Multivariate Analysetechniken zur Supersymmetriesuche am LHC

Bachelor-Arbeit

vorgelegt am: 13. September 2016

am Institut für Experimentalphysik an der Universität Hamburg

Name:	Phuoc Thien Le
Matrikelnummer:	6548524
Studiengang:	Physik
Erstgutachter:	Prof. Dr. Peter Schleper
Zweitgutachter:	Dr. Jory Sonneveld

Zusammenfassung

Diese Arbeit befasst sich mit einer Analyse des Zerfalls supersymmetrischer Teilchen, die am LHC Beschleuniger aus Gluon-Paarbildungen hervorgehen. Da die Wahrscheinlichkeit dieser Zerfallsprozesse von vielen Observablen abhängig ist, wird in dieser Arbeit eine multivariate Analyse entwickelt. Diese Arbeit knüpft an die Arbeit einer Analysegruppe der CMS Kollaboration an, welche für rein hadronische Endzustände Observablen der Jets und den fehlenden transversalen Impuls als wichtigste Observable für die Analyse verwendet. Die Optimierung der multivariaten Analyse stellt die Schwierigkeit der Analyse dar. Erst durch eine sorgfältige Wahl der verwendeten Observablen, sowie die Auswahl besonders effizienter Algorithmusparameter in der multivariaten Analyse, lässt sich eine Reduktion des Standardmodell Untergrundes gewährleisten und somit mögliche Supersymmetriezerfälle ausfindig machen.

Abstract

This thesis deals with the analysis of a the decay of supersymmetric particles, which is produced by a gluon pair production. Since the probability of these processes depend on many observables, a multivariate analysis will be developed in this thesis. This thesis is build on the work of the CMS collaboration, which uses for purely hadronicaly final states observables of the jets and the missing transverse momentum as the most important observables. The main difficulty of the analysis is the optimization of the multivariate analysis. Only with the right choice of observables and efficient algorithm parameters of the multivariate analysis a reduction of the standardmodel background is to be ensured and the supersymmetry process is detectable.

Inhaltsverzeichnis

1 Einleitung **1**

2 Grundlagen **3**

2.1 Theoretische Grundlagen . . . 3
2.1.1 Konventionen . . . 3
2.1.2 Standardmodell . . . 3
2.1.3 Kritik am Standardmodell . . . 4
2.1.4 Supersymmetrie . . . 5
2.2 Der LHC und das CMS-Experiment . . . 7
2.2.1 Der Large-Hadron-Collider . . . 7
2.2.2 Compact Muon Solenoid . . . 8
2.2.3 Innerer Spurdetektor . . . 9
2.2.4 Elektromagnetisches Kalorimeter . . . 10
2.2.5 Hadronisches Kalorimeter . . . 10
2.2.6 Solenoid-Magnet . . . 10
2.2.7 Myonenspektrometer . . . 10
2.3 Das Koordinatensystem und die Pseudorapidität η . . . 11
2.4 Kinematische Größen . . . 11
2.5 Signaturen am LHC . . . 12

3 Multivariate Analyse mit TMVA **15**

3.1 Kenngrößen . . . 15
3.2 Receiver-Operating-Characteristic Curve (ROC-Kurve) . . . 16
3.3 Verschiedene Methoden in TMVA . . . 17
3.4 Boosted Decision Trees . . . 18
3.4.1 Boosting . . . 19
3.4.2 BDT-Steuerparameter . . . 19

4 Analyse **20**

4.1 Gewichtung und Monte-Carlo-Simulationen . . . 21
4.2 Vorselektionen . . . 24
4.3 Optimierung der verwendeten Variablen . . . 26
4.4 Vergleich BDT mit anderen Methoden . . . 31
4.5 Optimierung des Trainings . . . 32
4.6 BDT-Ranking . . . 36

5 Resultate **37**

5.1 Daten . . . 37
5.2 Umgewichtung des BDT-Response . . . 38

5.3 Schnitt auf die BDT-Response 40
5.4 Vergleich mit der MHT-Analyse 43
5.5 Limitsberechnung 48

6 Zusammenfassung **50**

7 Ausblick **52**

1 Einleitung

Die Zusammensetzung des Universums war schon immer eine der großen Fragen der Menschheit. Das größte Ergebnis dieser Forschung ist das sogenannte Standardmodell. Dieses beschreibt alle bisher bekannten Elementarteilchen und deren Wechselwirkungen (schwache Wechselwirkung, starke Wechselwirkung und elektromagnetische Wechselwirkung) untereinander. Bestätigt wurde dieses Modell durch den Bau großer Teilchenbeschleuniger, wie dem Large Hadron Collider (LHC) in Europa. Mit Experimenten an diesen Teilchenbeschleunigern konnten mit einer hohen Genauigkeit Vorhersagen, die durch das Standardmodell hervorgingen, überprüft werden. Erst im Jahre 2012 wurde am LHC am CERN das Higgs-Boson entdeckt, was so dann das Standardmodell komplettiert hat. Jedoch stößt man in der Teilchenphysik auf Probleme, die durch das bisher bekannte Standardmodell noch nicht gelöst werden konnten. Beispielsweise kann die Dunkle Materie, deren Existenz in der Physik jedoch als sicher gilt, mit dem Standardmodell nicht erklärt werden. Zudem ist in dem Standardmodell die Gravitation noch nicht berücksichtigt worden, da diese im Vergleich zu den anderen fundamentalen Kräften sehr schwach ist. Daraus ergibt sich die Motivation zu neuen Modellen, die diese Probleme erklären sollen. Ein Modell, die diese Probleme erklären würde, ist die sogenannte Supersymmetrie [1]. Diese beschreibt eine weitere Symmetrie zwischen den Spins der Teilchen, woraus für jedes Teilchen im Standardmodell ein Supersymmetriepartner resultieren würde. Das Compact-Muon-Solenoid-Experiment (CMS) hat unter anderem die Aufgabe, diese Supersymmetrieteilchen zu finden. Das Modell der Supersymmetrie motiviert zu verschiedene Analysen der am LHC aufgenommenen Daten. Eine Analysetechnik ist dabei die sogenannte multivariate Analyse. Da Supersymmetrieprozesse von vielen Observablen abhängig sind, bietet sich eine multivariate Analyse an, der die verschiedenen Observablen in einem vieldimensionalen Phasenraum zusammen betrachtet und so Korrelationen zwischen diesen berücksichtigt. Das Problem bei der Suche nach Supersymmetrieteilchen in den aufgenommenen Daten ist, dass die meisten Ereignisse, die gemessen worden sind, nicht einem Supersymmetrieprozess entsprechen. Diese sogenannten Untergrundprozesse (Standardmodellprozesse) gilt es zu unterdrücken, damit die Ereignisse mit Supersymmetrieteilchen - in dieser Arbeit auch Signalevents genannt - in den Daten gefunden werden können. In der multivariaten Analyse wird ein Algorithmus trainiert, indem Events mit bekannter Klassifikation verwendet werden. Die resultierende Analyse gilt es auf die Daten anzuwenden, um die erwünschten Signalevents ausfindig zu machen. In Kapitel 3 geht es daher um die Optimierung des Trainingsprozesses. Im diesen Optimierungsprozess wird die Auswahl der Observablen und die Auswahl des Algorithmus getroffen, die das Training beeinflussen. In Kapitel 4 werden die Resultate des Trainings diskutiert und diese

mit einer vorhandenen Analyse verglichen. Insbesondere werden obere Grenzen für die Anzahl der Signalevents und die obere Grenze für den Wirkungsquerschnitt am LHC mit einer Schwerpunktsenergie von 13 TeV berechnet.

2 Grundlagen

Um die in dieser Arbeit durchgeführte Analyse verstehen zu können, werden in diesem Kapitel die benötigten Grundlagen geschafft. Diese Grundlagen sind in theoretische Grundlagen und experimenteller Aufbau unterteilt.

2.1 Theoretische Grundlagen

In diesem Kapitel werden die theoretischen Grundlagen dieser Arbeit beschrieben. Grundlage ist das Standardmodell, welches durch die Supersymmetrie erweitert wird und so viele physikalische Phänomene erklären könnte.

2.1.1 Konventionen

In der Hochenergiephysik ist es üblich physikalische Größen in natürlichen Einheiten anzugeben. Natürliche Einheiten vereinfachen in der Hochenergiephysik viele Notationen. In natürlichen Einheiten wird unter anderem das Plank'sche Wirkungsquantum $\hbar$ und die Lichtgeschwindigkeit c einheitenlos auf 1 gesetzt. Daher ändern sich auch die Einheiten aller physikalischen Observablen in Energiepotenzen. So haben Energie, Impuls und Masse zum Beispiel die Einheit $[E]$, Länge und Zeit die Einheit $[E]^{-1}$.

2.1.2 Standardmodell

Das Standardmodell der Teilchenphysik beschreibt die heutigen Erkenntnisse der Elementarteilchenphysik [2]. Darin enthalten sind alle nachgewiesenen Teilchen und deren Wechselwirkungen untereinander. Das Standardmodell erklärte bisher viele physikalische Phänomene und sagte insbesondere Teilchen voraus, die später nachgewiesen werden konnten, wie z.B. das Top-Quark. Jedoch kann das Standardmodell nicht jedes physikalische Phänomen erklären, was im Folgenden näher erläutert wird. In Abbildung 1 sind alle Teilchen des Standardmodells mit ihrer jeweiligen Masse, Spin und Ladung aufgelistet.

Unterteilt sind die Teilchen durch ihren Spin. Teilchen mit halbzahligen Spin sind Fermionen. Diese Teilchen sind Grundbausteine der bisher bekannten Materie. Im Gegensatz dazu werden Teilchen mit ganzzahligen Spin als Bosonen bezeichnet. Diese Teilchen wiederum sind für die Wechselwirkungen zwischen den Teilchen verantwortlich, die vor allem bei Zerfällen eine wichtige Rolle spielen, mit Ausnahme des Higgs-Bosons, dessen Vakuumerwartungswert für die Masse der Elementarteilchen verantwortlich ist. Die Materieteilchen sind zudem in drei Familien aufgeteilt. Die

erste Familie der Elementarteilchen sind Grundlage aller Atome des Periodensystems der Elemente, da Protonen und Neutronen aus Up-Quarks und Down-Quarks bestehen. Elektronen sind selber Elementarteilchen der ersten Familie. Materieteilchen der zweiten Familie und dritten Familie unterscheiden sich hauptsächlich in der größeren Masse der Teilchen der ersten Familie und sind daher auch instabil

Die fundamentalen Kräfte, die bisher beobachtet wurden, sind die schwache und die starke Wechselwirkung, die Gravitation und die elektromagnetische Kraft. Aufgrund der geringen Reichweite der schwachen und starken Wechselwirkung sind diese nur in subatomaren Größenordnungen beobachtbar. Die Gravitation ist jedoch die einzige fundamentale Kraft, die nicht im Standardmodell berücksichtigt ist. Die übrigen Wechselwirkungen werden von sogenannten Austauschteilchen getragen. Die schwache Wechselwirkung werden von W-Bosonen und Z-Bosonen, die starke Wechselwirkung von Gluonen und die elektromagnetische Wechselwirkung von Photonen getragen.

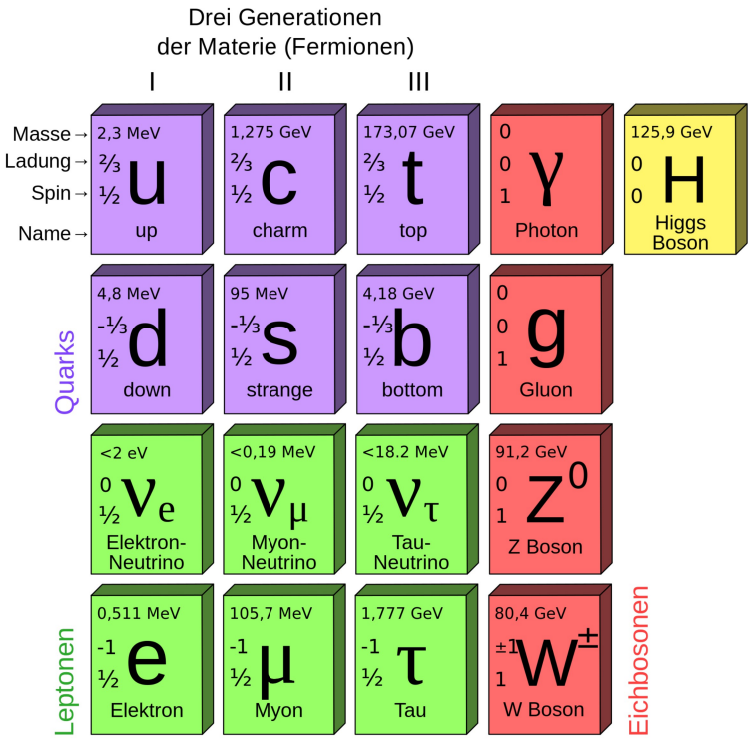


Abbildung 1: Standardmodell der Elementarteilchen [3].

2.1.3 Kritik am Standardmodell

Obwohl das Standard-Modell die meisten physikalischen Beobachtungen erklären kann, ergeben sich mit dem Standardmodell Probleme, die auf ein weiterführendes

Modell hindeuten [4]. Die Existenz von dunkler Materie lässt sich beispielsweise nicht mit dem Standardmodell erklären. Auch das Berücksichtigen der Gravitation fehlt im Standardmodell. Zudem ergeben sich auch Probleme der theoretischen Physik mit dem Standardmodell. Zum einen das Hierachieproblem, das auf den großen Unterschied in den Größenordnungen der fundamentalen Kräfte zurückgeht. Die elektromagnetische Wechselwirkung ist 10^{11} Mal so groß wie die schwache Wechselwirkung und die starke Wechselwirkung ist 10^{3} Mal so groß. Zum anderen ergeben sich mit dem Standardmodell freie Parameter. Diese 19 freien Parameter sind zwar experimentell bestimmt, jedoch ist deren physikalischer Ursprung unklar. Auch die experimentell gefundenen Neutrinomassen waren in der ersten Formulierung des Standardmodells vorhanden. Diese genannten Probleme motivieren neue Modelle der Teilchenphysik, welche im Grenzfall mit dem Standardmodell vereinbar sein sollte.

2.1.4 Supersymmetrie

Ein Modell, welches viele Probleme des Standardmodells lösen könnte, ist die sogenannte Supersymmetrie [4]. Diese beinhaltet eine weitere Symmetrie zwischen Fermionen und Bosonen im Standardmodell. So ergeben sich für alle Teilchen des Standardsmodells sogenannte Superpartner, welche sich im Idealfall ausschließlich in der Spinquantenzahl von den jeweiligen Standardmodellteilchen unterscheiden. Um eine einfache Kommunikation über die Supersymmetrieteilchen zu gewährleisten, gibt es Konventionen zur Bezeichnung dieser Teilchen. Supersymmetrieteilchen erhalten so ein Tilde ($\sim$) in der Bezeichnung. Zudem werden Supersymmetriepartner der Quarks und der Leptonen Squarks und Sleptonen genannt. Die restlichen Supersymmetrieteilchen erhalten eine ino-Endung. Der Supersymmetriepartner des Gluons heißt beispielsweise Gluino[1].

Eine ideale Symmetrie zwischen Fermionen und Bosonen kann jedoch nicht gegeben sein, da bisher kein Selektron entdeckt wurde, was mit der bisher in Beschleunigern verwendete Schwerpunktenergie erzeugt werden könnte. Daher spricht man von einer Symmetriebrechung, woraus folgt, dass die Massen der Supersymmetriepartner größer sein dürfen als die Masse der Standardmodell-Teilchen, was das Nichtentdecken des Selektrons erklären würde. Aus dieser Symmetriebrechung ergeben sich auch mögliche Mischzustände der Supersymmetriepartner. Beispielsweise spricht man bei einem Mischzustand von Spin1/2-Supersymmetriepartnern die elektrisch- und farbeneutral sind[2] von einem *Neutralino*. Ein Modell, welches

[1] Allgemein gilt: Fermion $\leftrightarrow$ Sfermion
Boson $\leftrightarrow$ Bosino.

[2] In dem minimalen supersymmetrischen Standardmodell sind auch Charginos (geladen) und Neu-

das Standardmodell mit den Theorien der Supersymmetrie vereinbart ist das **minimale supersymmetrische Standardmodell** (**MSSM**). Dabei handelt es sich um die minimalste Erweiterung des Standardmodells, welches zudem mit der Supersymmetrie vereinbar ist. In diesem Modell ergeben sich statt 19 freie Parameter[3] im Standardmodell 178 freie Parameter [5]. Einige im MSSM möglichen Wechselwirkungsprozesse würden einen spontanen Zerfall von Protonen ermöglichen. Da dies noch nicht beobachtet wurde, wird die R-Parität definiert. Diese besagt die Erhaltung einer neuen Quantenzahl P_R (R-Parität):

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

wobei B die Baryonenzahl, L die Leptonenzahl und s der Teilchenspin angibt.

Aus dieser Erhaltungsgröße ergeben sich einige Konsequenzen. Zum einem können SUSY-Teilchen nur paarweise ohne Mischung mit Standardmodell-Teilchen erzeugt werden. Zum anderen ergibt sich aus der R-Paritäts-Erhaltung das **LSP (Lightest Supersymmetric Particle)**, welches demnach stabil sein muss. Aus experimentellen und theoretischen Ergebnissen lässt sich die Masse des LSP einschränken, welches dadurch ein guter Kandidat für die dunkle Materie darstellen könnte.
Was außerdem die Supersymmetrie erklärt ist die Vereinigung der drei Kräfte. Im Standardmodell sind lediglich die elektromagnetische Wechselwirkung und die schwache Kraft vereinigt zu der elektroschwachen Wechselwirkung. Mit der Supersymmetrie lässt sich zudem die starke Wechselwirkung mit vereinigen, was auf ein gemeinsamen Ursprung dieser Kräfte deuten würde [1]. In Abbildung 2 ist die Vereinigung der drei Kräfte unter Berücksichtigung der Supersymmetrie dargestellt.

tralinos (ungeladen) möglich. (Mischzustände aus Supersymmetriepartern)
[3] 19 freie Parameter ergeben sich bei einer Nicht-Berücksichtigung der Neutrinomassen.

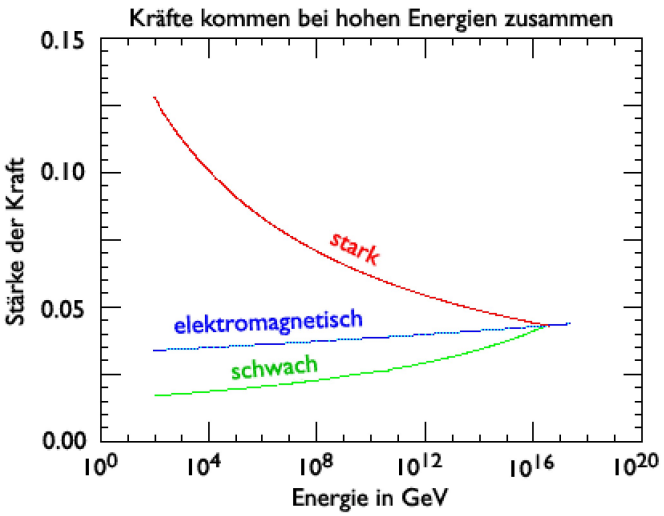


Abbildung 2: Vereinigung der drei fundamentalen Kräfte unter Berücksichtigung der Supersymmetrie bei hohen Energien [6].

2.2 Der LHC und das CMS-Experiment

Um Supersymmetrieteilchen detektieren zu können, werden Teilchenbeschleuniger benötigt, die den Supersymmetrieprozess aufgrund von Kollisionen von Teilchen hervorrufen sollen. Daher wird im folgenden Kapitel der experimentelle Aufbau des Kreisbeschleunigers **LHC (Large Hadron Collider)** beschrieben, welcher sich unter anderem mit der Suche von Supersymmetriepartnern beschäftigt. Genauer wird das CMS-Experiment des LHC's erläutert, welche diese Arbeit als Grundlage hat.

2.2.1 Der Large-Hadron-Collider

Der Large Hadron Collider (LHC) ist der weltweit größte Teilchenbeschleuniger [7]. Er befindet sich sich in einem Tunnel, der zwischen 50 und 175 Metern unter der Erdoberfläche am europäischen Teilchenphysikzentrum CERN in der Nähe von Genf liegt. Dieser ist ein 27 km langer Kreisbeschleuniger, der mit einer Schwerpunktsenergie von 13 TeV für Protonen und einer Luminosität von $10^{34}\text{cm}^{-2}\text{s}^{-14}$ der bislang leistungsstärkste Teilchenbeschleuniger ist [8]. Die Schwerpunktsenergie ist dabei die Energie am Kollisionsschwerpunkt des Beschleunigers. Die Luminosität gibt die Anzahl der Teilchenkollisionen pro Zeit und Fläche an und ist definiert als:

$$L = \frac{n \cdot N_1 \cdot N_2 \cdot f}{A}, \tag{2}$$

[4] Am 03.09 wurde ein Spitzenwert für die Luminosität von $1,33 \cdot 10^{34} \frac{1}{\text{s} \cdot \text{cm}^2}$ gemessen.

wobei N_1 und N_2 die jeweilige Anzahl der aufeinander treffenden Teilchenpakete, n die Anzahl der Pakete, f die Kollisionsfrequenz und A die Querschnittsfläche der Pakete ist. Die Gesamtzahl der Kollisionen in einem bestimmten Zeitraum $[t_1, t_2]$ beschreibt sich durch die integrierte Luminosität:

$$\mathcal{L} = \int_{t_1}^{t_2} L. \tag{3}$$

Diese wird in der Teilchenphysik üblicherweise in inverse barn (b^{-1}) angegeben[5].

Am LHC werden mehrere Experimente durchgeführt. Die vier größten Experimente sind CMS, ATLAS, ALICE und LHC-B (siehe Abbildung 3).

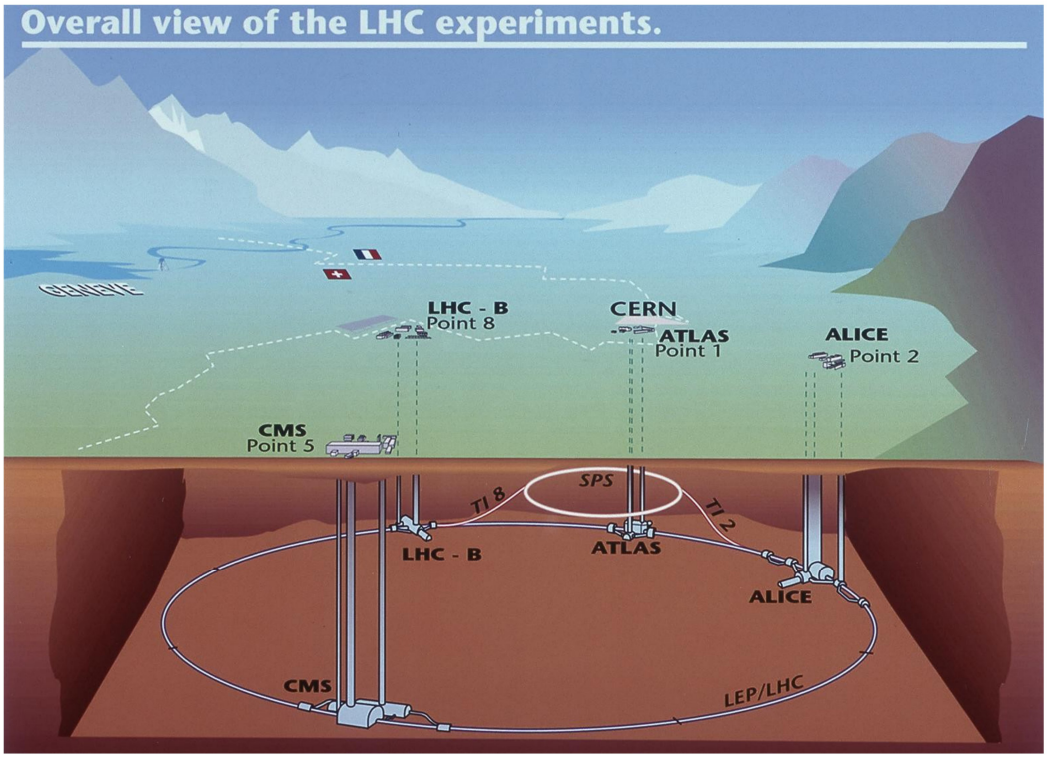


Abbildung 3: Eine schematische Skizze vom LHC mit den größten Experimenten CMS, ATLAS, ALICE, LHC-B [9].

2.2.2 Compact Muon Solenoid

Ein Experiment, welches beispielsweise eine Detektion von Higgs-Teilchen und SUSY Teilchen am LHC zur Aufgabe hat, ist **CMS (Compact Muon Solenoid)**.

[5]Das Barn b (engl. Scheune) ist eine Maßeinheit zur Angabe von Wirkungsquerschnitten in der Teilchenphysik. Es gilt:

$$1\ \mathrm{b} = 10^{-28}\ \mathrm{m}^2\ . \tag{4}$$

Das CMS Experiment hat es zudem als Aufgabe, Wissen zu erlangen zu neuen Phänomenen wie zum Beispiel Micro Black Holes und neue Zustände von sehr heißer dichter Materie. Diese Ergebnisse sollen einen weiteren Einblick und Verständnis zum Universum direkt nach dem Urknall liefern.

Der CMS Detektor ist schematisch in Abbildung 4 dargestellt. Der Detektor ist zylinderförmig konstruiert mit einer Länge von 21m und einem Durchmesser von 15m [10]. Das besondere am CMS ist, dass der Detektor um einen Solenoidmagneten gebaut ist, der ein Magnetfeld von $3,8$ T erzeugt[6]. Dieses Magnetfeld wird von einem Stahljoch beschränkt, welcher hauptsächlich das Gewicht des Detektors von 14000t ausmacht.

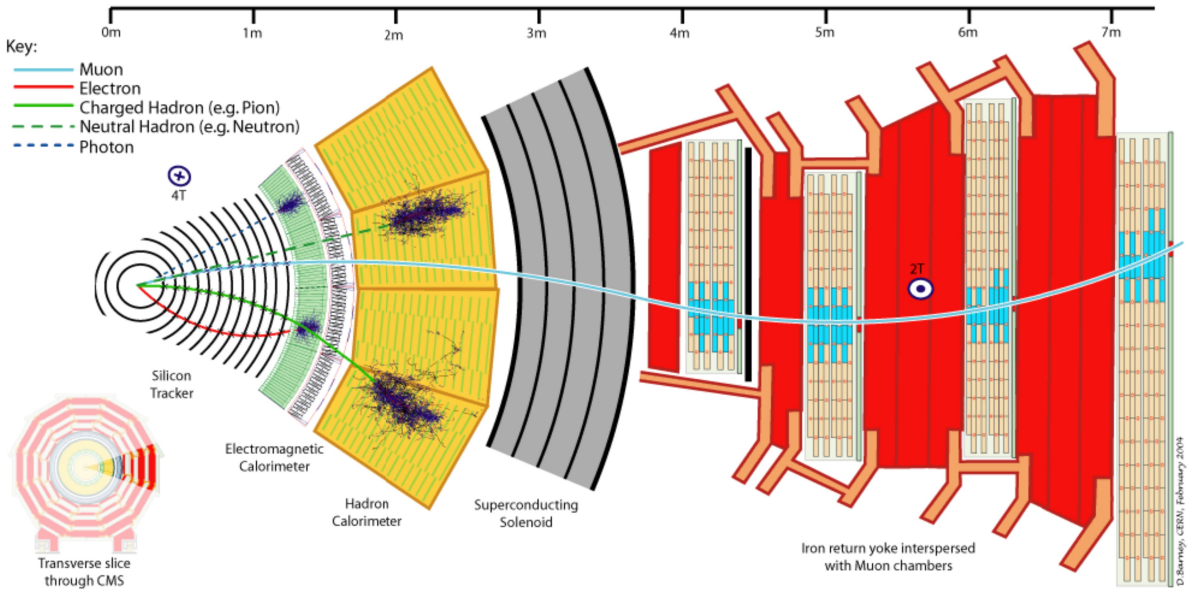


Abbildung 4: Eine schematische Skizze vom CMS-Experiment. Von links nach rechts: Innere Spurdetektor, Elektromagnetische Kalorimeter, Hadronische Kalorimeter, Solenoid-Magnet, Myonenspektrometer [11].

2.2.3 Innerer Spurdetektor

Mit dem inneren Spurdetektor aus Silizium lassen sich einige wichtige Größen der zu detektierenden Teilchen bestimmen. Dieser Spurdetektor ist innen ein Pixeldetektor, der die Teilchenposition in jedem Pixel mit einer Genauigkeit von 0,01mm bestimmen kann [12]. Weiter außen wird er durch einen Streifendetektor ergänzt. Mithilfe des Magnetfeldes des Solenoid-Magneten lässt sich die gekrümmte Teilchenbahn rekonstruieren und damit beispielsweise der Transversalimpuls und die Ladung bestimmen. Da der innere Spurdetektor sich nahe zu der Strahlenachse befindet, lassen sich mit dem auch kurzlebige Teilchen bestimmen.

[6] Dies entspricht einem Magnetfeld, welches ca. 100000 Mal größer ist als das Erdmagnetfeld.

2.2.4 Elektromagnetisches Kalorimeter

Um ebenfalls die Energie der Teilchen zu bestimmen, werden in dem CMS Detektor Kalorimeter benutzt. Im elektromagnetischen Kalorimeter befinden sich etwa 80.000 Kristalle aus Bleiwolframat als Szintillator-Material, die diese Funktion für elektromagnetisch wechselwirkenden Teilchen präzise erfüllen [7]. Die elektromagnetischen Kalorimeter sind zylindrisch aufgebaut und unterteilen sich in den Zentralbereich und in die zwei Endkappen. Sobald ein Teilchen in das Kalorimeter eindringt, wird die Energie des Teilchens über Bremsstrahlung und Paarbildungsprozesse in das Szintillator-Material gespeichert und es werden Photonen ausgesendet. Dieses optische Signal wird schließlich über Photodioden gemessen.

2.2.5 Hadronisches Kalorimeter

Hadronische Anteile des Zerfalls werden mit dem hadronischen Kalorimeter gemessen, welcher sich hinter dem ECAL befindet [7].

Indirekt über die Energie-Impuls-Bilanz lassen sich in dem hadronischen Kalorimeter Neutrinos detektieren. Die elektromagnetisch wechselwirkenden Anteile sind durch den ECAL bereits absorbiert. Um die maximale Detektionsrate zu erhalten, ist das hadronische Kalorimeter ebenfalls zylindrisch um den Interaktionspunkt aufgebaut.

2.2.6 Solenoid-Magnet

Es handelt sich um einen supraleitenden Solenoid-Magnet. Für eine entsprechende Supraleitung ist der Solenoid-Magnet auf eine Temperatur von $\mathrm{T} = -268,5°\mathrm{C} = 4,65°\mathrm{K}$ gekühlt. Dieser Solenoid-Magnet erzeugt ein Magnetfeld von 3,8 T, welcher für die Ablenkung der Teilchenbahn verantwortlich ist. Aufgrund dieser Ablenkung und der zuvor genannten Kalorimeter lassen sich wichtige Größen der Teilchen bestimmen, die in Kapitel 2.4 näher erläutert werden.

2.2.7 Myonenspektrometer

Die Myonenkammern sind für die Bestimmung des Impulses der Myonen im Zerfallsprozess da. Die durch den Solenoid-Magneten gekrümmte Bahn der Myonen dient als Grundlage zur Bestimmung des Impulses des Myon. Hauptsächlich gelangen Myonen in diese Kammern, da andere Teilchen üblicherweise bereits zerfallen sind oder von den Kalorimetern absorbiert worden sind.

2.3 Das Koordinatensystem und die Pseudorapidität η

Um die in der Arbeit verwendeten Variablen zu verstehen, muss die Konvention des Koordinatensystems für das CMS-Experiment erläutert werden. Das CMS-Experiment benutzt ein kartesisches rechtshändiges Koordinatensystem mit dem Kollisionspunkt als Ursprung [13]. Die x-Achse verläuft vom Kollisionspunkt zum Zentrum des LHC-Rings, die y-Achse steht orthogonal zur LHC-Ring-Ebene nach oben gerichtet und die z-Achse zeigt in Richtung der Protonenstrahlen.

Der Polarwinkel θ ergibt sich aus der positiven z-Achse und des Azimuthalwinkels aus der positiven x-Achse in der x-y-Ebene. Anstelle des Polarwinkels θ wird in der Teilchenphysik oft die Pseudorapidität η verwendet:

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right]. \tag{5}$$

2.4 Kinematische Größen

Im Folgenden werden die für die Analyse benutzen Variablen erklärt. Dies entspricht den Definitionen und Notationen, welche in der SUS-16-014 Analyse verwendet werden. Die abkürzte Notation der Variablen stehen jeweils in Klammern [14].
Die durch einen Prozess entstehenden Teilchen hadronisieren möglicherweise im Detektor und bilden ein sogenannten Teilchenschauer. Diese Teilchenschauer werden in der Teilchenphysik oftmals als *Jet* bezeichnet. Aufgrund von Vorselektionen erfüllen alle Jets in den Datensätzen die Bedingung

$$\vec{p}_{\mathrm{T}} > 30 \text{ GeV} \ . \tag{6}$$

Grundlegend ist die Variable `NJets`, welche die Anzahl der Jets pro Event angibt. Aufgrund der verschiedene Prozesse zwischen Untergrund und Signal findet man in der Anzahl der Jets eine gute Seperationskraft für das Training.
Eine in der Analyse verwendete Variable ist H_{T} (`HT`), welche hauptsächlich von den Jets der Zerfallsprodukte kommen. Dies entspricht die Summe der Skalare aller Vektoren $\vec{p}_{\mathrm{T}}$ der Jets.

$$H_{\mathrm{T}} = \sum_i |\vec{p}_T| \qquad \text{wobei } i \text{ der Index der Jets ist} \tag{7}$$

Eine besonders in der SUSY-Forschung wichtige Größe ist $H_{\mathrm{T}}^{\mathrm{miss}}$ (`MHT`), welche fol-

gendermaßen definiert ist.

$$H_T^{miss} = |\overrightarrow{H}_T^{miss}| = |\sum_i \vec{p}_{\mathrm{T},i}|, \qquad \text{wobei } i \text{ der Index der Jets ist.} \tag{8}$$

Diese Größe gibt den gesamten fehlenden Transversalimpuls der Jets an.

Zudem wird in der folgenden Analyse die Größe $\Delta\Phi_{H_T^{miss},j_i}$ (`DeltaPhi(i)`) verwendet. Diese gibt den Winkel zwischen $\vec{H}_{\mathrm{T}}^{\mathrm{miss}}$ und $\vec{p}_{\mathtt{Tjet}}$ an.
Zuletzt werden als Variablen noch die pro Event vier höchsten Skalare der $\vec{p}_{\mathrm{Tjet,i}}$ und die dazugehörigen Polarwinkel η_{i} der Jets verwendet.

2.5 Signaturen am LHC

Nun stellt sich die Frage, wie am LHC nach Supersymmetrie gesucht wird. Daher werden in diesem Kapitel die für die Analyse dieser Arbeit wichtigen Signaturen am LHC diskutiert.

Als Signal werden in dieser Arbeit die Produktion und der anschließende Zerfall von SUSY-Teilchen als Zerfallsprodukt bezeichnet. Prozesse, die zum Signal ähnliche Signaturen zeigen, werden als Untergrund bezeichnet.

In Abbildung 5 ist der sogenannte T1bbbb-Prozess dargestellt. T1bbbb ist dabei eine von dem CMS-Experiment genutzte Notation für einen bestimmten Prozess. Das T1 steht dabei für eine Gluino-Paarproduktion aufgrund der Protonenkollision im Experiment. Jedes Gluino zerfällt im Fall eines T1bbbb-Prozesses mithilfe der starken Wechselwirkung in ein Bottom-Antibottom-Paar ($b\bar{b}$) und nach der Supersymmetrie in ein Neutralino ($\tilde{\chi}_1^0$). Anstelle der Bottom-Quarks sind auch andere Quarks möglich. Dieses Bottom-Antibottom-Paar hadronisiert und bildet dann einen sogenannten Teilchenschauer, welcher in der Teilchenphysik und auch in dieser Arbeit als **Jet** bezeichnet wird. Also lässt sich vermuten, dass für den Signalprozess vier Jets zu erwarten sind. Jedoch können für einen Signalprozess auch mehr als vier Jets gemessen werden, da beispielsweise aufgrund von der sogenannten *Initial State Radiation* das Gluon ein weiteres Gluon abstrahlt, welches wiederum einen weiteren Jet erzeugt. Daher ist der b-Tag - die Anzahl der als Bottom- beziehungsweise Antibottom-Quark identifizierter Jet - eine wichtige Charakteristik dieses Zerfalls. Weil jedoch beispielsweise auch Jets von Gluonen als Bottom(Antibottom)-Quark vom Detektor identifiziert werden können, werden auch Signalprozesse mit mehr als

vier b-Tags auftreten. Daher ist in der Datensimulation die Effizienz des b-Tagging wichtig, welche der Anteil der als Bottom(Antibottom)-Quark identifizierten Jets von der Gesamtanzahl der Bottom(Antibottom)-Quarks angibt. Beispielsweise beträgt die Effizienz für b-Jets (als b-Tag falsch identifizierte Jets) für die Datensätze, die in dieser Arbeit verwendet werden, bei einem $p_T \approx 30$ GeV [7] 55% (1,6%) [14] . Oftmals werden für das Signal auch weniger als vier b-Jets gemessen, da diese keinen b-Tag erhalten haben aufgrund von Messungenauigkeiten.

Das Neutralino $\tilde{\chi}_1^0$ erhält durch die Energie und durch den Impuls des Gluinos einen transversalen Impuls $\vec{p_T}$. Da das Neutralino nicht messbar ist und daher nicht vom Detektor bestimmt werden kann, zeigt sich dieser Impuls als fehlender transversaler Impuls. Aus diesem Grund ist der fehlende transversale Impuls einer der wichtigsten Variablen dieser Analyse.

Problematisch wird das Identifizieren von Signalprozessen, wenn man die Untergrundprozesse miteinbezieht, die ähnliche Endzustandseigenschaften besitzen wie der Signalprozess. Bei den QCD-Multijet-Events beispielsweise kommt es zu fehlenden Transversalimpuls durch eine ungenaue Messung des Impulses der Jets und Hadronen. Bei dem Top-Antitop-Untergrund entstehen identifizierte b-tags, weil Top-Quarks mit einer hohen Wahrscheinlichkeit in Bottom-Quarks zerfallen und diese Jets als b-Jets identifiziert werden. Ein Beispiel eines solchen Zerfalles ist in Abbildung 6 dargestellt. In dieser Arbeit wird mit einer Vielzahl von Untergrundprozessen gearbeitet, welche in Kapitel 4.1 aufgelistet sind.

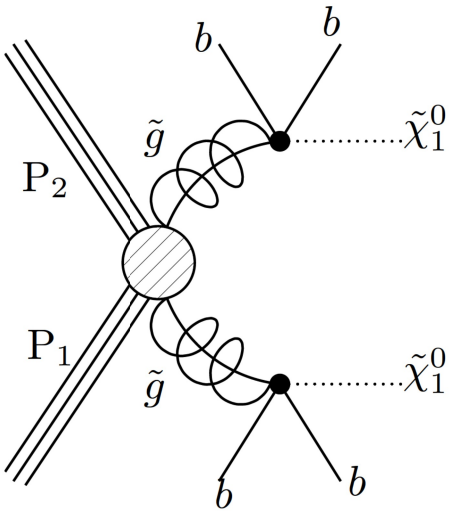


Abbildung 5: Feynman-Diagramm für den T1bbbb-Prozess [15].

[7] p_T ist der Transversalimpuls der Jets.

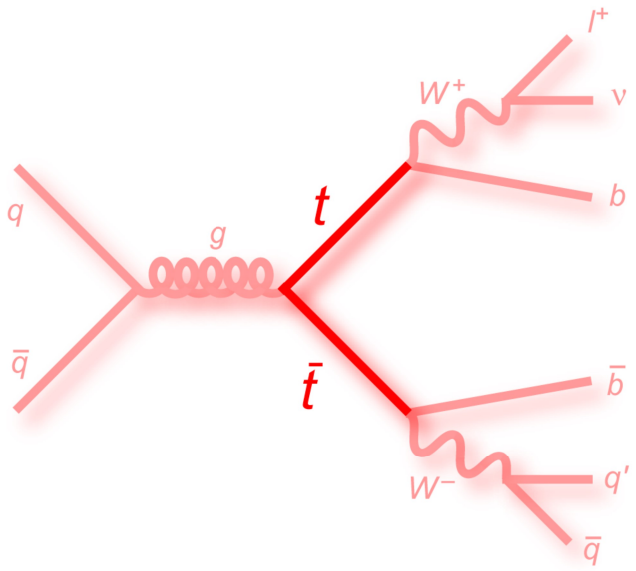


Abbildung 6: Feynman-Diagramm für einen Top-Antitop-Prozess mit Bottom-Jets.

Diese Arbeit ist angelehnt an die Arbeit der SUS-16-014-Gruppe, welche sich mit der voll hadronischen Supersymmetriesuche am CMS beschäftigt. Einen besonderen Schwerpunkt hat die voll hadronische Supersymmetriesuche dieser Arbeitsgruppe in der Berücksichtigung des fehlenden Transversalimpulses ($H_{\mathrm{T}}^{\mathrm{miss}}$). Die in dieser Arbeit verwendeten Datensätze und Monte-Carlo-Simulationen stammen ebenfalls von der SUS-16-014 Analyse. Diese CMS-Kollaboration betrachtet unter anderem den T1bbbb-Prozess. Außerdem verwenden sie zur Suche, die in Kapitel 2.4. Da Analysen nur in einer bestimmten Auflösung der Observablen durchgeführt werden können, teilt man die Observablen in Abschnitte, welche *Bins* genannt werden. Analysiert wird in der CMS-Kollaboration in 160 solcher Bins, welche durch die Observablen H_{T}, $H_{\mathrm{T}}^{\mathrm{miss}}$, N_{Jets}, b-Tags definiert werden [16]. Dabei gibt N_{Jets} die Anzahl an Jets und b-Tags die Anzahl an als Bottom-Quark identifizierte Jets an. Mithilfe dieser Bins werden sogenannte Kontrollregionen gebildet. Mit diesen Kontrollregionen für den Untergrund lassen sich die Abschätzungen innerhalb eines Vergleiches zwischen Daten und Untergrund beurteilen.
Im Gegensatz zur CMS-Kollaboration wird in der Analyse der Arbeit eine multivariate Analyse verwendet. Daher wird im Folgenden die multivariate Analyse erläutert. Zudem werden Grundlagen zum Toolkit geschafft, welche für diese Analyse verwendet wurde. Für die Analyse wendet die CMS-Kollaboration Schnitte auf die betrachteten Observablen einzeln an.

3 Multivariate Analyse mit TMVA

Ziel dieser Arbeit ist es, Multivariate Analysetechniken zu benutzen, um möglicherweise die Analyse zur Entdeckung von Supersymmetriepartner am CMS zu optimieren. Daher werden in diesem Kapitel die Grundlagen der Multivariaten Analyse und die Funktionsweise des **Toolkit for Multivariate Data Analysis (TMVA)** [17], welches am CERN entwickelt wurde, beschrieben, das vor allem in der Hochenergiephysik eine wichtige Rolle in der Datenanalyse einnimmt. Das Problem ist bei einer Datenmenge, die Daten zu finden, die dem Interesse des Forschers entspricht. Daher ist es ein wichtiger Schritt in der Datenanalyse, die Datenmenge in verschiedene Kategorien zu klassifizieren, wofür TMVA entwickelt wurde.

Im Folgenden werden Events entweder als Signal oder als Untergrund klassifiziert, wobei das Signal die gewünschten Daten darstellt und der Untergrund die restlichen Klassifikationen. In der Hochenergiephysik ergeben sich jedoch zu viele freie Parameter, die miteinander korrelieren können, welches eine Klassifikation über Schnitten auf den gebrauchten Observablen erschweren[8]. Daher hilft man sich mit einer Multivariaten Analyse, welche mehrere freie Parameter und ihre Korrelationen in einem vieldimensionalen Phasenraum zugleich untersucht. Diese Klassifikationen geschehen in einer multivariaten Analyse über verschiedene Algorithmen, die im Folgenden als Methoden bezeichnet werden.

3.1 Kenngrößen

TMVA beurteilt die Methoden anhand verschiedener Größen, die im Folgenden definiert werden.

Zum einen gibt es die Signalreinheit, welche den prozentualen Anteil der Signalevents beim Setzen eines Schnittwertes[9] angibt.

$$\text{purity} = \frac{S}{S+B} \ , \tag{9}$$

dabei ist S die Anzahl der Signalevents und B die Anzahl der Untergrundevents, welche nach dem Setzen des Schnittes übrig geblieben sind. Somit ergibt sich ein optimaler Schnittwert bei einem maximalen Wert für S und ein minimaler Wert für B. Dies führt zu einem optimalen Wert für die Signalreinheit bei 1.

[8] Üblicherweise werden Schnitte auf Observablen durchgeführt, um möglichst viele Signalevents von den Untergrundevents zu trennen (Hohe "Purity").

[9] In der Multivariaten Analyse werden üblicherweise Schnitte auf die von TMVA erstellte Variable durchgeführt.

Eine andere Größe für die Quantisierung der Methoden ist die Signifikanz

$$\text{significance} = \frac{S}{\sqrt{S+B}}, \tag{10}$$

welche ebenfalls nach dem Setzen eines Schnittwertes berechnet werden kann.

Die Signifikanz ist ebenfalls eine Größe, welche die Trennung der Signalevents von den Untergrundevents angibt. Beispielsweise erhält man eine hohe Signifikanz bei einer hohen Anzahl von übrig gebliebenen Signalevents und bei einer kleinen Zahl übrig gebliebener Untergrundevents. Die Benutzung der beiden Größen (Purity und Signifikanz) ist äquivalent.

TMVA arbeitet in zwei Phasen. Zuerst wird die Methode mithilfe von Datensätzen[10] mit bekannter Signal-Untergrund-Klassifikation trainiert. In dieser sogenannten Trainings-Phase werden die verwendeten Observablen definiert und die Methode ausgesucht. Mit diesen trainierten Algorithmus werden in der sogenannten Application-Phase die Resultate dieses Trainings analysiert und der Algorithmus auf Datensätze mit unbekannter Klassifikation angewendet.

3.2 Receiver-Operating-Characteristic Curve (ROC-Kurve)

Eine Möglichkeit die Performance einer trainierten Klassifikation zu beurteilen ist es, die ROC-Kurve der Klassifikation zu betrachten [18]. Für einen Punkt auf der ROC-Kurve wird in den Verteilungsfunktionen von Untergrund und Signal ein Schnittwert auf der betrachteten Variable ausgesucht. Aus diesem Schnittwert ergeben sich die *signal effiency* und die *background rejection*. Die *signal effiency* ist der Anteil der Signalevents, die nach dem Schnitt übrig geblieben sind S zu der Gesamtanzahl der Signalevents S_0:

$$\text{signal effiency} = \frac{S}{S_0} \tag{11}$$

Dagegen lässt sich die *background rejection* berechnen aus der Anzahl der Untergrundevents, die nach dem Schnitt übrig geblieben sind B und der Gesamtanzahl der Untergrundevents vor dem Schnitt B_0 mit:

$$\text{background rejection} = 1 - \frac{B}{B_0} \tag{12}$$

[10] In dieser Arbeit sind dies Monte-Carlo-Simulationen.

Für die ROC-Kurve wird die *signal effiency* auf die x-Achse und die *background rejection* auf die y-Achse abgebildet. Der Vergleich von verschiedenen ROC-Kurven kann mit einer Betrachtung des Integrals unter den ROC-Kurven erfolgen. Je größer das Integral desto besser erfolgt die Separation von Signal und Untergrund. Dabei kann das Integral einen maximalen Wert von 1 betragen.

Auf den ROC-Kurven lassen sich Arbeitspunkte mit gegebener signal effiency und background rejection auswählen. Die Performance ergibt sich durch die Höhe dieser Zahlen. Je größer diese sind, desto besser die Performance. Daher ist das ROC-Integral keine optimale Größe zur Quantisierung der Performance, da sich ROC-Kurven auch schneiden können.

Als Beispiel sind die ROC-Kurven der verschiedenen Methoden dieser Analyse in Abbildung 11 dargestellt.

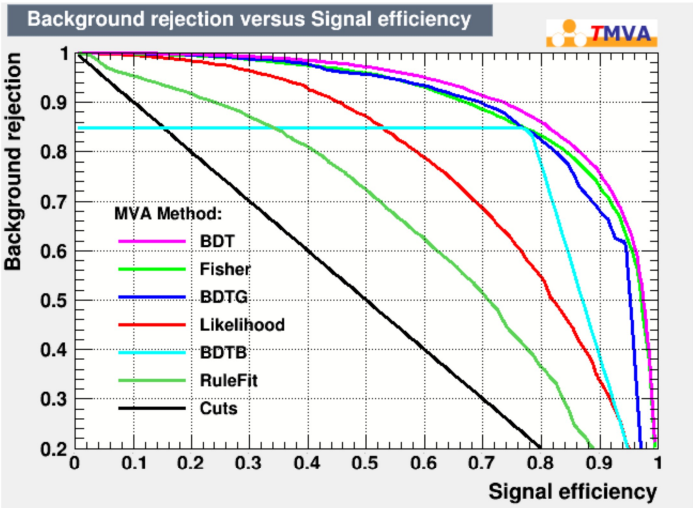


Abbildung 7: ROC-Kurve bei einer Auswahl von verschiedenen Methoden (siehe Kapitel 4.4. Diese Abbildung wurde mit TMVA erstellt.

Anhand dieser Abbildung lässt sich erkennen, dass die Methode `Cuts` aufgrund der linearen Form keine Separation hervorgerufen hat. Dies liegt möglicherweise an einer falschen Implementierung dieser Methode während des Trainings.

3.3 Verschiedene Methoden in TMVA

In TMVA stehen verschiedene Algorithmen zur Verfügung von dem einige in dieser Arbeit verwendet werden. Daher werden diese Methoden in diesem Kapitel erläutert. Eine genauere Erläuterung der Methoden lässt sich im TMVA-Manual [17] finden.

`BDT` **(Boosted Decision Trees)**: Bei der BDT-Methode werden Decision Trees

gebildet, die über Schnitte auf den Variablen, Klassifikationen erstellen. An jedem Knoten des Decision Trees wird ein Schnitt auf die Variablen ausgeübt und die Anzahl der Events aufgeteilt. Am Ende des Decision Trees sind dann sogenannte Blätter, die die Klassifikation des Events als Untergrund bzw. Signal angibt.

`RuleFit` **(Predictive learning via rule ensembles)**: In der Methode RuleFit behilft man sich mit einem Ensemble von sogenannten *Regeln*. Eine Regel ist dabei eine Kombination von Schnitten auf den Variablen. Diese Methode lässt sich mit der Methode BDT vergleichen. An jedem Knoten eines Decision Tree lässt sich so eine Regel ableiten (siehe Kapitel 3.4). Der RuleFit Output ist schließlich eine Linearkombination aller Regeln.

`Likelihood` **(Projective likelihood estimator)**: Diese Methode hat es als Ziel eine Wahrscheinlichkeitsdichtefunktion zu bilden, welche die Verteilungen der Variablen vom Signal und Untergrund wiedergeben sollen. Beispielsweise ist die Wahrscheinlichkeit, dass es sich um ein Signal bei einem gegebenen Event handelt, berechnet durch eine Multiplikation aller Wahrscheinlichkeitsdichten der Signalvariablen normalisiert auf die Summe der Wahrscheinlichkeiten von Signal und Untergrund. Jedoch werden bei dieser Methode die Korrelationen zwischen den Variablen nicht betrachtet.

`Fisher` **(linear Fisher discriminant analysis)**: Die Eventsklassifikation findet bei der `Fisher`-Methode in einem transformierten Variablenraum mit möglichst kleinen Korrelationen zwischen den Variablen statt. Diese Klassifikationen werden über die Mittelwerte der Verteilungsfunktionen von Signal und Untergrund ermittelt. Die Fisher-Methode erstellt mit dieser Klassifikation die sogenannte Fisher Diskriminante, die Signal und Untergrund separiert, wenn die Events von Signal und Untergrund auf dieser Diskriminanten projeziert werden.

Da sich in dieser Analyse herausstellen wird (siehe Kapitel 4.4), dass die BDT-Methode der beste Algorithmus für diese Analyse ist, wird diese Methode näher erläutert.

3.4 Boosted Decision Trees

Eine im TMVA verwendete Methode ist der Boosted Decision Tree (BDT). Decision Trees sind baumartig aufgebaute Klassifikationen. Der Baum beginnt an dem sogenannten Wurzelknoten aus dem wiederum eine Reihe von Knoten entspringen können [17]. Jeder Knoten stellt dabei einen Schnitt auf eine Variable dar. Am Ende

des Baums nach einer Zahl an Knoten sind die Blätter, die schließlich ein Event als Signal oder als Untergrund klassifizieren. Der Weg vom Wurzelknoten über die Knoten bis hin zu dem Blatt wird als Ast bezeichnet. Ein Ast stellt eine Reihe von Schnitten auf den Variablen dar. Die Knoten, welche Schnitte auf eine Variable repräsentieren, sind so ausgewählt, dass diese die beste Separation liefert. So werden an den Knotenpunkten die Datenmenge in zwei Teilmengen geteilt und diese als Daten für eine weitere Separation verwendet. Diese Separation über Knoten geschieht so lange, bis die Anzahl der Events einen bestimmten Wert unterschreitet. Dieser Wert lässt sich als Parameter unter den Namen `MinNodeSize` variieren[11]. Die statistische Instabilität stellt ein Problem eines Decision Trees dar. Das Trainieren einer Variablen aufgrund Fluktuationen in dieser Variablen würde das Trainieren für andere Variablen unterdrücken. Um dieses Problem zu lösen, lässt sich ein Forest aus Decision Trees erstellen. Beim Erstellen der weiteren Trees kommt das sogenannte Boosting zum Einsatz.

3.4.1 Boosting

Unter Boosting versteht man in der multivariaten Analyse eine Umgewichtung der Events beim Erstellen eines nächsten Trees. Bei dem im Folgenden benutzten Boosting, dem sogenannten Adaptive Boosting, wird falsch klassifizierten Events ein höheres Gewicht zugeschrieben als den richtig klassifizierten Events. Diese Umgewichtung findet bei jedem Erstellen eines neuen Trees statt. Das Gewicht der falsch klassifizierten Events werden mit einem Boost-Faktor α multipliziert der sich folgendermaßen ergibt:

$$\alpha = \frac{1 - \text{err}}{\text{err.}} \tag{13}$$

Dabei bezeichnet err der Anteil der falsch klassifizierten Events zur Gesamtzahl an Events. Schließlich werden die Gewichte aller Events so normiert, dass die Summe der Gewichte beim darauffolgenden Decision Tree erhalten bleibt.

3.4.2 BDT-Steuerparameter

Bei der BDT-Methode gibt es in TMVA viele Steuerparameter, welche das Training beeinflussen. Diese Steuerparameter ändern das Verhalten des Algorithmus hinsichtlich des Trainings. Im Folgenden werden die für die Optimierung des Trainings im Kapitel 4.5 wichtigen Steuerparameter erklärt. In der Tabelle 1 sind diese Steuerparameter mit einem Standardwert und einer Kurzbeschreibung aufgelistet[12].

[11] Die genannten Parameter sind nur in TMVA so definiert.
[12] Die genannten Parameter sind nur in TMVA so definiert.

Steuerparamter	Standardwert	Kurzbeschreibung
`NTrees`	800	Anzahl der Trees im Forest
`MaxDepth`	3	Maximale Tiefe der Trees
`MinNodeSize`	5%	Minimale Anteil der Training Events in einem Blatt
`nCuts`	20	Anzahl der Einteilungen einer Variable in einem Knoten
`AdaBoostBeta`	0.5	Lernrate beim Boosting

Tabelle 1: Für die Optimierung wichtige BDT-Steuerparameter mit einem Standardwert und einer Kurzbeschreibung.

Der Steuerparameter `NTrees` gibt die Anzahl der im Forest verwendeten Decision Trees an [17]. Die maximale Tiefe der Decision Trees ist durch den Parameter `MaxDepth` beeinflussbar. Unter der maximalen Tiefe ist die maximale Stufenlänge vom Anfangsknoten zum Blatt gemeint. Der minimale Anteil der Events von der Gesamtzahl der Events im Training in diesem Blatt wiederum ergibt sich aus dem Parameter `MinNodeSize`. In einem Knoten im Decision Tree wird der optimale Schnitt in einer Observablen gesucht. Dafür unterteilt der Algorithmus diese Variable in Abschnitte und findet so den optimalen Schnitt unter Betrachtung der Purity (siehe Kapitel 3). Die Anzahl dieser Abschnitte lässt sich durch den Parameter `nCuts` einstellen. Eine Größe, die beim Adaptive Boosting eine Rolle spielt, ist `AdaBoostBeta`. Dieser gibt die Lernrate beim Boosting an. Da das Boosting bei vielen aber kleinen Decisions Trees, die individuell eine geringe Separationkraft zeigen, am wirksamsten ist, ist "langsames Lernen"von Vorteil. Das Gewicht beim Boosting α wird mit dem AdaBoostBeta-Wert β modifiziert zu $\alpha \to \alpha^{\beta}$. Ein kleiner Wert für β verringert so den Boostfaktor und ermöglicht mehr Schritte beim Boosting.

4 Analyse

Die Suche nach den Supersymmetrieprozessen in den Datensätzen vom CMS erfolgt auch durch die CMS-Kollaboration SUS-16-014. Jedoch arbeitet diese Gruppe, indem die Observablen MHT, NJets, b-Jets und HT betrachtet werden. Eine andere Möglichkeit ist daher die Separationskraft von vielen Variablen in einer multivariaten Analyse zu benutzen. Dafür wird eine optimierte Klassifikation benötigt, die auf diesen Variablen trainiert ist. Es werden für das Training der Klassifikation Monte-Carlo-Simulationen für Signal und Untergrund verwendet, welche in diesem Kapitel beschrieben werden. Zudem werden die Performance verschiedener Methoden von TMVA für diese Analyse miteinander verglichen, um die bestmögliche Methode für diese Analyse zu bestimmen. Schließlich wird der Optimierungsprozess beschrieben, der zu einer optimierten Klassifikation führt.

4.1 Gewichtung und Monte-Carlo-Simulationen

In diesem Abschnitt wird das Gewicht, welches bei Monte-Carlo-Simulationen eine wesentliche Rolle spielen, definiert und die für das Training verwendeten Monte-Carlo-Simulationen[13] aufgelistet. Der Einfachheit halber sind die Namen der Dateien unverändert geblieben. Die SUS-16-014 Analyse verwendet anstelle von Monte-Carlo-Simulationen richtige Datensätze zur Abschätzung der Untergrundereignisse.

Gewichtung

Da bei den Monte-Carlo-Simulationen eine Anzahl von Events N_{sim} generiert wird, müssen, damit verschiedene Monte-Carlo-Simulationen vergleichbar sind, diese umgewichtet werden. Das Gewicht W berechnet sich folgendermaßen:

$$W = \frac{\sigma \ \cdot L}{N_{\text{sim}}} \ . \tag{14}$$

Dabei ist σ der Wirkungsquerschnitt des Prozesses und L die Luminosität.

Mithilfe des Gewichtes lässt sich die Anzahl der zu erwartenden Untergrundevents N_{erw} bei einer Messung bestimmen.

$$N_{\text{erw}} = N_{\text{sim}} \ \cdot W \ . \tag{15}$$

Die folgenden Monte-Carlo-Simulationen sind alle auf eine Luminosität L von:

$$L = 1 \text{ pb}^{-1} \ , \tag{16}$$

normiert (siehe Tabelle 2-8).

Monte-Carlo-Simulationen

Es folgt eine Auflistung aller Monte-Carlo-Simulationen, welche für die Analyse verwendet werden.

In Tabelle 2 sind die QCD-Untergründe gegeben. Am Wirkungsquerschnitt σ erkennt man durch einen Vergleich der anderen Untergründe in den Tabellen 3 bis 7, dass der QCD-Untergrund der größte Untergrund in dieser Analyse darstellt. Deutlich kleinere Untergründe stellen beispielsweise die Single-top Untergründe (siehe

[13] Diese Monte-Carlo-Simulationen stammen von der `Run2ProductionV9` wie das innerhalb der SUS-16-014 definiert ist.

Tabelle 5). Jedoch wurden auch diese benutzt, um den Untergrund bestmöglich wiederzugeben.

Dateiname	Gewicht	Anzahl der Events	σ (pb)
tree_QCD_HT-200to300	0,03096	22	1717000
tree_QCD_HT-300to500	0,006422	114	351300
tree_QCD_HT-500to700	0,0005138	894	31630
tree_QCD_HT-700to1000	0,0001496	2691	6802
tree_QCD_HT-1000to1500	7,925e-05	2821	1206
tree_QCD_HT-1500to2000	1,029e-05	6968	120,4
tree_QCD_HT-2000toInf	4,179e-06	9845	25,24

Tabelle 2: Verwendete Monte-Carlo-Simulationen für den QCD-Untergrund. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator MADGRAPH5_aMC@NLO 2.2.2 verwendet [14].

Dateiname	Gewicht	Anzahl der Events	σ (pb)
tree_WJetsToLNu_HT-100to200	5,507e-05	2887	1627,45
tree_WJetsToLNu_HT-200to400	2,18e-05	136697	435,24
tree_WJetsToLNu_HT-400to600	8,088e-06	137282	59,18
tree_WJetsToLNu_HT-600to800	3,916e-06	80867	14,58
tree_WJetsToLNu_HT-800to1200	8,904e-07	171763	6,66
tree_WJetsToLNu_HT-1200to2500	6,517e-06	6101	1,608
tree_WJetsToLNu_HT-2500toInf	1,539e-07	6834	0,03891

Tabelle 3: Verwendete Monte-Carlo-Simulationen für den W+Jets-Untergrund. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator MADGRAPH5_aMC@NLO 2.2.2 verwendet [14].

Dateiname	Gewicht	Anzahl der Events	σ (pb)
tree_ZJetsToNuNu_HT-100to200	1,845e-05	11315	344,3
tree_ZJetsToNuNu_HT-200to400	4,782e-06	1020958	95,23
tree_ZJetsToNuNu_HT-400to600	1,293e-05	152442	13,19
tree_ZJetsToNuNu_HT-600to800	5,7e-07	1026208	3,221
tree_ZJetsToNuNu_HT-800to1200	6,836e-07	417911	1,474
tree_ZJetsToNuNu_HT-1200to2500	9,819e-07	73140	0,3586
tree_ZJetsToNuNu_HT-2500toInf	2,022e-08	84496	0,008203

Tabelle 4: Verwendete Monte-Carlo-Simulationen für den Z+Jets-Untergrund. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator MADGRAPH5_aMC@NLO 2.2.2 verwendet [14].

Dateiname	Gewicht	Anzahl der Events	σ (pb)
tree_ST_s-channel	5,30484e-06	1704	3,34
tree_ST_t-channel_antitop	1,574e-05	1432	26,23
tree_ST_t-channel_top	1,344e-05	3553	44,07
tree_ST_tW_antitop	3,679e-05	1723	35,6
tree_ST_tW_top	3,566e-05	1880	35,6

Tabelle 5: Verwendete Monte-Carlo-Simulationen für den Single-top Untergrund. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator MADGRAPH5_aMC@NLO 2.2.2 verwendet [14].

Dateiname	Gewicht	Anzahl der Events	σ (pb)
tree_TTJets_DiLept	2,888e-06	45394	88,34
tree_TTJets_SingleLeptFromTbar	3,063e-06	282812	182,72
tree_TTJets_SingleLeptFromT	3,522e-06	245416	182,72

Tabelle 6: Verwendete Monte-Carlo-Simulationen für den TT+Jets-Untergrund. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator MADGRAPH5_aMC@NLO 2.2.2 verwendet [14].

Dateiname	Gewicht	Anzahl der Events	σ (pb)
tree_TTGJets	2,32087e-06	23088	3,697
tree_TTTT	2,20412e-08	10520	0,009103
tree_TTWJetsToLNu	1,56831e-06	2789	0,2043
tree_TTWJetsToQQ	9,32511e-07	4997	0,4026
tree_TTZToLLNuNu	1,35889e-06	24823	0,2529
tree_TTZToQQ	1,48916e-06	4506	0,5297
tree_WWTo1L1Nu2Q	1,78025e-05	14145	49,997
tree_WWTo2L2Nu	6.131e-06	1502	12,178
tree_WWZ	7,44626e-07	5136	0,1651
tree_WZTo1L1Nu2Q	9,51524e-07	84080	10,71
tree_WZTo1L3Nu	3,30295e-06	15140	3,058
tree_WZZ	2,54321e-07	7303	0,05565
tree_ZZTo2L2Q	3,3123e-07	4291	3,22
tree_ZZTo2Q2Nu	2,11934e-ö07	792591	4,04
tree_ZZZ	6,46527e-08	7008	0,01398

Tabelle 7: Verwendete Monte-Carlo-Simulationen für die Untergründe (Higher_Order). Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator `MADGRAPH5_aMC@NLO 2.2.2` verwendet [14].

Dateiname	Gewicht	Anzahl der Events	σ (pb)
tree_T1bbbb_1500_100	2,705e-07	33553	0,01419
tree_T1bbbb_1000_900	2,281e-06	24524	0,3254

Tabelle 8: Verwendete Monte-Carlo-Simulationen für den T1bbbb-Prozess. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Die erste Zahl 1500 bzw 1000 steht dabei für die Gluinomasse in [GeV] und die zweite Zahl 100 bzw. 900 für die Masse des Neutralinos [GeV]. Für die Simulation wurde der Generator `MADGRAPH5_aMC@NLO 2.2.2` verwendet [14].

4.2 Vorselektionen

Die in Kapitel 4.1 erwähnten Datensätze enthalten bereits einige Schnitte in den Variablen, um Untergrundevents zu selektieren. Diese Vorselektionen werden in diesem Kapitel erläutert. In der Tabelle 10 sind die Vorselektionen auf den Datensätzen aufgelistet [19].

Name der Selektion	Anwendbarkeitserkärung
HT	Fehlmessungen, Pileup und Abstrahlungen von weiteren Quarks und Gluonen ergeben Jets mit sehr kleinen HT. Dies wird mithilfe dieser Selektion unterdrückt.
MHT	Da auch MHT aufgrund von Messungenauigkeiten vorliegen kann, wird eine Grenze für MHT gesetzt.
MuonVeto	In T1bbbb ergeben sich keine Muonen, daher ist die Selektion anwendbar.
ElectronVeto	In T1bbbb ergeben sich keine Elektronen, daher ist die Selektion anwendbar.
IsoMuonTrackVeto	In T1bbbb ergeben sich keine isolierten Muonen, daher ist die Selektion anwendbar.
IsoElectronVeto	In T1bbbb ergeben sich keine isolierten Elektronen, daher ist die Selektion anwendbar.
IsoPionTrackVeto	In T1bbbb ergeben sich keine isolierten Pionen, daher ist die Selektion anwendbar.
DeltaPhi	Aufgrund von Fehlmessungen einzelner Jets kann es zu MET in Richtung dieser Jets kommen. Daher setzt man ein Limit auf DeltaPhi.

Tabelle 9: Aufgelistet sind die Selektionen mit einer für diese Analyse passende Anwendbarkeitserkärung.

Isoliert bezeichnet in diesem Fall, dass Teilchen in Abwesenheit anderer Teilchen detektiert werden. Das heißt, dass in der Nähe der Jets der Teilchen keine anderen Teilchen detektiert worden sind.

Name der Selektion	Selektion
NJet	`NJets` $\geq$ `2`
HT	`HT > 200`
MHT	`MHT > 100`
MuonVeto	`Muons->size() == 0`
ElectronVeto	`Electrons->size() == 0`
IsoMuonTrackVeto	`isoMuonTracks == 0`
IsoElectronTrackVeto	`isoElectronsTracks == 0`
IsoPionTrackVeto	`isoPionTracks == 0`
DeltaPhi	`DeltaPhi1>0,5; DeltaPhi2>0,5; DeltaPhi3>0,3; DeltaPhi4>0,3`

Tabelle 10: Auflistung der in den Datensätzen (siehe Kapitel 4.1) enthaltenen Vorselektionen. Die linke Spalte zeigt die von der MHT-Analyse-Gruppe verwendete Notationen und die rechte Spalte die konkrete Selektion.

Mithilfe dieser Selektionen werden Untergrundprozesse unterdrückt. Beispielsweise werden durch die Selektionen DeltaPhi hauptsächlich QCD Untergrundevents unterdrückt, da in diesen Prozessen gewöhnlich $H_{\mathrm{T}}^{\mathrm{miss}}$ in die Richtung der Jets zeigen [14].

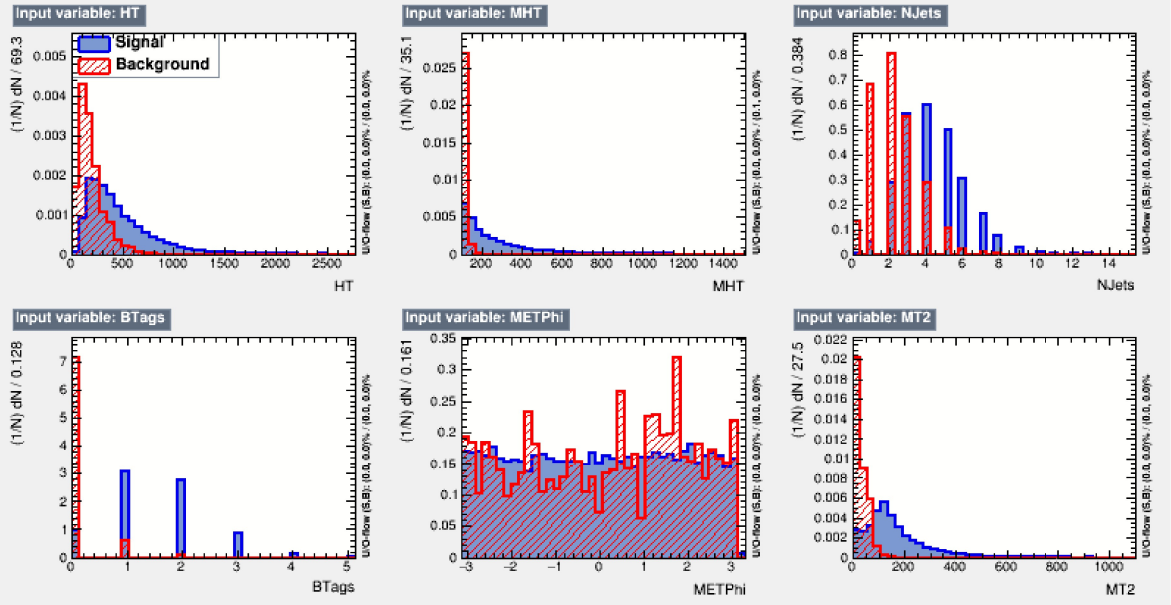
Aufgrund der isolierten Leptonen Selektionen werden Events ausgeschlossen, die hadronisch in Leptonen zerfallen. Aufgrund des Vetos könnten jedoch Events mit Leptonen als Zerfallsprodukt mit berücksichtigt werden. Da diese Vetos auf 0 gesetzt worden sind, ergibt sich so ein Veto nicht. Für den T1bbbb-Prozess sind diese Selektionen so gut wie vernachlässigbar, da isolierte Leptonen in diesem Prozess sehr selten auftreten [14].

Diese Vorselektionen sollen die Menge an Events in den Datensätzen deutlich reduzieren, damit es bei der Analyse zu weniger Laufzeitproblemen kommt. Die aufgelisteten Vorselektionen sind auf den betrachteten Prozess T1bbbb anwendbar, was in der Tabelle 4.2 aufgelistet ist.

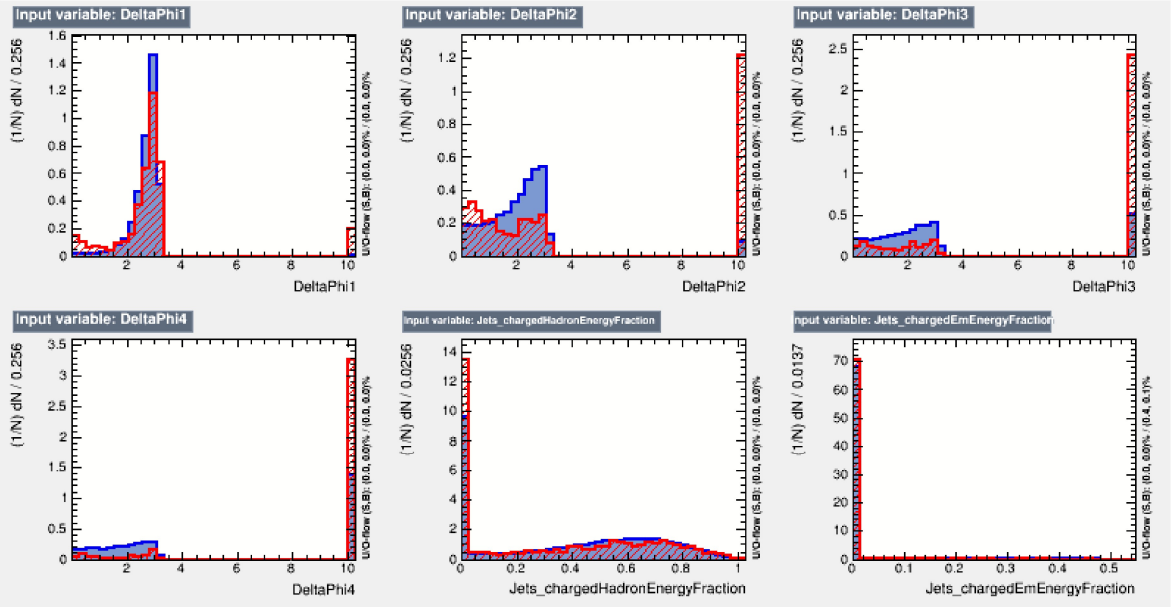
4.3 Optimierung der verwendeten Variablen

Am Anfang der Analyse wurde eine Vielzahl an Variablen verwendet, um die Separationskraft der Variablen zu beurteilen. Das Problem bei der multivariaten Analyse ist, dass das Verwenden von Variablen mit geringer Separationskraft die Analyse beeinträchtigt, da diese trotzdem mit berücksichtigt werden vom Algorithmus. Daher wird anhand der Verteilungsfunktionen der Variablen, eine Auswahl der Variablen getroffen. In Abbildung 9 sind diese Verteilungsfunktionen für verschiedene Variablen dargestellt. Diese Abbildung zeigt die Verteilungsfunktion von Untergrund und Events für verschiedene Variablen. Der Untergrund besteht aus allen QCD-Untergrundevents aus Tabelle 2, zudem von den Untergrundevents aus den Tabellen 3 - 7 wo jeweils nur der Datensatz genommen wurde, der den höchsten Wirkungsquerschnitt hatte[14]. Als Signal wurden die Datensätze aus der Tabelle 8 verwendet. Diese sind von TMVA so skaliert worden, dass das Signal sich in einer ähnlichen Größenordnung befindet wie der Untergrund und die Separationskraft vom Benutzer beurteilt werden kann. Die Notation der Variablen in den Abbildung ist in Tabelle 4.3 erläutert.

[14]Diese Konfiguration für den Untergrund wurde nur als Test verwendet.



(a)



(b)

Abbildung 8: Dargestellt sind die skalierten Verteilungsfunktionen verschiedener Variablen zur Beurteilung der Separationskraft. Die Bedeutung der Notation der Variablen ist in Tabelle ?? gegeben. In blau ist die Verteilung der Signalevents und in rot die Verteilung der Untergrundevents angegeben.

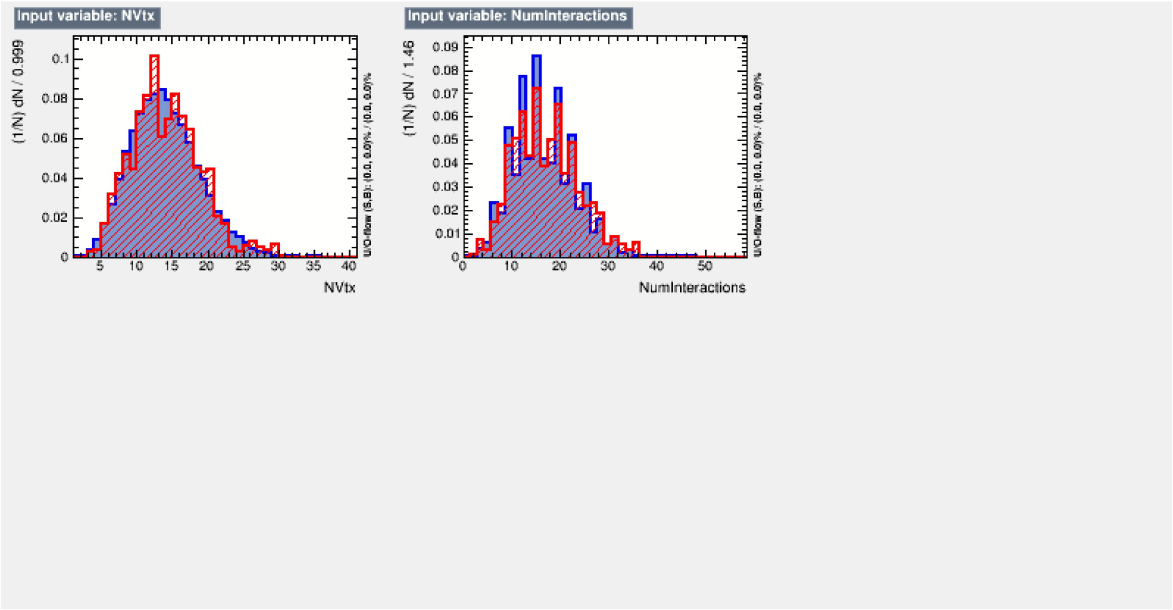


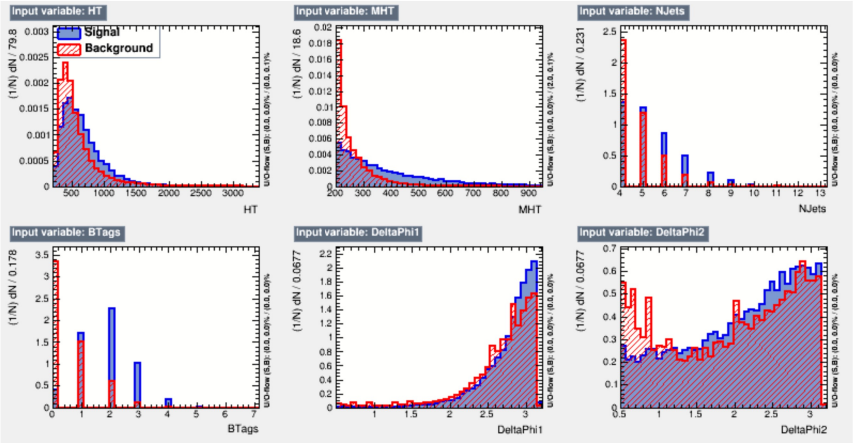
Abbildung 9: Dargestellt sind die skalierten Verteilungsfunktionen verschiedener Variablen zur Beurteilung der Separationskraft. Die Bedeutung der Notation der Variablen ist in Tabelle ?? gegeben. In blau ist die Verteilung der Signalevents und in rot die Verteilung der Untergrundevents angegeben.

Für die letztendliche Analyse wurden einige dieser Variablen nicht verwendet. Beispielsweise erkennt man an der Variable `Jets chargedEmEnergyFraction`, welche den Energieanteil der Jets angibt, die im elektromagnetischen Kalorimeter gemessen worden sind, dass diese Variable keine Separationskraft liefert, da das Signal sich nicht signifikant vom Untergrund unterscheidet. `MT2` ist eine Variable, welche von einer anderen Analysegruppe verwendet wird, weswegen sie in dieser Arbeit nicht weiterhin behandelt wird. Die Variable `NVtx` gibt die Anzahl der Vertices im Prozess an. Zudem erkennt in den Variablen `DeltaPhi` das Problem, dass `DeltaPhi` auf 10 gesetzt ist, wenn der Jet im Event nicht existiert. Daher ist die hinzugefügte Selektion `NJets` ≥ 4 eine Möglichkeit dieses Problem zu lösen.

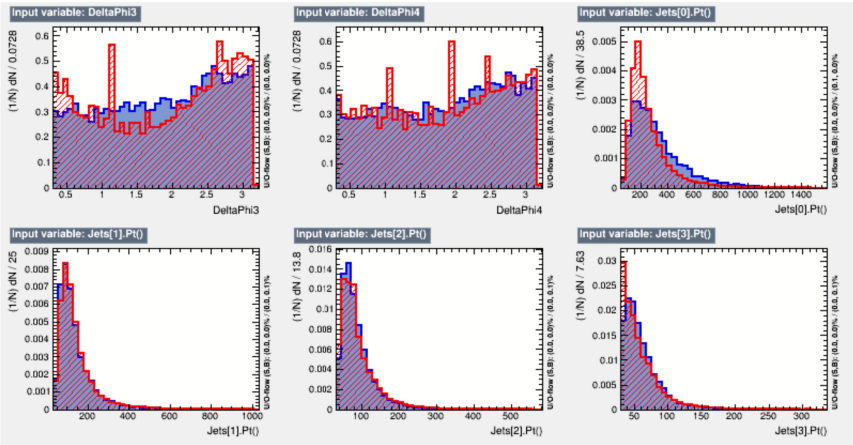
Nach der Beurteilung der Variablen werden für folgende Trainings- und Applikationsprozesse die Variablen benutzt, die in Tabelle ?? aufgelistet sind. Die Verteilungsfunktionen der letztendlich für das Training verwendeten Variablen sind in Abbildung 10 dargestellt.

H_{T}	HT
$H_{textT}^{\mathrm{miss}}$	MHT
N_{Jets}	NJets
B_{Tags}	BTags
$\Delta\Phi_1$	DeltaPhi1
$\Delta\Phi_2$	DeltaPhi2
$\Delta\Phi_3$	DeltaPhi3
$\Delta\Phi_4$	DeltaPhi4
$\mathrm{Jets}(p_{\mathrm{transversal}})_0$	Jets[0].Pt()
$\mathrm{Jets}(p_{\mathrm{transversal}})_1$	Jets[1].Pt()
$\mathrm{Jets}(p_{\mathrm{transversal}})_2$	Jets[2].Pt()
$\mathrm{Jets}(p_{\mathrm{transversal}})_3$	Jets[3].Pt()
$\mathrm{Jets}(\eta)_0$	Jets[0].Eta()
$\mathrm{Jets}(\eta)_1$	Jets[1].Eta()
$\mathrm{Jets}(\eta)_2$	Jets[2].Eta()
$\mathrm{Jets}(\eta)_3$	Jets[3].Eta()

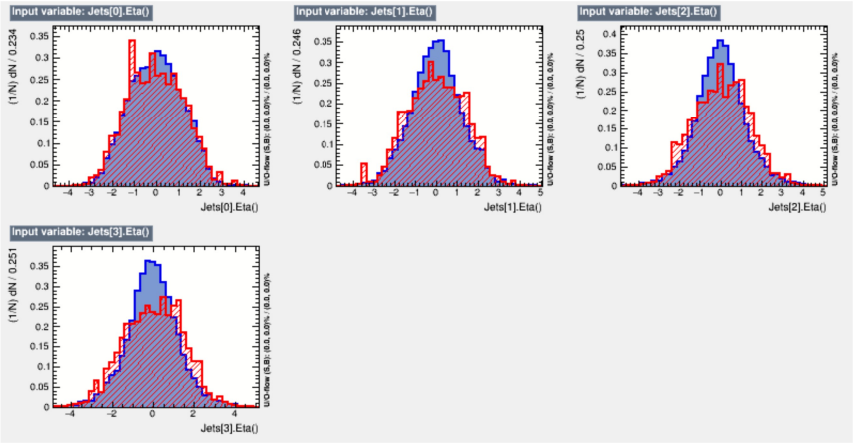
Tabelle 11: Variablen, die für die Trainingsprozesse und Applikationsprozesse verwendet wurden. Die linke Spalte sind die Variablen und physikalischer Notation und die rechte Spalte ist die Notation, welche im TMVA benutzt wurden. Dies entspricht auch die Notation, welche in der SUS-16-014 Analyse verwendet wird. Da die Jets vom größten transversalen Impuls zum kleinsten transversalen Impuls sortiert sind, ist die Nummerierung der Jets in der Tabelle gegeben.



(a)



(b)



(c)

Abbildung 10: Dargestellt sind die Verteilungsfunktionen der Variablen für das Training. Diese Funktionen sind skaliert, sodass die Separationskraft beurteilt werden kann. Die Bedeutung der Notation der Variablen ist in Tabelle **??** gegeben.

Erkennbar ist an den Verteilungsfunktionen, dass die Signalevents in jeweils kleinen Bereichen den Untergrund überwiegen. Diese Bereiche sorgen für eine hohe Separationskraft zwischen Signal und Untergrund. Jedoch muss man beachten, dass

diese Verteilungsfunktionen umskaliert worden sind.

4.4 Vergleich BDT mit anderen Methoden

Da es in TMVA viele verschiedene Methoden gibt, wird ein Vergleich einiger Methoden in diesem Kapitel folgen [17]. Die Bewertung findet über einen Vergleich der ROC-Kurven statt. Die hier verwendeten Methoden sind im Kapitel 3.3 erläutert worden.

Da die für das Training verwendeten Variablen ausgewählt worden sind, ist ein neuer Vergleich der verschiedenen Methoden in TMVA möglich.

Als Konfiguration wurden bei allen Methoden die Standardeinstellungen verwendet. Idealerweise hätte eine Optimierung der Steuerparameter für jede Methode für den Vergleich erfolgen müssen. Dies wurde aus zeitlichen Gründen jedoch nicht durchgeführt.

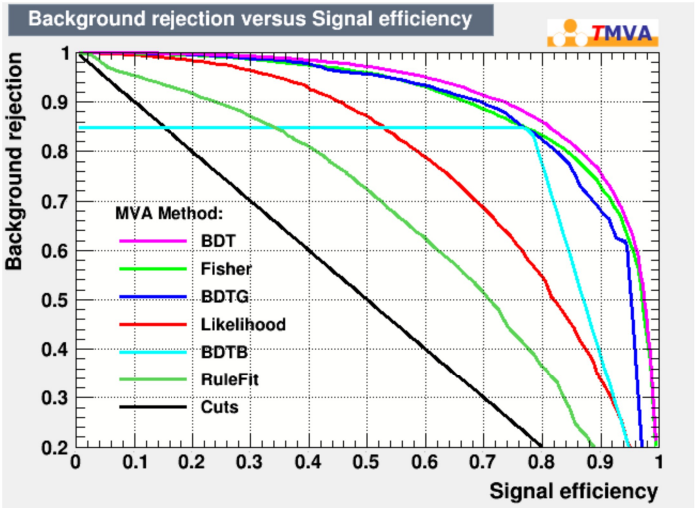


Abbildung 11: ROC-Kurve bei einer Auswahl von verschiedenen Methoden

Man erkennt anhand der ROC-Kurven, dass die Methoden `BDT (Adaptive Boosting)`; `RuleFit` und `BDTG (Gradient Boosting)` eine ähnliche Performance beim Training zeigen. Jedoch erscheint `BDT (Adaptive Boosting)` aufgrund des Funktionsverlaufes am stabilsten. Die Methode `BDTB (Bragging Boosting)` und `Cuts` erscheinen dagegen nicht richtig implementiert bzw. nicht richtig angewendet worden sein.

Wenn man alle Methoden betrachtet, erscheint `BDT` die beste Performance zu zeigen, weswegen im Folgenden `BDT` verwendet wird.

Insgesamt wurden für das Training die Variablen und die Methode ausgewählt. Nun folgt eine Optimierung der Steuerparameter der BDT-Methode.

4.5 Optimierung des Trainings

Da es Ziel dieser Arbeit ist Signalevents geringer Anzahl von den Untergrundevents hoher Anzahl zu trennen, ist die Optimierung des Trainings ein entscheidender Schritt in der Analyse. In TMVA stehen dafür mehrere Steuerparameter für die BDT-Methode zur Verfügung, welche in Kapitel 3.4.2 diskutiert wurden.

Für die Optimierung werden die Steuerparameter für die BDT-Methode variiert und die resultierenden ROC-Kurven und die Signal/Untergrund Plots miteinander verglichen. Erschwerend bei der Optimierung ist dabei das sogenannte *Overtraining*. Das Overtraining tritt auf, wenn das Training auch Fluktuationen in den Daten miteinbezieht. Das heißt auch die Fluktuationen mit separiert werden. Von daher ist beispielsweise die Wahl für eine zu hohe Anzahl der Decision Trees unpraktikabel. Zudem können bei der Variation der BDT-Variablen Laufzeitprobleme auftreten. Quantisiert wird das Overtraining mithilfe des *Kolmogorov-Smirnov-Tests*. Dieser ist in den BDT-Response Verteilungen in TMVA mit angegeben. Ein Beispiel für Overtraining ist in Abbildung 12 gegeben. In den sogenannten Overtraining-Check Abbildungen in TMVA wird in der BDT-Methode die von TMVA generierte Variable BDT-Response für Untergrund und Signal dargestellt. In diese Abbildung 12 wurde der Steuerparameter `NTrees` auf 1000 gesetzt, was in der folgenden Optimierung sich als zu hoch herausstellt. TMVA trennt die Anzahl der Events in zwei verschiedene Mengen auf. Die eine Menge wird Training Sample genannt und ist die Menge der Events, womit der Algorithmus trainiert wird. Damit der Algorithmus das Training beurteilen kann, wird mithilfe der anderen Eventmenge - die sogenannte Test Sample - der trainierte Algorithmus auf das Test Sample angewendet. Um Overtraining visuell darzustellen, wird die Anwendung des Algorithmus auf das Test Sample angewendet und dargestellt. Overtraining zeigt sich durch eine Abweichung der Test Sample Punkte zu den Training Sample Punkten. In Abbildung 12 erkennt man, dass dies bei einer BDT-Response von -0,2 bis 0 der Fall ist. Außerdem ist in der Abbildung 12 die Wahrscheinlichkeit von 1,8% gezeigt. Da diese Wahrscheinlichkeit sehr klein ist, ist Overtraining zu erwarten. Somit erhöht sich die Wahrscheinlichkeit für Overtraining.

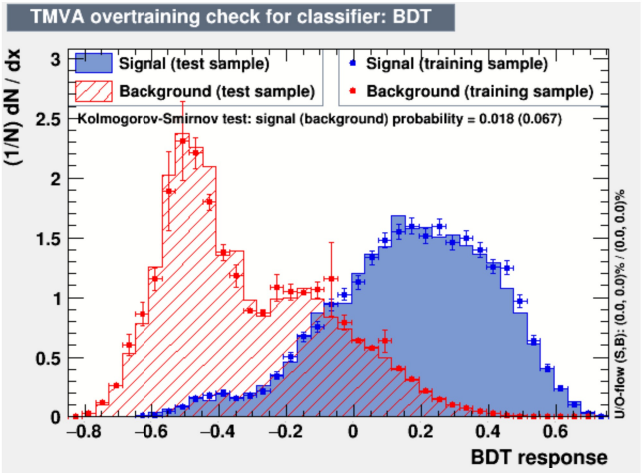


Abbildung 12: Ein Beispiel für Overtraining: Dargestellt ist die BDT-Response mit dem Steuerparameter `NTrees` = 1000.

Für den Optimierungsprozess wird die Funktion `factory->OptimizeAllMethods` verwendet, welcher in der TMVA-Version 4.2.0 nur für BDT mit Adaptive Boosting implementiert ist. Dieser automatischer Optimierungsprozess unterteilt die BDT-Steuerparameter `AdaBoostBeta`, `MaxDepth`, `MinNodeSize`, `NTrees` in Abschnitte und macht das Training für alle Kombinationen der Grenzen zwischen diesen Abschnitte. Quantisiert werden diese Trainingsprozesse mithilfe des ROC-Integrals, welcher optimalerweise 1 beträgt[15]. Das Ergebnis dieses Prozesses ist in den Abbildungen 13 dargestellt. Die Unterteilungen wurden durch den automatischen Optimierungsprozess wie folgt durchgeführt:

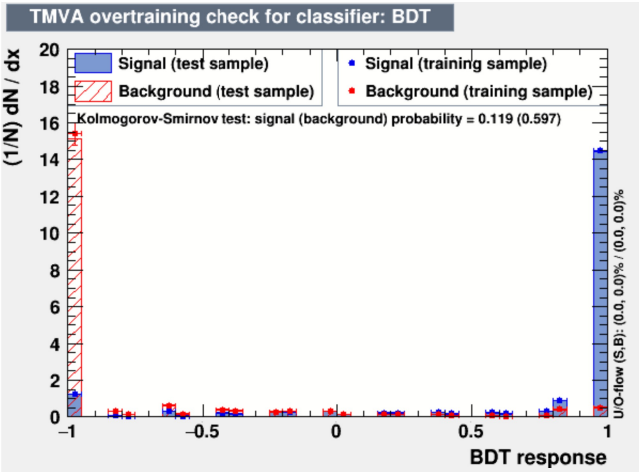
`AdaBoostBeta`: 0,2 | 0,4 | 0,6 | 0,8 | 1

`MaxDepth` : 2 | 3 | 4

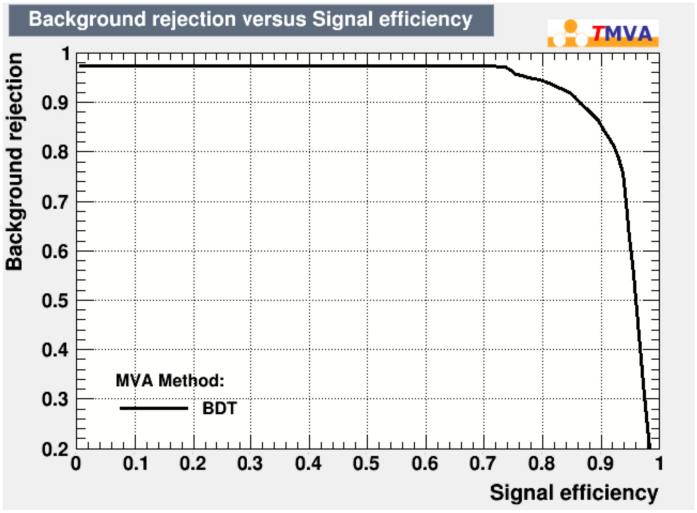
`MinNodeSize` : 1 | 1,12444 | 1,26436 | 1,42169 | 1,5986 | 1,79753 | 2,02121 | 2,27272 | 2,55553 | 2,87354 | 3,23111 | 3,63318 | 4,08529 | 4,59365 | 5,16527 | 5,80802 | 6,53076 | 7,34343 | 8,25722 | 9,28473 | 10,4401 | 11,7392 | 13,2 | 14,8426 | 16,6896 | 18,7664 | 21,1016 | 23,7274 | 26,68 | 30

`NTrees` : 10 | 257,5 | 505 | 752,2 | 1000

[15] Wie in Kapitel 3.2 diskutiert ist die Optimierung anhand des ROC-Integrals nicht optimal.



(a) Overtrainingcheck mit laut TMVA optimierten Einstellungen: `NTrees=10`, `AdaBoostBeta=0.6`, `MinNodeSize=1%`, MaxDepth=4.



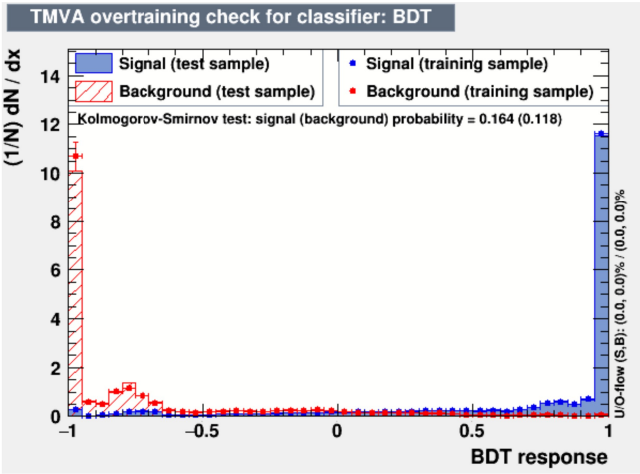
(b) ROC-Kurve mit laut TMVA optimierten Einstellungen: `NTrees=10`, `AdaBoostBeta=0.6`, `MinNodeSize=1%`, `MaxDepth=4`; ROC-Integral beträgt 0.922 .

Abbildung 13: Die Ergebnisse mit der Funktion `factory->OptimizeAllMethods` bestimmen optimierten Werte.

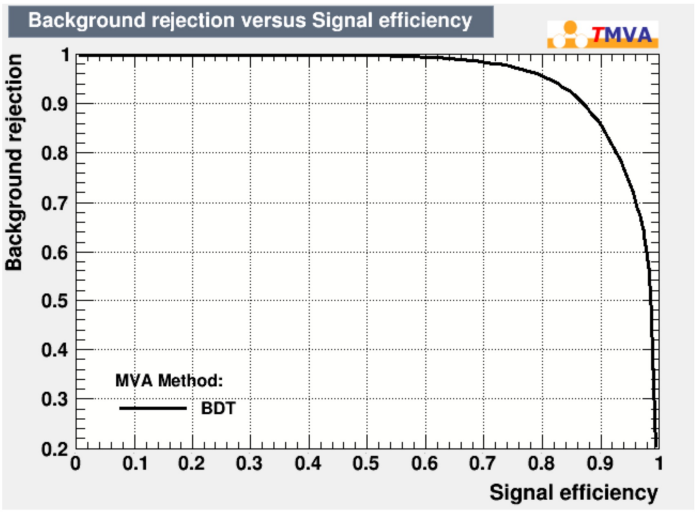
Da diese Einteilung in die Abschnitte willkürlich verläuft, wurde diese Einstellung nur als Referenzpunkt für den weiteren Optimierungsprozess verwendet. Vor dem Verwenden der automatischen Optimierung wurden die Steuerparameter variiert, um ein Gefühl für die BDT-Analyse zu erhalten. Der Steuerparameter mit der stärksten Auswirkung für das Training ist der Parameter `NTrees`. Hohe Werte für `NTrees` sorgten für ein besseres Ergebnis für die ROC-Kurve, jedoch trat Overtraining auf (siehe Abbildung 12). Die anderen Steuerparameter hatten eine geringere Auswirkung auf das Ergebnis. `MaxDepth` sollte nur nicht zu groß gewählt, da wieder Overtraining auftreten kann. Das Boosting ist bei kleineren Trees effektiver. Daher ist die Einteilung von `MaxDepth` in dem automatischen Optimierungsprozess gerecht-

fertigt. Dies trifft auch auf die Steuerparameter `AdaBoostBeta` und `MinNodeSize` zu. Jedoch erkennt man anhand der Einteilungen für `NTrees`, dass die Einteilung grob erfolgte. Daher wurde dieser Steuerparameter manuell verändert.

Ein besseres Ergebnis liefert so zum Beispiel bei `NTrees=50` (siehe Abbildungen 14).



(a) Overtrainingcheck mit den Einstellungen: `NTrees=50`, `AdaBoostBeta=0.4`, `MinNodeSize=1%`, `MaxDepth=4`.



(b) ROC-Kurve mit den Einstellungen: `NTrees=50`, `AdaBoostBeta=0.4`, `MinNodeSize=1%`, `MaxDepth=4`; ROC-Integral beträgt 0.953

Abbildung 14: Die Ergebnisse mit der Funktion `factory->OptimizeAllMethods` jedoch einer höheren Anzahl von Trees.

Im Gegensatz zu der Einstellung mit `NTrees = 10` erhält man in den Verteilungsfunktionen der BDT-Response eine kontinuierliche Funktion anstatt der vielen kleinen Peaks. Außerdem ist ein besseres Ergebnis bei einer erhöhten Anzahl an Trees zu erwarten. Bei 10 Trees kommt die Wirkung des Boostings nicht zum Einsatz. Das Boosting entfaltet erst bei einer höheren Anzahl an Trees die Separationskraft.

Da diese Einstellung bei `NTrees = 50` zudem ein besseres ROC-Integral ergibt, wird der Output des Trainings bei diesen Steuerparametern verwendet:

`NTrees=50, AdaBoostBeta=0.4, MinNodeSize=1%, MaxDepth=4`

Da das Training nun optimiert ist, lassen sich die verschiedenen Ergebnisse, die TMVA nach einem Training ausgibt, interpretieren. Zum einen gibt TMVA ein Ranking der im Training verwendeten Variablen nach ihrer Wichtigkeit für das Training an, welches im nächsten Kapitel erläutert wird.

4.6 BDT-Ranking

Der TMVA-Algorithmus bewertet zudem die Wichtigkeit der verwendeten Variablen und gibt ein sogenanntes Ranking an. Unter der BDT-Methode erfolgt dies unter Betrachtung der Häufigkeit der Variablen bei der Auswahl der Selektionen an jedem Knoten der Decision Trees. Im Folgenden wird diese Ordnung der Variablen nach der Wichtigkeit physikalisch erklärt. Die Ordnung der Wichtigkeit der Variablen ist für das optimierte Training, welches in Kapitel 4.5 diskutiert wurde, folgendermaßen gegeben:

Rank	Variable	Variable Importance
1	`BTags`	3,126e-01
2	`HT`	3,026e-01
3	`MHT`	2,789e-01
4	`Jets[2].Pt()`	1,740e-02
5	`Jets[1].Pt()`	1,323e-02
6	`NJets`	1,103e-02
7	`DeltaPhi2`	9,183e-03
8	`Jets[2].Eta()`	8,639e-03
9	`Jets[1].Eta()`	8,078e-03
10	`Jets[0].Eta()`	7,436e-03
11	`Jets[0].Pt()`	6,913e-03
12	`DeltaPhi1`	6,293e-03
13	`DeltaPhi3`	6,254e-03
14	`Jets[3].Pt()`	5,838e-03
15	`Jets[3].Eta()`	3,329e-03
16	`DeltaPhi4`	2,288e-03

Tabelle 12: Dargestellt ist die Wichtigkeit der Variablen für das optimierte Training bei der BDT-Methode (siehe Kapitel 4.5).

Man erkennt, dass die Wichtigkeit der ersten drei aufgelisteten Variablen um mehr als einen Faktor 10 höher ist als die anderen Variablen. Daher werden im Folgen-

den nur die ersten drei aufgelisteten Variablen diskutiert. Wie zu erwarten zeigt die Variable `BTags` eine hohe Separationskraft. Bei dem Signalprozess (siehe Abbildung 5) sind hauptsächlich nur vier `BTags` zu erwarten. Da die Untergrundprozesse mit ebenfalls Bottom-Quarks häufiger in einer niedrigeren Anzahl an Bottom-Quarks resultierten, ist dieses Ranking gerechtfertigt, weil es dann zu einer hohen Separationskraft in dieser Variablen kommt.

Nach dem Ranking haben H_{T} und $H_{\mathrm{T}}^{\mathrm{miss}}$ ebenfalls eine hohe Separationskraft. Weil die Neutralinos einen Anteil des Impulses der Gluinos erhalten und die Neutralinos nicht detektiert werden können, entsteht im Signalprozess fehlender Transversalimpuls $H_{\mathrm{T}}^{\mathrm{miss}}$. Die Untergrundprozesse erhalten auch fehlender Transversalimpuls durch nicht detektierte Neutrinos.

Die Variable H_{T} zeigt auch eine hohe Separationskraft. Im T1bbbb-Prozess sind vergleichsweise hohe Massen für die SUSY-Teilchen zu erwarten. Bottom-Quarks haben eine Masse von 4,18 GeV [20] und haben damit eine vergleichsweise kleine Masse, da die zu erwartenden SUSY-Teilchen-Massen in den Monte-Carlo-Simulation 100 GeV bzw. 900 GeV betragen. Aus diesem Grund ist zu erwarten, dass die SUSY-Teilchen einen hohen Anteil an Transversalimpuls der Gluinos erhalten. Dieser sorgt für höhere Werte von H_{T}.

Aufgrund dieses Rankings erkennt man, dass neben MHT auch HT und BTags Variablen mit hoher Separationskraft sind. Dies sind aber auch Variablen, die in der MHT-Analyse verwendet werden.

5 Resultate

Da das Training nun optimiert ist, sind nun die daraus resultierenden Ergebnisse zu berechnen. Im diesen Kapitel wird die aus der BDT-Analyse resultierende Variable - die BDT-Response - ausgelesen und die Separationskraft mit der MHT-Variable verglichen.

5.1 Daten

Weil ein trainierter Algorithmus vorhanden ist, können damit auch Datensätze mit unbekannter Klassifikation von Untergrund und Signal untersucht werden. Diese Datensätze werden im folgenden Daten genannt. Diese Daten stammen von einer Messreihe, die 2016 am LHC durchgeführt wurde. Die Luminosität beträgt bei diesen Daten $12,87\ \mathrm{fb}^{-1}$ [19].

5.2 Umgewichtung des BDT-Response

Da die Verteilungsfunktionen (siehe Abbildung 14a) so skaliert sind, dass die Verteilungsfunktionen vergleichbar sind, lässt sich die Separation der Signal- und Untergrundevents mithilfe der von TMVA gegebene Abbildung nicht deuten. Mit den Gewichten aus den Tabellen in Kapitel 4.1 wurden die Verteilung der BDT-Reponse gebildet. Da die Anzahl der Events in den Monte-Carlo-Simulationen willkürlich generiert wurden, ist die Gewichtung nötig, damit die verschienden Monte-Carlo-Simulationen vergleichbar werden. Das Ergebnis dieser Umgewichtung sieht folgendermaßen aus:

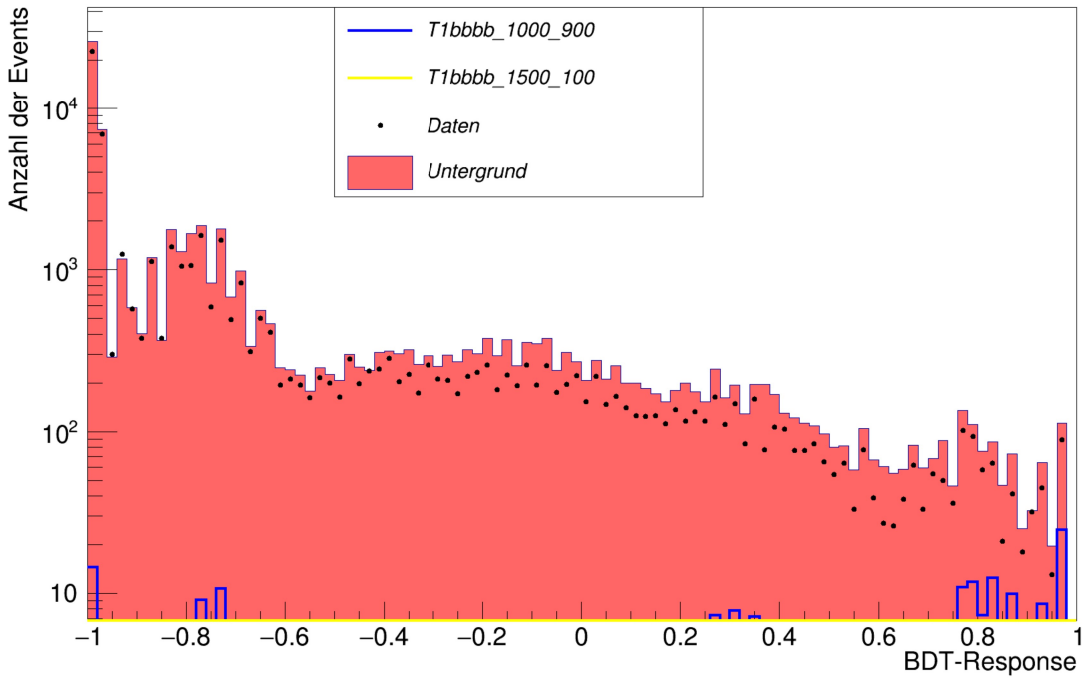
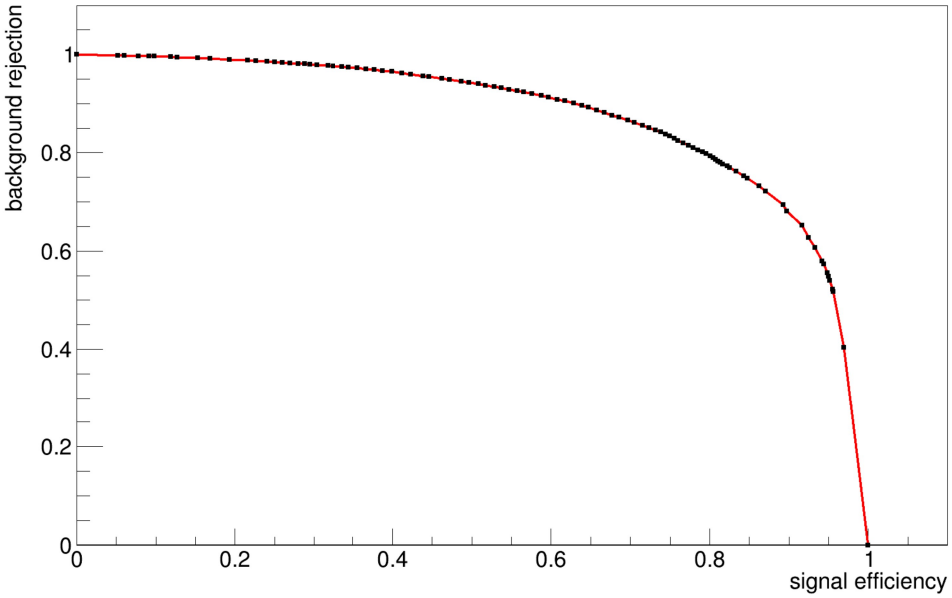
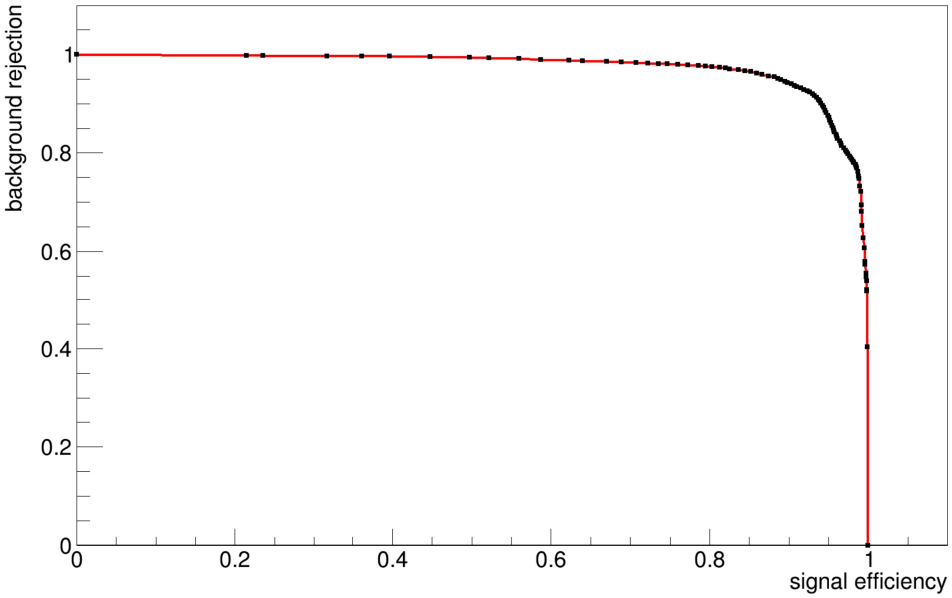


Abbildung 15: Darstellung der gewichteten BDT-Response. Die verschiedenen Untergrundtypen wurden zu einem Untergrund zusammen addiert.



(a) ROC-Kurven der in Abbildung 15 dargestellten Verteilung. Das Signal für die Berechnung der ROC-Kurve ist das mit kleinem Massensplit (siehe Tabelle 8).



(b) ROC-Kurve der in Abbildung 15 dargestellten Verteilung. Das Signal für die Berechnung der ROC-Kurve ist das mit hohen Massensplit (siehe Tabelle 8).

Abbildung 16: ROC-Kurven der in Abbildung 15 dargestellten Verteilung bei verschiedenen Signalen.

In der Abbildung 15 sind für die Untergrundevents die Datensätze aus den Tabellen 3 - 7 verwendet worden. Zudem ist die BDT-Response für Daten dargestellt. Zu erwarten ist, dass die Datenpunkte mindestens so groß sind wie die Untergrundverteilung, da die Daten aus Untergrundevents und möglicherweise Signalevents be-

stehen. Die generelle Verteilung der Datenpunkte stimmen mit dem Untergrund überein. Jedoch sind die Datenpunkte bis auf wenige Punkte stets kleiner als der Untergrund. Da die Anzahl der Events logarithmisch aufgetragen ist, ist die Differenz der Datenpunkt zum Untergrund höher als dies in der Abbildung zu erkennen ist.
Das, was nicht durch die von TMVA gegebene Darstellung der BDT-Response (siehe Abbildung 14a) erkenntlich war, ist, dass der Untergrund auch im hohen Bereich der BDT-Response die Signalevents deutlich überwiegen. Dies deutet auf eine nicht gelungene Separation hin. Die BDT-Response Verteilung des Signales mit dem hohen Massensplit ist in der Abbildung 15 aufgrund der Skala nicht erkennbar. Aufgrund des wesentlich kleineren Wirkungsquerschnittes des Signal mit dem hohen Massensplit zum Signal mit kleinem Massensplit (siehe Tabelle 8) ist diese Verteilung zu erklären. In den Abbildungen 16a und 16b sind die ROC-Kurven für jedes Signal einzeln berechnet worden. Man erkennt, dass die ROC-Kurve vom Signal mit hohen Massensplit eine bessere Performance liefert, da die ROC-Kurve kontinuierlich größer ist, als die ROC-Kurve vom Signal mit kleinem Massensplit. Da TMVA auch den optimalen Schnitt auf die BDT-Response angeben kann, wird dieser Schnitt im Folgenden analysiert.

5.3 Schnitt auf die BDT-Response

Um die erwünschte Separation zwischen Signalevents und Untergrundevents zu erreichen, muss auf die neue BDT-Variable (BDT-Response) geschnitten werden. Ein optimaler Schnitt ist dann erreicht, wenn die Signifikanz an diesem Schnitt am größten ist. TMVA bietet nach dem Training dafür ein Signifikanzplot, mithilfe dessen dieser optimale Schnittwert bestimmt werden kann. Dafür benötigt TMVA die gewichtete Anzahl der Signal- und Untergrundevents. Diese Zahl wurde mit den Wirkungsquerschnitten σ der Monte-Carlo-Simulationen (siehe Kapitel 4.1) bestimmt.

$$N_{\text{gew}} = \sum_{i=0}^{n} \sigma_i \quad i = \text{ Datensätze aus Kapitel 4.1 } . \tag{17}$$

Dabei ist die gewichtete Anzahl auf eine Luminosität L von 1 pb^{-1} normiert. Dies ist gerechtfertigt, da für den Signifikanzplot nur das Verhältnis zwischen Signal- und Untergrundevents wichtig ist. So ergibt sich ein optimaler Schnittwert auf die BDT-Response bei **0,9715**. Der Signifikanzplot mit gewichteter Anzahl der Events mit dem optimalen Schnittwert ist in Abbildung 17 gegeben.

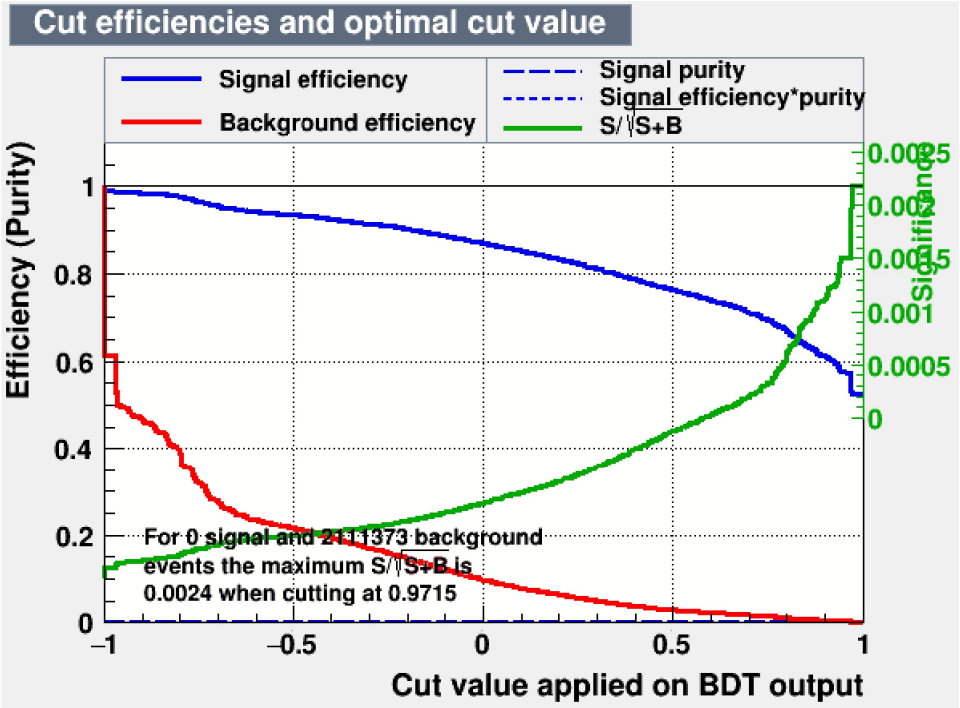


Abbildung 17: Signifikanz und Effizienzen für das optimierte Training und angegeben optimalen Schnittwert bei einer gewichteten Anzahl von Events.

Nun folgt ein gewichtetes Histogramm von MHT mit implementierten BDT-Schnitt bei 0,9715. Also werden nur die Events betrachtet die eine BDT-Response aufweisen von

$$\text{BDT Response} = 0,9715 \tag{18}$$

Da die Daten und die Monte-Carlo-Simulationen auf verschiedene Luminositäten normiert sind, muss für ein Vergleich eine Umgewichtung stattfinden. Die Daten sind dabei auf eine Luminosität von 12,87 fb^{-1} und die Monte-Carlo-Simulationen auf eine Luminosität von 1 pb^{-1} normiert. Also findet für folgende Plots eine Skalierung von 12870 auf den Monte-Carlo-Simulationen statt.

Bei der Betrachtung der Abbildung 14 fällt auf, dass dieser berechnete Schnittwert mit den Erwartungen übereinstimmt. Bei einem Schnitt auf die BDT-Response bei 0,9715 ist zu erwarten, dass hauptsächlich Untergrundevents weggeschnitten werden und die Anzahl der Signalevents sich nicht stark reduziert.

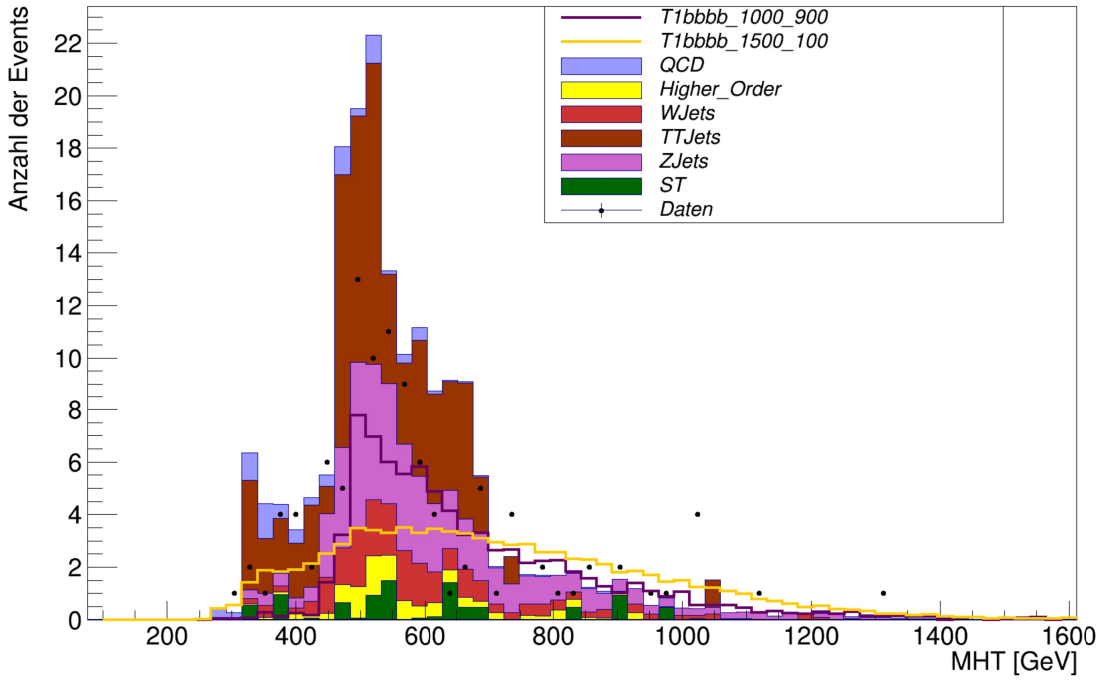


Abbildung 18: Fehlende transversale Impuls (MHT) bei einem Schnitt auf der BDT-Response bei 0,9715. Die Farben stellen je ein Untergrundtypen dar und die Signalevents sind mit Linien gekennzeichnet.

An der Abbildung 18 sieht man, dass im hohen MHT-Bereich ab 700 GeV die Signalevents die Untergrundevents dominieren. Dies verhält sich auch bei den Datenevents so. Aufgrund dieser Erkenntnisse zeigt sich, dass die BDT-Response hauptsächlich Untergrundevents im hohen MHT-Bereich von den Signalevents separiert hat. Da das Signal mit dem hohen Massensplit zwischen Gluon und Neutralino hohe Werte für MHT beinhaltet, sind diese Signalevents vergleichsweise zu den Signalevents mit kleinen Massensplit stärker separiert worden. Die aus diesen Verteilungen resultierenden Signaleffizienzen und obere Grenzen für die Anzahl der Signalevents in den Daten (Limits) werden im Kapitel 5.5 diskutiert.

5.4 Vergleich mit der MHT-Analyse

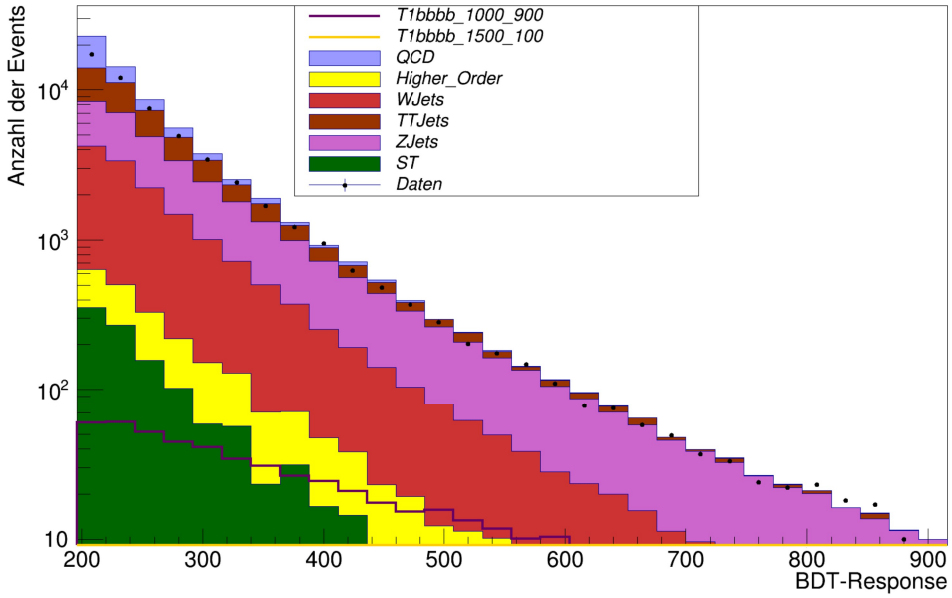
Nun stellt sich die Frage ob die BDT-Response eine eine bessere Separation liefert als die MHT-Variable. Daher werden die Resultate der BDT-Analyse mit der MHT-Analyse verglichen. Als MHT-Analyse wird im Folgenden eine Analyse über einfache eindimensionale Schnitte auf der MHT-Variablen bezeichnet. Bei der Analyse der CMS-Kollaboration SUS-16-014 werden weitere Vorselektionen durchgeführt [14], um die signalähnliche Events weiterhin einzugrenzen. Die Vorselektionen werden auch auf die MHT-Analyse angewendet und werden im folgenden Baseline-Schnitte genannt und sind wie folgt definiert.

- $N_{\text{jet}} \geq 4$, wobei die Jets die Bedingung $|\eta| < 2{,}4$ erfüllen müssen. Diese Selektion ist durch die Betrachtung des T1bbbb-Prozesses gerechtfertigt.

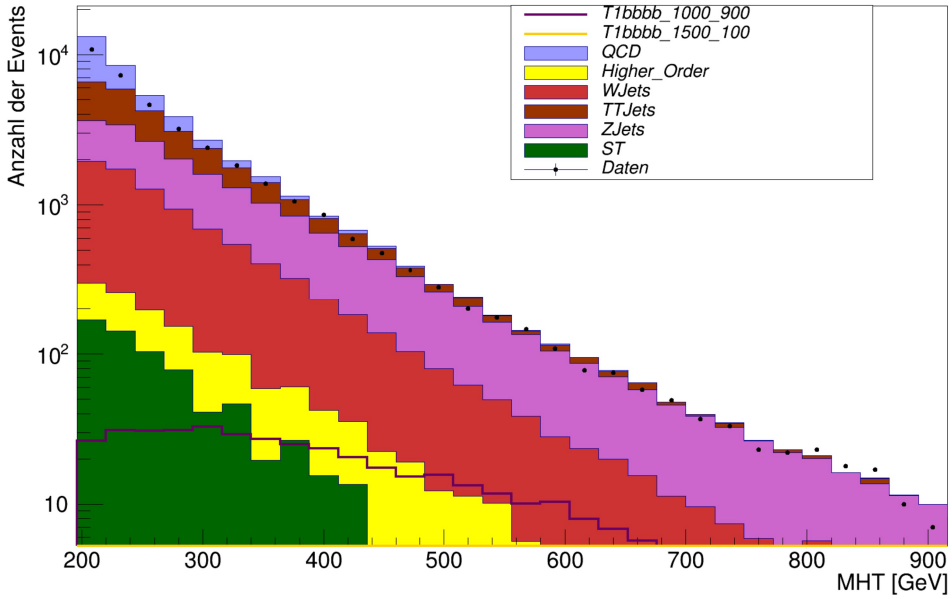
- $H_{\text{T}} > 200$ GeV, wobei die Jets die Bedingung $|\eta| < 2{,}4$ erfüllen müssen.

- $\Delta\Phi_{H_{\text{T}}^{\text{miss}},j_i} > 0,5 (> 0,3)$ für die zwei höchsten p_{T} Jets j_1 und j_2 (für die nächsten zwei höchsten p_{T} Jets j_3 und j_4).

Um die BDT-Analyse mit der MHT-Analyse vergleichen zu können werden im Folgenden MHT-Plots diskutiert. Als erstes folgt ein MHT-Plot (siehe Abbildung 19a) nur mit den Vorselektionen auf den Monte-Carlo-Simulationen als Vergleichsplot(siehe Kapitel 4.2). Zudem ist die dazugehörige ROC-Kurve für Schnitte nur auf der MHT-Variablen für jedes Signal in Abbildungen 20 dargestellt, um dies mit der BDT-Analyse (siehe Abbildungen 16) vergleichen zu können.

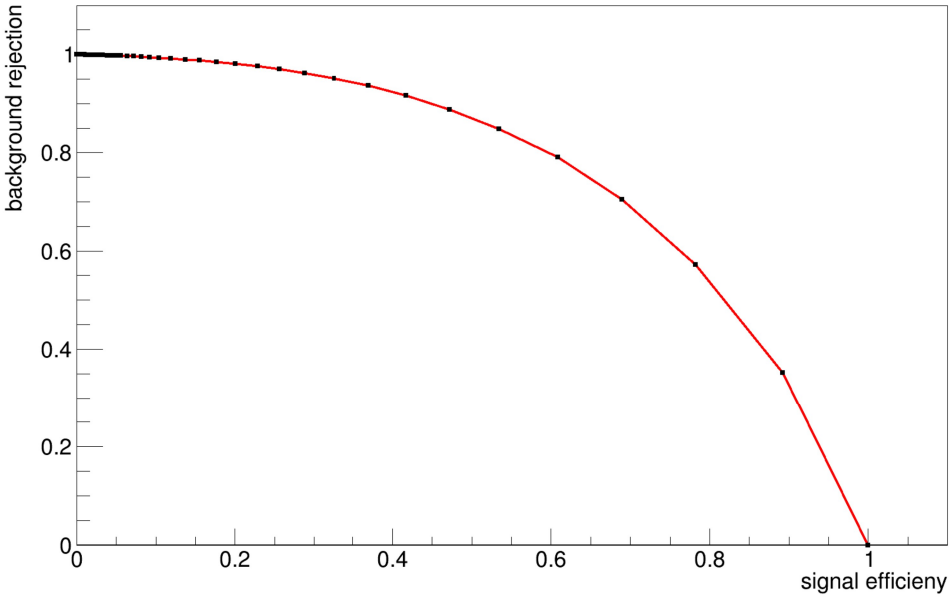


(a) Fehlender transversaler Impuls (MHT) dargestellt nur mit den Vorselektionen aus Kapitel 4.2 in den Monte-Carlo-Simulation.

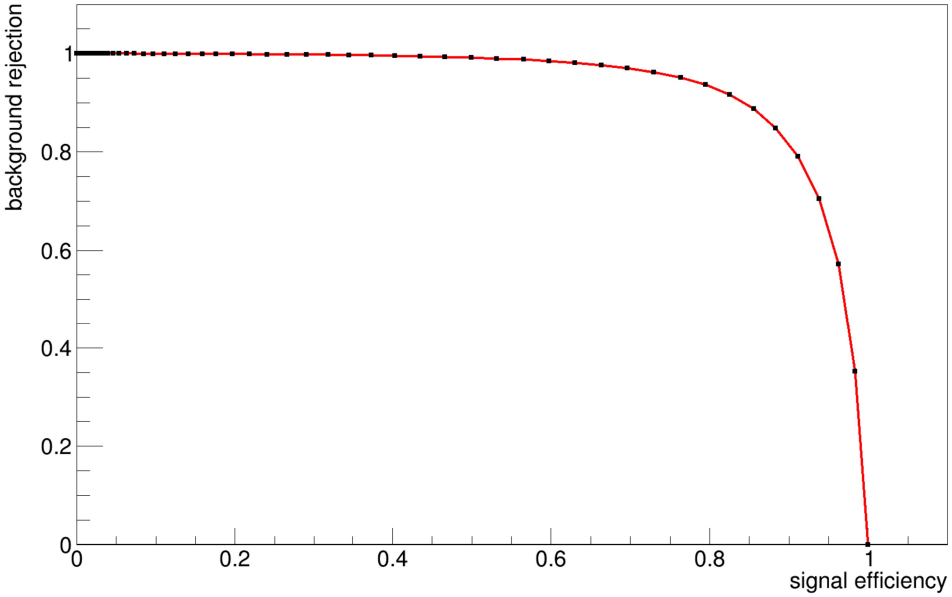


(b) Fehlender transversaler Impuls (MHT) mit den Baseline-Selektionen.

Abbildung 19: Fehlender transversaler Impuls (MHT) dargestellt bei verschiedenen Selektionen. Die Farben stellen je ein Untergrundtypen dar und die Signalevents sind mit Linien gekennzeichnet.



(a) ROC-Kurve für das Signal mit kleinem Massensplit (siehe Tabelle 8).

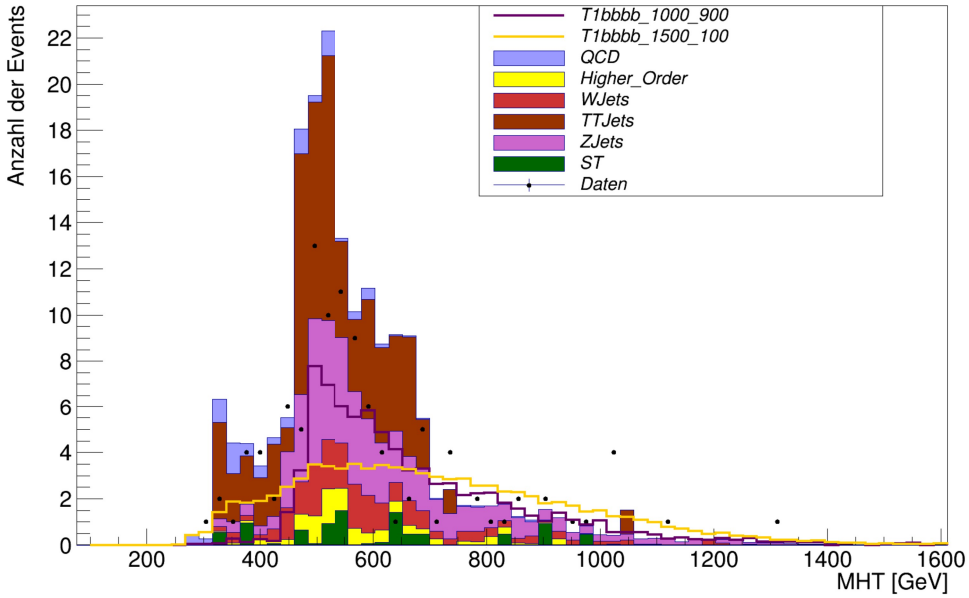


(b) ROC-Kurve für das Signal mit großen Massensplit (siehe Tabelle 8).

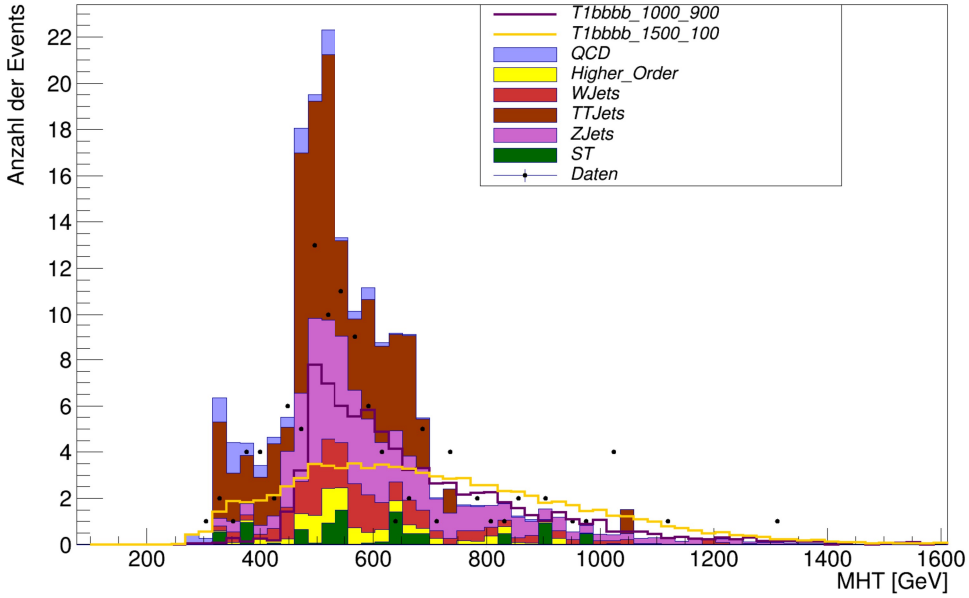
Abbildung 20: ROC-Kurve der MHT-Analyse nur mit Vorselektionen (siehe Abbildung 19a) bei verschiedenen Signalen.

Erkennbar ist, dass die Untergrundverteilung von MHT die Verteilung des Signales deutlich überwiegen. Nur im hohen MHT-Bereich fallen die Anzahl der Untergrundereignisse auf sehr kleine Werte und das Signal überwiegt in diesem Bereich. Anhand dieser Abbildung lässt sich zudem die Wichtigkeit der verschiedenen Untergrundprozesse abschätzen. Der QCD Untergrund nimmt den größten Anteil des

Untergrund dar. Außerdem hat der Z+Jets-Untergrund einen großen Anteil des Untergrundes. Weniger wichtig sind dagegen die Higher-Order (siehe Tabelle 7) Prozesse, die auch in gemessenen Daten seltener auftreten. Es lässt sich vermuten, dass das Signal mit einem hohen Massensplit zwischen Gluino und Neutralino in der MHT-Variable im hohen MHT-Bereich vom Untergrund separiert ist. Dies ist an dieser Abbildung aufgrund der hohen Anzahl an Untergrundevents kaum erkennbar. Dies zeigt sich aber durch die dazugehörige ROC-Kurve. Im Gegensatz dazu überwiegen beim Untergrund die Signalevents mit kleinen Massensplit. Da die Energie des Gluons von 1000 GeV hauptsächlich an das Neutralino mit 900 GeV übergeben wird, ergeben sich für MHT nur kleinere Werte als in dem Signal mit hohen Massensplit. Dies entspricht der Separation nach dem Schnitt auf die BDT-Response in der Abbildung 18. Zudem ist der Wirkungsquerschnitt des Signales mit hohem Massensplit wesentlich kleiner als der Wirkungsquerschnitt des Signal vom kleinen Massensplit (siehe Tabelle 8). Daher erkennt man keine Ausprägung des Signales mit hohen Massensplit in der Abbildungen 19.
Die Abbildung 19b zeigt dagegen den fehlenden transversalen Impuls MHT mit den zuvor diskutierten Baseline-Schnitten.
Bei einem Vergleich der MHT-Verteilung mit implementierten Baseline-Selektionen (siehe Abbildung 19 oder 21) mit der MHT-Verteilung ohne weitere Selektionen erkennt man, dass die Baseline-Selektionen vergleichsweise wenige Events weggeschnitten hat. Das Besondere ist jedoch, dass der Datenpunkt bei einem MHT von 1100 GeV (siehe Abbildung 21) eine erhöhte Ausprägung hat. Dies spricht für signalähnliche Events in den Daten bei hohen Werten von MHT. Jedoch ist die Anzahl der Datenevents an diesem Punkt 4. Dies könnten auch statistische Fluktuationen sein, da die übrigen Datenevents nicht gut zum Untergrund passen.



(a) Fehlender transversale Impuls (MHT) bei einem Schnitt auf der BDT-Response bei 0,9715.



(b) Fehlender transversale Impuls (MHT) mit den Baseline-Selektionen und bei einem optimalen Schnitt auf der BDT-Respone bei 0,9715.

Abbildung 21: Fehlender transversale Impuls (MHT) bei verschiedenen Selektionen. Die Farben stellen je ein Untergrundtypen dar und die Signalevents sind mit Linien gekennzeichnet.

Da sich die Verteilungen und Anzahl der Events zwischen den Abbildungen 19b und 21a stark ändert, ist klar, dass der Schnitt auf die BDT-Response mehr Events weggeschnitten hat als die Baseline-Selektionen. Bei der Betrachtung der Abbildung 19b, in dem nur die Baseline-Selektionen durchgeführt wurden, bestätigt sich diese

Vermutung, da die Anzahl der Events von Untergrund und Signal in der Baseline-Selektionen höher ist als die Anzahl bei dem BDT-Schnitt.

Insgesamt erhalten wir das Resultat, dass der BDT-Algorithmus hauptsächlich die Signalevents bei hohen Werten von MHT von den Untergrundevents separiert ist.

Bei einem Vergleich der ROC-Kurven der BDT-Analyse in Abbildung 16 mit den ROC-Kurven der MHT-Analyse in Abbildung 20 erkennt man, dass die Performance der BDT-Analyse generell besser ist als die MHT-Analyse, da die ROC-Kurve der BDT-Analyse stets größer ist als die der MHT-Analyse. Damit zeigt sich, dass die Multivariate Analyse eine bessere Performance liefert als einfache Schnitte auf der MHT-Variablen.

5.5 Limitsberechnung

Mithilfe dieser Ergebnisse lassen sich zudem Limits für die Anzahl der Signalevents in den Monte-Carlo-Simulationen bestimmen. Diese Limits geben die maximale Anzahl der Signalevents in den Monte-Carlo-Simulationen wieder. Im folgenden Kapitel werden die Limits für verschiedene Schnittwerte berechnet und diskutiert. Dafür wird die Anzahl der Datenevents nach den jeweiligen Schnitten S sowie die Anzahl der Untergrundevents mit jeweiligen Fehler benötigt. Für den Fehler wird eine Poisson-Verteilung für die Untergrundevents angenommen, weswegen sich der Fehler mit

$$\Delta N = \sqrt{N} \tag{19}$$

berechnen lässt. Diese Limits werden bei verschiedenen Schnitten auf MHT und gegebenfalls auf die BDT-Response R_{BDT} berechnet. MHT-Schnitte wurden hinzugefügt, um dies mit den Ergebnissen der SUS-16-014 Analyse vergleichen zu könenn. Die Berechnung der Limits erfolgt über ein Tool namens SModelS, welches ein Framework namens MadAnalysis5 verwendet [21]. Dies wurde entwickelt um Ausschlussgrenzen zu bestimmen. Die Limits werden für die Signalevents berechnet und geben die maximale Anzahl von Signalevents in den Daten an. Dabei wird ein Konfidenzniveau von 95%[16] angenommen. Zudem wird die Signaleffizienz bei gegebenen Schnitten berechnet. Daraus lassen sich die Limit für den Wirkungsquerschnitt $\sigma_{\mathrm{signal}}^{\mathrm{Limit}}$ bestimmen mit:

[16]Dies ist bei vergleichbaren Experimenten auch der Fall.

$$\sigma_{\text{signal}}^{\text{Limit}} = \frac{N_{\text{signal}}^{\text{Limit}}}{L \cdot \epsilon_{\text{signal}}} \ . \tag{20}$$

Dabei gibt $N_{\text{signal}}^{\text{Limit}}$ das Limit für die Anzahl der Signalevents in den Daten an und ϵ_{signal} die Signaleffizienz.

Die Ergebnisse sind in der Tabelle 14 dargestellt.

$H_{\text{T}}^{\text{miss}}$ [GeV]	R_{BDT}	$N_{\text{Untergrund}}$	N_{Daten}	$N_{\text{signal}}^{\text{Limit}}$	ϵ_{signal}	$\sigma_{\text{signal}}^{\text{Limit}}$ [fb]
-	$>$ 0,9715	132,682 ± 11,519	102	17,744	0,060	22,585
$>$ 300	-	5400,49 ± 73,49	5069	89,780	0,417138	16,723
$>$ 500	-	804,282 ± 28,360	768	57,515	0,1554	28,758
$>$ 750	-	133,079 ± 11,536	134	33,847	0,0441881	59,516
$>$ 800	-	95,753 ± 9,785	93	26,454	0,03301	62,268
$>$ 300	$>$ 0,9715	167,736 ± 12,951	103	3,368	0,995252	0,263
$>$ 500	$>$ 0,9715	76,593 ± 8,752	48	11,312	0,673119	1,306
$>$ 750	$>$ 0,9715	14,923 ± 3,863	14	11,303	0,218042	4,028
$>$ 800	$>$ 0,9715	11,467 ± 3,386	12	11,253	0,167275	5,227
$>$ 900	$>$ 0,9715	6,420 ± 2,534	8	9,832	0,10263	7,444

Tabelle 13: Die Anzahl der Untergrundevents und die Anzahl der Datenevents nach bestimmten Schnitten mit berechneten Limit für die Signalevents mit kleinen Massensplit und Wirkungsquerschnitt. Die angegebenen Fehler bei der Anzahl der Untergrundevents sind statistische Fehler bei der Annahme einer Poisson-Verteilung.

$H_{\text{T}}^{\text{miss}}$ [GeV]	R_{BDT}	$N_{\text{Untergrund}}$	N_{Daten}	$N_{\text{signal}}^{\text{Limit}}$	ϵ_{signal}	$\sigma_{\text{signal}}^{\text{Limit}}$ [fb]
-	$>$ 0,9715	132,682 ± 11,519	102	17,744	0,23541	5,857
$>$ 300	-	5400,49 ± 73,49	5069	89,780	0,825457	8,451
$>$ 500	-	804,282 ± 28,360	768	57,515	0,56557	7,902
$>$ 750	-	133,079 ± 11,536	134	33,847	0,265728	9,897
$>$ 800	-	95,753 ± 9,785	93	26,454	0,218411	9,411
$>$ 300	$>$ 0,9715	167,736 ± 12,951	103	3,368	0,949542	0,276
$>$ 500	$>$ 0,9715	76,593 ± 8,752	48	11,312	0,6963	1,262
$>$ 750	$>$ 0,9715	14,923 ± 3,863	14	11,303	0,327153	2,685
$>$ 800	$>$ 0,9715	11,467 ± 3,386	12	11,253	0,269344	3,246
$>$ 900	$>$ 0,9715	6,420 ± 2,534	8	9,832	0,174415	4,380

Tabelle 14: Die Anzahl der Untergrundevents und die Anzahl der Datenevents nach bestimmten Schnitten mit berechneten Limit für die Signalevents mit hohen Massensplit und Wirkungsquerschnitt. Die angegebenen Fehler bei der Anzahl der Untergrundevents sind statistische Fehler bei der Annahme einer Poisson-Verteilung.

Zu erwarten ist, dass die Anzahl der Datenevents stets höher ist als die Anzahl der Untergrundevents, da die Daten Untergrundevents und Signalevents wiederspiegeln.

Dies ist bei den durchgeführten Schnitten nicht immer der Fall. Daher lässt sich vermuten, dass die Vorhersage der Untergrundkomposition nicht optimal gewählt worden ist. Zudem wurden für die Abschätzung des Untergrundes Monte-Carlo-Simulationen verwendet, die die Untergrundprozesse nicht optimal wiedergeben. Die größte relative Differenz ergibt sich bei einem Schnitt bei $H_{\mathrm{T}}^{\mathrm{miss}} > 500$ und BDT-Response $> 0{,}9715$. Diese Differenz ist jedoch unsignifikant und eventuell durch statistische Fluktuationen hervorgerufen. Ein Überschuss an Untergrundevents ist auch in den Daten der MHT-Analysegruppe zu finden[17]. Anhand der Limits auf den Wirkungsquerschnitten erkennt man zudem, dass der BDT-Schnitt bei 0,9715 eine größere Einschränkung für den Wirkungsquerschnitt hervorruft.

6 Zusammenfassung

Das Ziel dieser Arbeit ist es, die eine Analyse mit der Variablen MHT zur Suche im T1bbbb-Prozess mithilfe einer multivariaten Analyse zu verbessern. Die Arbeit der SUS-16-014-Analyse behandelt dabei hauptsächlich den fehlenden transversalen Impuls MHT, die Anzahl der Bottom Jets BTags sowie die skalare Summe der Jets HT in deren Analyse, indem die Analyse jeweils bin-weise geschieht. Im Gegensatz dazu werden in einer multivariaten Analyse die Daten in Kategorien in diesen Variablen unterteilt, die mehrere Variablen zugleich berücksichtigen. Es konnte gezeigt werden, dass die multivariate Analyse bessere Ergebnisse liefert als die Analyse über einfache Schnitte auf die MHT-Variable. Jedoch werden in dieser Arbeit Monte-Carlo-Simulationen verwendet, die die richtigen Klassifikationen nicht optimal wiedergeben. Daher ist diese Analyse nicht mehr der SUS-16-014-Analyse zu vergleichen. Die SUS-16-014-Analyse schneidet dreidimensional auf die erwähnten Variablen und benutzt Kontrollbereiche in den Datensätzen zur Abschätzung von Untergrund- und Signalevents. Im Folgenden wird beurteilt wie die Analyse dieser Arbeit abgelaufen ist. Für die multivariate Analyse wurde ein optimierter Algorithmus benötigt. Mithilfe der Verteilungsfunktionen der BDT-Response ließ sich so die Separationkraft der Observablen abschätzen. Diese Separationskraft wurde anhand eines Rankings der Observablen von TMVA nach dem Training überprüft. Bei dieser Optimierung wurde die Auswahl der verwendeten Observablen und die Steuerparameter des BDT Algorithmus berücksichtigt. Daraufhin wurde die Separationskraft der BDT-Analyse mit der Separationskraft einer MHT-Analyse verglichen. Dabei zeigte sich eine bessere Performance bei beiden Signalpunkten bei der BDT-Analyse. Dies deutet auf eine Optimierung der Analyse im Vergleich mit einer Analyse mit der MHT-Variablen. Schließlich wurden in Kapitel 5.5 für die beiden Massenpunkte in den Signalen Limits berechnet, welche als Vergleichswerte weiterer Analysen

[17] Diese Information stammt von einer CMS-internen Quelle.

benutzt werden können. Die berechneten Limits dieser Analyse lassen sich mit den Limits der SUS-16-014 Analyse vergleichen. Der mit der SUS-16-014 Analyse ermittelter Limit beträgt bei einer Gluinomasse von 1500 GeV bzw. 1000 GeV [22]:
$\sigma^{\text{Limit}}_{\text{signal}/1500} = 14,19 \text{ fb} \pm 22,7296\%$ bzw.
$\sigma^{\text{Limit}}_{\text{signal}/1000} = 325,388 \text{ fb} \pm 16,758\%$
Die in dieser Analyse berechneten Limits sind deutlich geringer als die von der SUS-16-014 berechneten Limits. Dies deutet auf eine erhöhte Selektion in der BDT-Analyse.

7 Ausblick

Diese Arbeit bietet mehrere Optimierungsmöglichkeiten und Ansätze zur Weiterführung der Analyse, welche in diesem Kapitel diskutiert werden. Bei der Auswahl der verwendeten Variablen bestehen Optimierungsmöglichkeiten. Diese Auswahl erfolgte durch einen Vergleich der Untergrundverteilungen mit den Signalverteilungen in diesen Variablen. Die hierfür betrachteten Variablen wurden anhand der bisherigen Veröffentlichungen der CMS-Kollaboration und durch physikalischen Betrachtungen ausgewählt. Natürlicherweise wäre eine Betrachtung von noch mehr Variablen in Zukunft möglich.
Eine weitere Optimierungsmöglichkeit besteht in der verwendeten Methode. Die Auswahl der Methode erfolgte durch einen ROC-Kurven-Vergleich bei Standardeinstellungen in den Steuerparameter der jeweiligen Methode. Optimaler Weise hätte man jede einzelne Methode optimiert und dann ROC-Kurven vergleichen müssen. Da jedoch die automatische Optimierung durch TMVA nur in BDT implementiert war, hätte dies den Rahmen dieser Arbeit überschritten. Interessant wäre eine Analyse der RuleFit Methode, da diese Regeln für den Benutzer angibt, wie auf den Variablen geschnitten werden sollen. Diese Methode ist so durchsichtiger als die BDT-Methode, da die genaue Komposition der BDT-Response unklar ist.
Weiterhin wäre eine Optimierung dieser Arbeit durch das Einbinden weiterer Datensätze möglich gewesen. Für die Analyse hätten mehr Monte-Carlo-Simulationen verwendet werden können. Dies hätte eine bessere Monte-Carlo-Statistik zur Folge gehabt. Das Auslesen der BDT-Response erfolgte jedoch mit einer sehr langen Laufzeit. Beispielsweise hat das Auslesen des größten Datensatzes eine Woche gebraucht.
Fortführen kann man diese Arbeit durch Betrachtungen weitere Signalpunkte. In dieser Analyse wurden nur zwei Massenpunkte betrachtet (kleiner Massensplit und hoher Massensplit zwischen den supersymmetrischen Teilchen). Durch Betrachtung weiterer Massenpunkten ließen sich so Limits für viele Massenpunkte berechnen und so die Masse der Supersymmetrieteilchen im T1bbbb-Zerfall weiterhin eingrenzen. Außerdem wurde in der Analyse nicht betrachtet, ob das Training für jeden einzelnen Signalpunkt einzeln eine bessere Performance insgesamt geliefert hätte als beide Signalpunkte zusammen. Zudem wäre eine Einbindung von komplizieren Signalmodellen eine Möglichkeit diese Analyse fortzusetzen. In dieser Analyse wurden Monte-Carlo-Simulationen mit vereinfachenden Modellen T1bbbb für das Signal verwendet. Ein kompliziertes Modell für das Signal sind beispielsweise pMSSM-Punkte[18]. Dieses Model geht von weiteren Annahmen aus, die die Anzahl der zum Standardmodel zusätzlichen Parameter auf 19 reduziert und somit die Suche nach Supersymmetrie-

[18]Phenomenological Minimal Supersymmetric Standard Model

teilchen vereinfachten, falls die Annahmen sich als richtig erweisen.

Literatur

[1] Ian Aitchison. *Supersymmetry in Particle Physics (An Elementary Introduction)*. Cambridge University Press, 2007.

[2] Dr.Christian Autermann; Prof. Dr. Caren Hagner. Universität Hamburg Vorlesung: Kern- und Teilchenphysik, 2015.

[3] https://de.wikipedia.org/wiki/Standardmodell#/media/File:Standard_Model_of_Elementary_Particles-de.svg.

[4] Martin Reitz. Einführung in die Supersymmetrie (SUSY). http://desy.de/~troms/teaching/WiSe1011/talks/Reitz_SUSY_handout.pdf, 2010.

[5] Peter Schleper. Physik jenseits des Standard-Modells Supersymmetrie. http://www.desy.de/~schleper/lehre/BSM/SS_2007/Supersymmetrie.pdf, 2007.

[6] http://particleadventure.org/german/frameless/images/merging_forces.gif.

[7] https://www.lhc-facts.ch/index.php?page=cms.

[8] http://weltderphysik.de/gebiet/teilchen/experimente/teilchenbeschleuniger/lhc/lhc-im-ueberblick/.

[9] http://www.weltmaschine.de/sites/site_weltmaschine/content/e161/e570/9906026-A4-at-144-dpi%5B1%5D.jpg.

[10] http://home.cern/about/experiment/cms.

[11] https://www.phys.ksu.edu/reu2014/wabehn/index_files/image005.gif.

[12] http://weltderphysik.de/gebiet/teilchen/experimente/teilchenbeschleuniger/lhc/lhc-experimente/cms/cms-detektor.

[13] http://arxiv.org/pdf/1403.2286.pdf.

[14] The CMS Collaboration. Search for supersymmetry in the multijet and missing transverse momentum final state in pp collisions at 13 TeV. http://arxiv.org/pdf/1602.06581v2.pdf, 2016.

[15] http://twiki.cern.ch/twiki/pub/CMSPublic/SUSYSMSSummaryPlots7TeV/T1bbbb_feyn.pdf.

[16] Simon Kurz. Susy searches with jets and missing transverse momentum: Lost lepton background and results with 12.9/fb data.

[17] A. Hoecker; P. Speckmayer; J. Stelzer; J. Therhaag; E. von Toerne; H. Voss. Tmva 4 toolkit for multivariate data analysis with root users guide. `http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf`, 2013.

[18] Olaf Behnke, Kevin Kröninger, Thomas Schörner-Sadenius, and Gregory Schott, editors. *Data analysis in high energy physics*. Wiley-VCH, Weinheim, Germany, 2013.

[19] `https://twiki.cern.ch/twiki/bin/viewauth/CMS/RA2b13TeVProduction`.

[20] Particle Data Group. `http://pdg.lbl.gov/2014/tables/rpp2014-sum-quarks.pdf`.

[21] `http://smodels.hephy.at/wiki`.

[22] `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/SUSYCrossSections13TeVgluglu`.

Abbildungsverzeichnis

1 Standardmodell der Elementarteilchen [3]. 4
2 Vereinigung der drei fundamentalen Kräfte unter Berücksichtigung der Supersymmetrie bei hohen Energien [6]. 7
3 Eine schematische Skizze vom LHC mit den größten Experimenten CMS, ATLAS, ALICE, LHC-B [9]. 8
4 Eine schematische Skizze vom CMS-Experiment. Von links nach rechts: Innere Spurdetektor, Elektromagnetische Kalorimeter, Hadronische Kalorimeter, Solenoid-Magnet, Myonenspektrometer [11]. 9
5 Feynman-Diagramm für den T1bbbb-Prozess [15]. 13
6 Feynman-Diagramm für einen Top-Antitop-Prozess mit Bottom-Jets. 14
7 ROC-Kurve bei einer Auswahl von verschiedenen Methoden (siehe Kapitel 4.4. Diese Abbildung wurde mit TMVA erstellt. 17
8 Dargestellt sind die skalierten Verteilungsfunktionen verschiedener Variablen zur Beurteilung der Separationskraft. Die Bedeutung der Notation der Variablen ist in Tabelle **??** gegeben. In blau ist die Verteilung der Signalevents und in rot die Verteilung der Untergrundevents angegeben. 27
9 Dargestellt sind die skalierten Verteilungsfunktionen verschiedener Variablen zur Beurteilung der Separationskraft. Die Bedeutung der Notation der Variablen ist in Tabelle **??** gegeben. In blau ist die Verteilung der Signalevents und in rot die Verteilung der Untergrundevents angegeben. 28
10 Dargestellt sind die Verteilungsfunktionen der Variablen für das Training. Diese Funktionen sind skaliert, sodass die Separationskraft beurteilt werden kann. Die Bedeutung der Notation der Variablen ist in Tabelle **??** gegeben. 30
11 ROC-Kurve bei einer Auswahl von verschiedenen Methoden 31
12 Ein Beispiel für Overtraining: Dargestellt ist die BDT-Response mit dem Steuerparameter `NTrees` = 1000. 33
13 Die Ergebnisse mit der Funktion `factory->OptimizeAllMethods` bestimmen optimierten Werte. 34
14 Die Ergebnisse mit der Funktion `factory->OptimizeAllMethods` jedoch einer höheren Anzahl von Trees. 35
15 Darstellung der gewichteten BDT-Response. Die verschiedenen Untergrundtypen wurden zu einem Untergrund zusammen addiert. . . . 38
16 ROC-Kurven der in Abbildung 15 dargestellten Verteilung bei verschiedenen Signalen. 39

17 Signifikanz und Effizienzen für das optimierte Training und angegeben optimalen Schnittwert bei einer gewichteten Anzahl von Events. . . . 41
18 Fehlende transversale Impuls (MHT) bei einem Schnitt auf der BDT-Response bei 0,9715. Die Farben stellen je ein Untergrundtypen dar und die Signalevents sind mit Linien gekennzeichnet. 42
19 Fehlender transversaler Impuls (MHT) dargestellt bei verschiedenen Selektionen. Die Farben stellen je ein Untergrundtypen dar und die Signalevents sind mit Linien gekennzeichnet. 44
20 ROC-Kurve der MHT-Analyse nur mit Vorselektionen (siehe Abbildung 19a) bei verschiedenen Signalen. 45
21 Fehlender transversale Impuls (MHT) bei verschiedenen Selektionen. Die Farben stellen je ein Untergrundtypen dar und die Signalevents sind mit Linien gekennzeichnet. 47

Tabellenverzeichnis

1 Für die Optimierung wichtige BDT-Steuerparameter mit einem Standardwert und einer Kurzbeschreibung. 20

2 Verwendete Monte-Carlo-Simulationen für den QCD-Untergrund. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator `MADGRAPH5_aMC@NLO 2.2.2` verwendet [14]. 22

3 Verwendete Monte-Carlo-Simulationen für den W+Jets-Untergrund. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator `MADGRAPH5_aMC@NLO 2.2.2` verwendet [14]. 22

4 Verwendete Monte-Carlo-Simulationen für den Z+Jets-Untergrund. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator `MADGRAPH5_aMC@NLO 2.2.2` verwendet [14]. 23

5 Verwendete Monte-Carlo-Simulationen für den Single-top Untergrund. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator `MADGRAPH5_aMC@NLO 2.2.2` verwendet [14]. 23

6 Verwendete Monte-Carlo-Simulationen für den TT+Jets-Untergrund. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator `MADGRAPH5_aMC@NLO 2.2.2` verwendet [14]. 23

7 Verwendete Monte-Carlo-Simulationen für die Untergründe (Higher_Order). Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Für die Simulation wurde der Generator `MADGRAPH5_aMC@NLO 2.2.2` verwendet [14]. 24
8 Verwendete Monte-Carlo-Simulationen für den T1bbbb-Prozess. Diese Anzahl der Events ist die Anzahl nach den Vorselektionen aus der Tabelle 10. Die Werte für das Gewicht und die Anzahl der Events sind direkt aus den Dateien übernommen. Der Wirkungsquerschnitt stammt von einer CMS-internen Webseite. Die erste Zahl 1500 bzw 1000 steht dabei für die Gluinomasse in [GeV] und die zweite Zahl 100 bzw. 900 für die Masse des Neutralinos [GeV]. Für die Simulation wurde der Generator `MADGRAPH5_aMC@NLO 2.2.2` verwendet [14]. . . . 24
9 Aufgelistet sind die Selektionen mit einer für diese Analyse passende Anwendbarkeitserkärung. 25
10 Auflistung der in den Datensätzen (siehe Kapitel 4.1) enthaltenen Vorselektionen. Die linke Spalte zeigt die von der MHT-Analyse-Gruppe verwendete Notationen und die rechte Spalte die konkrete Selektion. 25
11 Variablen, die für die Trainingsprozesse und Applikationsprozesse verwendet wurden. Die linke Spalte sind die Variablen und physikalischer Notation und die rechte Spalte ist die Notation, welche im TMVA benutzt wurden. Dies entspricht auch die Notation, welche in der SUS-16-014 Analyse verwendet wird. Da die Jets vom größten transversalen Impuls zum kleinsten transversalen Impuls sortiert sind, ist die Nummerierung der Jets in der Tabelle gegeben. 29
12 Dargestellt ist die Wichtigkeit der Variablen für das optimierte Training bei der BDT-Methode (siehe Kapitel 4.5). 36
13 Die Anzahl der Untergrundevents und die Anzahl der Datenevents nach bestimmten Schnitten mit berechneten Limit für die Signalevents mit kleinen Massensplit und Wirkungsquerschnitt. Die angegebenen Fehler bei der Anzahl der Untergrundevents sind statistische Fehler bei der Annahme einer Poisson-Verteilung. 49
14 Die Anzahl der Untergrundevents und die Anzahl der Datenevents nach bestimmten Schnitten mit berechneten Limit für die Signalevents mit hohen Massensplit und Wirkungsquerschnitt. Die angegebenen Fehler bei der Anzahl der Untergrundevents sind statistische Fehler bei der Annahme einer Poisson-Verteilung. 49

Danksagung

Hiermit bedanke ich mich bei Prof. Dr. Schleper für die Möglichkeit dieser Arbeit. Auch für die Gespräche, die die Analyse motiviert haben möchte ich mich bedanken Aufgrund dieser Arbeit habe ich sehr viel gelernt, wie eine Datenanalyse im Bereich der Teilchenphysik funktioniert. Meinen besonderen Dank gilt Dr. Jory Sonneveld für eine sehr gute Betreuung auch bei Abwesenheit in der Arbeitsgruppe. Für die Beantwortung meiner zahlreichen kleinen Fragen möchte ich mich zudem bei Dr. Anastasia Karavdina und Simon Kurz bedanken.

Erklärung

Hiermit bestätige ich, dass die vorliegende Arbeit von mir selbständig verfasst wurde und ich keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe und die Arbeit von mir vorher nicht einem anderen Prüfungsverfahren eingereicht wurde. Die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium. Ich bin damit einverstanden, dass die Bachelorarbeit veröffentlicht wird.