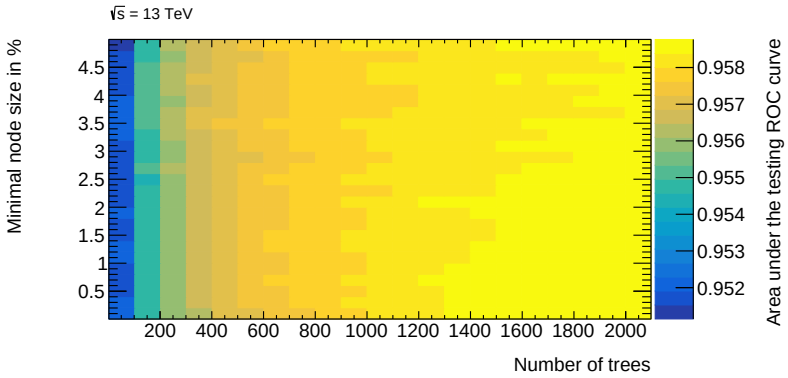


(c) Significance of the reference model for BDT trainings.

Figure 11: Kolmogorov-Smirnov-Score and significance of the reference model as a function of the number of trees and the number of cut values.

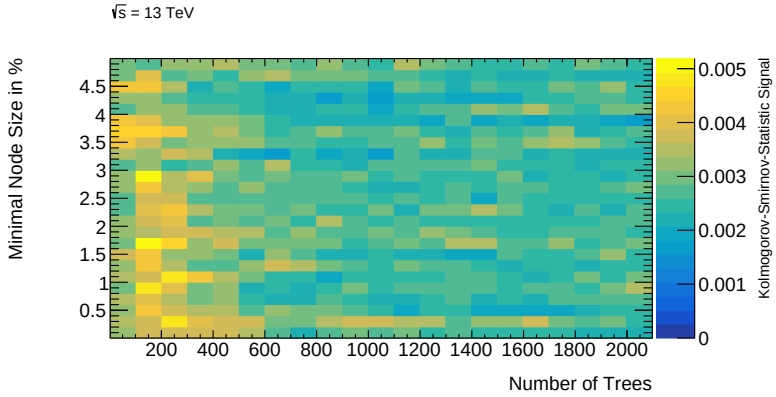


(a) Area under the Training-ROC-Curves.

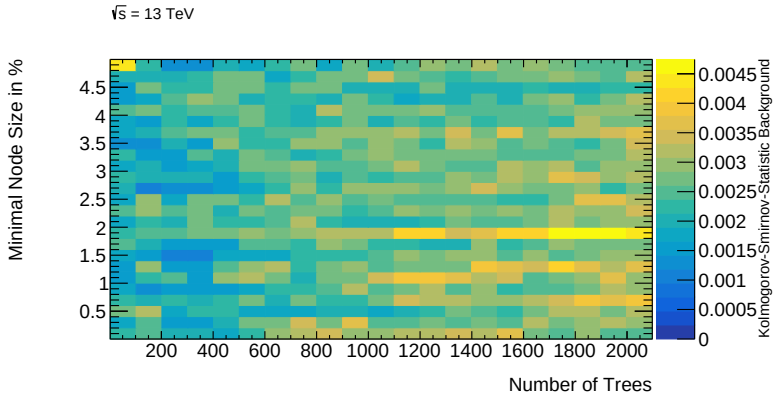


(b) Area under the Testing-ROC-Curves.

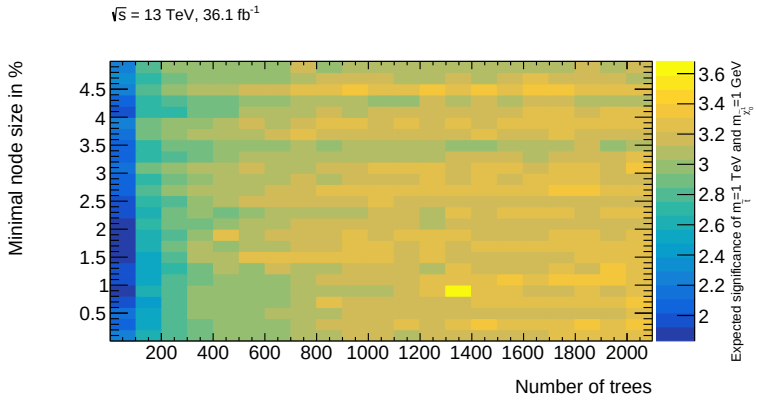
Figure 12: Area under ROC-Curves obtained from performing a $\tilde{t}_1$-BDT training with high-mass $\tilde{t}_1$ models with varied training parameters number of trees and minimal node size for MVA testing and training events.



(a) Signal Kolmogorov-Smirnov-Score.

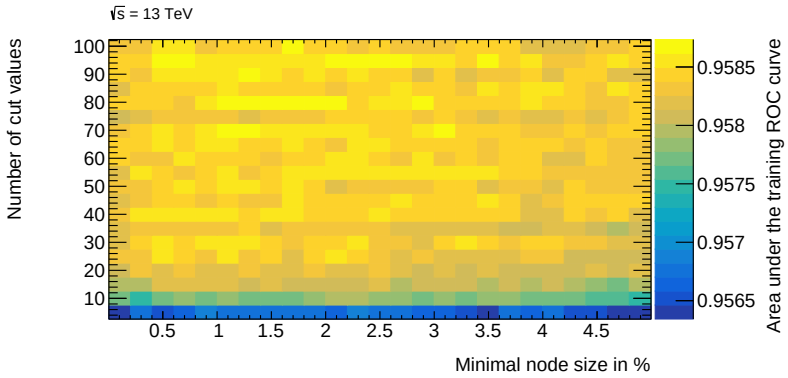


(b) Background Kolmogorov-Smirnov-Score.

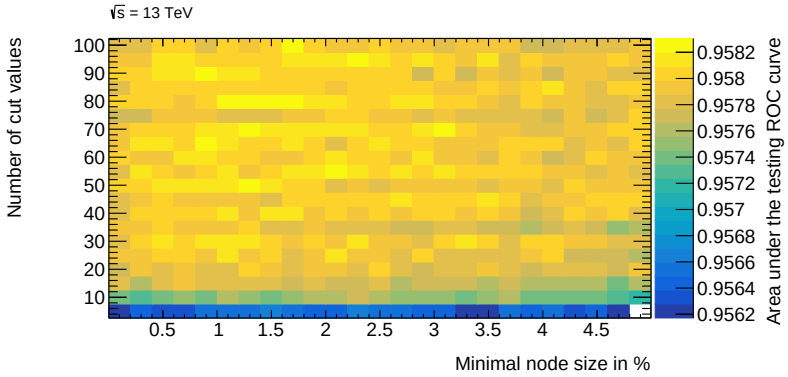


(c) Significance of the reference model for BDT trainings.

Figure 13: Area under ROC-Curves obtained from performing a $\tilde{t}_1$-BDT training with high-mass $\tilde{t}_1$ models with varied training parameters maximal number of trees and minimal node size for MVA testing and training events.

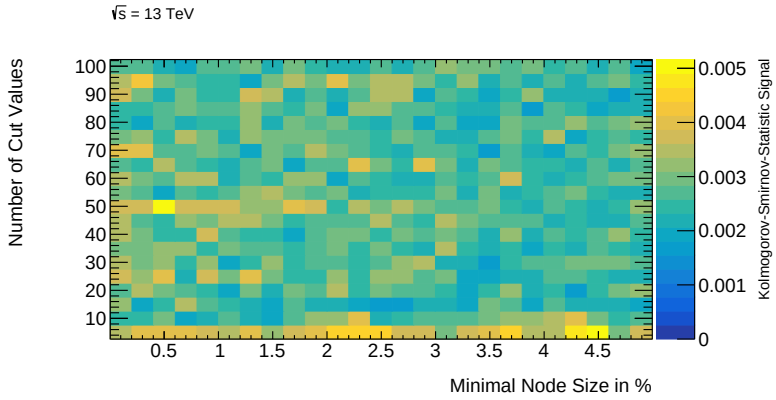


(a) Area under the Training-ROC-Curves.

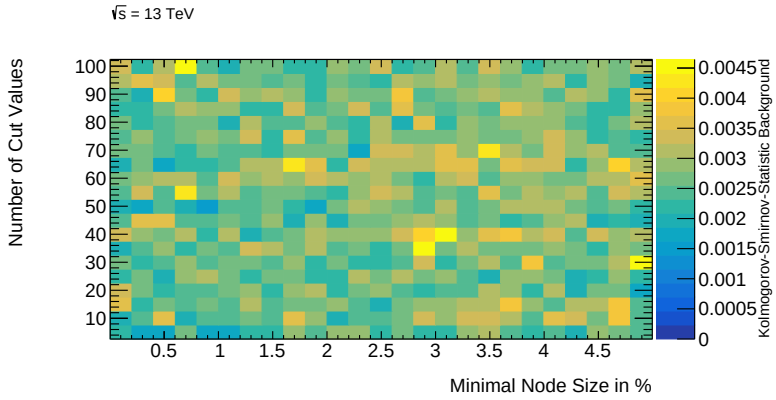


(b) Area under the Testing-ROC-Curves.

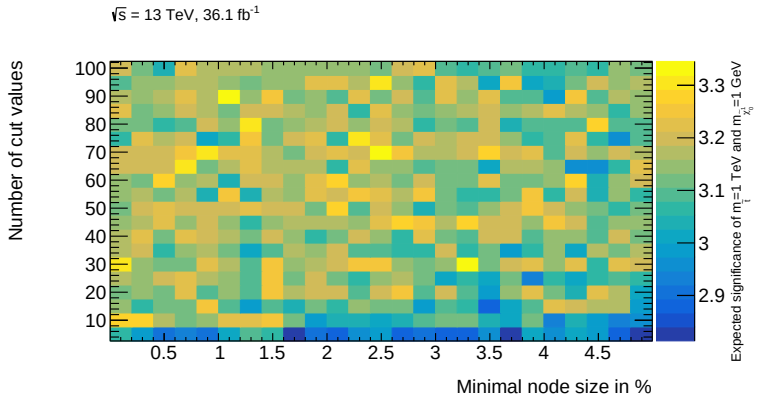
Figure 14: Area under ROC-Curves obtained from performing a $\tilde{t}_1$-BDT training with high-mass $\tilde{t}_1$ models with varied training parameters minimal node size and number of cuts MVA testing and training events.



(a) Signal Kolmogorov-Smirnov-Score.



(b) Background Kolmogorov-Smirnov-Score.



(c) Significance of the reference model for BDT trainings.

Figure 15: Kolmogorov-Smirnov-Score and significance of the reference model as a function of the minimal node size and the number of cut values.

3 Neural Networks

3.1 The BFGS algorithm

An alternate way of optimizing the weights is the BFGS-algorithm. Instead of using the steepest descent, minima of the error function E are approximated at each iteration with a quasi Newton method. Therefore, one has to look for local minima satisfying

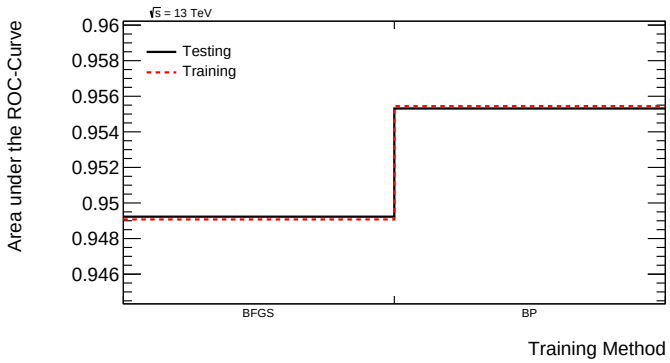
$$\nabla_{\mathbf{w}} E(\mathbf{w}) = 0. \tag{8}$$

The initial weights $\mathbf{w}^{(0)}$ are uniformly distributed in the interval $(-2, 2)$ which is equally distributed as the initial weights of the BP-algorithm. Starting from the weight vector $\mathbf{w}^{(q)}$ at the q^{th} iteration, the $q+1^{st}$ vector needs to be calculated. The search direction shall be labeled $\mathbf{p}^{(q)}$. First, the Newton equation must be solved, with $H^{(q)}$ being the Hessian matrix:

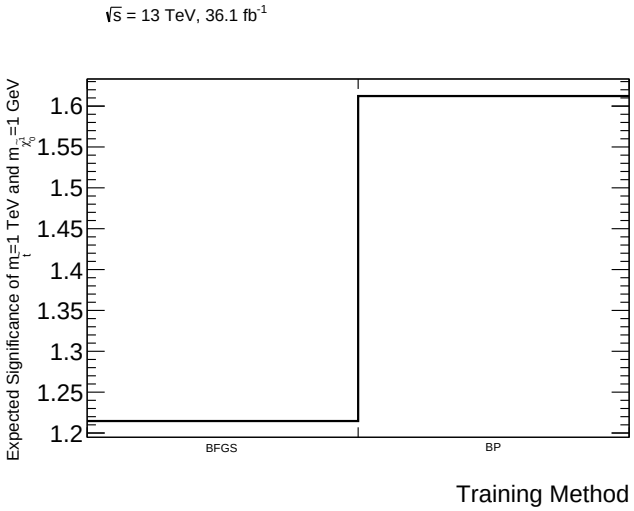
$$H^{(q)} \cdot \mathbf{p}^{(q)} = -\nabla_{\mathbf{w}^{(q)}} E\left(\mathbf{w}^{(q)}\right). \tag{9}$$

So in order to solve Equation 9, the Hessian matrix has to be inverted. This requires an enormous amount of computing resources. Therefore, the inverse Hessian matrix is approximated. There are various possibilities of approximating the inverse Hessian matrix of the BFGS algorithm (Equation 9). The approximation matrix, which is used by TMVA, can be found in [61].

3.2 Parameter Plots

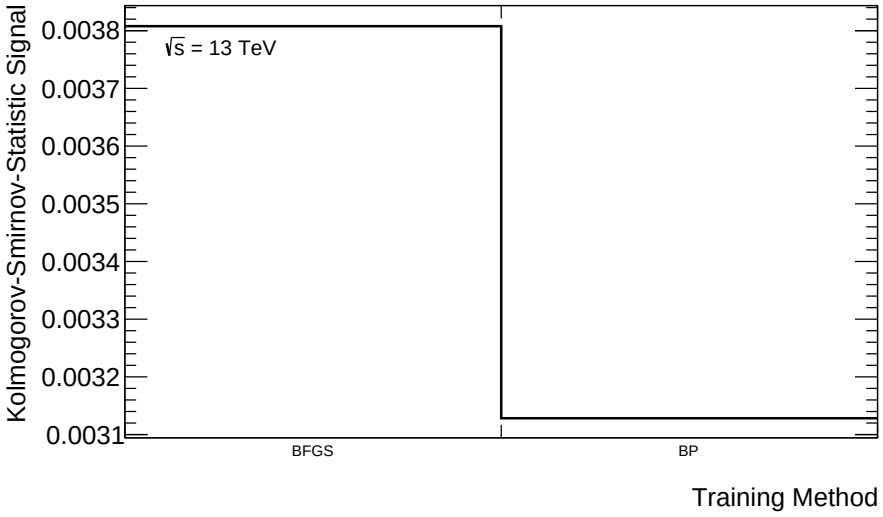


(a) Area under the ROC-Curves for both Training and Testing.

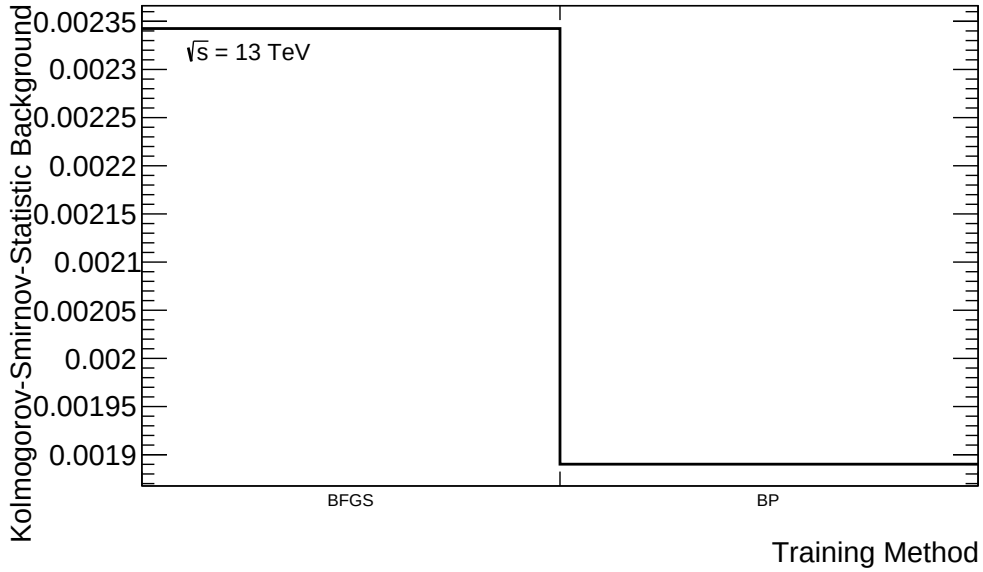


(b) Significance of the reference model for BDT trainings.

Figure 16: Area under the ROC-Curves and significance of the reference model as a function of the training method

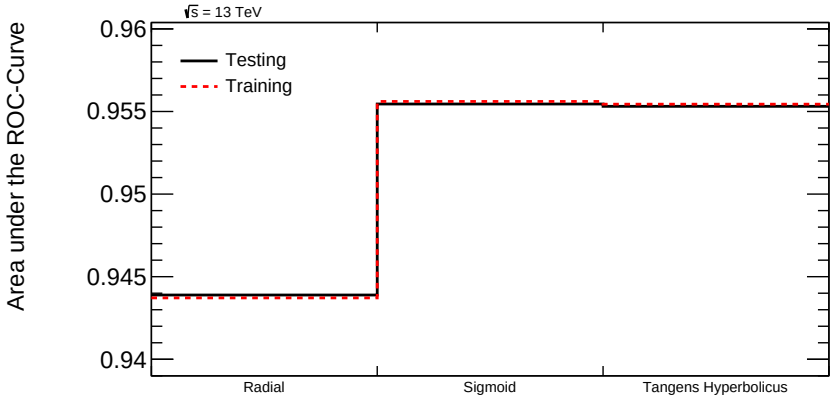


(a) Signal Kolmogorov-Smirnov-Score.

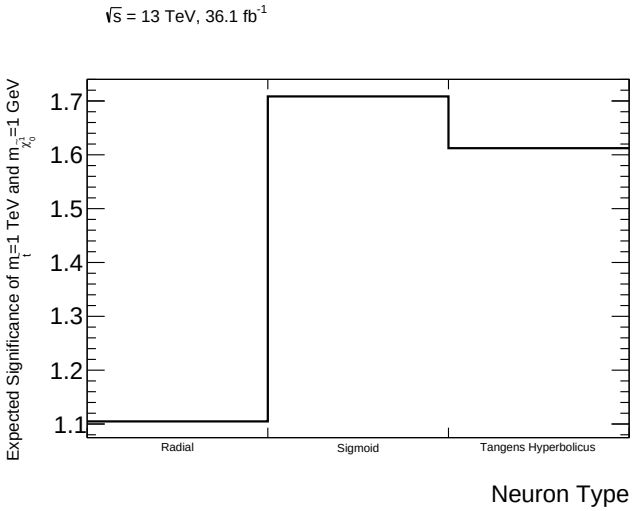


(b) Background Kolmogorov-Smirnov-Score.

Figure 17: Kolmogorov-Smirnov-Score as a function of the training method.

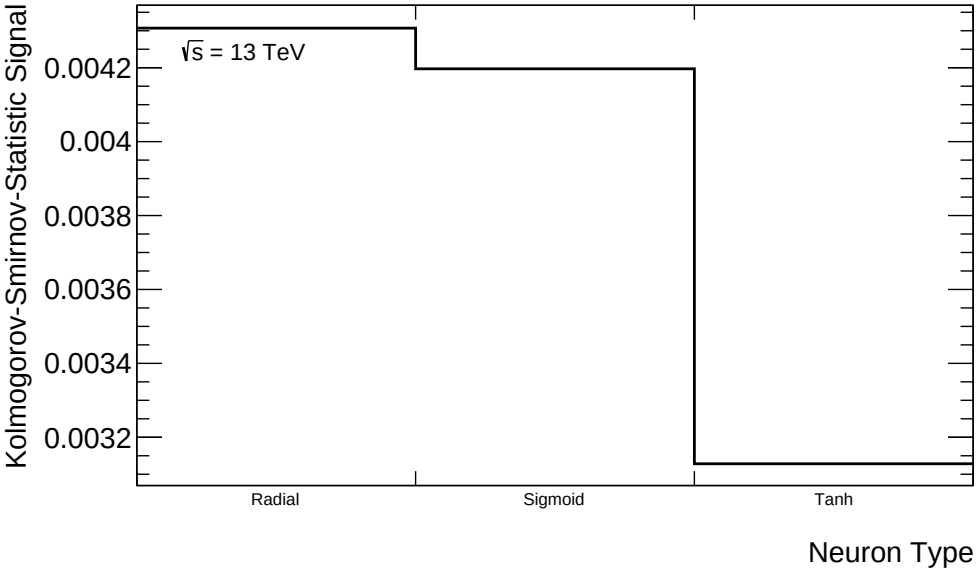


(a) Area under the ROC-Curves for both Training and Testing.

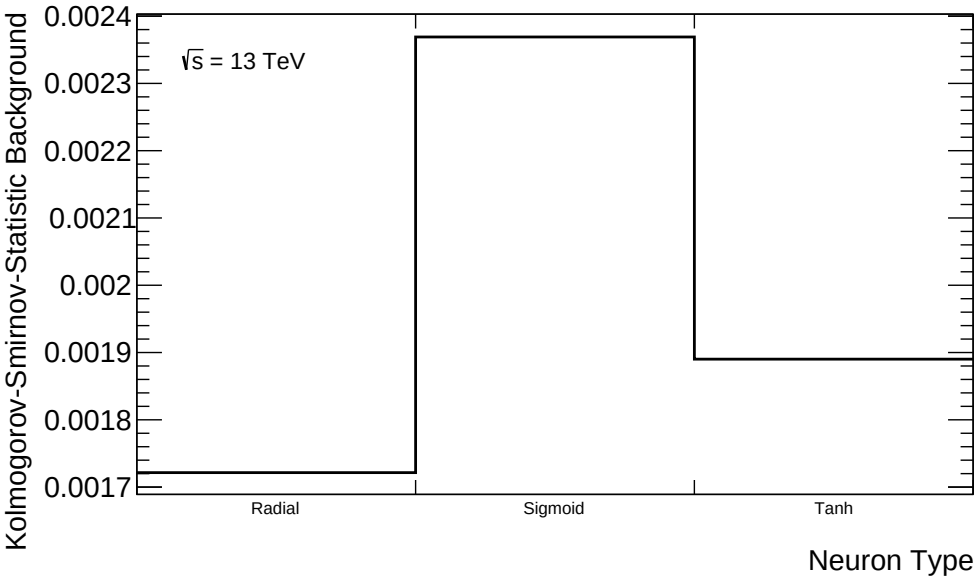


(b) Significance of the reference model for BDT trainings.

Figure 18: Area under the ROC-Curves and significance of the reference model as a function of the neuron type

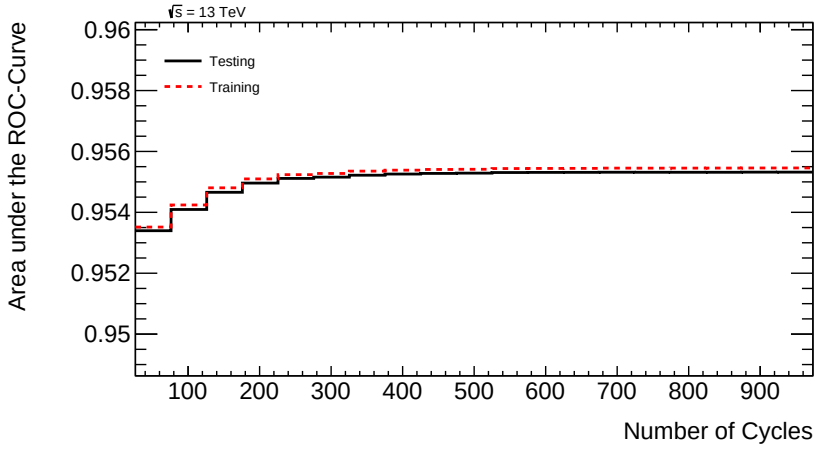


(a) Signal Kolmogorov-Smirnov-Score.

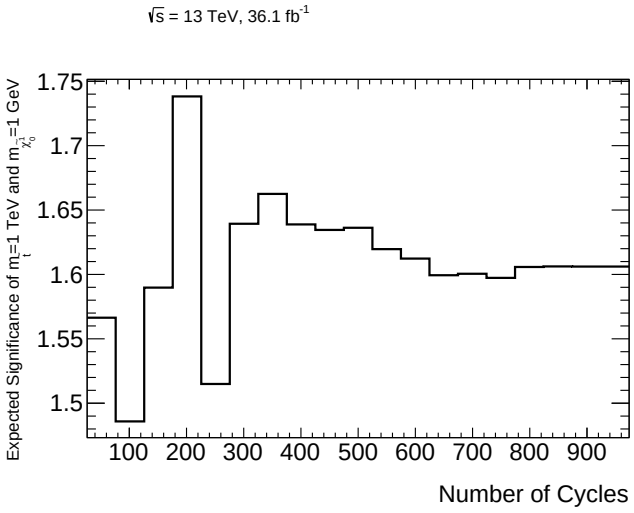


(b) Background Kolmogorov-Smirnov-Score.

Figure 19: Kolmogorov-Smirnov-Score as a function of the neuron type.

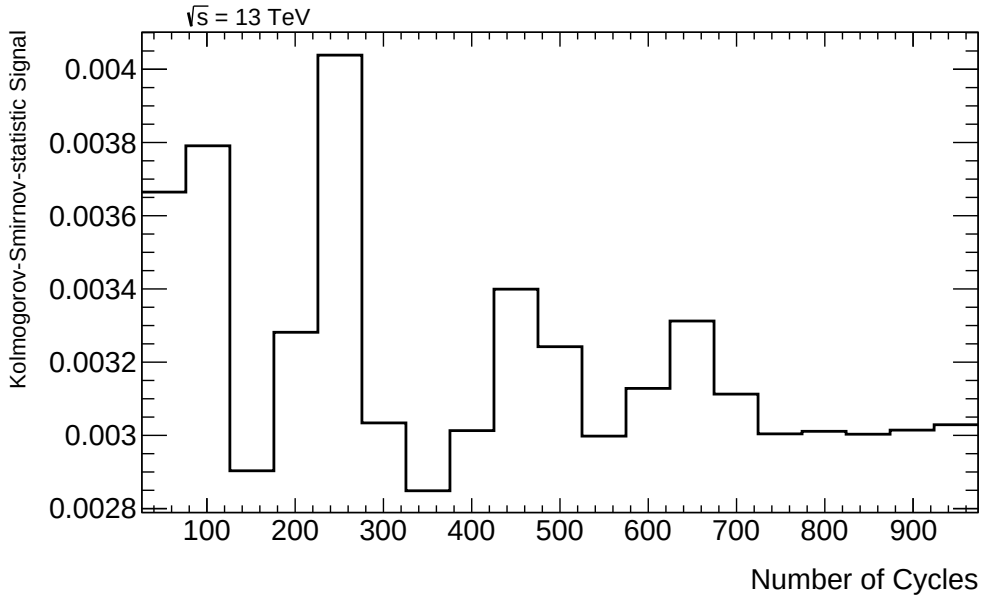


(a) Area under the ROC-Curves for both Training and Testing.

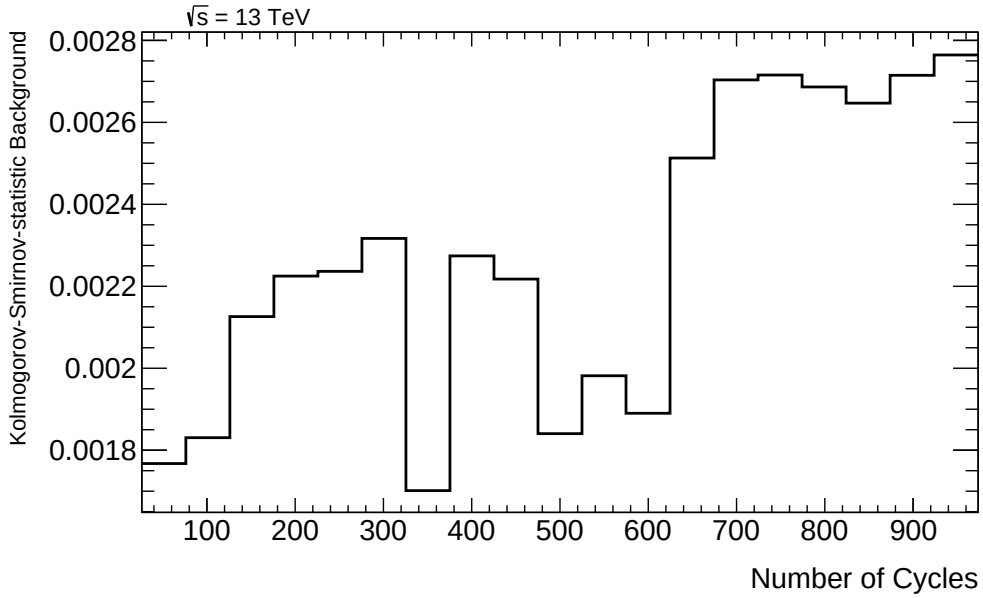


(b) Significance of the reference model for BDT trainings.

Figure 20: Area under the ROC-Curves and significance of the reference model as a function of the number of cycles.

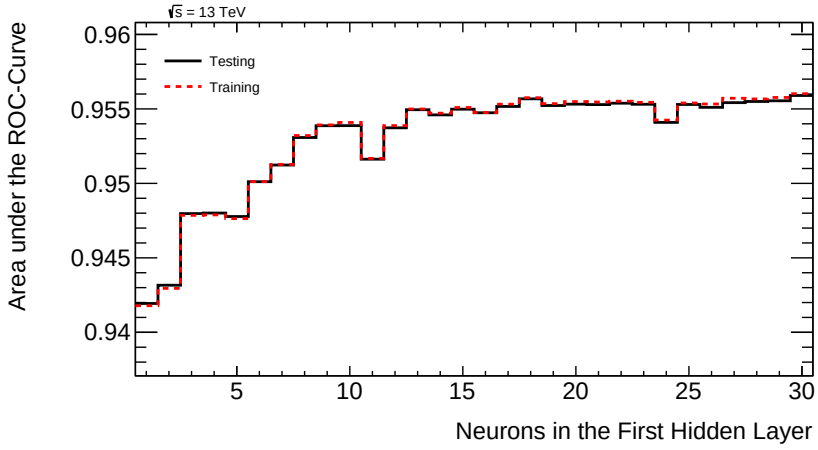


(a) Signal Kolmogorov-Smirnov-Score.

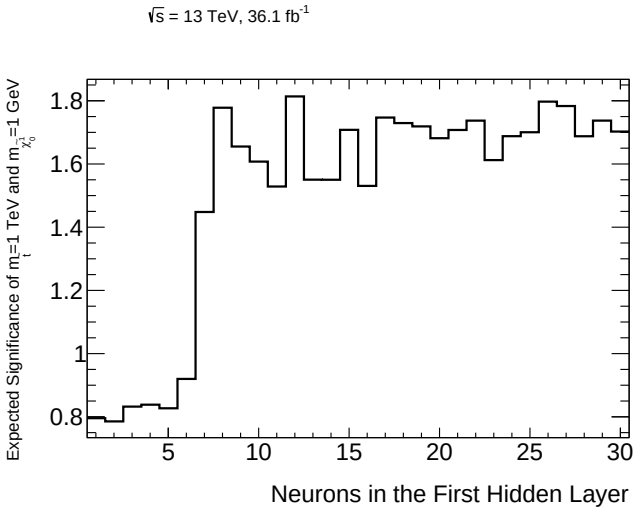


(b) Background Kolmogorov-Smirnov-Score.

Figure 21: Kolmogorov-Smirnov-Score as a function of the number of cycles.

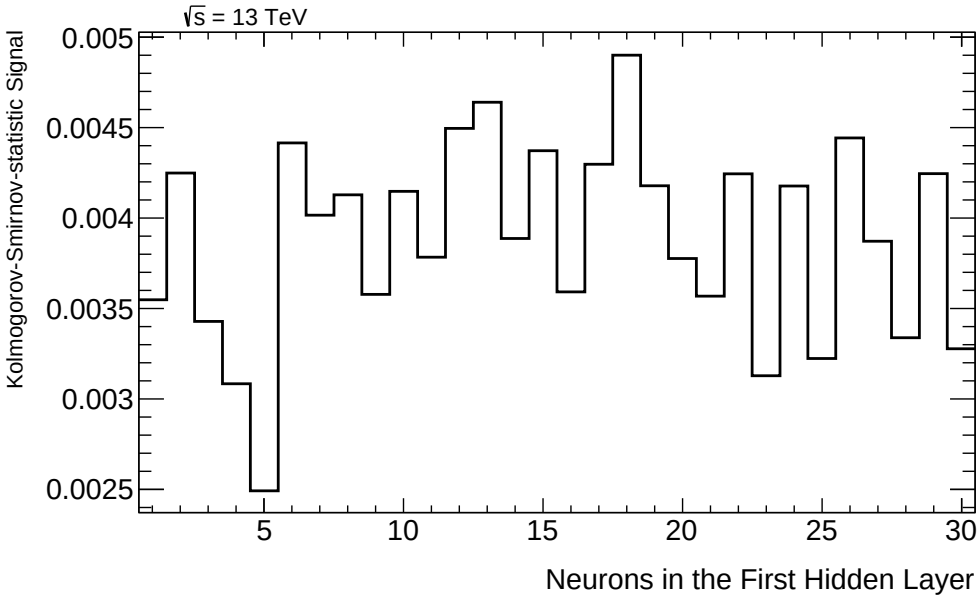


(a) Area under the ROC-Curves for both Training and Testing.

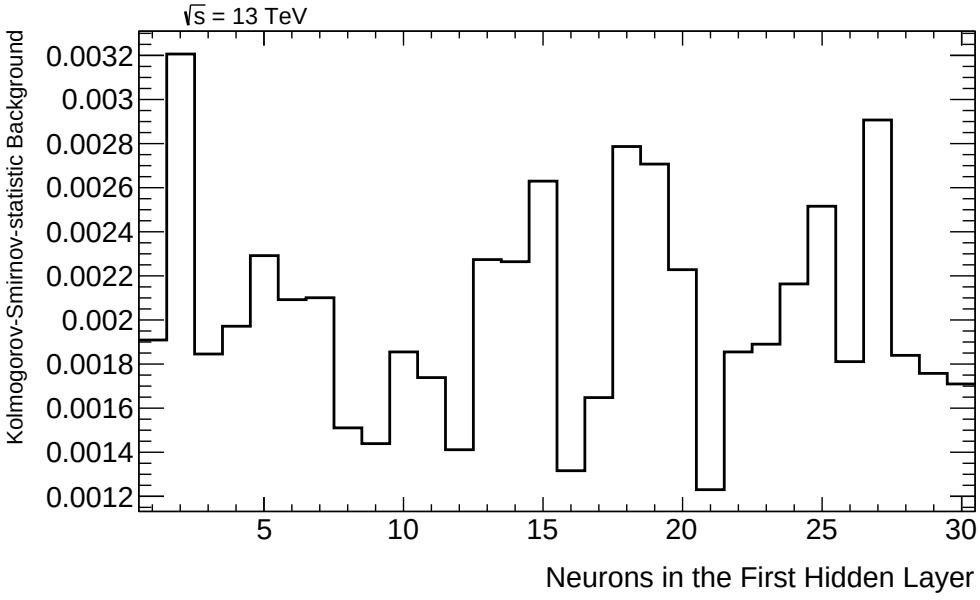


(b) Significance of the reference model for BDT trainings.

Figure 22: Area under the ROC-Curves and significance of the reference model as a function of the number of neurons in the first and only hidden layer.

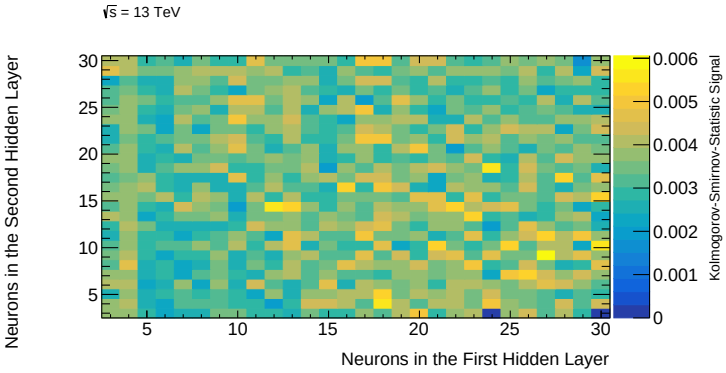


(a) Signal Kolmogorov-Smirnov-Score.

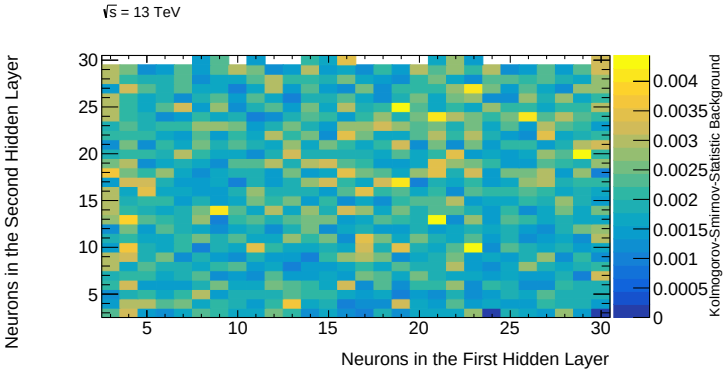


(b) Background Kolmogorov-Smirnov-Score.

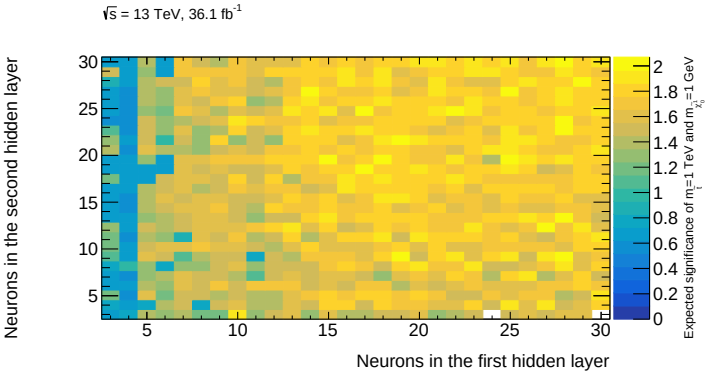
Figure 23: Kolmogorov-Smirnov-Score as a function of the number of neurons in the first and only hidden layer.



(a) Signal Kolmogorov-Smirnov-Score.

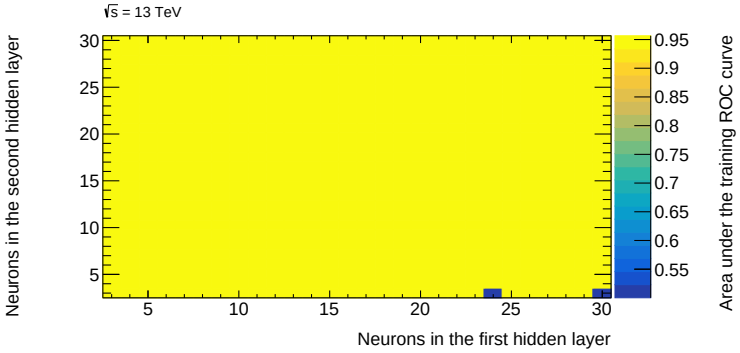


(b) Background Kolmogorov-Smirnov-Score.

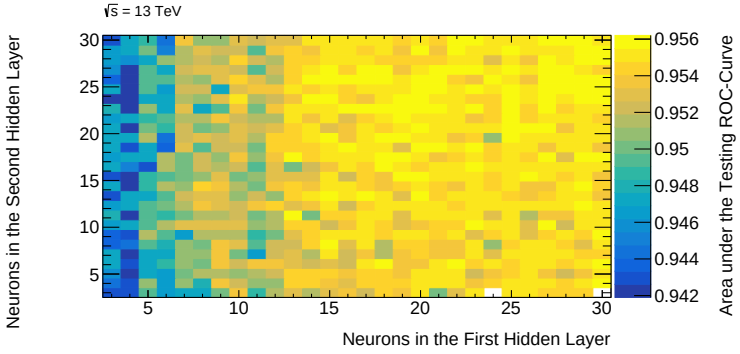


(c) Significance of the reference model for BDT trainings.

Figure 24: Kolmogorov-Smirnov-Score and significance of the reference model as a function of the number of neurons in the first and in the second hidden layer for a Back-Propagation neural network with two hidden layers.



(a) Area under the Training-ROC-Curves.



(b) Area under the Testing-ROC-Curves.

Figure 25: Area under the ROC-Curves in dependecy of the number of nodes in the first and in the second hidden layer for a Back-Propagation neural network with two hidden layers.

Bibliography

[1] ATLAS Collaboration, G. Aad et al., *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. **B716** (2012) 1–29, arXiv:1207.7214 [hep-ex].

[2] CMS Collaboration, S. Chatrchyan et al., *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. **B716** (2012) 30–61, arXiv:1207.7235 [hep-ex].

[3] F. Zwicky, *Die Rotverschiebung von extragalaktischen Nebeln*, Helv. Phys. Acta **6** (1933) 110–127. [Gen. Rel. Grav.41,207(2009)].

[4] Planck Collaboration, P. Ade et al., *Planck 2013 results. I. Overview of products and scientific results*, Astron. Astrophys. **571** (2014) A1, arXiv:1303.5062 [astro-ph.CO].

[5] S. Glashow, *Partial-symmetries of weak interactions*, Nuclear Physics **22** no. 4, (1961) 579 – 588.

[6] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19** (1967) 1264–1266.

[7] Wikipedia, *Standard Model*. https://en.wikipedia.org/wiki/Standard_Model.

[8] G. Kane, *Modern Elementary Particle Physics*. Cambridge University Press, 2017.

[9] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*, Phys. Lett. **12** (1964) 132–133.

[10] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (1964) 508–509.

[11] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (1964) 321–323.

[12] ATLAS Collaboration, M. Aaboud et al., *Measurement of inclusive and differential cross sections in the $H \to ZZ^* \to 4\ell$ decay channel at 13 TeV with the ATLAS detector*, Tech. Rep. ATLAS-CONF-2017-032, CERN, Geneva, May, 2017.

[13] ATLAS Collaboration, M. Aaboud et al., *Measurement of the Higgs boson coupling properties in the $H \to ZZ^* \to 4\ell$ decay channel at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Tech. Rep. ATLAS-CONF-2017-043, CERN, Geneva, Jul, 2017.

[14] ATLAS Collaboration, M. Aaboud et al., *Measurements of Higgs boson properties in the diphoton decay channel with 36.1 fb^{-1} pp collision data at the center-of-mass energy of 13 TeV with the ATLAS detector*, Tech. Rep. ATLAS-CONF-2017-045, CERN, Geneva, Jul, 2017.

[15] ATLAS Collaboration, M. Aaboud et al., *Measurement of the Higgs boson mass in the $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ channels with $\sqrt{s}$=13TeV pp collisions using the ATLAS detector*, Tech. Rep. ATLAS-CONF-2017-046, CERN, Geneva, Jul, 2017.

[16] ATLAS Collaboration, M. Aaboud et al., *Evidence for the $H \to b\bar{b}$ decay with the ATLAS detector*, `arXiv:1708.03299 [hep-ex]`.

[17] Super-Kamiokande Collaboration, Y. Fukuda et al., *Evidence for Oscillation of Atmospheric Neutrinos*, Phys. Rev. Lett. **81** (1998) 1562–1567.

[18] SNO Collaboration, Q. Ahmad et al., *Measurement of the Rate of $\nu_e + d \to p + p + e^-$ Interactions Produced by ^8B Solar Neutrinos at the Sudbury Neutrino Observatory*, Phys. Rev. Lett. **87** (2001) 071301.

[19] SNO Collaboration, Q. Ahmad et al., *Direct Evidence for Neutrino Flavor Transformation from Neutral-Current Interactions in the Sudbury Neutrino Observatory*, Phys. Rev. Lett. **89** (2002) 011301.

[20] A. Zee, *Quantum Field Theory in a Nutshell*. Nutshell handbook. Princeton Univ. Press, Princeton, NJ, 2003.

[21] N. Köhler, *Personal Communication*.

[22] J. Haffner, *The CERN accelerator complex. Complexe des accélérateurs du CERN*. Oct, 2013. General Photo.

[23] G. Aad et al., *The ATLAS Experiment at the CERN Large Hadron Collider*, Journal of Instrumentation **3** no. 08, (2008) S08003.

[24] J. Pequenao and P. Schaffner, "An computer generated image representing how ATLAS detects particles." Jan, 2013.

[25] ATLAS Collaboration, M. Aaboud et al., *Performance of the ATLAS Trigger System in 2015*, Eur. Phys. J. **C77** no. 5, (2017) 317, `arXiv:1611.09661 [hep-ex]`.

[26] ATLAS Collaboration, M. Aaboud et al., *Search for a Scalar Partner of the Top Quark in the Jets+E_T^{miss} Final State at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Tech. Rep. ATLAS-CONF-2017-020, CERN, Geneva, Apr, 2017.

[27] ATLAS Collaboration, G. Aad et al., *Reconstruction of hadronic decay products of tau leptons with the ATLAS experiment*, Eur. Phys. J. **C76** no. 5, (2016) 295, arXiv:1512.05955 [hep-ex].

[28] ATLAS Collaboration, G. Aad et al., *Performance of the ATLAS Inner Detector Track and Vertex Reconstruction in the High Pile-Up LHC Environment*,.

[29] M. "Aaboud and others ", *Vertex Reconstruction Performance of the ATLAS Detector at $\sqrt{s} = 13$ TeV*, Tech. Rep. ATL-PHYS-PUB-2015-026, CERN, Geneva, Jul, 2015.

[30] M. Cacciari, G. P. Salam, and G. Soyez, *The Anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063, arXiv:0802.1189 [hep-ph].

[31] W. Lampl et al., *Calorimeter Clustering Algorithms: Description and Performance*, Tech. Rep. ATL-LARG-PUB-2008-002. ATL-COM-LARG-2008-003, CERN, Geneva, Apr, 2008.

[32] M. Cacciari and G. P. Salam, *Pileup subtraction using jet areas*, Phys. Lett. **B659** (2008) 119–126, arXiv:0707.1378 [hep-ph].

[33] LHCb Collaboration, B. Pal, *Measurements of b hadron lifetimes and effective lifetimes at LHCb*, arXiv:1309.5911 [hep-ex].

[34] ATLAS, *b-tagging in dense environments*, Tech. Rep. ATL-PHYS-PUB-2014-014, CERN, Geneva, Aug, 2014.

[35] ATLAS Collaboration, M. Aaboud et al., *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*, Tech. Rep. ATL-PHYS-PUB-2016-012, CERN, Geneva, Jun, 2016.

[36] ATLAS Collaboration, G. Aad et al., *Electron efficiency measurements with the ATLAS detector using 2012 LHC proton-proton collision data*, Eur. Phys. J. **C77** no. 3, (2017) 195, arXiv:1612.01456 [hep-ex].

[37] D0 Collaboration, V. Abazov et al., *Observation of Single Top Quark Production*, Phys. Rev. Lett. **103** (2009) 092001, arXiv:0903.0850 [hep-ex].

[38] CDF Collaboration, T. Aaltonen et al., *First Observation of Electroweak Single Top Quark Production*, Phys. Rev. Lett. **103** (2009) 092002, arXiv:0903.0885 [hep-ex].

[39] C. G. Lester and D. J. Summers, *Measuring masses of semiinvisibly decaying particles pair produced at hadron colliders*, Phys. Lett. **B463** (1999) 99–103, arXiv:hep-ph/9906349 [hep-ph].

[40] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv:1405.0301 [hep-ph].

[41] T. Sjöstrand, S. Mrenna, and P. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** no. arXiv:0710.3820. CERN-LCGAPP-2007-04. LU TP 07-28. FERMILAB-PUB-07-512-CD-T, (2007) 852–867. 27 p.

[42] D. J. Lange, *The EvtGen particle decay simulation package*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **462** no. 1, (2001) 152 – 155. BEAUTY2000, Proceedings of the 7th Int. Conf. on B-Physics at Hadron Machines.

[43] L. Lönnblad and S. Prestel, *Merging Multi-leg NLO Matrix Elements with Parton Showers*, JHEP **03** (2013) 166, arXiv:1211.7278 [hep-ph].

[44] R. Ball et al., *Parton distributions with LHC data*, Nuclear Physics B **867** no. 2, (2013) 244 – 289.

[45] ATLAS Collaboration, G. "Aad and others", *ATLAS Run 1 Pythia8 tunes*, Tech. Rep. ATL-PHYS-PUB-2014-021, CERN, Geneva, Nov, 2014.

[46] W. Beenakker et al., *Stop production at hadron colliders*, Nuclear Physics B **515** no. 1, (1998) 3 – 14.

[47] W. Beenakker et al., *Squark and Gluino Hadroproduction*, International Journal of Modern Physics A **26** no. 16, (2011) 2637–2664, arXiv:1105.1110 [hep-ph].

[48] W. Beenakker et al., *Supersymmetric top and bottom squark production at hadron colliders*, JHEP **08** (2010) 098, arXiv:1006.4771 [hep-ph].

[49] C. Borschensky et al., *Squark and gluino production cross sections in pp collisions at $\sqrt{s}$ = 13, 14, 33 and 100 TeV*, Eur. Phys. J. **C74** no. 12, (2014) 3174, arXiv:1407.5066 [hep-ph].

[50] S. Alioli et al., *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **06** (2010) 043, arXiv:1002.2581 [hep-ph].

[51] T. Sjostrand, S. Mrenna, and P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, JHEP **05** (2006) 026, arXiv:hep-ph/0603175 [hep-ph].

[52] H.-L. Lai et al., *New parton distributions for collider physics*, Phys. Rev. **D82** (2010) 074024, arXiv:1007.2241 [hep-ph].

[53] P. Z. Skands, *Tuning Monte Carlo Generators: The Perugia Tunes*, Phys. Rev. **D82** (2010) 074018, arXiv:1005.3457 [hep-ph].

[54] T. Gleisberg et al., *Event generation with SHERPA 1.1*, Journal of High Energy Physics **2009** no. 02, (2009) 007.

[55] NNPDF Collaboration, R. Ball et al., *Parton distributions for the LHC Run II*, JHEP **04** (2015) 040, arXiv:1410.8849 [hep-ph].

[56] ATLAS Collaboration, G. Aad et al., *The ATLAS Simulation Infrastructure*, Eur. Phys. J. **C70** (2010) 823–874, arXiv:1005.4568 [physics.ins-det].

[57] S. Agostinelli et al., *Geant4—a simulation toolkit*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **506** no. 3, (2003) 250 – 303.

[58] ATLAS Collaboration, M. Beckingham et al., *The simulation principle and performance of the ATLAS fast calorimeter simulation FastCaloSim*, Tech. Rep. ATL-PHYS-PUB-2010-013, CERN, Geneva, Oct, 2010.

[59] A. Martin et al., *Parton distributions for the LHC*, Eur. Phys. J. **C63** (2009) 189–285, arXiv:0901.0002 [hep-ph].

[60] ATLAS Collaboration, G. "Aad and others", *Summary of ATLAS Pythia 8 tunes*, Tech. Rep. ATL-PHYS-PUB-2012-003, CERN, Geneva, Aug, 2012.

[61] A. Höcker et al., *TMVA: Toolkit for Multivariate Data Analysis*, PoS **ACAT** (2007) 040, arXiv:physics/0703039.

[62] R. E. Schapire, *The strength of weak learnability*, Machine Learning **5** no. 2, (1990) 197–227.

[63] Y. Freund and R. E. Schapire, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, Journal of Computer and System Sciences **55** no. 1, (1997) 119 – 139.

[64] F. Rosenblatt and C. A. Laboratory, *The perceptron: a theory of statistical separability in cognitive systems (Project Para)*. Report no. VG-1196-G-1. Cornell Aeronautical Laboratory, 1958.

[65] W. H. Delashmit and M. T. Manry, *Recent Developments in Multilayer Perceptron Neural Networks*, in *Proceedings of the 7th Annual Memphis Area Engineering and Science Conference, MAESC*. 2005.

[66] K. Weierstraß, *Über die analytische Darstellbarkeit sogenannter willkürlicher*

Funktionen einer reellen Veränderlichen., Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin. (1885).

[67] N. E. Cotter, *The Stone-Weierstrass Theorem and Its Application to Neural Networks*, Trans. Neur. Netw. **1** no. 4, (1990) 290–295.

[68] R. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series)*. John Wiley & Sons, 1999.

[69] T. Fawcett, *An Introduction to ROC Analysis*, Pattern Recogn. Lett. **27** no. 8, (2006) 861–874.

[70] J. Linnemann, *"Measures of Significance in HEP and Astrophysics"*, `physics/0312059`.

[71] ATLAS Collaboration, M. Aaboud et al., *Top-quark mass measurement in the all-hadronic $t\bar{t}$ decay channel at $\sqrt{s}$ = 8 TeV with the ATLAS detector*, `arXiv:1702.07546 [hep-ex]`.

[72] ATLAS Collaboration, M. Aaboud et al., *Measurement of the W-boson mass in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, `arXiv:1701.07240 [hep-ex]`.

[73] A. X. El-Khadra and M. Luke, *The Mass of the b quark*, Ann. Rev. Nucl. Part. Sci. **52** (2002) 201–251, `arXiv:hep-ph/0208114 [hep-ph]`.

[74] P. Jackson, C. Rogan, and M. Santoni, *Sparticles in motion: Analyzing compressed SUSY scenarios with a new method of event reconstruction*, Phys. Rev. **D95** no. 3, (2017) 035031, `arXiv:1607.08307 [hep-ph]`.

[75] J. H. Friedman, *Greedy function approximation: A gradient boosting machine.*, Ann. Statist. **29** no. 5, (2001) 1189–1232.

List of Figures

List of Tables