

**Arbeit zur Erlangung des akademischen Grades  
Bachelor of Science**

**Verbesserung der Energiregression bei  
CTA**

Lars Möllerherm  
geboren in Mettingen

2018

Lehrstuhl für Experimentelle Physik Vb  
Fakultät Physik  
Technische Universität Dortmund

Erstgutachter: Prof. Dr. Dr. Wolfgang Rhode  
Zweitgutachter: Prof. Dr. Kevin Kröninger  
Abgabedatum: 21. September 2018

## **Kurzfassung**

In dieser Arbeit wird die Energierekonstruktion von hochenergetischen kosmischen Photonen, welche durch das Cherenkov Telescope Array(CTA) beobachtet werden, optimiert. Bisher wird für jedes Teleskop eine eigenständige Energieschätzung durch Algorithmen des überwachten maschinellen Lernens durchgeführt. Da bei CTA die Möglichkeit besteht, dass mehrere Teleskope den selben Schauer messen, wird durch das Zusammenfassen der einzelnen Ergebnisse die Energieschätzung verbessert. Eine einfache oder gewichtete Mittelung über die Schätzungen liefert keinen Qualitätsgewinn, eine verschachtelte und eventspezifische Regression jedoch führt auf eine Verringerung des relativen Fehlers und auf eine Verbesserung der Energieauflösung. Der große Zielbereich der Analyse führt auf einen Qualitätsverlust bei niedrigen Energien, was durch eine geeignete bijektive Transformation geändert wird, wodurch die Schätzung in diesem Energiebereich verbessert werden kann. Die getesteten Methoden führen auf eine Verringerung des relativen Fehlers und der Energieauflösung um 75 % in weiten Teilen des Sensitivitätsbereichs von CTA.

## **Abstract**

In this thesis, a method for the optimization of the energy reconstruction for high energy cosmic photons, which are detected by the Cherenkov Telescope Array will be tested. Currently there is a prediction by supervised machine learning regressors for every single telescope. Because of the array structure of CTA, the analysis can be optimized by summarizing the results of these telescopes. There is no performance gain in the simple or the weighted average over each prediction, but a nested model, which predicts every event, causes an improvement of the relative error and for the energy resolution of CTA. Because of the estimator's wide number range, there is a large performance loss for low energies. This can be managed by a transformation of the output, which ensures a performance improvment. These methods improve the bias of the relative error and the energy resolution by 75 % in a large energy range of CTA.

# Inhaltsverzeichnis

<b>1 Einleitung</b>	<b>1</b>
<b>2 Theoretische Grundlagen</b>	<b>2</b>
2.1 Gammaastronomie . . . . .	2
2.2 Cherenkov Teleskope Array (CTA) . . . . .	4
2.3 Maschinelles Lernen . . . . .	5
2.4 Random Forest Regressor . . . . .	6
2.5 Energierekonstruktion . . . . .	8
2.6 Modellevaluation . . . . .	9
<b>3 Ergebnisse</b>	<b>12</b>
3.1 Energierekonstruktion mit Hilfe eines Random Forest Regressors . . . . .	12
3.2 Optimierung durch Mittelwerte und geeignete Gewichte . . . . .	15
3.3 Verschachtelung von Regressionsverfahren . . . . .	18
3.4 Transformation der Energie . . . . .	21
<b>4 Zusammenfassung und Ausblick</b>	<b>24</b>
4.1 Fazit der Arbeit . . . . .	24
4.2 Perspektiven . . . . .	25
<b>Literatur</b>	<b>26</b>

# 1 Einleitung

Auf die Erdatmosphäre trifft eine große Menge von hochenergetischer kosmischer Strahlung, welche uns viel über unverstandene Prozesse im Universum verrät. Die Gammastrahlung gelangt auf direktem Weg zur Erde, wodurch ihr Entstehungsort erfasst werden kann. Die Energie der Strahlung liegt weit über der bisher erreichten Schwerpunktsenergie des LHC von ca. 13 TeV [4] und weckt ein besonderes Interesse an der Entdeckung neuer Physik. Als Beispiel dient die Suche nach Zerfällen der Dunklen Materie oder die Untersuchung der physikalischen Gesetze in extremer Umgebung. Auch ein Großteil der Beschleunigungsprozesse von Teilchen im Universum sind noch nicht erklärt. Um diese Fragen zu beantworten, muss die Gammastrahlung kosmischer Quellen genau untersucht werden.

Da CTA jährlich 3,7 PB [6] an Rohdaten verarbeiten muss, wird eine computergesteuerte Datenanalyse genutzt und das Feld des maschinellen Lernens liefert die richtigen Werkzeuge. Eine direkte Beobachtung mithilfe von Satelliten erlaubt es nicht genügend Ereignisse der hochenergetischen Strahlung zu beobachten. Um eine ausreichende Fläche zu beobachten, wird eine indirekte, erdgebundene Beobachtung über das atmosphärische Schauer und das entstehende Cherenkov-Licht gewählt. Diese Lichtblitze werden mithilfe von Cherenkov-Teleskopen gemessen oder im Fall von CTA mit einer Gruppe von Teleskopen.

Um die Ursprungsenergie des Photons zu ermitteln, muss sie mithilfe von Random Forest Regressoren rekonstruiert werden. Diese Random Forest Algorithmen gehören zum Bereich des überwachten maschinellen Lernens und können mithilfe von Trainingsdatensätzen, die in Monte-Carlo-Simulationen entstehen, trainiert werden.

CTA besitzt mehr als 100 Cherenkov-Teleskope, wobei mehrere Teleskope das selbe Schauer sehen können, jedoch wird momentan für jedes Teleskop eine eigenständige Energierekonstruktion durchgeführt. Da jedes Schauer nur ein Primärteilchen besitzt, müssen die Ergebnisse der einzelnen Teleskope zusammengefasst werden oder eine Schätzung für jedes Schauerereignis und nicht für jedes Teleskop durchgeführt werden. Da der Sensitivitätsbereich von CTA über vier Größenordnungen geht, könnte eine Transformation diesen Zielbereich verkleinern und damit dem Random Forest die Schätzung erleichtern.

## 2 Theoretische Grundlagen

Um die Qualität der Energieschätzung zu verbessern, muss ausgearbeitet werden, welche Anforderungen an die Analyse gestellt werden, um zu neuen Erkenntnissen in der Astroteilchenphysik zu gelangen. Außerdem muss der Random Forest Algorithmus genau verstanden werden, um Möglichkeiten des Qualitätsgewinns zu erarbeiten, wobei am Ende eine sinnvolle Evaluierung der Qualität vorgenommen werden muss.

### 2.1 Gammaastronomie

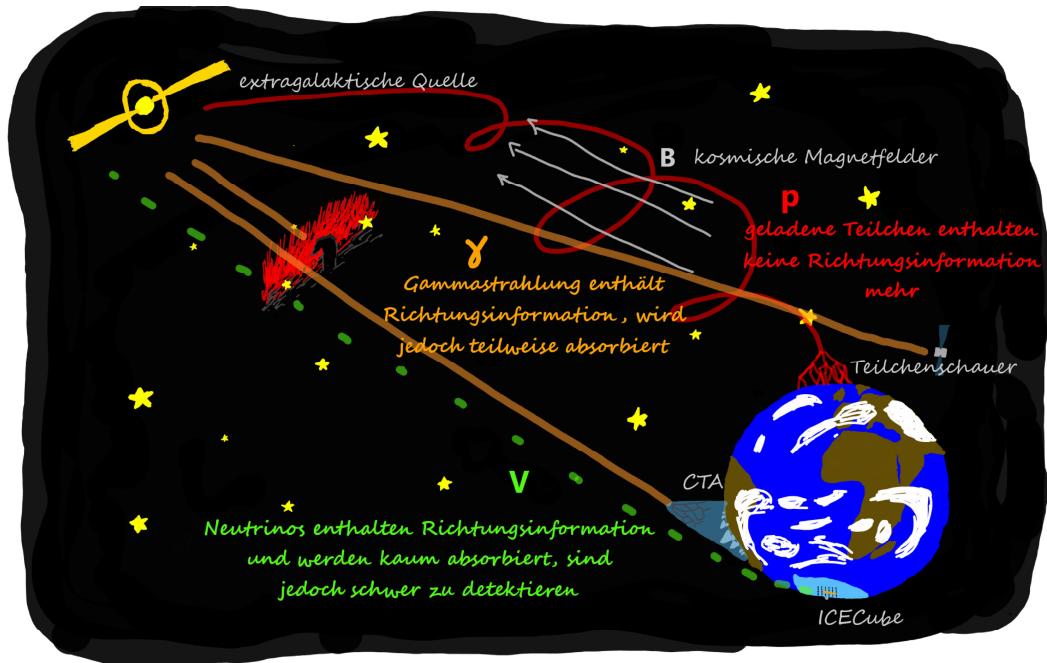
Im Universum gibt es zahlreiche Prozesse, bei denen hochenergetische Teilchen entstehen, oder auf hohe Energien beschleunigt werden. Bei diesen Teilchen handelt es sich zum großen Teil um Protonen oder leichte Atomkerne, aber auch Elektronen sind Bestandteil der kosmischen Strahlung. Ein großer Teil der Beschleunigung geschieht in Druckwellen, wie sie bei Sternexplosionen vorkommen, dabei spielt das Modell der Fermi-Beschleunigung erster und zweiter Ordnung eine große Rolle.

Für den Teilchenfluss der kosmischen Strahlung direkt nach der Fermibeschleunigung in Schockfronten gilt  $\Phi(E) \propto E^{-2}$ . Wenn die Strahlung mit dem extragalaktischen Plasma wechselwirkt, wird das Spektrum um  $E^{-\frac{1}{3}}$  steiler. Da es auch das Plasma der Milchstraße durchqueren muss, um im Sonnensystem gemessen werden zu können, wird der Teilchenfluss der kosmischen Strahlung durch  $\Phi(E) \propto E^{-2.7}$  [1, S. 5] beschrieben.

Jedoch erklären diese Prozesse nicht den gemessenen Energiefluss von ultrahochenergetischen Teilchen, denn bei Energien von  $3 \cdot 10^{15}$  eV tritt ein erstes „knee“ im Spektrum auf, welches nicht durch die Fermibeschleunigung erklärt wird. Die Phänomene die Materie bis auf diese Energien beschleunigen, sind noch nicht vollständig erforscht. Ein möglicher Prozess wäre die Beschleunigung durch elektrische Potentialunterschiede oder der Zerfall von Dunkler Materie.

Diese hochenergetische Teilchenstrahlung erzeugt durch Wechselwirkungen oder Zerfälle Gammastrahlung. Wichtige Prozesse bei der Erzeugung hochenergetischer Photonen sind die Wechselwirkungen von Photonen und Elektronen, wie die inverse Comptonstreuung, bei der das geladene Teilchen Energie auf das Photon überträgt.

Aber auch die Annihilation, welche den Umkehrprozess der Paarerzeugung darstellt und aus einem Elektron-Positron-Paar ein Photon-Paar erzeugt, und die Bremsstrahlung, bei der ein Photon bei der Impulsänderung geladener Teilchen abgestrahlt wird, spielen eine Rolle.



**Abbildung 2.1:** Skizzierter Weg der kosmischen Strahlung zur Erde. Die Gamma- und Neutrinostrahlung gelangt auf direktem Weg zur Erde, im Gegensatz zur geladenen Teilchenstrahlung, die durch kosmische Magnetfelder abgelenkt werden. Die Gammastrahlung wird durch Gaswolken teilweise absorbiert. Die Neutrinostrahlung ist jedoch schwer zu detektieren (angelehnt an [9]).

Abbildung 2.1 stellt den Weg der kosmischen Strahlung zur Erde dar, wobei die Photonen nicht mit kosmischen Magnetfeldern wechselwirken und somit deren Ursprungsrichtung Rekonstruiert werden kann. Jedoch wechselwirken die Photonen mit interstellarem Staub, der zwischen der Erde und der Quelle ist, wodurch ein Teil der Gammastrahlung nicht zur Erde gelangt.

Diese hochenergetischen Photonen können entweder mithilfe von Satelliten im Welt- raum oder mithilfe von Cherenkov-Teleskopen auf der Erdoberfläche beobachtet werden. Aufgrund der hohen Energien der Gammastrahlung messen Satelliten diese mithilfe von Szintillationszählern, dessen Detektionsfläche aufgrund technischer

Umsetzbarkeit begrenzt ist. Energiereiche Strahlung sorgt jedoch dafür, dass in der Erdatmosphäre durch die Wechselwirkung mit den Luftmolekülen ein hochrelativistisches Teilchenschauer aus überlichtschnellen geladenen Teilchen entsteht.

Da sich bei Geschwindigkeiten über der Lichtgeschwindigkeit des Mediums die durch die Polarisierung entstandenen elektromagnetischen Wellen nicht mehr destruktiv überlagern, entstehen Cherenkov-Blitze mit einer Dauer von ca. 150 ns [2], welche sich kegelförmig mit einem Winkel von

$$\cos(\theta) = \frac{1}{n\beta} \quad (2.1)$$

ausbreiten und von den Cherenkov Teleskopen am Boden beobachtet werden können. Diese Art der Beobachtung macht es möglich eine größere Fläche zu observieren, wodurch die Wahrscheinlichkeit ein hochenergetisches Photon zu messen größer wird. Der Winkel  $\theta$  hängt von dem Brechungsindex  $n$  der Luft ab, welcher von der Feuchtigkeit und Dichte der Luft abhängt und somit höhenabhängig ist. Das kontinuierliche Spektrum der Cherenkov-Strahlung besitzt eine zur Frequenz proportionale Intensität im sichtbaren Bereich und wird daher als bläulich wahrgenommen. Die Kamera des Teleskops löst aus, wenn eine bestimmte Anzahl an Photonen des Schauers registriert werden. Da sowohl hochenergetische Gammastrahlung als auch geladene Teilchenstrahlung Schauer in der Atmosphäre erzeugen, gibt es einen Untergrund, der separiert werden muss.

## 2.2 Cherenkov Teleskope Array (CTA)

Das geplante Cherenkov Teleskop Array wird von einer internationalen Kollaboration von 210 Instituten aus 32 Ländern [5] geführt und bildet den nächsten Schritt in der Hochenergiegammaastronomie. Mit einer Gesamtanzahl von 108 Teleskopen hat das Array nach Simulationen zur Folge in seinem Hauptenergiebereich eine Sensitivität von 0,1 % des Energieflusses des Krebsnebels, wodurch es ungefähr zehn Mal sensitiver als das HESS-Experiment ist [17]. Da der Teilchenfluss  $\Phi$  der kosmischen Strahlung dem Potenzgesetz  $\Phi \propto E^{-2.7}$  [1, S. 5] folgt, treffen bei einer Energie  $E$  von 1 TeV noch  $1/(m^2 s)$  Teilchen auf die Erdatmosphäre. Um dennoch genug Ereignisse zu beobachten, muss eine möglichst große Effektive Fläche observiert werden, weshalb der Standort in Chile eine Fläche von  $\approx 4 \text{ km}^2$  abdeckt [7]. Durch drei verschiedene Teleskopgrößen kann CTA Photonen mit Energien von 30 GeV bis 300 TeV detektieren, was es ermöglicht, die verschiedenen Beschleunigungsprozesse im Universum zu untersuchen. Zu den Arten gehören das LST (Large Sized Telescope) mit einer Spiegelgröße von 23 m, das MST (Medium Sized Telescope) mit einer Größe von 11,5 m oder 9,7 m und das SST (Small-Sized Telescope), welches

eine Größe von 4,3 m oder 4,0 m besitzt [8]. Die hohe Sensitivität und die niedrige Energieuntergrenze ermöglichen die Entdeckung neuer Quellen mit einer starken Rotverschiebung, die nur bei niedrigen Energien sichtbar sind, da die höherenergetische Gammastrahlung mit dem extragalaktischen Hintergrundlicht wechselwirkt und somit nicht beobachtet werden kann. Die große Anzahl an Teleskopen führt zusätzlich auf eine bessere Winkelauflösung, was entscheidend bei der Multiwellenlängen-Beobachtung von Quellen ist. Diese Multiwellenlängen-Beobachtung ist entscheidend für das vollständige Verständnis der Beschleunigungsprozesse.

## 2.3 Maschinelles Lernen

Aufgrund des geringen Teilchenfluxes bei hohen Energien, muss die Anzahl an beobachteten Ereignissen bei modernen Experimenten stark ansteigen, wodurch eine händische Analyse unmöglich wird. Daher werden Algorithmen des maschinellen Lernens trainiert, die diese Aufgabe übernehmen. Das maschinelle Lernen wird als Teilgebiet der künstlichen Intelligenz verstanden. Hierbei lernen Algorithmen aus Datensätzen, indem sie verschiedene Optimierungsverfahren nutzen, um eine Fehlerfunktion zu minimieren. Ein trainierter Algorithmus kann im Anschluss Vorhersagen über neue Datenpunkte treffen.

Der Bereich des maschinellen Lernens wird in das überwachte Lernen, bei dem der Algorithmus vor einer Vorhersage mit einem Datensatz, bei dem das Ergebnis und die Eingangsdaten bekannt sind, trainiert wird, und das unüberwachte Lernen, bei dem der Algorithmus Muster in den Eingangsdaten sucht, gegliedert. Zwei große Aufgabengebiete im Bereich des überwachten Lernens sind die Regression und die Klassifikation. Die Regression bildet auf die reellen Zahlen ab und die Klassifikation auf  $N$  Klassen, womit die Regression als Grenzfall  $N \rightarrow \infty$  der Klassifikation verstanden werden kann.

Das Modell der Regressionsanalyse benutzt die abhängige Variable  $\mathbf{y}$  die über eine Funktion  $f(\mathbf{X}, \boldsymbol{\theta})$  von der Variable  $\mathbf{X}$  abhängt, um den Parameter  $\boldsymbol{\theta}$  so zu optimieren, dass für  $\hat{\mathbf{y}} = f(\mathbf{X}, \boldsymbol{\theta}) + L(\hat{\mathbf{y}}, \mathbf{y})$  der Fehler  $L(\hat{\mathbf{y}}, \mathbf{y})$  minimiert wird. Wenn  $\boldsymbol{\theta}$  aus  $k$  Parametern besteht und  $(\mathbf{X}_i, y_i)$   $N$  Tupel sind, müssen drei Fälle unterschieden werden. Im ersten Fall gilt  $k > N$ , was zu einem unterbestimmten System führt, in dem es nicht genug Datenpunkte gibt, um alle Parameter vorherzusagen, wodurch viele Regressionsmethoden zu keinem Ergebnis führen. Bei  $k = N$  existiert genug Information um ein lineares System exakt zu lösen. Im letzten Fall gilt  $k < N$ , was das System überbestimmt werden lässt, wodurch mehreren Lösungen existieren und es wird die Lösung gewählt, die  $L(\hat{\mathbf{y}}, \mathbf{y})$  minimiert.

Bei der linearen Regression wird angenommen, dass der Trainingsdatensatz das Problem vollständig repräsentiert und  $L(\hat{\mathbf{y}}, \mathbf{y})$  keinen Trend und keine Korrelation besitzen. Eine weitere Annahme muss sein, dass, wenn die Variable  $\mathbf{X}_i$  einen Vektor darstellt, für diese Unkorreliertheit und lineare Unabhängigkeit gilt. Wenn der Fehler von  $\mathbf{X}$  eine nicht konstante Varianz besitzt, muss dies durch eine gewichtete Methode korrigiert werden.

## 2.4 Random Forest Regressor

Eine Methode des überwachten Lernens, welche für die Regression verwendet werden kann, stellt der Random Forest Algorithmus (RF) dar. Dieser Algorithmus baut einen Wald aus mehreren möglichst unkorrelierten Entscheidungsbäumen auf, die eigenständige Vorhersagen treffen, über die abschließend gemittelt wird.

Ein Entscheidungsbau wird aufgebaut, indem der Datensatz in Teildatensätze aufgeteilt wird, wobei ein gewähltes Kriterium optimiert wird. Dieses Kriterium kann die Gini Unreinheit oder der Informationsgewinn sein, bei der Regression wird jedoch häufig die Varianzreduktion verwendet. Bei dieser Optimierung wird in jedem Schritt der mittlere quadratische Fehler

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} (y_i - c_m)^2 \quad (2.2)$$

jedes Teildatensatzes  $m$  minimiert, mit

$$c_m = \frac{1}{N_m} \sum_{i \in N_m} \hat{y}_i \quad (2.3)$$

als Mittelwert der Vorhersage  $\hat{y}_i$  für jeden Datenpunkt  $i$  und  $y_i$  als wahren Wert. Dies wird rekursiv wiederholt und somit der Baum ausgebaut, bis der Algorithmus ein Abbruchkriterium erfüllt. Dieses Abbruchkriterium kann eine vorher festgelegte maximale Tiefe des Baumes sein, eine minimale Größe des Datensatzes, der getrennt werden soll, oder eine minimale Größe des getrennten Datensatzes. Nach Erreichen der Abbruchbedingung bildet  $c_m$  des letzten Schrittes die endgültige Vorhersage.

Bei Entscheidungsbäumen gibt es eine Vielzahl von Umsetzungen. Die aktuellsten Arten sind der C5.0 und der CART Algorithmus [20, S. 1]. Die Besonderheit am C5.0 Algorithmus stellt, im Gegensatz zum CART Algorithmus, die nicht notwendige binäre Trennung des Datensatzes dar, jedoch kann mit ihm keine Regression durchgeführt werden. Im scikit-learn-Framework [15] wird eine CART Implementierung des Entscheidungsbäumes benutzt, welche zur Regression fähig ist und die Varianzreduktion als Optimierungskriterium nutzt.

Die Vorteile eines Entscheidungsbaumes sind die Interpretierbarkeit, die Zeitkomplexität von  $\Omega(n \log(n))$  bei  $n$  Datenpunkten und die einfache Datenpräparation. Außerdem besteht keine Anfälligkeit gegenüber unbedeutenden Attributen. Jedoch besteht eine Gefahr des Übertrainierens, was bei einem zu großen Ausbau des Baumes dazu führt, dass der Trainingsdatensatz nachgebildet wird und die Vorhersagen für einen unabhängigen Testdatensatz unpräzise werden. Des Weiteren kann es zu einer Verzerrung in den Vorhersagen kommen, wenn der Trainingsdatensatz eine Verzerrung aufweist. Entscheidungsbäume besitzen keine Stabilität gegenüber Änderungen im Datensatz, was zu einer hohen Varianz der Ergebnisse unterschiedlicher Bäume führt. Da der Entscheidungsbaum zu den gierigen Algorithmen gehört, welche die Entscheidung aufgrund des derzeitig besten Gewinns treffen, findet dieses Verfahren schnell ein Optimum, jedoch nicht immer das globale Extremum und somit nicht die optimale Lösung. Die letzten beiden Nachteile können durch Erweiterungen des Entscheidungsbaumes behoben werden.

Eine Möglichkeit ist die zufällige und unabhängige Auswahl der Teildatensätze. Dazu kann unter anderem das Adaboost Verfahren oder das Bagging verwendet werden, wobei der RF das Bagging verwendet. Dass in SCIKIT-LEARN verwendete Bagging funktioniert, indem aus dem Datensatz  $N$  Stichproben der Größe  $M$  gezogen werden und für die  $N$  Entscheidungsbäume verwendet werden. Die  $N$  Ergebnisse werden am Ende gemittelt oder zusätzlich mit der Genauigkeit des jeweiligen Ergebnisses gewichtet. Um die Größe der Datensätze nicht zu verkleinern, ist es möglich die Datensätze mit Bootstrapping künstlich zu vergrößern, wobei Datenpunkte durch andere Datenpunkte des Datensatzes zufällig ersetzt werden, anstatt sie zufällig auszusortieren. Zusätzlich zum Bagging können die Attribute in einem Umfang, der festgelegt werden kann, zufällig gezogen werden. Hierdurch verlieren die Entscheidungsbäume an Genauigkeit, die Korrelation des Ergebnisses verringert sich jedoch, was Verzerrungs-Varianz-Dilemma genannt wird. Die Varianz wird soweit minimiert, dass es zu keiner Überanpassung kommt und die Verzerrung möglichst klein bleibt [16, S. 2].

Wenn die Anzahl der Entscheidungsbäume in einem RF erhöht wird, konvergiert der generalisierte Fehler

$$PE = P_{X,y}(mg(X, y) < 0) \quad (2.4)$$

gegen

$$P_{X,y}(P_\theta(h(X, \theta) = y) - \max_{j \neq y} P_\theta(h(X, \theta) = j) < 0) \quad (2.5)$$

und es kann durch eine Vergrößerung des Waldes nicht zum Übertraining kommen [3, S. 7]. Bei diesem Theorem bildet

$$mg(X, y) = av_k I(h_k(X) = y) - \max_{j \neq y} av_k I(h_k(X) = j) \quad (2.6)$$

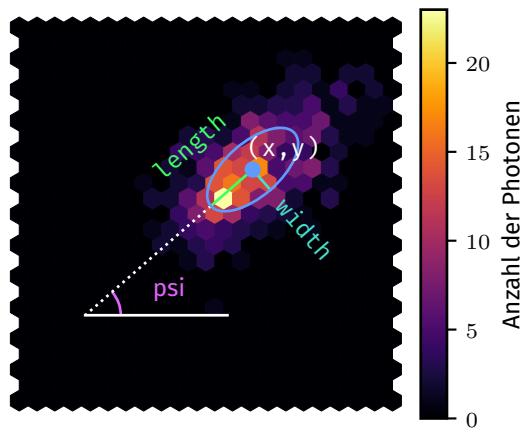
die Gewinn Funktion,  $h_k(X)$  die Vorhersage des  $k$ -ten Entscheidungsbaum und  $I(\cdot)$  die charakteristische Funktion, welche eine 1 ergibt, wenn der Entscheidungsbaum das geforderte Ergebnis liefert und eine 0 wenn nicht. Wenn  $mg(X, y) > 0$  gilt, sagt der Entscheidungsbaum das richtige Ergebnis vorher.

Durch Kreuzvalidierung kann das Modell auf Übertraining untersucht werden. Hierbei wird der Datensatz aufgeteilt, um den Algorithmus mit einem Teil zu trainieren und mit dem anderen unabhängigen Teil zu testen. Die einfache Kreuzvalidierung stellt die in SCIKIT-LEARN implementierte Methode dar, bei der der Datensatz in  $k$  Teildatensätze geteilt wird und jeder dieser Datensätze einmal als Validierungsdatensatz verwendet wird und  $k - 1$  Datensätze zum Training dienen.

## 2.5 Energerekonstruktion

Um die in Kapitel 2.1 erwähnten Bilder auszuwerten, muss zunächst die Kamera kalibriert werden, da die Funktionsweise der elektronischen Komponenten stark von äußeren Bedingungen beeinflusst wird. Der nächste Schritt stellt das Extrahieren der wichtigen Information aus dem Kamerabild in Form der Hillasparameter dar. Es werden die Momente der Verteilung im Kamerabild bestimmt, wobei die ersten Momente als  $x$ -und  $y$ - Koordinate des Mittelpunktes einer Ellipse und die zweiten Momente als Achsen  $w$  und  $L$  der Ellipse dargestellt werden. Weitere Parameter sind die Polarkoordinaten  $r$  und  $\phi$  des Ellipsenmittelpunkts oder der Rotationswinkel  $\psi$  der Ellipsenhauptachse, welcher relativ zur Verbindungsline zwischen Ellipsenmittelpunkt und Kameramittelpunkt gemessen wird. Die geometrischen Hillasparameter sind in Abbildung 2.2 dargestellt. Ein weiterer wichtiger Parameter bildet die totale Intensität des Bildes und da CTA aus mehreren unterschiedlichen Teleskopen besteht, bekommt die Anzahl der Teleskope, die den gleichen Schauer gesehen haben, und welche Art von Teleskop dieses Schauer gesehen hat, eine große Bedeutung für die anschließende Signal Separation und Energie Schätzung. Diese Parameter werden genutzt um die Untergrundschauer von den photoninduzierten Schauern zu trennen. Diese Aufgabe wird mit einer Klassifizierungsmethode des maschinellen Lernens gelöst.

Für die Schätzung der Photonenergie werden Regressionsmethoden verwendet, wobei der RF sich als stabilster Algorithmus erweist [2]. Die in Kapitel 2.3 aufgeführten Annahmen für eine erfolgreiche Regressionsanalyse, sind bei diesem Problem bestmöglich erfüllt. Die Trainingsdaten werden durch Monte Carlo Simulationen erstellt, die die Wechselwirkung der Primär- und Sekundärteilchen mit der Atmosphäre und die Reaktion des Teleskops auf das Schauerlicht simuliert. Dadurch repräsentiert der Trainingsdatensatz das Problem bestmöglich, jedoch führen mögliche systematische



**Abbildung 2.2:** Schematische Darstellung der Hillasparameter, die das Kamerabild der Schauer charakterisieren. Diese Parameter werden verwendet, um die Energie des primären Teilchens zu schätzen. [10]

Fehler in der Simulation zu einer Verzerrung. Darüber hinaus werden die Parameter mit der größtmöglichen Präzision gemessen, um den Fehler der Parameter zu minimieren, jedoch sind aufgrund der Berechnung der Hillasparameter, welche in [14, S. 102] genauer beschrieben werden, diese nicht linear unabhängig. Auch die Varianz des Fehlers geht aufgrund der unterschiedlichen Sensitivitäten der Teleskope nicht homogen über den ganzen Energiebereich, was jedoch durch eine Gewichtung ausgeglichen werden könnte.

## 2.6 Modellevaluation

Um zu erkennen, ob das weiterentwickelte Modell eine genauere Vorhersage liefert als das Bisherige, muss die Qualität des RFs beurteilt werden können.

Für einen ersten Überblick über die Genauigkeit des Algorithmus, wird die Wahrheit gegen die Vorhersage in einem zweidimensionalen Histogramm aufgetragen, welches als Migrationsmatrix bezeichnet wird. Dabei hat die Diagonale die Bedeutung der exakten Vorhersage und eine geringe Streuung der gefüllten Bins um diese deutet auf einen guten Schätzer hin. Die Aussagekraft dieser Abbildung hängt von der Anzahl der Bins ab, daher wird ein Raster von  $300 \times 300$  verwendet.

## 2 Theoretische Grundlagen

---

Ein mögliches Maß, um die Anpassungsgüte von Regressionsmodellen beurteilen zu können, stellt der Determinationskoeffizient dar, welcher durch

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (2.7)$$

definiert wird. Wobei  $\hat{y}_i$  die Schätzwerte,  $\bar{y}$  der Mittelwert und  $y_i$  die Messwerte darstellen. Der Determinationskoeffizient nimmt den Wert 1 an, wenn das Modell die Messpunkte exakt beschreibt und bei dem Wert 0 ist die Schätzung so gut wie die Verwendung des Mittelwerts als Schätzung. Wenn jedoch  $R^2 \leq 0$  ist, kann das Modell als unbrauchbar eingestuft werden, da die Attribute  $X$  keine Information zur Lösung des Problems einbringen.

Dieser Koeffizient besitzt Grenzen der Interpretierbarkeit, da er nur eine Aussage über das gesamte Modell macht und nicht über die Genauigkeit in einzelnen Wertebereichen. Außerdem reagiert er empfindlich gegenüber Trends, die  $R^2$  senken, obwohl dies nicht bedeutet, dass das Modell die Abhängigkeiten des Problems nicht gut beschreibt. Des Weiteren muss die gleiche Anzahl an Datenpunkten vorliegen, um Modelle miteinander zu vergleichen.

Eine Aussage über die Genauigkeit des Algorithmus kann auch über den relativen Fehler getroffen werden. Dabei spielen der Mittelwert und die Hälfte des Interquartilen Abstandes (IQA) von 68,26 %, der durch

$$\text{IQA} = \frac{Q_{84} - Q_{16}}{2} \quad (2.8)$$

definiert wird, eine wichtige Rolle. Dabei stellt  $Q_{84}$  das 84,13 % Quantil und  $Q_{15}$  das 15,87 % Quantil dar, welche durch die NUMPY-Bibliothek [12] bestimmt werden, die das  $q$ -te Quantil als den  $q \cdot N$ -ten Wert eines geordneten Datensatzes der Größe  $N$  definiert. Der Mittelwert ist ein Maß für die Verzerrung und der IQA ist ein Maß für die Auflösung des Schätzers, wobei dies für unterschiedliche Energie-Bins untersucht wird, da aufgrund der energieabhängigen Statistik auch diese Werte über das Energiespektrum hinweg variieren.

Ein weiteres Maß stellt der Mittlere quadratische Fehler

$$\text{mse} = \frac{1}{N} \sum_i (y_i - \hat{y}_i)^2 \quad (2.9)$$

dar. Wenn dieser Fehler kleiner wird, verringert sich auch der relative Fehler, jedoch ist die Größe der Verringerung abhängig vom Energiebereich. Wie in Kapitel 2.4 beschrieben, wird der RF anhand dieses Kriteriums optimiert.

Die Frage ob der relative Fehler oder der mse minimiert werden soll, hängt von der angestrebten Analyse ab. Soll ein Energiespektrum erstellt werden, um ein Potenzgesetz zu analysieren, bietet sich die Minimierung des relativen Fehlers an, da die Energie logarithmisch aufgetragen wird. Ist jedoch die Untersuchung eines Linienspektrums, wie bei der Suche nach Zerfällen der Dunklen Materie, die Absicht der Analyse, dann sollte der absolute Fehler in dem zu analysierenden Energiebereich minimiert werden.

Um die Frage des Informationsgehalts der Attribute zu beantworten, kann zum Beispiel die Wichtigkeit mithilfe von Selektionshäufigkeit, Gini-Wichtigkeit oder die Wichtigkeit der Permutations-Genauigkeit bestimmt werden. Die in SCIKIT-LEARN implementierte Methode nutzt die Rangordnung der Attribute in den einzelnen Entscheidungsbäumen, indem die früh und häufig genutzten Attribute als wichtiger klassifiziert werden[18]. Diese Wichtigkeit kann für jeden Baum ausgelesen werden und als Boxplot dargestellt werden. Dabei stellt die durchgezogene Linie den Median dar, die Enden des Kastens das 0,25-und das 0,75-Quartil, die Antennen das 0,125-und das 0,875-Quantil und die Punkte stellen Ausreißer, die außerhalb des 0,75-IQA liegen, dar. Wenn eine unterschiedliche Zahl an Kategorien oder ein unterschiedlich großer Wertebereich der Attribute vorliegt, besitzt Bootstrapping mit Ersetzen der Datenpunkte und die Attribut Selektion bei CART Algorithmen eine Verzerrung, welche sich auf die Bestimmungsmethoden der Wichtigkeit überträgt und somit das Ergebnis verfälschen kann [19].

## 3 Ergebnisse

Aufgrund der Architektur von CTA bietet sich eine Mittelwertbildung über das Array an, um die Qualität des Schätzers zu verbessern. Zusätzlich scheint aufgrund der Tatsache, dass die Energie des primären Teilchens proportional zur Schauergröße ist und das ganze Array eine bessere Aussage über die Größe treffen kann als einzelne Teleskope, eine Zusammenfassung der Information aller Teleskope sinnvoll. Da das in Kapitel 2.4 beschriebene Kriterium sich auf den absoluten Fehler und nicht auf den relativen Fehler bezieht, muss der Tatsache, dass die richtige Schätzung großer Energien für den Algorithmus einen größeren Gewinn bedeutet, durch Transformationen entgegengesteuert werden, um eine gute Energieauflösung in allen Energiebereichen zu erlangen.

### 3.1 Energierkonstruktion mit Hilfe eines Random Forest Regressors

Um ein Vergleichsergebnis zu erhalten, wird zunächst ein Random Forest Regressor, wie er in Kapitel 2.4 beschrieben wird, verwendet, der eine Energieschätzung für jedes Teleskop vornimmt. Für das Training werden nur Gamma Ereignisse verwendet und somit eine erfolgreiche Signal-Untergrund Trennung vorausgesetzt. Für ein realitätstreueres Testen werden diffuse und punktgerichtete Gamma-Simulationsdaten verwendet, da in der Realität die Signal-Extraktion nicht zwischen punktgerichteten Photonen und Photonen, die in der Atmosphäre entstehen oder zu der allgemeinen kosmischen Strahlung gehören, unterscheiden kann.

Als Attribute werden die nicht richtungsabhängigen Hillasparameter verwendet, dazu gehören Intensität, Länge, Breite, Schiefe und Wölbung sowie die totale Intensität, die von allen Teleskopen aufgenommen wird. Zusätzlich werden die Anzahl der ausgelösten Teleskope, SST, MST und LST als Attribute genutzt, sowie die Identifikationsnummer des Teleskops, welche für das LST 1, für das MST 2 und für das SST 3 ist.

Des Weiteren werden die skalierte Länge

$$SW = \frac{w - \langle w \rangle}{\sigma_w} \quad (3.1)$$

und Breite

$$SL = \frac{l - \langle l \rangle}{\sigma_l} \quad (3.2)$$

verwendet, wobei der Mittelwert über alle Trainingsdatenwerte genommen wird. Diese Methode nennt sich Scaled Cuts Technik. [14, S. 104]

Der ganze Datensatz besteht aus 3 322 938 Datenpunkten, wovon 78 % punktgerichtete Photonen und 22 % diffuse Photonen sind, die in 33 % Trainings- und 66 % Testdatensatz aufgeteilt werden. Die Aufteilung geschieht jedoch mithilfe der Ereignisnummer, damit bei der Trennung keine Ereignisse getrennt werden. Diese Datenpunkte stellen Teleskop-Ereignisse dar. Jedes Schauer stellt ein Array-Ereignis dar, was von mehreren Teleskopen gemessen werden kann und somit mehrere Teleskop-Ereignisse beinhalten kann.

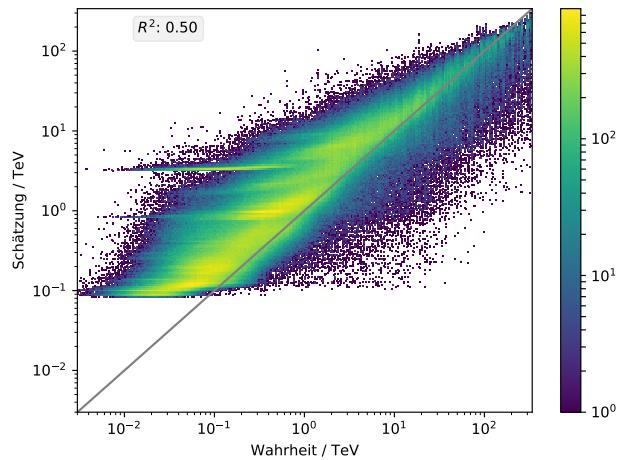
Bei dem Training wird ein Random Forest verwendet, der zur Verhinderung des Übertrainings eine maximale Tiefe von 10 Ebenen besitzt. Jeder Baum trainiert mit  $\sqrt{N}$  Attributen, wobei  $N$  Attribute zur Verfügung stehen, um Korreliertheit der Bäume zu vermeiden, was zu gleichen Baumstrukturen führen würde und somit zu bevorzugten Ergebnissen. Außerdem besteht der Wald aus 100 Bäumen, was mit einer größeren Rechenleistung vergrößert werden kann, ohne das es, wie in Kapitel 2.4 erklärt, zum Übertraining kommt. Der Algorithmus nutzt das Kriterium der Varianz-Reduktion. Da die gering gewählte Tiefe ein Übertraining bereits verhindert und die Ereigniszahl ausreichend groß ist, werden die Hyperparameter der minimalen Blatt- und Trenggröße auf ihrer Grundeinstellung von 1 und 2 gelassen.

Dieser trainierte Entscheidungswald wird mit 2 193 140 Teleskop-Ereignissen getestet und liefert eine Qualität, wie sie in Abbildung 3.1 zu sehen ist. Dabei ist eine Überschätzung der Wahrheiten zu beobachten sowie feine horizontale Linien, die auf eine Korreliertheit der Entscheidungsbäume hindeuten. Der  $R^2$ -Wert von 0,50 deutet auf eine bessere Beschreibung des Modells als der bloße Mittelwert hin. Die Korreliertheit und die Verzerrung können durch eine geeignete Wahl der Hyperparameter minimiert werden. Es ist zu betonen, dass durch die Schätzung auf Teleskop-Ereignissen dieser RF mehrere Vorhersagen für das selbe Schauer liefert.

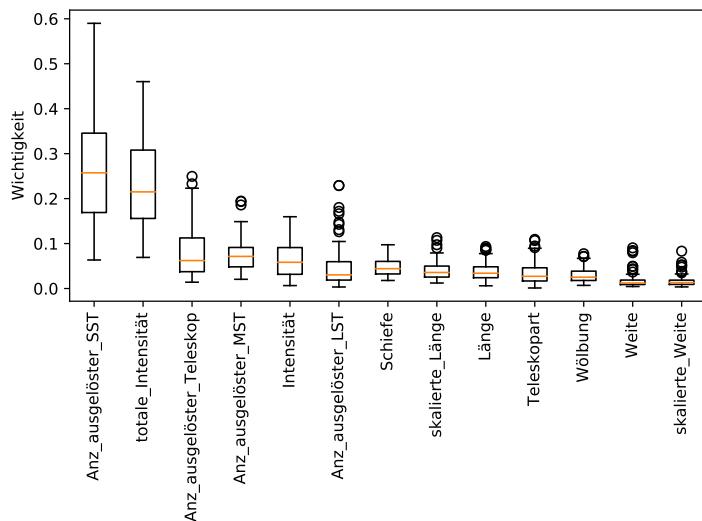
Wie in Abbildung 3.2 dargestellt, besitzen eventspezifische Attribute wie die Anzahl der ausgelösten Teleskope und die totale Intensität die größte Wichtigkeit und teleskopspezifische Attribute wie die Hillasparameter liefern keinen großen Informationsgewinn. Eine Erklärung liefert die Tatsache, dass die in dem Schauer deponierte Energie ein wichtiger Parameter für die Energieschätzung ist und die Hillasparameter im Gegensatz zu den eventspezifischen Parametern von dem Abstand des Teleskops zum Schauer abhängen. Aufgrund der Unterschiedlichkeit der Attribute kann die Wichtigkeit nach Kapitel 2.6 eine Verzerrung besitzen und Attribute wie die Intensität überschätzt, sowie Attribute wie die Teleskopart unterschätzt werden.

### 3 Ergebnisse

---



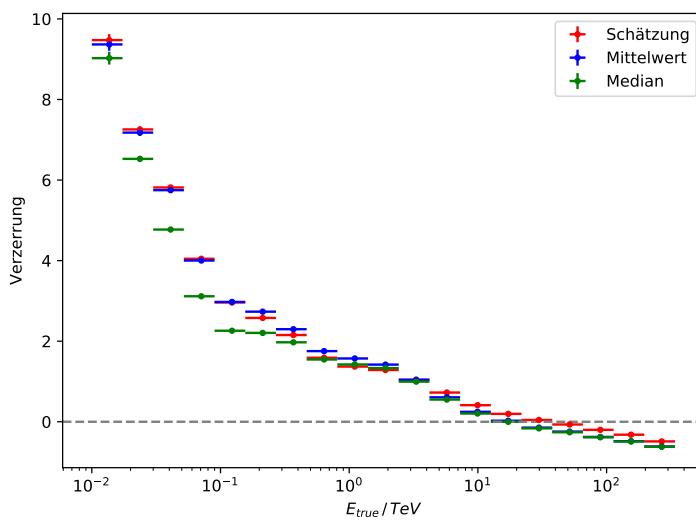
**Abbildung 3.1:** In diesem Graphen ist die vorhergesagte Energie gegen die Wahrheit aufgetragen, wobei Streuung und Verzerrung um die Diagonale zu erkennen sind.



**Abbildung 3.2:** Darstellung der Wichtigkeit der genutzten Attribute bei der ersten Vorhersage als Kastengrafik. Die Anzahl der ausgelösten SST und die totale Intensität stellen die wichtigsten Attribute dar und die Breite der Hillasellipse trägt kaum zur Schätzung bei.

## 3.2 Optimierung durch Mittelwerte und geeignete Gewichte

Um eine physikalisch sinnvolle Aussage über die Energie des Primärteilchens treffen zu können, müssen die Schätzungen der Teleskop-Ereignisse zusammengefasst werden und somit ein Ergebnis für jedes Array-Ereignis gefunden werden. Die Energieschätzung der einzelnen Teleskope bei einem Ereignis variiert, obwohl sie den gleichen Schauer beobachten. Dies liegt an den unterschiedlichen Arten, Blickwinkeln und Abständen der Teleskope, was dazu führt, dass jedes Teleskop unterschiedlich viel Information über den Schauer erfasst. Wenn die variierenden Ergebnisse zufällig und nach dem Zentralen Grenzwertsatz der Statistik normalverteilt um den wahren Wert liegen [11, S. 10], verbessert eine Mittelwertbildung das Ergebnis. Daher wird als Schätzer das arithmetische Mittel und der Median untersucht.

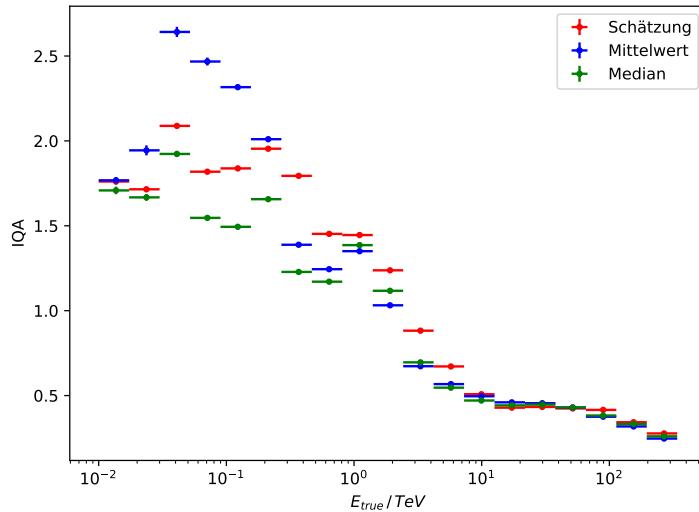


**Abbildung 3.3:** Darstellung der Verzerrung für verschiedene Energiebereiche. Es sind die Mittelwerte der relativen Fehler für die Schätzung der Teleskop-Ereignisse, die Vorhersage nach der Mittelwertbildung und der Median der Vorhersagen aufgetragen.

Die zusammengefassten Schätzungen führen auf die Verzerrungen und IQAs, die in Abbildung 3.3 und Abbildung 3.4 zusehen sind. Bei niedrigen Energien führt das arithmetische Mittel auf keine geringere Verzerrung und auch der Median bringt nur eine leichte Verbesserung. Dies liegt daran, dass eine Verzerrung die Annahme verletzt, dass die Ergebnisse um den wahren Wert liegen, stattdessen liegen sie um einen verschobenen Wert. Das arithmetische Mittel kann eine Verzerrung nicht verringern. Nur bei Ereignissen mit einer geringen Verzerrung sorgt das Mitteln für

### 3 Ergebnisse

---



**Abbildung 3.4:** Darstellung des IQA des relativen Fehlers für verschiedene Energiebereiche. Aufgetragen sind die Schätzung der Teleskop-Ereignisse, die Vorhersage nach der Mittelwert Bildung und der Median der Vorhersagen.

eine bessere Qualität, was zum einen bei dem IQA für große Energien zu beobachten ist und bei dem gesamten gemittelten quadratischen Fehler, der von  $102,94 \text{ TeV}^2$  für die erste Schätzung auf  $65,31 \text{ TeV}^2$  nach einer Mittelwertbildung fällt.

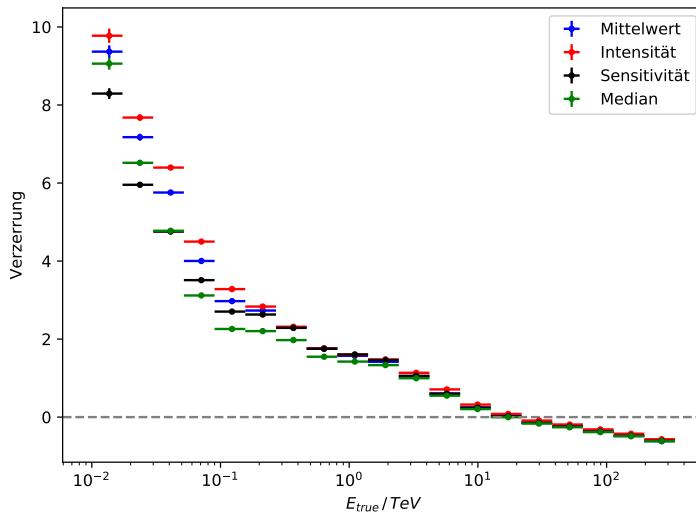
Das der Median die Verzerrung verringert liegt daran, dass es bei dem Beobachten von Schauern vorkommen kann, dass einzelne Teleskope im Gegensatz zu anderen besser positionierten Teleskopen nur einen geringen Teil des Schauers beobachten und somit weniger Information des Schauers besitzen, was zu einer falsche Vorhersage führt. Da dies meist nur auf vereinzelte Teleskope zutrifft, liegt der Median der Teleskope richtig. Dies zeichnet sich auch durch einen gemittelten quadratischen Fehler von  $66,61 \text{ TeV}^2$  aus, welcher besser als der Fehler der ersten Schätzung ist und vergleichbar mit dem Fehler der Mittelwertbildung ist.

Eine weitere Möglichkeit um das Problem, dass nicht alle Teleskope den Schauer gleich gut sehen, zu beheben, stellt das Mitteln mit einem Gewicht, welches die Sichtbarkeit des Schauers beschreibt, dar. Ein Indiz auf den gesehenen Anteil des Schauers stellt die beobachtete Intensität dar, wobei eine hohe Intensität eine gute Sichtbarkeit bedeutet und somit die Intensität direkt als Gewicht verwendet wird.

Zum Anderen besitzen die in Kapitel 2.2 beschriebenen Teleskope eine energieabhängige Sensitivität, wobei es einen Energiebereich gibt, indem eine volle Sensitivität

### 3.2 Optimierung durch Mittelwerte und geeignete Gewichte

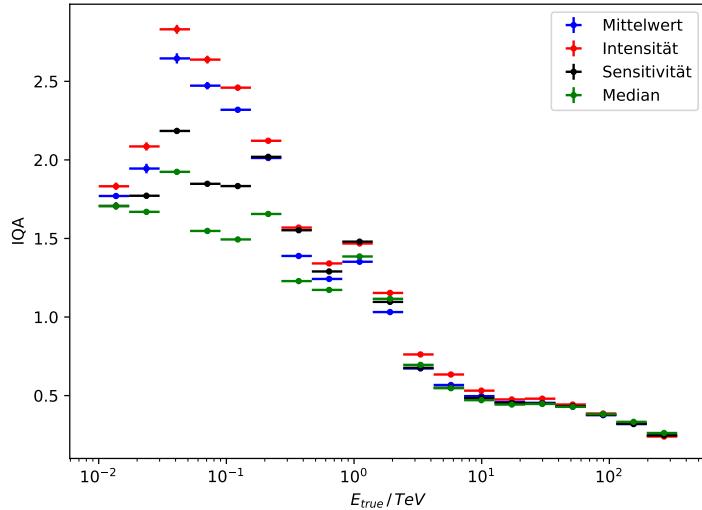
herrscht und einen, indem eine Teilsensitivität herrscht. Das LST besitzt eine Teilsensitivität bei  $20\text{ GeV} - 3\text{ TeV}$  und eine volle Sensitivität bei  $20\text{ GeV} - 150\text{ GeV}$ . Das MST ist teilsensitiv bei  $80\text{ GeV} - 50\text{ TeV}$  und vollsensitiv bei  $150\text{ GeV} - 5\text{ TeV}$  und das LST hat seinen Teilsensitivitätsbereich bei Energien zwischen  $1\text{ TeV} - 300\text{ TeV}$  und seine Haupt sensitivität bei  $5\text{ TeV} - 300\text{ TeV}$ . [8] Wenn die geschätzte Energie im vollsensitiven Bereich liegt, wird ein Gewicht von 2 angelegt, wenn sie im teilsensitiven Bereich liegt, ein Gewicht von 1 und wenn die Energie in keinem Sensitivitätsbereich liegt, wird ein Gewicht von 0.1 angelegt.



**Abbildung 3.5:** Darstellung der Verzerrung für die arithmetische Mittelung, den Median und für eine gewichtete Mittelung mit der Intensität oder der Sensitivität als Gewicht.

Das Gewichten führt auf eine Energieschätzung, die in Abbildung 3.5 und Abbildung 3.6 zu sehen ist. Beide Gewichte führen nicht auf das gewünschte Ergebnis, wobei die Gewichtung mit der Intensität auf eine Verschlechterung der Qualität führt. Die Sensitivität scheint zwar ein Indikator für die Sichtbarkeit zu sein, jedoch scheint die Höhe der Gewichtung nicht ausreichend zu sein, um die Fehlschätzungen ausreichend zu korrigieren. Durch die Energieabhängigkeit der Intensität scheint dieses Attribut als Gewicht ungeeignet zu sein. Jedoch verbessert sich der gemittelte quadratische Fehler auf  $58,68\text{ TeV}^2$  im Vergleich zu der Gewichtung mit der Sensitivität, wo er bei  $64,72\text{ TeV}^2$  bleibt.

Welche der Methoden die bessere Qualität liefert, hängt von der Analyse ab. Soll ein Energiespektrum untersucht werden, wird Wert auf einen geringen relativen

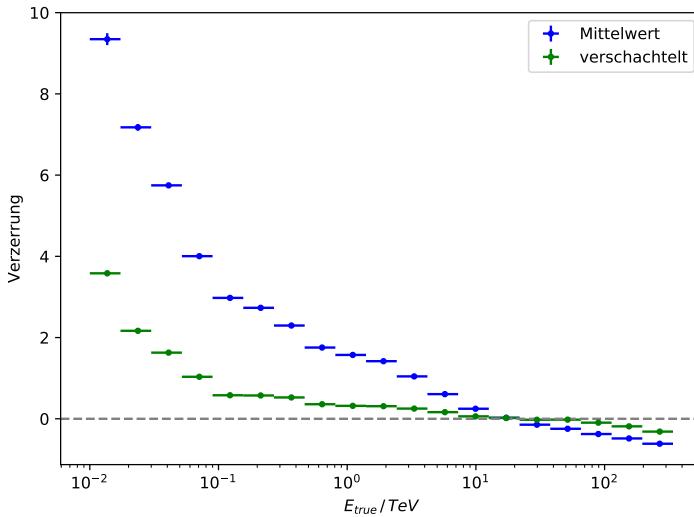


**Abbildung 3.6:** Auftragen des IQA des relativen Fehlers für verschiedene Energiebereiche, jeweils für den arithmetischen Mittelwert, den Median und für die mit der Intensität oder der Sensitivität gewichtet gemittelten Vorhersage.

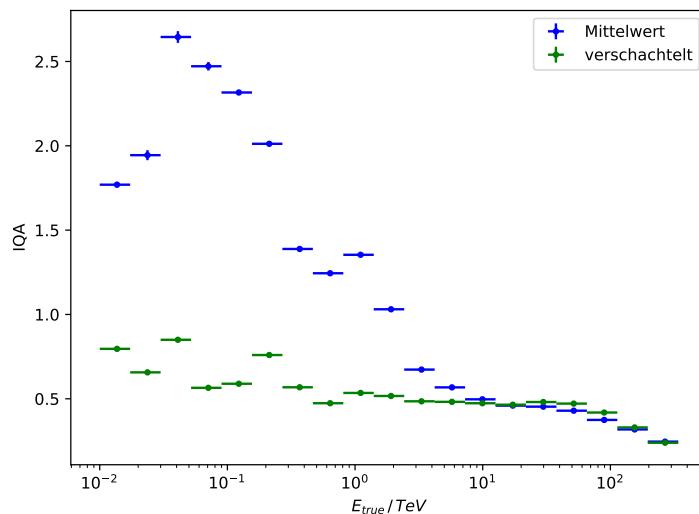
Fehler gelegt, was das Nutzen des Medians nahe legt. Werden jedoch Linienspektren untersucht, kann der Mittelwert in einigen Energiebereichen die richtige Wahl sein, da dieser auf einen geringeren absoluten Fehler führt. Der Median sorgt für eine Verbesserung in niedrigen Energiebereichen, jedoch führt er zu einer Verschlechterung bei Energien von  $(1 - 10) \text{ TeV}$ . In diesem Bereich besitzt CTA jedoch die größte Statistik, womit der Qualitätsgewinn infrage gestellt wird, was durch den gleichbleibenden gemittelten quadratischen Fehler bestätigt wird.

### 3.3 Verschachtelung von Regressionsverfahren

Schon die Wichtigkeit der eventspezifischen Attribute in Abbildung 3.2 deutet darauf hin, dass die zusammengefassten Attribute eines Ereignisses mehr Information beinhalten, als die teleskopspezifischen Attribute. Daher wird nach der ersten Schätzung ein zweiter Random Forest trainiert, der mithilfe von eventspezifischen Attributen Schätzungen für jedes Ereignis abgibt. Zu den Attributen gehören die Anzahl der ausgelösten Teleskope sowie SST, MST und LST, die totale Intensität, die arithmetisch gemittelte Schätzung des ersten Waldes, die Mittelwerte und Standardabweichungen der skalierten Größen und die Mittelwerte, Standardabweichungen, Maximal- und Minimalwerte der Schätzungen von den SST, MST und LST.



**Abbildung 3.7:** Abbildung der Verzerrung nach einer Verschachtelung von zwei Random Forests. Es ist eine deutliche Verbesserung gegenüber der Mittelwertbildung zu erkennen.

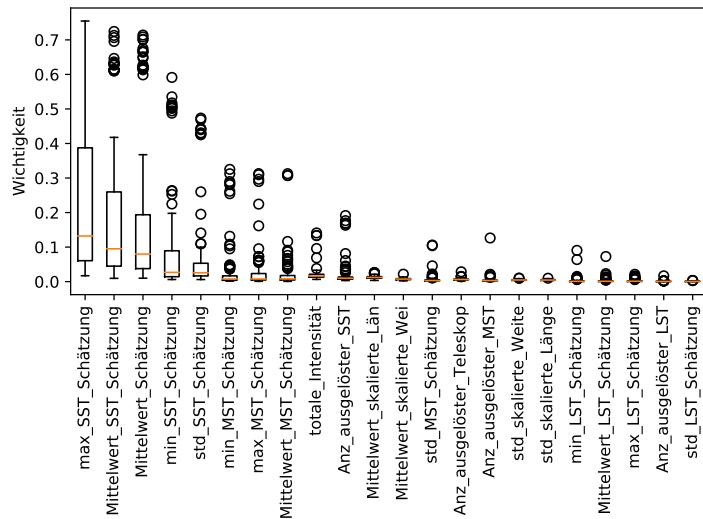


**Abbildung 3.8:** Abbildung des IQA für den zweiten Random Forest. Die Verschachtelung führt auf eine starke Verringerung des IQA bei niedrigen Energien.

### 3 Ergebnisse

Der Random Forest wird mit 498 081 Ereignissen trainiert und getestet. In Abbildung 3.7 und Abbildung 3.8 ist die Qualität des zweiten Schätzers dargestellt und der gemittelte quadrierte Fehler fällt auf  $37,18 \text{ TeV}^2$ . Die Verschachtelung der Entscheidungswälder führt auf eine verbesserte Qualität in allen Energiebereichen, sowohl bei der Verzerrung als auch bei dem IQA, wodurch ein Übertraining durch den zweiten Wald ausgeschlossen werden kann.

Die genutzten Attribute haben eine starke Korrelation aufgrund der Gewinnung der Attribute aus der gleichen Information. Dies ist durch die Anzahl an Ausreißer in Abbildung 3.9 zu erkennen. Eine genauere Untersuchung des Aufbaus der Entschei-



**Abbildung 3.9:** Abbildung der Wichtigkeit der Attribute im zweiten Random Forest. Die Informationen der SST und die Schätzung des ersten Random Forest sind bedeutende Attribute und die hohe Fluktuation deutet auf eine Korreliertheit der Attribute hin.

dungsbäume ergibt, dass die erste Separation der Datensätze häufig mithilfe der gemittelten Schätzung geschieht, um sich in den Unterbäumen auf die für diesen Energiebereich interessanten Attribute zu konzentrieren. Die Anzahl der ausgelösten SST stellt zum Beispiel einen guten Parameter für hohe Energien dar, da solche Photonen große Schauer erzeugen, die zu großen Cherenkov-Kegeln führen. Eine geringe Anzahl an ausgelösten SST stellt jedoch kein Kriterium für eine niedrige Photonenenergie dar, weil auch die Möglichkeit besteht, dass nur der Rand des Schauers beobachtet wird.

Um die Frage zu beantworten, wie der Qualitätsgewinn im Vergleich zum Zeitaufwand steht, wird zunächst die Komplexität für das Ausbauen des Entscheidungswaldes

untersucht, welche im Mittel bei

$$\Theta(MK\tilde{N} \log^2(\tilde{N})) \quad (3.3)$$

liegt [13, S. 96]. Dies gilt für Random Forests mit  $M$  Bäumen, die mit  $K$  zufällig gezogenen Attributten und  $N$  Datenpunkten vollständig ausgebaut werden. Da das Bootstrapping in SCIKIT-LEARN Datensätze erzeugt, die gleich groß wie der Originaldatensatz sind, gilt  $\tilde{N} = N$ . Der Datensatz des zweiten Entscheidungswaldes besitzt  $N_2 \approx \frac{1}{5}N_1$  Datenpunkte, da im Mittel  $\approx 5$  Teleskope auslösen. Damit stellt

$$\Theta\left(MK\frac{1}{5}\tilde{N}_1 \log^2\left(\frac{1}{5}\tilde{N}_1\right)\right) \quad (3.4)$$

den zusätzlichen Zeitaufwand dar. Da der Zeitaufwand der Schätzung, welcher

$$\Theta\left(M \log\left(\frac{1}{5}N\right)\right) \quad (3.5)$$

beträgt [13, S. 98], für die angestrebte Echtzeitanalyse von größerem Interesse ist, steht der Qualitätsgewinn noch besser da. Außerdem werden die Entscheidungsbäume nicht vollständig ausgebaut, wodurch der Zeitaufwand deutlich verringert wird, jedoch können beide Entscheidungswälder nicht parallelisiert werden, wodurch definitiv ein Zeitverlust entsteht.

## 3.4 Transformation der Energie

In Abbildung 3.1 ist ein Abschneiden für niedrige Energien durch den Schätzer zu beobachten. Durch die logarithmische Skala scheint es ein drastischer Schnitt zu sein, jedoch ist der Fehler, der entsteht, wenn alle Energien  $E_\gamma < 0,1 \text{ TeV}$  auf  $0,1 \text{ TeV}$  geschätzt werden, gering. Daher konzentriert sich der Algorithmus darauf, die großen Energien richtig zu schätzen. Dies führt jedoch auf einen großen relativen Fehler für kleine Energien, wie er in Abbildung 3.3 zu sehen ist und somit auf eine schlechte Energieauflösung.

Eine bijektive Transformation auf  $\mathbb{R}^+$  mit

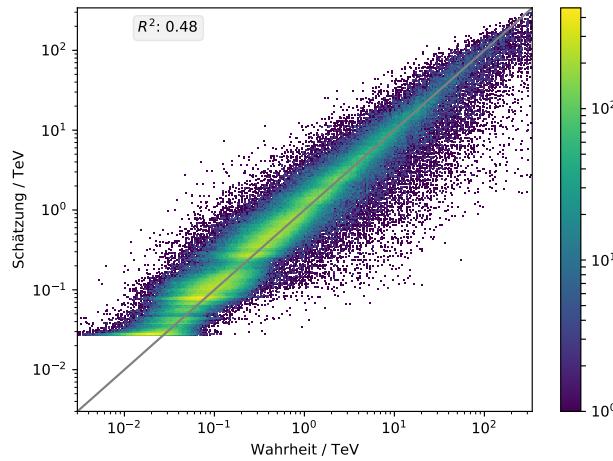
$$E_{\text{trafo}} = \ln(E_\gamma + 3) \quad (3.6)$$

führt auf eine kleinere Zielmenge des Schätzers, wodurch die Qualität weniger stark energieabhängig ist. Eine anschließende Rücktransformation mit

$$E_\gamma = \exp(E_{\text{trafo}}) - 3 \quad (3.7)$$

### 3 Ergebnisse

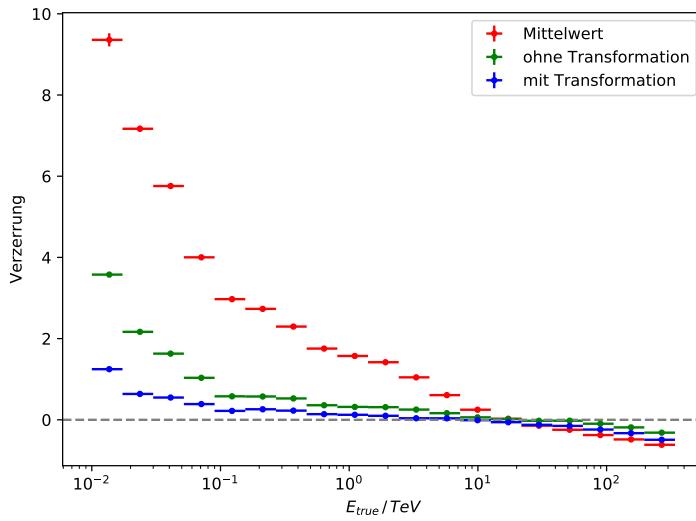
---



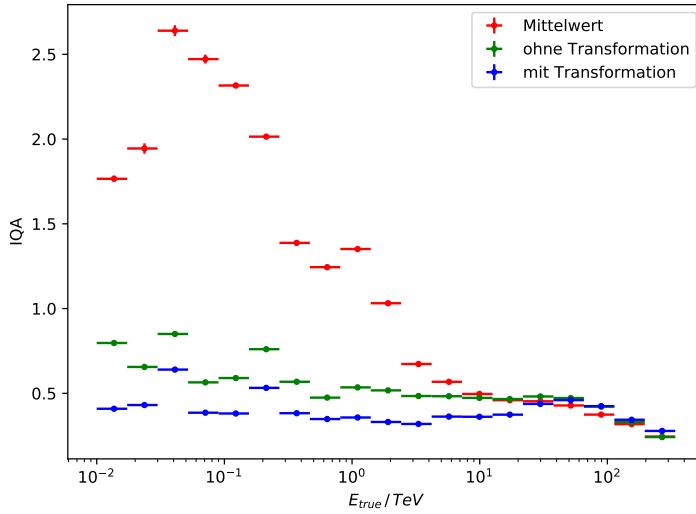
**Abbildung 3.10:** Abbildung der geschätzten Energie durch einen auf die transformierte Energie trainierten Random Forest. Der Schätzer schneidet fast eine Größenordnung später ab als der Schätzer aus Kapitel 3.1.

führt wieder auf die richtige Energie.

Es wird das verschachtelte Modell aus Kapitel 3.3 benutzt, wobei beide Schätzer auf die transformierte Energie trainiert werden. Abbildung 3.10 zeigt, dass der Schätzer beinahe eine Größenordnung später abschneidet und die Transformation keinen Genauigkeitsverlust in anderen Energiebereichen zufolge hat. Der Schätzer scheint mehr Wert auf einen geringen Fehler bei kleinen Energien zu legen. Schon das Verschachteln der Algorithmen sorgt für ein Herabsetzen der Grenze, da durch das frühe Trennen der Energiebereiche im Entscheidungsbaum, dem Algorithmus die Möglichkeit gegeben wird, sich auf die Verbesserung der niedrigen Energiebereiche zu konzentrieren. Die Deutung von Abbildung 3.11 und Abbildung 3.12 zeigt, dass die Transformation die Qualität im Vergleich zum Algorithmus aus 3.3 verbessert. Sowohl der IQA als auch die Verzerrung des relativen Fehlers werden mithilfe der Transformation bei niedrigen Energien geringer. Bei  $E_\gamma > 30 \text{ TeV}$  muss jedoch eine Verschlechterung der Verzerrung und Auflösung in Kauf genommen werden, welche jedoch eine große Verschlechterung bezogen auf den absoluten Fehler bedeutet, der von  $37,18 \text{ TeV}^2$  auf  $53,65 \text{ TeV}^2$  steigt.



**Abbildung 3.11:** Darstellung der Verzerrung bei einer verschachtelten Schätzung mit und ohne Transformation der Art (3.6) sowie die Mittelwertbildung ohne Transformation. Die Transformation führt zu einer Verringerung des relativen Fehlers in niedrigen Energiebereichen.



**Abbildung 3.12:** Abbildung der Energieauflösung nach einer Transformation des Zielbereichs im Vergleich zu der untransformierten gemittelten Schätzung und der verschachtelten Schätzung. Die Transformation führt auf eine Verbesserung in niedrigen Energiebereichen.

## 4 Zusammenfassung und Ausblick

Durch die in dieser Arbeit genutzten Methoden kann die Performance der Energierkonstruktion deutlich verbessert werden. Sowohl die Verzerrung als auch die Auflösung des relativen Fehlers können auf 75 % gesenkt werden.

### 4.1 Fazit der Arbeit

Durch das Bilden des arithmetischen Mittels über aller Teleskope ergibt sich kein Performancegewinn, da die Schätzungen eine Verzerrung besitzen, die durch eine Mittelwertbildung nicht verbessert werden kann. Die Bildung des Median hingegen ermöglicht eine Verbesserung, da einzelne Schätzungen, die möglicherweise eine falsche Schätzung aufgrund der geringen Sicht des Teleskops auf den Schauer abgeben, herausgefiltert werden. Eine geeignete Gewichtung bei der Mittelwertbildung, die diesen Effekt beim arithmetischen Mittel erreichen könnte, wurde nicht gefunden, da die Fehlschätzungen über einen Bereich von fünf Größenordnungen gehen können. Eine Schätzung mithilfe eines zweiten Random Forests, der eventspezifische Informationen und eine Schätzung des ersten Algorithmus bekommt, liefert einen enormen Performancegewinn, da durch die erste Schätzung eine Trennung vorgenommen werden kann, die eine anschließende energiespezifischere Analyse ermöglicht. Das Problem des großen Zielbereichs wird mithilfe einer logarithmischen Transformation gelöst, wodurch eine weitere Verbesserung der Auflösung und der Verzerrung verzeichnet wird. Eine verschachtelte Analyse mit einem transformierten Zielbereich, die jedoch einen nicht parallelisierbaren Zeitaufwand bedeutet, führt auf eine Energieschätzung mit einer ungefähren Auflösung von 0.4 in weiten Teilen des Energiebereichs. Die Ergebnisse dieser Arbeit wurden mit Daten erzeugt, auf die noch keine Schnitte gesetzt sind. Solche Schnitte führen dazu, dass Ereignisse, die schwer zu schätzen sind, aus der Analyse genommen werden, wodurch die Performance weiter verbessert werden kann.

Die abschließende Frage, welche Methode den größten Gewinn bedeutet, kann nicht beantwortet werden. Die Methode sollte auf das Ziel der Analyse angepasst werden. Soll ein gesamtes Energiespektrum untersucht werden, so liefert die verschachtelte und mit (3.6) transformierte Methode den besten Performancegewinn. Strebt die Analyse jedoch eine Untersuchung eines Linienspektrums an, so sollte der quadrierte

Fehler minimiert werden, was mithilfe der verschachtelten Methode gelingt, jedoch die untersuchte Transformation eine Verschlechterung für große Energien bedeutet.

## 4.2 Perspektiven

Eine deutliche Verbesserung wird die Hinzunahme des Abstandes von Teleskop und Schauer und die Schätzung des ersten Wechselwirkungspunktes des primären Teilchens mit der Atmosphäre als zusätzliche Attribute bewirken, da dies wichtige charakteristische Merkmale für die Energieschätzung sind. Außerdem dürften die Hillasparameter wichtiger werden, da durch die Distanz zum Schauer die Hillasparameter ein Indiz auf die Größe des Schauers geben können. Dadurch wird die Performance der einfachen Schätzung näher an die des verschachtelten Waldes herankommen. Diese Attribute stehen zum Zeitpunkt der Arbeit jedoch nicht funktionsfähig zur Verfügung.

Zusätzlich wäre es von Interesse den Abstand als Gewicht zu testen, da dieser ein direktes Indiz auf die Sichtbarkeit liefert. Vielleicht liefern andere Gewichte, Kombinationen von Gewichten oder skalierte Gewichte einen Performancegewinn.

Von Interesse wäre es andere Transformationen zu testen, die den Zielbereich weiter verkleinern und für eine gleichverteilte Statistik in allen Energiebereichen sorgen. Vielleicht können auch unterschiedliche Transformationen angewendet werden, je nachdem welcher Energiebereich für die Beobachtung von Interesse ist und somit die beste Performance für jede Analyse individuell herausgeholt werden kann.

Ein großer Grund für den Erfolg des verschachtelten Modells und der Transformationsmethode ist die Tatsache, dass für das Ausbauen der Entscheidungsbäume ein Kriterium genutzt wird, welches den absoluten Fehler minimiert, wobei das Anforderungsprofil von CTA eine Minimierung des relativen Fehlers verlangt. Das Entwickeln von anderen Kriterien oder das Nutzen von anderen Regressionsmethoden, die die Minimierung des relativen Fehlers anstreben, würden die Analyse anforderungsorientiert verbessern.

## Literatur

- [1] Pijushpani Bhattacharjee und Günter Sigl. „Origin and propagation of extremely high-energy cosmic rays“. In: *Phys. Rep.* 327.3 (2000), S. 109–247. ISSN: 0370-1573. DOI: [https://doi.org/10.1016/S0370-1573\(99\)00101-5](https://doi.org/10.1016/S0370-1573(99)00101-5). URL: <http://www.sciencedirect.com/science/article/pii/S0370157399001015>.
- [2] Christian Bockermann et al. „Online Analysis of High-Volume Data Streams in Astroparticle Physics“. In: *Machine Learning and Knowledge Discovery in Databases*. Hrsg. von Albert Bifet et al. Cham: Springer International Publishing, 2015, S. 100–115. ISBN: 978-3-319-23461-8.
- [3] Leo Breiman. „Random Forests“ In: *Machine Learning* 45.1 (2001), S. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.
- [4] *CERN Accelerating science*. URL: <https://home.cern/about/accelerators> (besucht am 15.07.2018).
- [5] *CTA Consortium*. URL: <https://www.cta-observatory.org/about/cta-consortium/> (besucht am 15.07.2018).
- [6] *CTA Data Management: Data processing, archiving and access*. URL: <https://www.cta-observatory.org/project/technology/data/> (besucht am 15.07.2018).
- [7] *CTA Performance*. URL: <https://www.cta-observatory.org/science/cta-performance/#1472563397821-893dc9a7-f7ec> (besucht am 15.07.2018).
- [8] *CTA Technology*. URL: <https://www.cta-observatory.org/project/technology/> (besucht am 15.07.2018).
- [9] *DESY: Astroteilchenphysik*. URL: [https://astro.desy.de/index\\_ger.html1](https://astro.desy.de/index_ger.html1) (besucht am 15.08.2018).
- [10] *GitHub Gist: MaxNoe/header-matplotlib.tex*. URL: <https://gist.github.com/MaxNoe/fb1e2042e0d8e1e38b3d2ab4c976cba2> (besucht am 17.08.2018).
- [11] Michael Grabe. *Grundriss der Generalisierten Gauß'schen Fehlerrechnung*. ger. Physics. Berlin ; Heidelberg [u.a.]: Springer, 2011, XIV, 191 S. ISBN: 978-3-642-17821-4.
- [12] Eric Jones, Travis Oliphant, Pearu Peterson et al. *SciPy: Open source scientific tools for Python*. 2001. URL: <http://www.scipy.org/> (besucht am 15.07.2018).

- 
- [13] Gilles Louppe. „Understanding Random Forests: From Theory to Practice“. arXiv:1407.7502. Diss. University of Liege, Belgium, Okt. 2014.
  - [14] Mathieu de Naurois. „Very High Energy astronomy from H.E.S.S. to CTA. Opening of a new astronomical window on the non-thermal Universe“. Diss. Ecole Polytechnique, 2012. URL: <http://tel.archives-ouvertes.fr/tel-00687872>.
  - [15] Fabian Pedregosa et al. „Scikit-learn: Machine Learning in Python“. In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830.
  - [16] Philipp Probst, Marvin Wright und Anne-Laure Boulesteix. „Hyperparameters and Tuning Strategies for Random Forest“. In: *ArXiv e-prints* (Apr. 2018). arXiv: 1804.03515 [stat.ML].
  - [17] Thomas Schweizer et al. „Cherenkov Telescope Array: The next-generation ground-based gamma-ray observatory“. In: *Proceedings, 30th International Cosmic Ray Conference (ICRC 2007): Merida, Yucatan, Mexico, July 3-11, 2007*. Bd. 3. [3,1313(2007)]. 2007, S. 1313–1316. arXiv: 0709.2048 [astro-ph].
  - [18] *scikit learn: Ensemble methods: Feature importance evaluation*. URL: <http://scikit-learn.org/stable/modules/ensemble.html#feature-importance-evaluation> (besucht am 15.08.2018).
  - [19] Carolin Strobl et al. „Bias in random forest variable importance measures: Illustrations, sources and a solution“. In: *BMC Bioinformatics* 8.1 (Jan. 2007), S. 25. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-25.
  - [20] Loh Wei-Yin. „Classification and regression trees“. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1 (), S. 14–23. DOI: 10.1002/widm.8. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.8>.

## Danksagung

Ich möchte mich an dieser Stelle bei allen bedanken, die mich während meiner Bachelorarbeit unterstützt und motiviert haben. Auch meinem Erstgutachter Herrn Prof. Dr. Dr. Rhode und meinem Zweitgutachter Herrn Prof. Dr. Kröninger möchte ich für die Mühe danken und für die Möglichkeit ein Teil dieses spannenden Forschungsbereiches zu sein. Des Weiteren möchte ich dem ganzen Lehrstuhl für Experimentelle Physik Vb für die schöne Arbeitsatmosphäre und die große Hilfsbereitschaft danken.

Ein besonderer Dank gilt meinem Betreuer Herrn Kai Brügge für seine Unterstützung und fachliche Expertise. Er hat mir ermöglicht an einem so spannenden Thema zu forschen und hat mir mit der Beantwortung all meiner Fragen und mit kritischem Hinterfragen den ersten Kontakt mit der wissenschaftlichen Arbeit perfekt gemacht. Daher möchte ich mich für seine Geduld und Mühe bedanken.

Außerdem muss ich meinen Kommilitonen Felix Kratz, Julian Hochhaus und Niko Salewski danken, die meine Arbeit Korrektur gelesen haben und dafür gesorgt haben, dass meine Arbeit dem Standard der deutschen Rechtschreibung genügt.

## **Eidesstattliche Versicherung**

Ich versichere hiermit an Eides statt, dass ich die vorliegende Abschlussarbeit mit dem Titel „Verbesserung der Energiregression bei CTA“ selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

---

Ort, Datum

---

Unterschrift

## **Belehrung**

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50 000 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden (§ 63 Abs. 5 Hochschulgesetz –HG–).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird ggf. elektronische Vergleichswerkzeuge (wie z. B. die Software „turnitin“) zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen.

---

Ort, Datum

---

Unterschrift